# Study of modularity in molecular, morphological and linguistic evolution using networks methods

Jananan Pathmanathan

Université Pierre et Marie Curie

Ecole doctorale Complexité du Vivant

*Evolution Paris Seine - UMR CNRS 7138*

*Equipe Adaptation, Intégration, Réticulation et Evolution*

# Study of modularity in molecular, morphological and linguistic evolution using networks methods

Par Jananan PATHMANATHAN

Thèse de doctorat en Biologie Evolutive

Dirigée par Eric BAPTESTE et Philippe LOPEZ

Présentée et soutenue publiquement le 23 Octobre 2017

Devant un jury composé de :

| | |
|---|---|
| Dr Eric BAPTESTE (DR CNRS, Université Pierre et Marie Curie, France) | Encadrant |
| Pr Didier CASANE (PR, Université Paris Diderot, France) | Rapporteur |
| Pr Tal DAGAN (PR, Université de Kiel, Allemagne) | Examinatrice |
| Pr Christophe DESSIMOZ (PR, Université de Lausanne, Suisse) | Rapporteur |
| Pr Michel MORANGE (PR, Université Pierre et Marie Curie, France) | Examinateur |
| Pr Philippe LOPEZ (PR, Université Pierre et Marie Curie, France) | Encadrant |
| Pr Mark RAGAN (PR, Université du Queensland, Australie) | Examinateur |
| Dr Hélène TOUZET (DR CNRS, Université Lille 1, France) | Examinatrice |

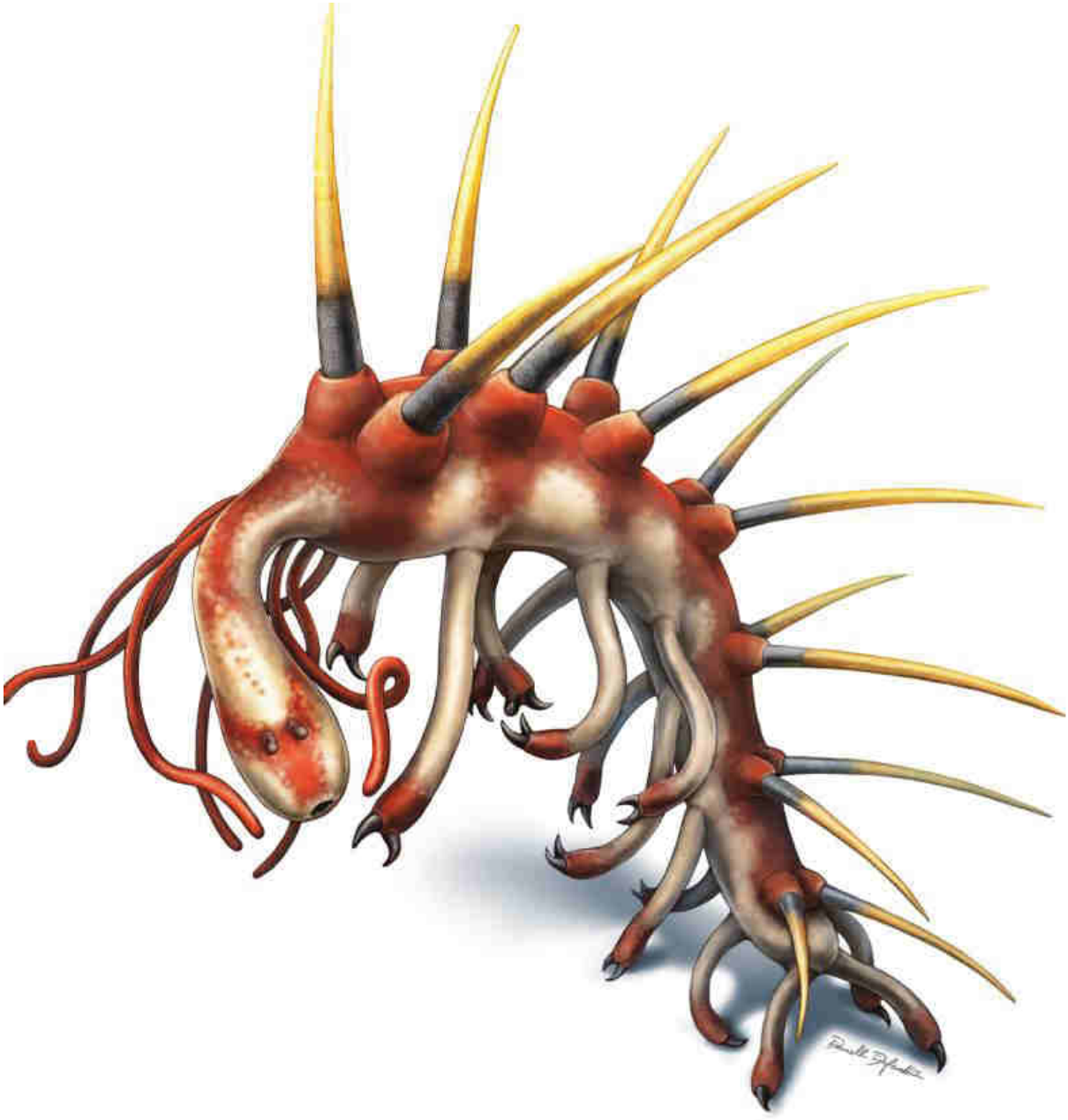
***கற்றது கைமண் அளவு, கல்லாதது உலகளவு!***

*("Kattrathu Kai Mann Alavu, Kallathathu Ulagalavu")*

*One of my favorite Tamil proverbs !*

*"What you have learned is a mere handful;*
*What you haven't learned is the size of the world"*

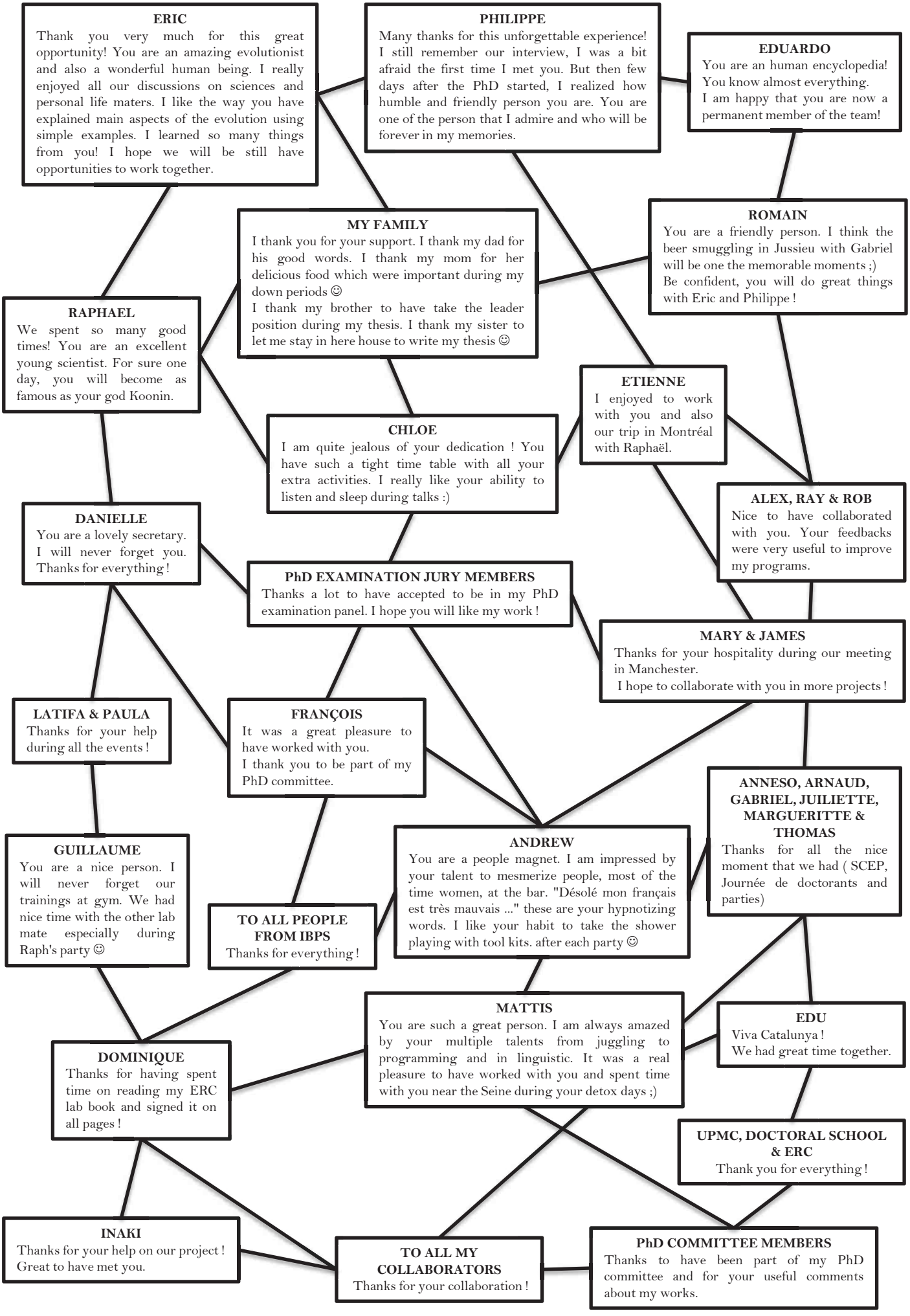
*"Known is a drop unknown is an ocean"*



*Reconstruction of Hallucigenia sparsa by Danielle Dufault.*

# Acknowledgements

I think writing the acknowledgments should be the hardest part of the thesis. Especially for me who grown up in the Tamil culture, I need to be careful not to forget anybody and the main problem is deciding which person to thank first! One thing that I have learned during my PhD is that "Networks are cool". So I decided to construct an undirected network of acknowledgments just behind this page. With this network, people can start reading from any node. Nodes are linked randomly !


ERIC
Thank you very much for this great opportunity! You are an amazing evolutionist and also a wonderful human being. I really enjoyed all our discussions on sciences and personal life maters. I like the way you have explained main aspects of the evolution using simple examples. I learned so many things from you! I hope we will be still have opportunities to work together.
PHILIPPE
Many thanks for this unforgettable experience! I still remember our interview, I was a bit afraid the first time I met you. But then few days after the PhD started, I realized how humble and friendly person you are. You are one of the person that I admire and who will be forever in my memories.
EDUARDO
You are an human encyclopedia! You know almost everything.
I am happy that you are now a permanent member of the team!
MY FAMILY
I thank you for your support. I thank my dad for his good words. I thank my mom for her delicious food which were important during my down periods ☺
I thank my brother to have take the leader position during my thesis. I thank my sister to let me stay in here house to write my thesis ☺
ROMAIN
You are a friendly person. I think the beer smuggling in Jussieu with Gabriel will be one the memorable moments ;)
Be confident, you will do great things with Eric and Philippe !
RAPHAEL
We spent so many good times! You are an excellent young scientist. For sure one day, you will become as famous as your god Koonin.
ETIENNE
I enjoyed to work with you and also our trip in Montréal with Raphaël.
CHLOE
I am quite jealous of your dedication ! You have such a tight time table with all your extra activities. I really like your ability to listen and sleep during talks :)
ALEX, RAY & ROB
Nice to have collaborated with you. Your feedbacks were very useful to improve my programs.
DANIELLE
You are a lovely secretary. I will never forget you. Thanks for everything !
PhD EXAMINATION JURY MEMBERS
Thanks a lot to have accepted to be in my PhD examination panel. I hope you will like my work !
MARY & JAMES
Thanks for your hospitality during our meeting in Manchester.
I hope to collaborate with you in more projects !
LATIFA & PAULA
Thanks for your help during all the events !
FRANÇOIS
It was a great pleasure to have worked with you.
I thank you to be part of my PhD committee.
ANNESO, ARNAUD, GABRIEL, JUILIETTE, MARGUERITTE & THOMAS
Thanks for all the nice moment that we had ( SCEP, Journée de doctorants and parties)
GUILLAUME
You are a nice person. I will never forget our trainings at gym. We had nice time with the other lab mate especially during Raph's party ☺
ANDREW
You are a people magnet. I am impressed by your talent to mesmerize people, most of the time women, at the bar. "Désolé mon français est très mauvais ..." these are your hypnotizing words. I like your habit to take the shower playing with tool kits. after each party ☺
TO ALL PEOPLE FROM IBPS
Thanks for everything !
MATTIS
You are such a great person. I am always amazed by your multiple talents from juggling to programming and in linguistic. It was a real pleasure to have worked with you and spent time with you near the Seine during your detox days ;)
EDU
Viva Catalunya !
We had great time together.
DOMINIQUE
Thanks for having spent time on reading my ERC lab book and signed it on all pages !
UPMC, DOCTORAL SCHOOL & ERC
Thank you for everything !
INAKI
Thanks for your help on our project ! Great to have met you.
TO ALL MY COLLABORATORS
Thanks for your collaboration !
PhD COMMITTEE MEMBERS
Thanks to have been part of my PhD committee and for your useful comments about my works.

# TABLE OF CONTENTS

# LIST OF FIGURES

# I. INTRODUCTION

*Image from Clément Escoffier (source: SlideShare.net)*

# I.1 MODULARITY

Most evolving objects in biology can be subdivided into smaller parts, each of which can be affected by natural selection. For instance, proteins are composed of protein domains, genomes are a set of genes and communities are made of individual organisms. As such, one can say that they have a modular organization. An object is modular if it can be divided into multiple sets of strongly interacting parts that are relatively autonomous with respect to each other (Clune et al. 2013; Melo et al. 2016). During my thesis, I especially focused on studying the modularity of various evolving objects (genes, genomes, organisms and languages), using the analogy of jigsaw puzzles. For example, genomes can be considered as a puzzle of genes. These puzzles can be divided in two categories: simple and composite puzzles (Figure 1). Simple puzzles are composed of pieces with the same phylogenetic origin (Figure 1A), whereas composite puzzles are composed of pieces with distinct origins (Figure 1B). For genomes, acquisition of foreign genes is the consequence of a variety of processes, the most well-known being Lateral Gene Transfer.

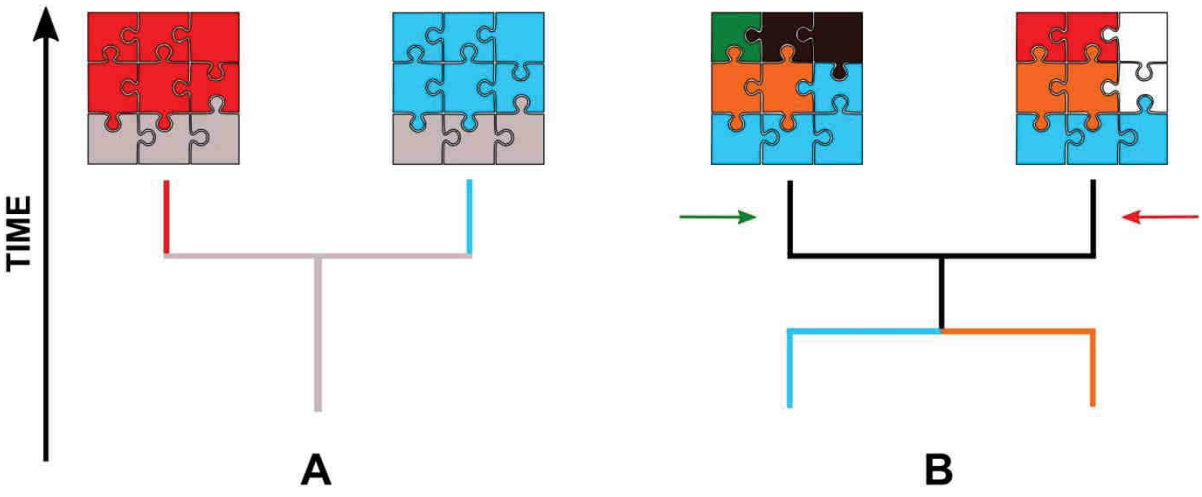



**Figure 1: Evolution of puzzles.**
(A) Evolution of simple puzzles with a common ancestor and (B) evolution of composite puzzles with distinct origins. Colors represent the phylogenetic origin of the component. On the left is a classical scenario where a grey ancestor undergoes a speciation, followed by the acquisition of red and blue synapomorphies. On the right is a more complex scenario, where blue and orange ancestor combine into an object that speciates and acquires black synapomorphies, and afterwards undergoes transfers from green and red objects (arrows). The resulting objects are composite.

Classical tree methods are ill-suited to describe the evolution of objects of the latter kind. Since composite objects are made of elements from different origins (and thus having different histories), their evolution cannot be described by a tree that describes gradual divergence from a unique ancestor. Networks however allow for a better representation of modular entities, where each node represents a puzzle and each edge represents their relation, better capturing their evolutionary history (Figure 2).

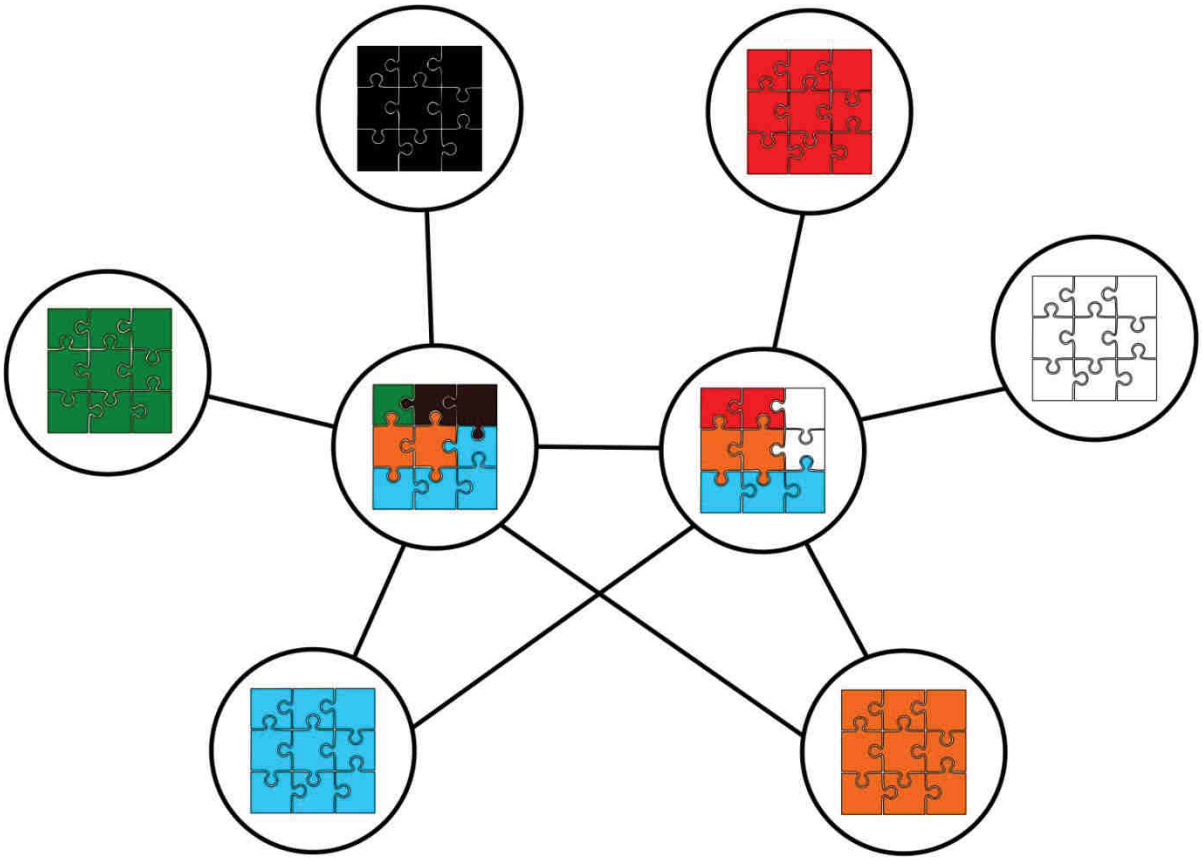


**Figure 2: Network of puzzles.**
Network of puzzles showing the relation between each entities. Nodes represent evolving modular objects, an edge is drawn between two nodes if the corresponding objects share components.

Moreover, pieces (e.g. genes) composing the puzzle (e.g. genomes) can be subject to different evolutionary processes (e.g. duplication, tinkering by fusion or fission, *de novo* evolution, mutation, losses) (Figure 3). During my thesis, I also developed and used network methods to detect and analyze the origin of these modular elements.

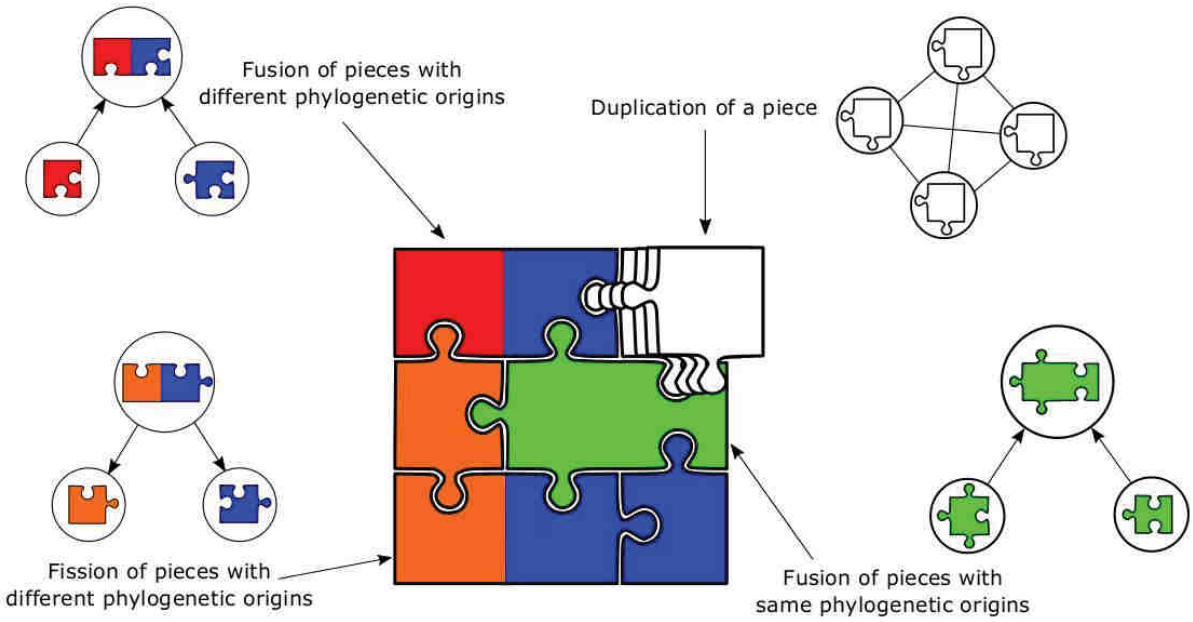



**Figure 3: Representation of evolutionary processes affecting the puzzle pieces.**
Different phylogenetic origins are differently colored.

## I.2 NETWORKS AND BIG DATA

First of all, we need to be aware that the concept of a network is not new in biology. They are used as a convenient representation of patterns of interaction between appropriate biological elements such as chemical reactions in cells or neuronal connections in the brain (Newman 2010).

A network is a system of elements that are connected (or not) to one another. It can be represented by entities modeled as nodes and their connections as edges (Figure 4). The nodes can represent units at all levels of the biological hierarchy, from genes and proteins to neurons and organs and limbs, and from individuals in a population to species in a community (Proulx et al. 2005). Edges usually represent some kind of interaction between nodes, including transcriptional control, biochemical interaction, energy flow and species interactions.

The study of a system by a network model requires two steps (Brandes et al. 2013). The first is to abstract the phenomenon in the form of a network, that means to define entities which will constitute the nodes and the relations between these entities which will constitute the edges. The second step consists in building the network and then analyzing it, using tools issued from graph theory.

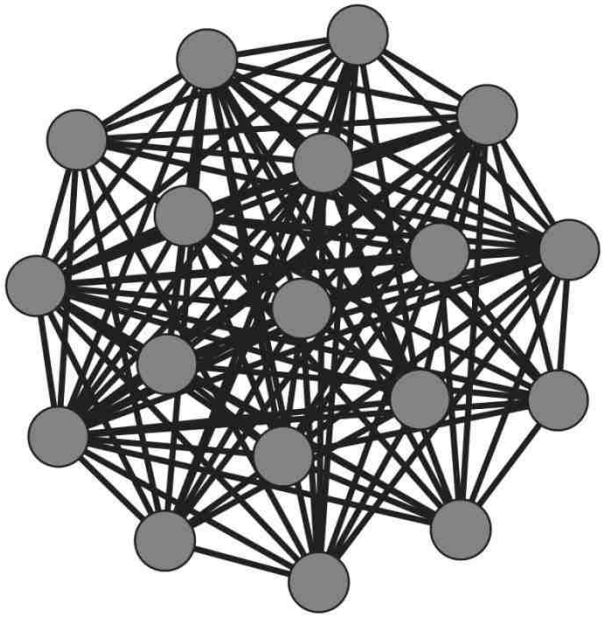


**Figure 4: Representation of a network**.
In this network nodes are in grey connected by an edge in black.

A huge advantage of networks in evolutionary studies is that they are relatively fast to compute. Starting from a set of objects, one only has to compute which pairs are connected, according to a given rule. As such, building a network has a very convenient quadratic complexity, allowing for the simultaneous study of a very large number of objects, typically up to millions. This property is especially desirable when studying the evolution of biological sequences, since the post-genomic era is characterized by an accelerated accumulation of molecular sequences with a considerable genetic diversity, from genome and metagenome sequencing projects (Sharpton 2014; Dolinski and Troyanskaya 2015; Eisenstein 2015).

## I.2.1 Data avalanche

After the discovery of the first sequencing methods, major ambitious projects such as the human genome sequencing project in 1990, were launched. The human genome sequencing project costed nearly three billion dollars and lasted 13 years. Since then, the aim has been not only to improve the sequencing methods by reducing their time but also their cost. This evolution has led to the reduction of the human labor force. Although still reserved for laboratories with considerable resources, the evolution of sequencing methods tends towards democratization.

Metagenomics is one of the disciplines that have emerged from sequencing. Its purpose is to simultaneously analyze the whole population of micro-organisms from a given medium instead of looking at a single species or clones. Thus, using high-throughput

sequencing methods, genes of interest can be isolated. The advantage is a greater speed of the analysis and the principle is also relevant from an ecological point of view. The important information is to know which genes, and therefore what biological functions, exist in the environment. Metagenomics has led to the study of the metagenomes of marine (Kennedy et al. 2010; Ma et al. 2012; Kodzius and Gojobori 2015), terrestrial (Daniel 2005; Delmont et al. 2011; Nesme et al. 2016) or even internal environments of animals (Mandal et al. 2015; Wang et al. 2015). The study of the genomes of organisms living in the environment can be useful in understanding the composition of an ecosystem. The TARA oceans project is a non-profit initiative. This project involves the sampling of planktonic samples and the core drilling of a selection of important coral colonies (Sunagawa et al. 2015). Planktonic organisms form the basis of oceanic ecosystems. The analysis of the samples, using high-throughput sequencing methods, will allow to know the study of the diversity and the geographical distribution of planktonic and coral species in order to better preserve them. The analysis of metagenomic data from TARA Oceans is expanding scientific databases and knowledge (Mitchell et al. 2016). Since 2015, several milestone articles were published on the TARA Oceans and studies related to ocean microbes (Zhang and Ning 2015; Gimmler et al. 2016).

Equally ambitious is the TerraGenome project (Vogel et al. 2009). The aim of this project is to complete the sequencing of the genome of all soil microorganisms. A colossal challenge, which opens up countless perspectives. Currently, 70% of the antibiotics on the market are derived from soil bacteria. These bacteria represent only a tiny fraction of the total bacterial biodiversity. On average, each gram of soil contains one billion bacterial cells, which is an almost inexhaustible reservoir of new bioactive molecules to discover. Results are thus expected in the understanding of bacterial mechanisms of adaptation and evolution in underground ecosystems. Metagenomics is also useful for exploring the bacterial "ecosystem" in humans.

In 2008, the METAHIT project was setup with the aim to characterize the genes and functions of microbes in human intestinal flora, as well as to understand their impact on our health. The first results observed by the sequencing of the genomes of the microorganisms made it possible to identify 1,150 species of bacteria, many of which were unknown until then (Qin et al. 2010). Many if the sequenced genes are genes needed by bacteria for the use of complex carbohydrates, the synthesis of vitamins and amino acids, as well as survival in the hostile environment of the intestine. The symbiosis between man and intestinal flora is particularly important for human physiology. Indeed, there is a relationship between the state

of the intestinal flora and certain chronic diseases of the intestine (Guinane and Cotter 2013; Zhang et al. 2015). The intestinal flora is composed of a group of bacteria common to all individuals, as well as a group specific to each. Studies on this second group would help to understand the cause of intestinal diseases or the tendency to obesity in some people (Duranti et al. 2017).

All these metagenomic projects are producing enormous amount of data that should be exploited in evolutionary biology (see article in Annex 1). This phenomenon is often described as a deluge of sequences, as a powerful flow difficult to channel. Indeed, the analysis of the information contained in such quantities of data poses many practical difficulties. The computer is now indispensable as a tool at all stages, from the sequencing of the nucleotides and the assembly of the fragments produced. Gene detection is based on necessarily automated statistical models. The study of these genes, once detected, uses methods that are extremely computationally intensive. Bioinformatic tools and associated databases for handling those datasets have been developed for the scientific community (Kim et al. 2013).

A fundamental challenge is the interpretation of this huge amount of data to elucidate new proteins functions, three-dimensional structures and evolutionary origin. Classical computational approaches heavily rely on homology-based annotation transfer, using tools such as BLAST (Altschul et al. 1990), HMMs (Yoon 2009), multiple alignments (Edgar and Batzoglou 2006) and motif finding algorithms (Bailey et al. 2009). So to study a set of new genetic sequences, biologists usually start to compare them with already known sequences and group them. For example, they can create groups of homologous genes (inherited genes in different species from a common ancestor) to study the evolutionary history of genes or create groups of orthologous genes (homologous genes where a gene diverges after a speciation event) for functional annotation (Pearson 2013). Along with the evolution of high-throughput sequencing technology, new methods based on network approaches have been introduced to analyze the rapid influx of these massive datasets of molecular sequences. That is why another key aspect of this thesis has focused on the development of new *in silico* approaches, which extend the exploitation of these large molecular data sets, based on networks.

## I.2.2 Sequence Similarity Networks

In the late 1990s, networks of sequences based on their similarity, known as "Sequence Similarity Networks" (SSNs) started to represent an attractive approach to enhance multiple sequence alignments and phylogenetic trees (Atkinson et al. 2009). One of the earliest formal and heuristic uses of SSNs was to define the COG groups of homologous families and facilitate prediction of the functions of large numbers of genes based on homology (Tatusov et al. 1997; Tatusov et al. 2000). SSNs are undirected graphs, where each node represents a unique sequence and each edge represents the similarity between connected sequences. This is the abstraction of a sequence similarity network, which in practice can be constructed in several ways. Sequence similarity searches can be performed by alignment tools, considering the sequence set as both a request and a target. In general, BLAST is the most commonly used tool for this purpose. BLAST returns all the local alignments with high similarity found between pairs of sequences. The construction of SSNs is based on the descriptors of these alignments such as the E-value and the percentage of identity (Figure 5). This output can already be interpreted as a network, where each line is an edge between a target sequence and a query sequence. This output needs to be filtered in order to keep useful information for the SSN.

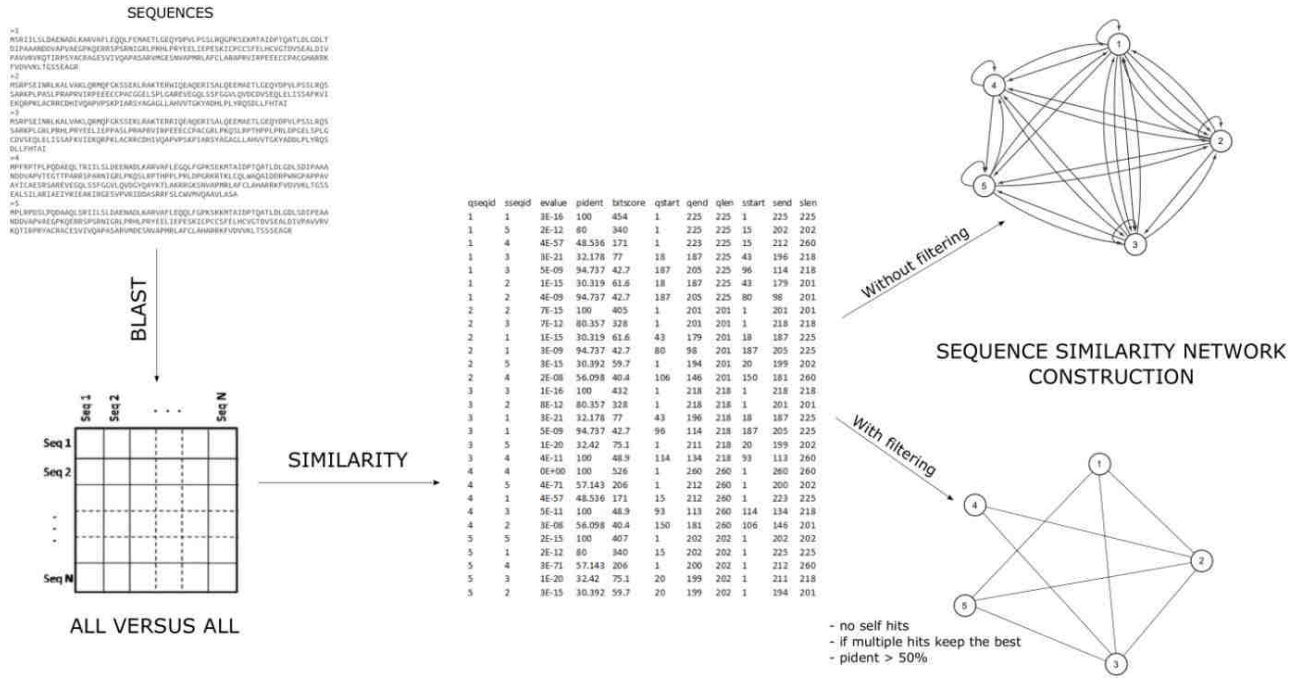



**Figure 5:** Constructing a simple sequence similarity network.

As all sequences are compared against themselves (all versus all), we obtain self alignment information for each sequence. Self-hits are not informative and should be deleted. For a given comparison between two sequences, the alignment, score and E-value are not symmetric. The BLAST score between a pair of sequences can vary depending on which sequence is used as the query. It is also possible that given a pair of sequences, the alignment is present in one direction and not in the other. If the E-values associated with the comparisons are on either side than the threshold limit given in argument to BLAST. This asymmetry has no biological meaning, and it is therefore convenient to symmetrize the network by considering the best match of each pairwise comparison. In order to build the SSN, it is then common to annotate the undirected edge by the descriptors of this alignment (Percentage of identity, length, etc.). Finally, there may be multiple alignments at distinct locations along a pair of sequences, for various reasons related to the evolution of the sequences (variable divergence rhythms, insertion of non-homologous regions) or the BLAST algorithm (excluding regions of low complexity). In that situation, we will only keep the best alignment based on E-value.

### I.2.3 Computational bottleneck of SSN

Although networks allow the study of large datasets, there are some major computational bottlenecks that need to be overcome in the construction of a SSN. The most expensive step is often the comparison of sequences with alignment tools which produce a hypothesis of homology. Alignment is the first and most important step in the network analysis. This fundamental procedure attempts to infer which series of individual characters or patterns within sequences are homologous, that is to say, share a common evolutionary origin. The alignments may contain errors depending on the nature of the data and may have huge downstream effects (Rosenberg 2005).

Since the 1970s and the seminal work of Needleman and Wunsch, more than hundred alignment programs have been developed (Rosenberg 2009). However, this field still needs more exploration. We can divide the alignment algorithms into two categories: global alignment and local alignment. Global alignment attempts to align the entire sequence, end-to-end. It was introduced by Needleman and Wunsch and was the first alignment procedure (Needleman and Wunsch 1970). Global alignment is well suited for comparing closely related sequences having approximately the same length. Nevertheless, this assumption may be incorrect in molecular evolution involving sequence rearrangement and shuffling. In this

situation, local alignment is an alternative to global alignment. The local alignment attempts to align subsections of the sequences without considering the alignment of rest of the sequence regions. The subsections may be part or all of the sequences. These local alignment tools, used to find conserved patterns between sequences, are appropriated for aligning more divergent or distantly related sequences. Although the first local alignment approaches were introduced by Sankoff (Sankoff 1972) and Sellers (Sellers 1974), the most commonly used procedure is a modification of the Needleman-Wunsch algorithm proposed by Smith and Waterman (Smith and Waterman 1981).

From the mid 1980s, local alignment tools like FASTA (Pearson and Lipman 1988) were developed in the aim of database searching rather than a simple sequence comparison (Pearson 2013). In 1990, Altschul *et al.* published an article about their alignment tool called **B***asic* **L***ocal* **A***lignment* **S***earch* **T***ool* (or BLAST) providing flexible and fast alignments involving large sequence databases. BLAST, considered as the reference among alignment tools, is the most popular and most widespread approach with more than 65,000 citations of the original paper (Altschul et al. 1990). BLAST uses a "seed-extension" approach. A seed is short word (k-mer) of k letters. First, all identical or very similar k-mers between two sequences are identified. Secondly, these short subsequences matches between the sequences are extended by measuring the similarity score at each extension. The seed extension is stopped when the score decreases, and the best score alignment obtained during the extension is retained. Since BLAST, the seeding technique became central in the theory of sequence alignment. Unlike Smith-Waterman and Needleman-Wunsch strategy which compare sequences base by base, the seeding and extending approach significantly increases speeds but cannot be guaranteed to find the optimal alignment (Altschul et al. 1990). Recently, new variants of seed-extension approach have been implemented using flexible-length seeds on a reduced amino acid alphabet like Tachyon (Tan et al. 2012), PAUDA (Huson and Xie 2014), PSimScan (Kaznadzey et al. 2013), RAPsearch2 (Zhao et al. 2012), Lambda (Hauswedell et al. 2014), UBLAST (Edgar 2010), DIAMOND (Buchfink et al. 2015) and MMseqs (O'Driscoll et al. 2015).

There is a fundamental difference between the biological and computational goals of alignment algorithms, respectively homology and optimization. A computationally optimal solution is not always biologically correct (Kumar 1996; Nei et al. 1998; Takahashi and Nei 2000). Computing similarities against very large datasets or databases is almost impossible in a single workstation in a feasible time using exhaustive sensitivity settings. Thus, much effort

has been put on the improvement of existing programs to use high-performance computing (HPC) environments, such as clusters, grids, graphics processing units and clouds, together with parallelism techniques. For example, HBLAST (O'Driscoll et al. 2015) or HAMOND (Yu et al. 2017) are the parallelized versions of BLAST and DIAMOND using the Hadoop framework for computer clusters.

## I.3 AIMS OF THE THESIS

As stated above, the exponential increase of the number of available sequences, in particular from metagenomic studies, requires new comparative methods to explore the diversity of large datasets, in a way that also accounts for the complexity of the evolving entities (i.e. their modularity).

The main subject of this thesis was thus the study of the modular evolution of genes. The modular nature of genes, i.e. the fact that genes are comprised of various components, such as introns, exons, domains, is well known (Gilbert 1978). The remodelling of these modular genes by shuffling, fusion and fission of genetic fragments, as well as *de novo* DNA synthesis, contributes to the creation and diversification of gene families. These processes differ from mechanisms where sequences progressively diverge by accumulating point mutations (substitution, insertion, deletion) within a gene family. They are problematic for the construction of gene phylogenies and, as much as possible, are eliminated when studying the evolution of genes using trees. This recognized modularity complexifies the study of molecular evolution, requiring the development of specific strategies to characterize genes features, i.e. to identify the components of the genes and to decipher the rules of these components' associations. Before the start of this thesis, sequence similarity networks had proven to be an important tool to identify homologous gene families and to provide a useful analytical framework to study the impact of combinatorial processes on molecular evolution, such as recombination, fusion or fission.

Further, in my thesis, I show that similarity networks are adapted to capture and analyze evolutionary history of modular entities beyond genes, such as organisms morphology and languages.

In Chapter I, I explain why network-based methods are starting to be used to complement phylogenetic analyses in studies in molecular evolution. With my colleagues, I

contributed to write a book chapter on the different kinds of networks based on sequence similarity that have been introduced to tackle a wide range of biological questions, including sequence similarity networks, genome networks and bipartite graphs, and a guide for their construction and analyses.

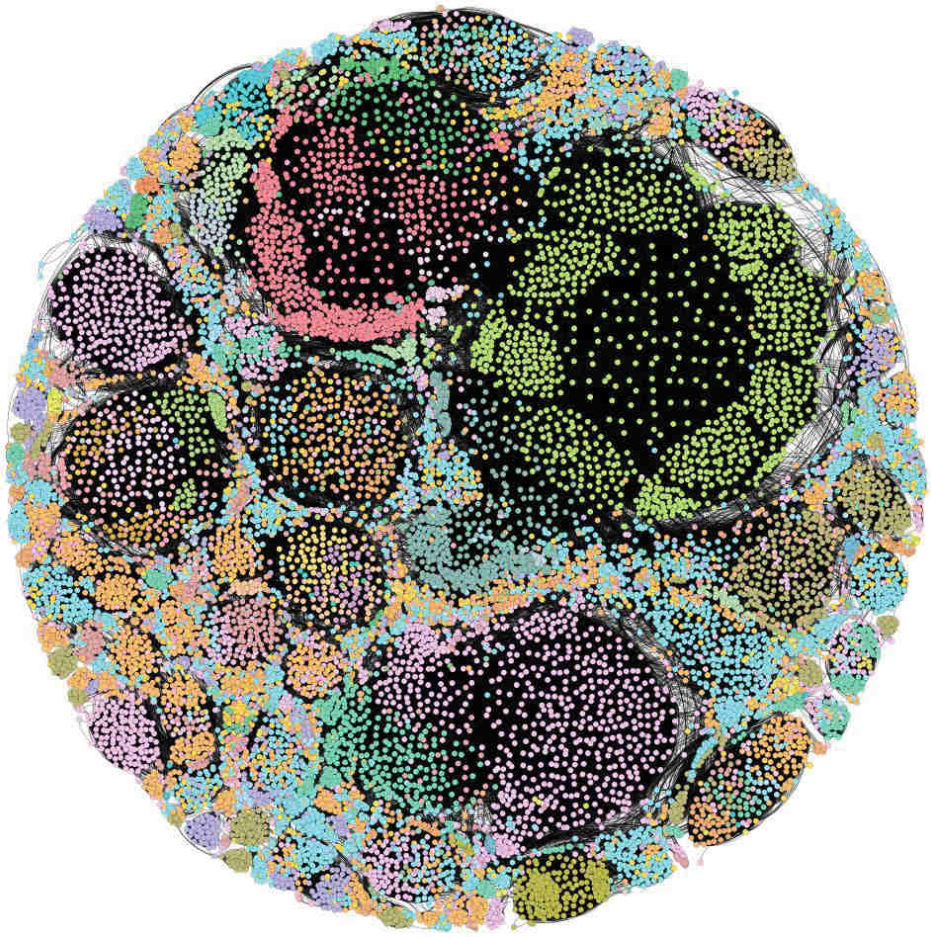
In Chapter II, I introduce case studies that show how networks based approaches can be used to study the modularity in molecular, morphological and linguistic evolution.

First, I explain the benefit of using networks to study gene remodeling (Chapter II.1). I introduce CompositeSearch, one of the software that I developed during my thesis, for the detection of composite genes and composite gene families. I applied CompositeSearch to analyze the distribution and impact of remodeled genes in plasmids, in eukaryotes (to study the transition of unicellularity to multicellularity, in collaboration with Pr Iñaki Ruiz-Trillo), and in microbiomes from polluted environments.

Second, I introduce a new approach, developed with a palaeontologist (Pr Pierre-Olivier Antoine) and ecologist and biostatistician (Pr François-Joseph Lapointe), to study the evolution of organisms morphology (Chapter II.2). Organisms are modular at one or more levels of organization, e.g. interconnected regulatory, metabolic, protein-protein interaction and genetic or developmental networks (Wake 2008; Mitra et al. 2013). Beyond the molecular level, organisms can also be seen as networks of morphological components, whose organization stems from that of underlying molecular networks. With Dr Etienne Lord (former Postdoc in Pr F-J Lapointe Lab), we developed Component-Grapher, a tool using network approaches and applied it on palaeontological and extant morphological data to analyse the co-occurrence relationships between organismal traits during the evolution of panarthropods since the Cambrian, and the evolution of rhinocerotid mammals.

Finally, I investigated the important evolutionary processes in biology and in linguistics with our linguist collaborator Dr Mattis List (Chapter II.3), and we identified specific and common processes in these disciplines. We showed that network-based methods can also be used to detect non-tree like aspects of language history, like compound words, which are similar to composite genes or words borrowing similar to horizontal gene transfer in language. We also designed a case study, using networks in linguistics for the reconstruction of one aspect of language evolution, i.e. phonemes in Old Chinese pronunciation.

# II. NETWORK-THINKING IN EVOLUTION



*Sequence similarity network of complete virus genomes*

# II.1 NETWORKS: A COMPLEMENTARY METHOD TO PHYLOGENETIC ANALYSIS OF EVOLUTION

An evolutionary biologist is interested in how processes affecting evolution have produced the diversity of genes, genomes, organisms, species and communities that are observed today.

A classical approach to study these processes is the reconstruction of phylogenetic trees of genes, genomes, organisms and species (Figure 6); an outcome from the crucial Darwin works on theory of evolution by natural selection, published in his book "On the Origin of Species" (1859) (Darwin 1859). The theory of evolution is based on the idea that all living organisms evolve from earlier forms by modification and divergence according to a tree process as a result of natural selection (Lewontin 1970). It is commonly assumed that such evolving units present a few necessary conditions for evolution by natural selection, namely (*i*) phenotypic variation among members of an evolutionary unit, (*ii*) a link between phenotype, survival, and reproduction (i.e., differential fitness), and (*iii*) heritability of fitness differences (individuals resemble their relatives more than unrelated individuals) (Bapteste et al. 2012).

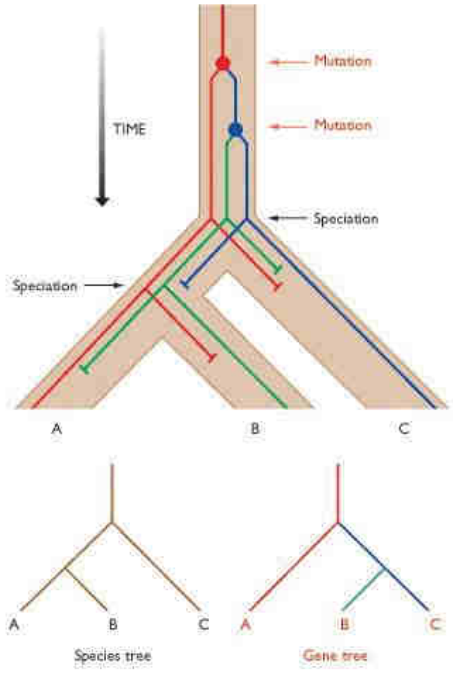



**Figure 6: A gene tree can have a different branching order from a species tree.**
In this example, the gene has undergone two mutations in the ancestral species, the first mutation giving rise to the 'blue' allele and the second to the 'green' allele. Random genetic drift in association with the two subsequent speciations results in the red allele lineage appearing in species A, the green allele lineage in species B and the blue allele lineage in species C. Molecular phylogenetics based on the gene sequences will reveal that the red-blue split occurred before the blue-green split, giving the gene tree shown on the right. However, the actual species tree is different, as shown on the left. Based on Li W-H (1997) Molecular Evolution. Sinauer, Sunderland, MA. (Brown 2002)

Like Darwin, scientists believed that evolution was a slow and gradual process. However, the tree model is not enough to explain the evolutionary history of life on Earth (Nutman et al. 2016). Besides the tree-like process, other processes called nongradual, involving combinatorial (e.g. recombination, fusion, fission) and introgressive (e.g. integration of foreign genetic element in to a genome by HGT) exist and cannot be represented accurately by a tree (Dagan et al. 2008; Halary et al. 2010; Corel et al. 2016).

Saltational processes, such as recombination events, fusion, fission or lateral gene transfer (or horizontal gene transfer), are found at different levels of biological organization (Figure 7) Network-based methods have been described to be a well suited approach to detect, analyze and visualize the vertical and horizontal relationships at the genomic level and in several genomes at the same time (Corel et al. 2016). Network approaches are increasingly used to complement phylogenetic analysis in molecular evolution, comparative genomics, classification and ecological studies (Halary et al. 2013). For example, their suitability for investigating introgressive events have enhanced our understanding of the chimeric origin of genes in the eukaryotic proteome (Thiergart et al. 2012; Alvarez-Ponce et al. 2013), the flow of genes between prokaryotes and their mobile genetic elements (Halary et al. 2010; Dagan 2011; Kloesges et al. 2011; Popa et al. 2011; Jaffe et al. 2016) and gene sharing across mobile elements to study the transfer of resistance factors (Fondi and Fani 2010; Tamminen et al. 2012). Networks have also been used to describe complex biological systems, including inferring the “social networks” of biological life forms (Halary et al. 2010), producing maps of genetic diversity (Cheng et al. 2014), detecting distant homologues (Park et al. 1997; Bolten et al. 2001; Bapteste et al. 2012) and exploring gene and genome rearrangements (Jachiet et al. 2013; Meheust et al. 2016).

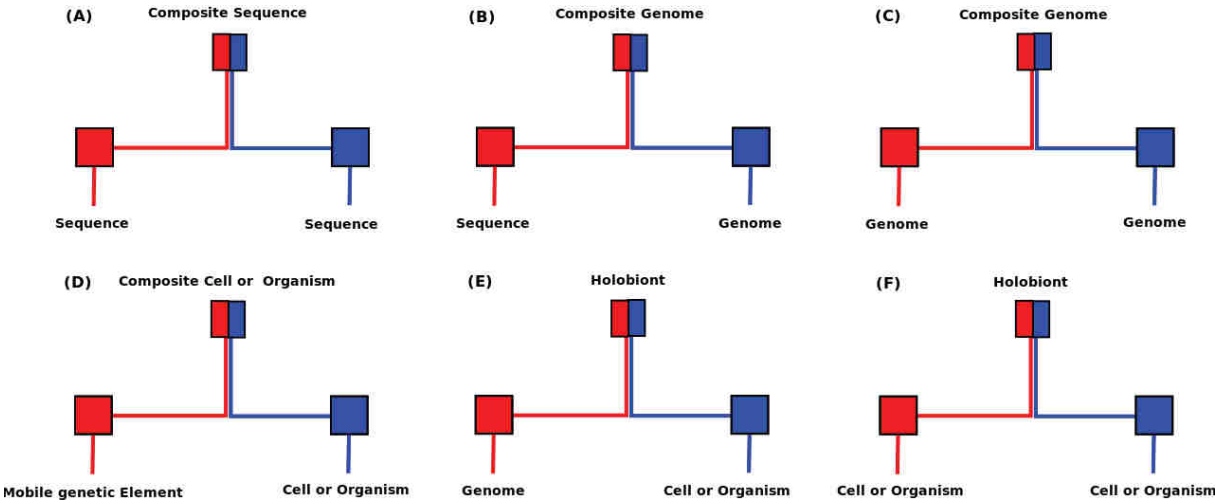



**Figure 7: Several illustrations of mosaicism through merging events.**
(A) Composite genes result from the fusion of different gene domains. (B) Composite genomes can result from the introgression of a gene into a genome, or (C) from the introgression of a genome into a genome. (D) Composite organisms can arise from the introgression of a mobile genetic element. Holobionts result from the introgression of a genome (E) or of another cell (F) into a cell. (Corel et al. 2016)

The revolution in DNA sequencing has been a major advance for evolutionists, giving them new opportunities to investigate these diverse kinds of questions with molecular data; however they also present challenges in terms of the scale of the analyses. Consequently, development of new methods for the construction and analysis of networks has been necessary. In the book chapter "*The Methodology Behind Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution*", we present the different kinds of networks based on sequence similarity that have been introduced to tackle a wide range of biological questions, including sequence similarity networks, genome networks and bipartite graphs, and a guide for their construction and analyses. This book chapter has been submitted to the editor Anisimova Maria for the book "Evolutionary Genomics: statistical and computational methods" (Humana Press, Springer).

# *The Methodology Behind Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution.*

Andrew K. Watson[1], Romain Lannes[1], Jananan S. Pathmanathan[1], Raphaël Méheust[1], Slim Karkar [1+], Philippe Colson[2,3], Eduardo Corel[1], Philippe Lopez[1], Eric Bapteste[1].

[1] Sorbonne Universités, UPMC Université Paris 6, Institut de Biologie Paris-Seine, F-75005, Paris, France; [+] Author is now with: Rutgers, the State University of New Jersey, CCIB, Rutgers-Camden.

[2] Fondation Institut Hospitalo-Universitaire Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Tione, Assistance Publique-Hôpitaux de Marseille, Marseille, France.

[3] Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM63;CNRS 7278 ; IRD 198;INSERM U1095, Aix-Marseille Univ., Marseille, France

# Summary/Abstract

In the post genomic era, large and complex molecular datasets from genome and metagenome sequencing projects expand the limits of what is possible for bioinformatic analyses. Network based methods are increasingly used to complement phylogenetic analysis in studies in molecular evolution, including comparative genomics, classification and ecological studies. Using network methods, the vertical and horizontal relationships between all genes or genomes, whether they are from cellular chromosomes or mobile genetic elements, can be explored in a single expandable graph. In recent years, development of new methods for the construction and analysis of networks has helped to broaden the availability of these approaches from programmers to a diversity of users. This chapter introduces the different kinds of networks based on sequence similarity that are already available to tackle a wide range of biological questions, including sequence similarity networks, genome networks and bipartite graphs, and a guide for their construction and analyses.



# Key Words

Sequence similarity network, evolution, lateral gene transfer (LGT), metagenomics, gene remodelling, ecology.

# Introduction

An evolutionary biologist is interested in how processes governing evolution have produced the diversity of genes, genomes, organisms, species and communities that are observed today. For example, a biologist interested in the eukaryotes may wonder what symbiotic partners have contributed to their origins and evolution. Eukaryotic nuclear genomes are chimeric in nature, encoding many genes acquired from their alpha-proteobacterial endosymbiont *(1–3)*. However in recent years it has been proposed that the ongoing gain of genes by both microbial *(4–6)* and multicellular Eukaryotes *(7, 8)* via lateral gene transfer (LGT) has continued to contribute to eukaryotic evolution, though to a lesser extent than prokaryotes *(9)*. A biologist interested in prokaryotes may wish to investigate lateral gene transfer to explore the extent and kinds of genes transferred between bacteria, archaea and their mobile genetic elements *(10–14)*. These transfers are important for understanding the accessory genomes of prokaryotes *(15–17)*. Further, studying gene transfers in real bacterial communities from different environments can help to test the effect of LGT on ecology and evolution of communities *(18)*. Given the prevalence of introgression *(9–11, 19)*, one interesting question is whether gene transfer has led to the formation of novel fusion genes that combine parts of genes originating from separate domains of life *(20)*. An ecologist may wish to analyse the distribution of genes and species in the environment *(21)*. A metagenome analyst may need to overcome an additional challenge is exploring the nature of the large proportion of sequences in metagenome sequence projects that have with little or no detectable similarity to characterised sequences in order to study the "microbial dark matter" *(22)*.

High-throughput sequencing technology presents new opportunities to investigate these diverse kinds of questions with molecular data; however they also present challenges in terms of the scale of the analyses. Consequently, a number of network based methods have recently been developed to expand the toolkit available to molecular biologists *(23)*, and these have already made major contributions to our understanding molecular evolution. Networks have been used to shed light on the nature of the "microbial dark matter" *(24)* and used in ecological studies to explore the geographical distribution of organisms or genes *(25, 26)* or the evolution of different lifestyles *(27)*. Their suitability for investigating introgressive events have been used to enhance our understanding of the chimeric origin of genes in the eukaryotic proteome *(28, 29)*, the flow of genes between prokaryotes and their mobile genetic elements *(30–35)* and gene sharing across mobile elements to study the transfer of resistance

factors *(14, 36)*. Networks have also been used to classify highly mosaic viral genomes *(37, 38)* and identify gene families *(39, 40)*.

While the generation and analysis of networks was previously limited to biologists with programming experience, tools have recently been developed to simplify the process and broaden the availability of network analyses of molecular sequence data. This chapter introduces the different kinds of networks that are already available to biologists and a guide to how these networks can be constructed and analysed for a large range of applications in molecular evolution. More precisely, this chapter will focus on three kinds of network and the types of analyses that are possible using these networks: sequence similarity networks, genome networks, and multipartite graphs *(23)*.

# Sequence similarity networks (SSNs)

Sequence similarity networks are the bread and butter of network based molecular sequence analyses, with a huge range of applications in molecular biology. The use of SSNs for molecular sequence analysis first came to the fore in the late 1990s and early 2000s, when SSNs were suggested as a way to analyse the rapid influx of new molecular sequence data due to advances in sequencing technology and cost, as well as to predict gene functions and protein-protein interactions *(39, 41–43)*. One of the earliest formal and heuristic uses of SSNs was to define the COG groups of homologous families and facilitate prediction of the functions of large numbers of genes based on homology *(39, 40)*. The need for efficient computation and analyses for large biological databases still pervades; however more recently SSNs have been increasingly appreciated as useful approaches to describe complex biological systems, including inferring the “social networks” of biological life forms *(30)*, producing maps of genetic diversity *(27)*, detecting distant homologues *(44–46)* and exploring gene and genome rearrangements *(47, 48)*.

A SSN is a graph in which each node is a sequence and edges connect any two nodes that are similar at the sequence level above a certain threshold (e.g. coverage, percent identity and E-value) as determined by their pair-wise alignment (Box 1) (Figure 1). While the principle behind SSN construction is simple, the expression of similarity data in this structure can enable the use of powerful algorithms for graph analyses to study complex biological phenomena. Construction of a SSN is also frequently the starting point in a diversity of further graph analyses. A SSN can be constructed directly from fasta formatted sequence files using pipelines, such as EGN *(49)*, the updated and faster performing EGN2 (forthcoming),

or PANADA *(50)*. Visualisation of networks can be performed with programs such as Cytoscape *(51)* or Gephi *(52)*, both of which also have a range of internal tools and external plugins for network analysis. While these programs are useful for the visualisation and analysis of relatively small networks, it can be difficult to load large and complex networks with a lot of edges (e.g. ≥50,000 edges). In these cases the iGraph library offers an extremely powerful and well supported implementation of a broad range of commonly used methods for both complex graph generation and analysis in R, Python and C++ *(53)*. However, using iGraph requires knowledge of programming in at least one of these languages. An additional package for network analysis in Python is NetworkX *(54)*. It is our goal here to further generalise network approaches by explaining how evolutionary biologists with less programming knowledge could analyse their data. A list including many of the tools and programs available for SSN generation is available at https://omictools.com.

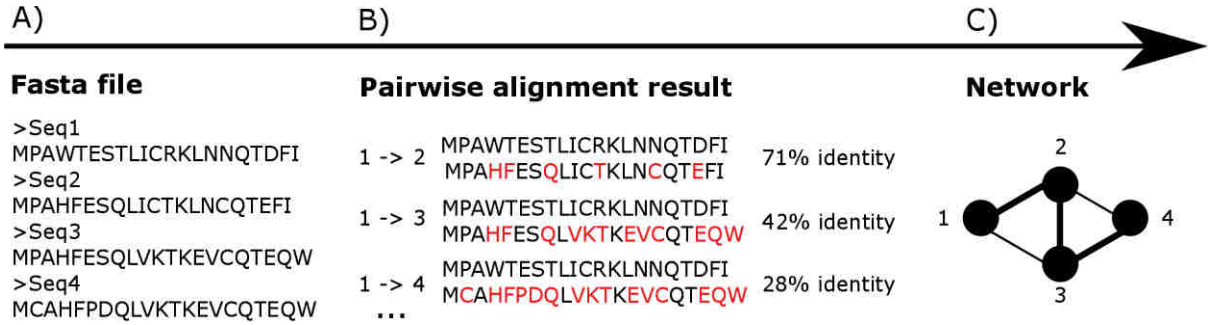



**Figure 1: Constructing a simple sequence similarity network:** A set of sequences (protein or DNA) in fasta format (A) are aligned in pairs using alignment tools (such as BLAST). These alignments (B) are scored with metrics such as the percentage identity between two sequences (the number of identical nucleotides / amino acids displayed above) or the E-value of the alignment. In the resulting network, sequences are represented as nodes. Two sequence nodes are joined with an edge if they can be aligned above a define threshold, with the weight of the edge often based on percentage identity or E-value.

## Box 1: How to build your own sequence similarity network

1) **Dataset assembly:** The first and most important step of SSN construction is the assembly of a dataset of sequences relevant to your biological question, usually in fasta format. This can be used as the initial input for wizards such as EGN or EGN2 *(49)*, which can fully automate the process. The nature of the dataset is highly dependent on the question, so here we focus on the practicalities of database assembly. To construct the similarity network all sequences in the dataset are aligned against one another in a similarity search. This similarity search is often the time limiting step in an analysis, and the total number of searches required is quadratic to the number of sequences in the dataset. For large datasets it is useful to benchmark the alignment using a subset of the data to estimate the timescale for the alignment. Large datasets can generate huge outputs, not only due to the number of sequences but also the length of their identifier. One way to reduce the output

size is to replace each sequence name in the fasta file with a unique integer. The use of integers will reduce disk space use and the memory consumption for any software used to analyse the sequence data.

2) **Similarity search**: To generate a sequence similarity network all sequences must be aligned against one another in an all versus all search, in which the dataset of sequences is searched against a database including the same sequences. For gene networks, the alignment is usually done with fast pairwise aligner such as BLAST *(55, 56)* as implemented in EGN *(49)*. Filters are often used to remove low-complexity sequences from the search, as these can cause artefactual hits (BLAST options --seg yes, -soft-masking true). The BLAST method of alignment will be the focus of future discussion in this chapter, however alternatives are available including BLAT *(57)* (also implemented in EGN), SWORD *(58)*, USEARCH *(59)* and DIAMOND *(60)*. These alternatives generally include an option to produce a "BLAST" style tabulated output, making them compatible with programs commonly used in network analyses.

   Within alignment tools like BLAST it is possible to assign set thresholds, such as the maximum E-value of the alignment to retain only significant hits or to output only the best alignment for a pair of sequences (BLAST option –max_hsps 1), drastically reducing the size of the output. It is not recommended to set minimal thresholds for some parameters (such as % sequence identity) unless required due to memory constraints so that you can generate networks from a single sequence alignment with different thresholds for comparison (e.g. comparison of a 30% similarity threshold to a 90% threshold, where edges will only be drawn between highly similar genes).

   Note: It may be intuitive to use additional CPUs to speed up the alignment process, however in BLAST it can be more efficient to split the query file and launch multiple searches on separate cores instead of using the BLAST multithreading option. The pairwise alignment step is generally the most time limiting part of generating a SSN, so benchmarking should be used to establish the optimal settings for the pairwise and/or determine the feasibility of a project given the size of the dataset and the available computational resources.

3) **Filtering similarity search results**: In an all versus all similarity search any given query sequence will have a self-hit in the corresponding database. For example with sequences A and B: a self-hit is query sequence A matching to sequence A in the database, cases of

which must be removed prior to network construction (**Erreur ! Source du renvoi introuvable.**). When query sequence A in a similarity search is aligned with sequence B in the database, often the reciprocal result is also identified (an alignment between query sequence B and sequence A in the database). These are called reciprocal hits; while the sequences involved are identical, the alignments and scores are not. Retaining both hits would generate two different edges between the same two nodes in a SSN, so generally only the best results from reciprocal hits are retained, based on a score such as the E-value (**Erreur ! Source du renvoi introuvable.**). Finally, a single query sequence may be significantly aligned multiple times in different positions of the same sequence in the database, however for SSN construction only the best BLAST hit is generally retained (**Erreur ! Source du renvoi introuvable.**). The selection of the best BLAST hit is again generally often based on the E-value (corresponding to the BLAST -max_hsps 1 option). Removing multiple hits against the same sequence allows the generation of an undirected network where a single edge connects two nodes, representing the best possible alignment between these nodes.

4) **Thresholding and Network construction**: Constructing a SSN from a BLAST output is conceptually simple; an edge is created between two sequences (nodes) that have been aligned in the sequence similarity search. It is common to apply thresholding criteria such as minimal % ID and/or coverage and/or maximal E-value to determine whether an edge is drawn between two sequences in the network (Figure 1). There are different ways to calculate the % coverage of an alignment. This could be based on the coverage of a single sequence in the alignment, selecting either the query or the database sequence in each alignment, or the longest or shortest sequence in each alignment. Alternatively both (mutual coverage) can be used, retaining an alignment when both values are above a given threshold. Edges above the thresholding criteria can be assigned a weight based on these criteria, producing a weighted sequence similarity network that retains information of the properties of the alignment between two sequences (Figure 1). It is often useful to construct and compare several SSNs with variable stringencies defining the edges between sequences, for example, to optimise gene family detection within the SSN (discussed below).

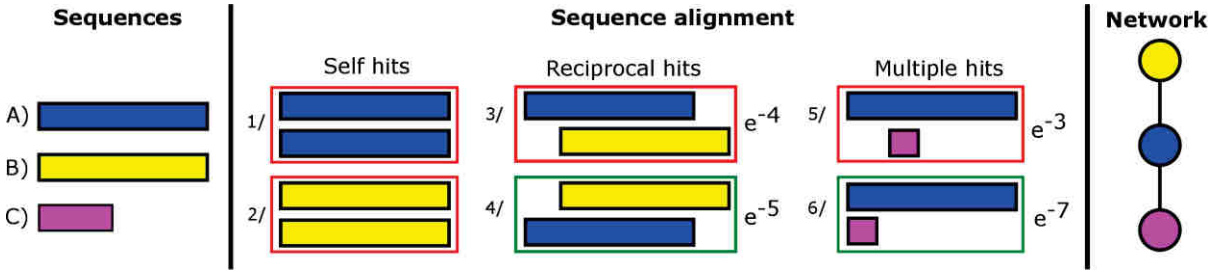



**Figure 2: Filtering sequence similarity results for network construction:** In the output of an all against all sequence similarity search there are a number of features that are often filtered out prior to network construction. Self hits (1/ and 2/), where like sequences are paired in a sequence alignment, are not informative to network construction and are removed (highlighted by the red box surrounding the alignments). In cases where there are reciprocal hits (3/ and 4/) between two sequence then only the alignment with the highest E-value is retained (highlighted with a green box around the retained alignment) to ensure only one edge representing the best possible alignment connects any two nodes in the network, The same is true for cases where a sequence has multiple hits against another sequence, such as when it aligns to another sequence in multiple positions (5/ and 6/).

## Exploiting sequence similarity networks for identification of gene families

A gene family is usually defined as a group of sequences that are similar at the sequence level, indicative of homology and potentially of shared functions, however there is no uniform way to define this similarity *(61, 62)*. One of the early contributions of SSNs in molecular sequence analysis was in the construction of the COG database of homologous protein sequences *(39, 40)*. This study attempted to define gene families based on similarity at the sequence level using the results of sequence similarity searches. Within the results of an all *versus* all BLAST search, groups of at least three proteins encoded by different genomes that were more similar to each other than they were to other proteins found in the same genomes were defined as a likely orthologous gene family. Orthologous gene families are group of genes in different genomes that show sequence similarity, likely as a result of their shared evolutionary history.

The idea of using graphs to identify gene families is now a core part of many graph-based analyses. Members of a gene family aggregate in a sub-network in a SSN. These sub-networks are called connected components (CCs) at these defined thresholds, i.e. clusters of nodes connected by edges either directly or indirectly (via intermediate nodes) (Figure 3). The size (number of nodes and edges in a CC) and density (the proportion of potential connections between all nodes in a CC that are actually connected by edges in the graph) of CCs will depend on the thresholds used for constructing the SSN as well as the relationships between sequences in the network. For example, for a given dataset at a given mutual coverage threshold, a threshold of 90% sequence identity will identify a large number of small connected components that only include highly similar genes, while at threshold of 30% sequence identity there will be fewer but larger connected components including genes

with more variation in sequence similarity. Commonly used thresholds for detecting homologous gene families are an *E-value* ≤ e-5, mutual coverage ≥ 80% and a percentage of identity ≥ 30% *(23)*.

CCs are often detected in a SSN using the Depth-First Search (DFS) algorithm; however there are also other approaches for the detection of gene families based on the idea of detecting "communities" *(63)*. In some cases, a CC can be further separated into communities of sequences that share more similarity to one another than to other sequences in the CC, thus are more highly linked in the SSN (Figure 3). Communities are commonly identified by using graph clustering algorithms such as Louvain *(64)*, MCL *(65)* or OMA *(66)*, however different clustering algorithms will result in different outputs. The Louvain weighted method is widely used because it is simple to implement and scales very well to large graphs (Figure 3, Figure 4) *(64)*. MCL is a strong deterministic algorithm that has been implemented, for example, in tribeMCL *(65)* and orthoMCL *(67)*. A potential drawback of MCL is that it requires user specification of the "inflation index", a parameter which controls cluster granularity (or "tightness"). A high inflation index increases the tightness of clustering; producing a larger number of clusters that are smaller on average than those that would be obtained clustering the same dataset using a low inflation index. Selecting an appropriate inflation index is not trivial and requires optimisation *(65)*.

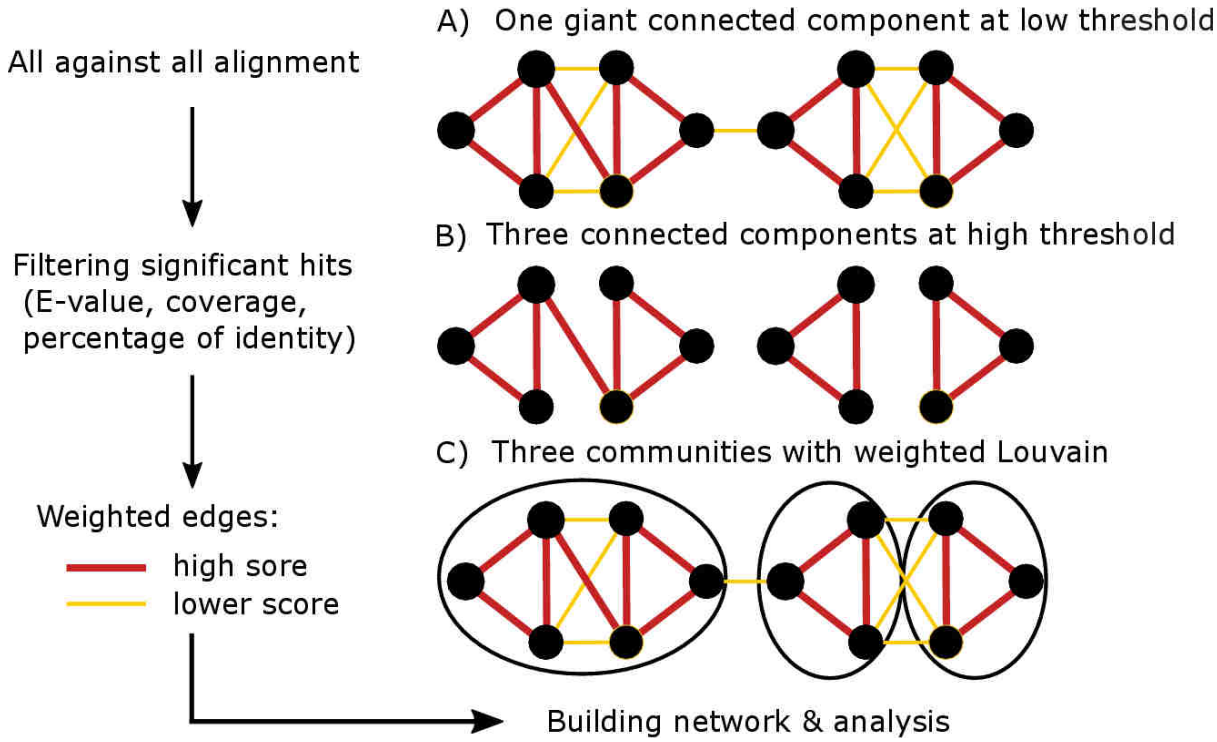



**Figure 3 : Louvain community detection in a sequence similarity network.** The network is assembled from the results of an all *versus* all alignment, as previously described**.** Edges can be weighted by E-value, percentage of identity or bitscore. For the purpose of simplification we consider strong or weak weights rather than actual value. A) A giant connected component at relaxed threshold. B) Three connected components at a more stringent threshold. C) Three communites with Louvain clustering algorithm, taking into account edge weights.

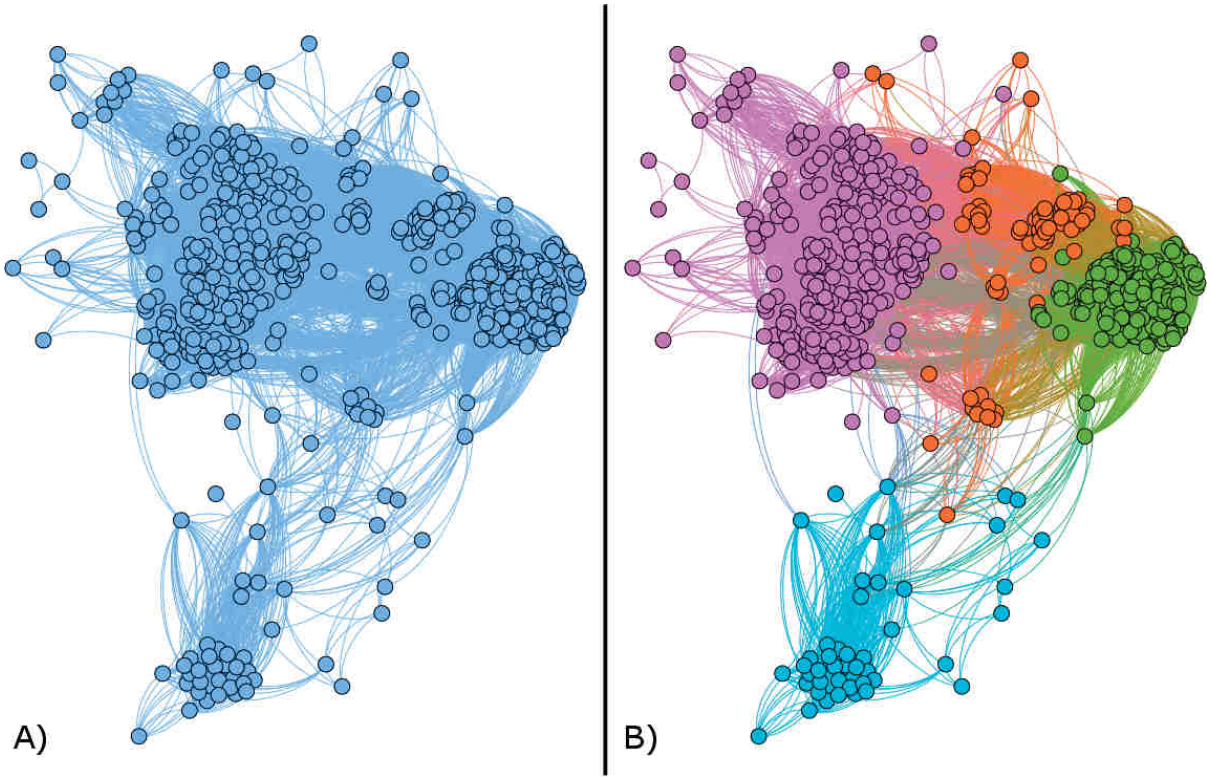



**Figure 4: Giant connected component before and after community detection**. A) A single giant connected component from a sequence similarity network. B) The same giant connected component after application of a community detection algorithm. Node colours correspond to the newly assigned communities.

A number of the above approaches have been used to compile additional databases of orthology that can act as useful reference datasets. OMA is a program that uses graph based algorithms and exact Smith-Waterman alignments to identify orthology between genes *(68–71)*. OMA is also available as a web browser *(72)* including a database of orthologues that, in 2015, included more than 2000 genomes and more than 7 millions proteins *(66)*. SILIX is a software package *(73)* that aims at building families of homologous sequences by using a transitive linkage algorithm and HOGENOM *(74)* is a database that contains families inferred by SILIX for 7 millions proteins.

In addition to clustering genes into families, valuable information can be extracted from the connected components using network metrics. Highly conserved sequences tend to form CCs where most of the nodes are connected to each other by edges, while sequences from more divergent families will tend to form more sparsely interconnected CCs. This information can be easily assessed for each component using the clustering coefficient. Conserved families will have a clustering coefficient close to 1, even for stringent thresholds. Identifying such conserved families can be useful to produce multiple sequence alignments (MSA) needed for phylogenetic reconstruction, but SSNs have also been demonstrated to unravel relationships between distant homologues by linking distantly related sequences together *(24, 29, 45)*. In a

SSN, two distant sequences A and C which do not share similarity according to BLAST can be linked together due to a sequence B which shows similarity to both A and C.

The idea of distant homology has been particularly illuminating regarding chimeric organisms such as eukaryotes who carry genes inherited from a bacterial ancestor and from an archeal ancestor *(29)*. A common way to analyse sequence similarity networks is to identify certain "paths" of interest, for example, the shortest possible paths between two nodes. This notion describes the path between two nodes in a connected component that minimises the sum of the edge weights. Alvarez-Ponce *et al.* used this approach to explore the topology of connected components in a SSN including the complete proteomes of 14 eukaryotes, 104 prokaryotes (including archaea and bacteria), 2,389 viruses and 1,044 plasmids. 899 CCs contained sequences from all three domains and of these 208 contained eukaryotic sequences that were not directly similar to one another, but only linked to one-another via a "eukaryote-archaea-bacteria-eukaryote" shortest path. These are putatively distant homologues in Eukaryotes that were present in both the Archaeal-host for the mitochondrial endosymbiont and the alphaproteobacterial endosymbiont, with both copies subsequently retained in Eukaryotes, and as such strong evidence for the chimeric origin of eukaryotes *(29)*. This adds to previous studies to demonstrate the utility of networks in the study of ancient evolutionary relationships including the origin of eukaryotes *(28)* or rooting the tree of life *(75)*. Simple path analysis for a network is possible using existing plugins within visualisation tools such as Cytoscape *(51)* and Gephi *(52)*.

## Exploiting SSNs to identify signatures of "tinkering" and gene fusion

When discussing identification of gene families we have focused on networks where edges are drawn between protein sequences that show a high enough similarity across their entire length, defined by a high mutual coverage threshold (e.g. 80%). Sequence similarity can also be partial, for example following gene remodelling or "tinkering" *(76)* producing new combinations of gene domains via gene fusion and fission events, or through the *de novo* sequence synthesis of gene extensions, adding to existing sequences. The term "Rosetta Stone sequence" was coined to define the formation of a new fusion protein in a species as the result of the fusion of two proteins that are found separate in another species, with authors originally predicting that these fusions could occur between proteins that physically interact in a common structural complex *(77)*. One of the earliest applications of sequence similarity searches to identify fusion proteins was an attempt to predict pairs of proteins that may physically interact in an organism based on whether they could be identified as a single

"composite" fusion protein in another organism *(41)*. Beyond predicting protein-protein interactions, this kind of gene remodelling and recycling of existing gene parts has the potential to contribute to the expansion of functional diversity in genomes, creating new and unique combinations of domains and functions *(48, 76, 78–82)*. Similarity search based screens have been implemented to identify composite genes and genome rearrangements in a range of prokaryotes *(83–85)*, eukaryotes *(78, 86–88)* and viruses *(89)*.

Early attempts to identify composite genes were based on the output of sequence similarity searches, but without formalising the results of search methods into a graph structure. The first attempt to formalise the problem of identifying "composite" genes in networks was the "Neighbourhood Correlation" approach, aiming to distinguish genuine multi-domain proteins sharing common ancestry (homologues) from novel multi-domain proteins that share domains due to insertions *(90)*. The later development of the FusedTriplets and MosaicFinder tools attempted to unify existing graph based methods for detection of "composite" gene detection *(47)*. FusedTriplets is a graph based implementation of the traditional gene centred method for composite gene identification, originally introduced by Enright *et al*. 1999, with additional cross-checks on the absence of similarity between the two component genes contributing to a composite gene based on varying thresholds *(47, 91)*. MosaicFinder is a gene family centred approach which will only identify highly conserved composite gene families that form "minimal clique separators" (Figure 5) *(47)*. This graph topology implies that MosaicFinder may fail to detect divergent (e.g. ancient or fast evolving) composite gene families which will tend to form "quasi-cliques" without perfect separation. CompositeSearch (forthcoming: available at http://www.evol-net.fr/index.php/en/downloads) is a new program designed to overcome this limitation by identifying both conserved and divergent composite gene families (Box 2).

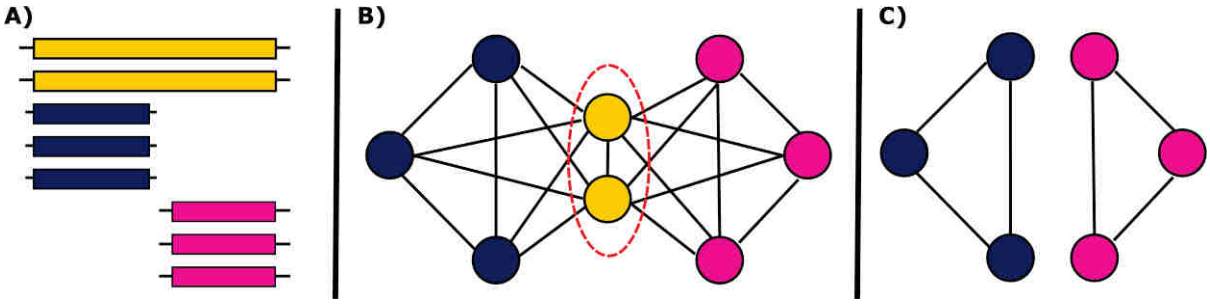



**Figure 5: Composite gene identification using "minimal clique separators".** A) A multiple sequence alignment of composite genes (yellow) with two components (blue and magneta). B) The sequence similarity network corresponding to the multiple sequence alignment. The composite genes (yellow) are a minimal clique separator for the network. Their removal (shown in C) decomposes the network to the two separate component families.

## Box 2: How to identify composite genes using composite search.

1) **BLAST search and filtering**: All versus all BLAST search filtered as described in "How to build your own sequence similarity network".
2) **CompositeSearch:** Composite search takes a filtered BLAST output and a list of genes within as the initial input. Two search algorithms are implemented: "fastcomposites" detects a list of potential composite genes. "Composites" additionally detects potential composite and component gene families. Additional options are included to filter the network based on a number of standard metrics (e.g. E-value, sequence similarity, mutual coverage) and set the maximum overlap allowed between different components aligned on the same potential composite gene. The definition of a maximum overlap allows adjustment for the tendency of BLAST to produce overhanging alignments *(91)*. The output includes a node, edge and information file including information on number of nodes, edges and family connectivity from family detection. Two outputs are included for composite gene detection, a "composites" file with detailed information on each predicted composite gene in fasta format, and "compositesinfo", summarising the data. Similarly two files provide detailed information on composite gene families and a summary of composite gene families.
3) **Filtering results**: By default compositeSearch outputs all possible composite gene families, alongside a number of different scores and measures designed to help to filter these results for more confident cases.

Recent studies have explored composite gene formation as a source of innovation by "tinkering" *(76)* during major evolutionary transitions. These can be especially interesting when exploring genome evolution following introgression, raising the possibility of formation of new composite genes using components with different evolutionary origins *(20, 48, 92)*. For example, the gain of a cyanobacterial endosymbiont at the origin of photosynthetic eukaryotes was accompanied by the transfer of whole cyanobacterial genes to its new host genome, with gene functions related to the role of the plastid *(93–95)*. Identification of composite genes related to the origin of photosynthetic eukaryotes unravelled novel symbiogenetic composite genes; unique fusions of genes encoded in the nucleus of photosynthetic eukaryotes that included components derived from the endosymbiont. As with whole genes transferred to the nucleus, several of these components had predicted functions related to the role of the plastid, including redox regulations and light response *(48)*.

## Exploiting SSNs for ecological studies

Ecological studies increasingly involve the assembly, analysis and comparison of large metagenome datasets. In addition to identification of functions and organisms associated with a particular environment, these studies enable the investigation of important hypotheses in microbial ecology at the level of organism or function, such as the often quoted hypothesis that "everything is everywhere, but the environment selects" from Bass Becking: the idea that microbial lineages are limitlessly dispersible in the environment, but the environmental conditions will select for certain lineages and control their distribution rather than any specific geographical separation *(21)*.

Networks are useful for these kinds of ecological studies because existing graph algorithms can be used to investigate the structure of the network. When investigating gene (or genome networks) it is possible to distinguish nodes by labelling them based on their properties, such as categories for taxonomic or environmental origins (Figure 6). A simple way to represent this visually is to colour nodes based on these properties in Cytoscape or Gephi. A formal way to explore the relationships between node properties is to use networks metrics such as conductance *(96)*, modularity *(64)* and assortativity coefficient (normalised modularity) *(97)*. Assortativity and conductance are different metrics that attempt to answer the same type of question: do nodes labelled as belonging to a particular category, such as environmental origin, tend to be connected with other nodes labelled as belonging to the same category? More precisely, conductance quantifies whether a given group of nodes shares more edges between themselves than with the rest of the nodes. A conductance of 0 implies that the graph is isolated, while a conductance close to 1 implies more connections are shared between that group of nodes and other nodes than are shared within the group of nodes. Assortativity is a measure of the preference for labelled nodes in a network to attach to other nodes with identical labels. Normalised assortativity values range between -1 and 1, where 0 indicates random distribution of labels within the network, 1 indicates that nodes with labels of the same type tend to be connected in the network, and -1 indicates that nodes with labels of different types tend to be connected in the network. A detailed description of the algorithms used in these calculations can be found in *(98)*.

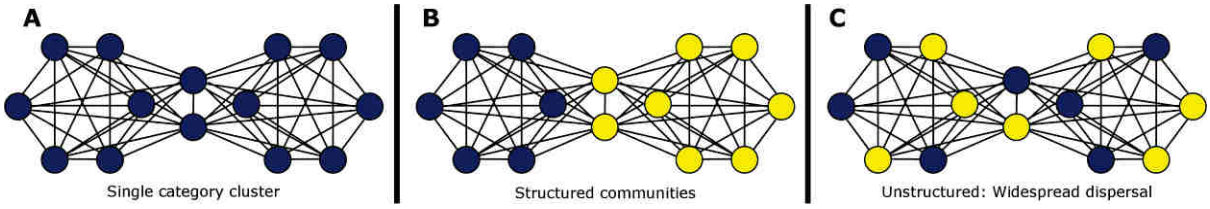



**Figure 6: Exploring distribution of annotations in sequence similarity networks.** In this example nodes within a single connected component are assigned two colours, blue and yellow, corresponding to their having a different categorical annotation (E.g. originating from a different environmental source). Using the example of environmental source, genes in cluster A would all have the same environmental source (blue), indicating an environment specific cluster of genes. Genes in cluster B are found in two different environmental sources (blue and yellow); however nodes of the same type are preferentially linked to each other in the network than to genes from different environmental sources. This would result in a positive assortativity coefficient approaching 1 for environment, and a low conductance score, suggesting a strong environmental community structure. Genes in cluster C are also found in two different environmental sources; however there is no clear pattern for the distribution of genes with regard to environment. This network would have an assortativity approaching 0 and a high conductance score.

## Assortativity as a tool to study geographical and habitat distributions of microbes and genes

Forster *et al.* used assortativity (among other network statistics, including the previously discussed shortest path analysis) to explore the geographical dispersion patterns of marine ciliates in a network generated from Ciliate SSU-rDNA sequences *(25)*. Sequences were clustered into two different levels of gene family – CCs, and Louvain communities (LCs) as described in the section. Sequences were assigned categorical labels based on their geographical point of origin (eight locations) or habitat of origin (three habitats) and assortativity was calculated. If sequences, and thus species, are broadly distributed across geographical categories then assortativity of SSU-rDNA sequences labelled with these geographical categories would be low because similar sequences would be found in different environments. Contrarily, if similar sequences tend to be from the same geographical category, indicative of endemism, then assortativity of sequence geographical origin will be high (Figure 6). The majority of CCs and LCs showed a positive assortativity for geographical origin, higher than expected by chance, indicative of geographical community structure as opposed to global dispersal of Ciliates. Similar approaches were used by Fondi *et al.* and applied to a collection of environmental metagenome samples to test the "everything is everywhere" hypothesis at the gene pool and functional level. Gene pools were more strongly associated with a particular ecological niche than with specific geographical location, supporting the idea that microbial genes are found everywhere but the environment selects for them *(26)*.

## Conductance in the comparison of lifestyles and evolutionary histories

Conductance is used to explore the clustering of pairs of different node categories in a graph connected component. In a study by Cheng *et al.* the proteomes of 84 prokaryote genomes

were categorised into four broad redox groups based on their lifestyle, methanogens, obligate anaerobes, facultative anaerobes and obligate aerobes *(27)*. For each CC in a pan-proteome sequence similarity network including all 84 genomes, the conductance was calculated for pairs of redox categories and compared to values obtained following random relabeling of the components. The distributions of conductance values for methanogens and for obligate anaerobes groups indicated that the sequences in these groups have features distinct from those in other groups; and that anaerobes and aerobes tend to be dissimilar and networks that are more isolated from one another than expected by chance.

An additional example of the use of conductance is in exploring the propensity of a gene family to lateral gene transfer. Within a network of archaeal and bacterial genes, CCs showing a low conductance for both archaeal and bacterial sequences indicate that the bacterial and archaeal genes within the corresponding families are structured in two separate and conserved groups (Figure 6). Structuring gene families in to two groups would indicate that there was little or no evidence for lateral gene transfer between archaea and bacteria within this particular gene family. This kind of gene family is rare, with only 86 gene families from 40,584 (0.2%) meeting this criteria *(24)*.

## SSNs in remote homologue identification: Shedding light on the microbial dark-matter

Up to 99% of microbial species are not cultivable and thus have not been studied in isolated culture. Analysis of high-throughput sequencing and metagenomics datasets has shed light on these uncultivable organisms, often referred to as the "microbial dark-matter" *(99)*, and in some cases enabled the reconstruction of draft genomes *(100–104)*. A considerable portion of most metagenome studies have predicted ORFs showing no detectable similarity to any known proteins, termed metaORFans *(105)*. These can represent 25% -85% of the total ORFs identified in metagenomes *(22)*. Identifying distant homologues of ORFans may help to predict their functions and begin to unravel the microbial dark-matter. Recent work by Lopez *et al.* in 2015 probed the microbial diversity of metagenome datasets from a range of environments including the human gut microbiome, identifying homologues of genes from 86 ancient gene families that are distributed across archaea, bacteria and eukaryotes. The majority of these gene families included environmental homologues that were highly divergent from any of their cultured homologues, and many branched deeply with the phylogenetic tree of life, highlighting our limited understanding of diverse elements of the microbial world and hinting at the existence of yet unknown major divisions of life *(24)* (Figure 7).

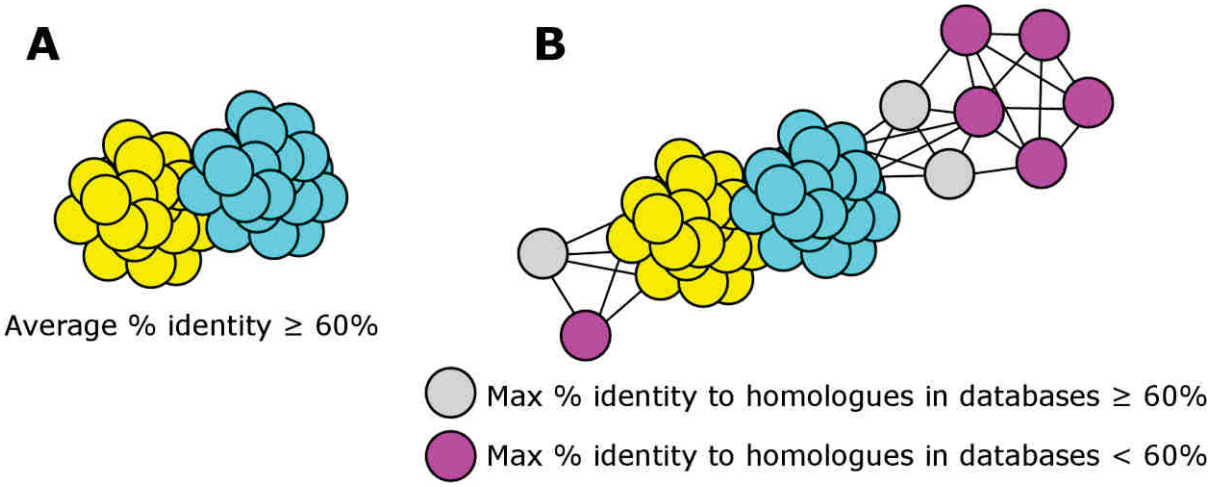



**Figure 7: Remote homologue detection to help characterise the microbial dark matter.** A) A hypothetical highly conserved cluster of genes from genomes present in sequence databases, where the average % of identity is high (≥60%). B) The same cluster after addition of divergent environmental sequences to the network. Environmental sequences in gray are more similar to those already identified from genome surveys (≥60% max identity) so are connected directly to the conserved gene cluster in the network. More divergent sequences in pink have <60% maximum identity to their homologues in the database. Many of these are only identified as linked to the sequences from the conserved database via intermediate gray nodes – the idea of "transitive homology".

## Exploiting SSNs to analyse classifications

Metagenomic and genomic data are providing scientists with a tantalizing amount of sequence data, casting the analysis of the extent of biodiversity as a major research theme in biology *(106–110)*. In theory, existing organismal and viral classifications are invaluable tools to structure and analyze this biodiversity. However, the way taxonomical classifications are constructed raises questions about their naturalness and their actual application scope *(38, 110–118)*, in particular regarding genetic diversity surveys. There are three major reasons for this. First, organismal and viral diversity is still largely undersampled, which means that existing classifications are incomplete *(109, 110)*. Therefore, taxonomically unassigned sequences cannot be readily used in class-based genetic diversity surveys, since this dark matter remains outside existing classes. Second, classifications are constructed using different features (i.e. for viruses, a mix of phylogenetic, morphological, and structural criteria, such as replication properties in cell culture, virion morphology, serology, nucleic acid sequence, host range, pathogenicity, epidemiology or epizootiology), therefore their classes do not necessarily offer immediate proxies for quantifying genetic diversity *per se*. Third, evolutionary processes responsible for both genetic and organismal diversity are diverse, and they operate at different tempos and modes in different lineages *(46, 113, 119–131)*. As a result, genetic diversity within classes and between classes can be heterogeneous, meaning that existing classifications may lack efficiency to discriminate, predict or compare taxa on genetic bases, potentially hampering diversity studies, a profound practical issue at a time where the analysis of metagenomic sequences is becoming a priority in biology.

Addressing these challenges is notably crucial for viral studies. Recently, the Executive Committee of the ICTV *(132)* proposed that network analyses methods that create similarity metrics based on the detection of homologous genes and their genetic divergence constitute a valuable strategy to assist classification of viruses. Consistently, basic network properties and metrics (Table 1) can quantify (i) whether genetic diversity is consistent within and between the classes of existing classifications, and (ii) describe what classes are the most homogeneous and distinctive in terms of genetic diversity. Three criteria can be used to estimate intra-class genetic heterogeneity (Figure 8 A, B, C). First, the average edge weights (measured as % of identity, PID) between pairs of sequences from genomes of the same class provides a trivial measure of intra-class genetic diversity. Second, the average proportion of Conserved Canonical Connections between sequences from the same connected component and from the same taxonomic class can be exploited (CCC, i.e. in each connected component of the SSN, the total number of edges connecting sequences of a given class i (intra-group edges, denoted $E_{ii}$) divided by the theoretical maximal number of possible edges between sequences of that class in the connected component $CCC(i) = 2*E_{ii}/ (N_i \times (N_i-1))$ where Ni is the number of sequences of class i present in the connected component.). CCC ranges between 0 and 1. Within a connected component, if all pairs of sequences from the same class are directly connected, CCC equals 1, since all these sequences are more conserved than a given %ID threshold (e.g. >20 % ID and > 50% mutual cover). By contrast, low CCC are observed when sequences from genomes from the same class lack cohesive evolution; for example, when some related sequences evolved so fast that they show less than the minimal similarity required to be directly connected to their homologs in the graph. Third, the genetic consistency of a class can be estimated by 1) identifying what cluster of sequences was present in the largest number of genomes of the class, and then 2) by quantifying the proportion (in %) of the class members harboring that most ubiquitous cluster (maxCore%). When maxCore% of a class is < 100%, it means that, for this dataset, there is no gene family shared by all members of that class (i.e. no core genes). The SSN structure can also serve to estimate the genetic distinctiveness of each class, i.e. whether sequences from a given class are more similar to one another than they are to sequences from other classes (Figure 8 D, E). Such sequences could be used as classificatory features to assign members to the class. In a SSN, this property translates to a low ratio of inter-class edges over intra-class edges and is measured by conductance (Figure 8 D). Likewise, the proportion of clusters comprised exclusively of sequences from one class, a diagnostic features of the class, provides an estimate of the class genetic distinctiveness. Genetically highly distinct classes have a high % of such exclusive clusters. Based on these network measures, inter-classes genetic

heterogeneity can simply be diagnosed by contrasting estimates of genetic consistency for all the above measures for each class. There is inter-class heterogeneity within a classification when the mean PID, mean CCC, maxCore%, DRC, and % of exclusive components differ between classes.

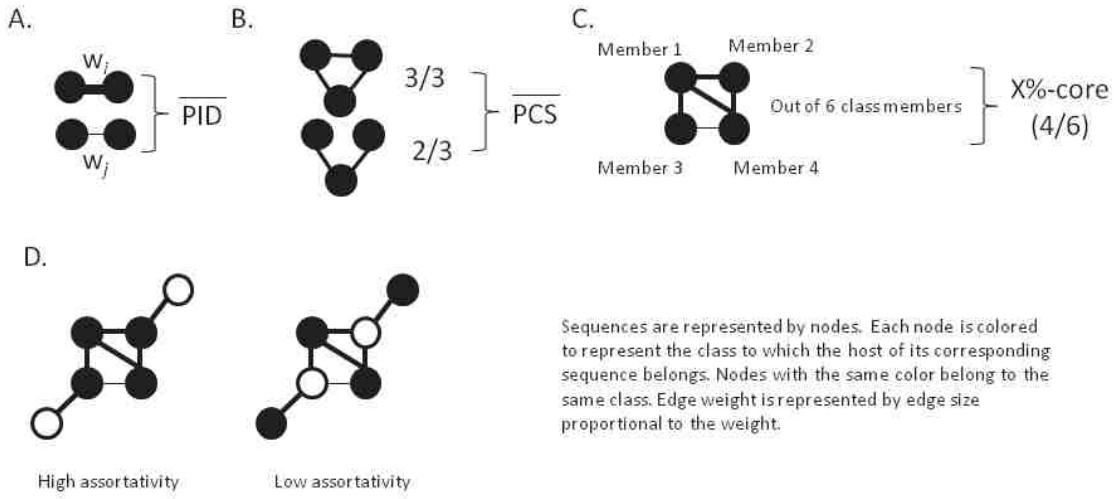



**Figure 8: Intra- and inter- classes heterogeneity measurements in weighted similarity networks.** Sequences are represented by nodes. Each node is colored to represent the taxonomic class to which its host belongs. Nodes with the same color belong to the same class. Edge weight is represented by edge size proportional to the weight. Subgraphs correspond to clusters of sequences. Direct neighbors have a greater similarity than the threshold set to allow such connections. PID, average edge weights (% identity) between two sequences from genomes of the same class; CCC, average proportion of genetic conservation between sequences from the same cluster and from the same taxonomic class; maxCore%, conductance and %-exclusive components correspond to the estimates used to assess genetic consistency of classes.

| **'Ideal' classes** | **Not ideal classes** |
|---|---|
| Low intra-class genetic diversity<br>(high average PID) | High intra-class genetic diversity<br>(low average PID) |
| High genetic cohesion<br>(high average CCC) | Low genetic cohesion<br>(low average CCC) |
| Core components<br>(high maxCore%) | No core components<br>(low maxCore%) |
| Obvious genetic distinctiveness<br>(high conductance difference with random groups) | Limited genetic distinctiveness<br>(conductance similar to random groups) |
| Exclusive pangenome<br>(high % of exclusive CC) | No exclusive pangenome<br>(low % of exclusive CC) |

**Table 1: Schematic properties of two extreme kinds of taxonomic classes with respect to their genetic diversity:** The 3 top properties inform about genetic diversity within classes (intra-class genetic diversity). The last 2 properties inform about the genetic distinctiveness (core and signature genes) of the classes. Inter-classes genetic heterogeneity identifies when genetic diversity of a class is not comparable with genetic diversity of another class in the classification. CCC, average proportion of genetic conservation between sequences from the same cluster and from the same taxonomic class; PID, average edge weights (% identity) between two sequences from genomes of the same class.

Application of this approach to a curated dataset of 3,058 classified viruses (all viral sequences available at the NCBI in November 2012, and sequences from Mimiviridae from URMITE laboratory, Marseille, France) classified according to 3 different schemes ((i) the *International Committee on Taxonomy of Viruses (133)*, (ii) the Baltimore classification that classified viruses according to the nature of their genome and their replicative strategy *(134)*, and (iii) a classification into five monophyletic classes of viruses and selfish genetic elements as demonstrated by *(135)*. The network was built by an all-against-all BLAST thresholds set

at an E-value of < 1e-5, a mutual coverage > 50%, and a mininum %ID ≥ 20%. This protocol produced 13,819 CCs, and their analysis with the described metrics indicated that viral classes are genetically heterogeneous (Table 2), and also unraveled some class-specific widespread (maxCore%) genes (available on https://figshare.com/s/0b7428ea3c1b3a03d657.) and signature genes (available on https://figshare.com/s/0b7428ea3c1b3a03d657) for these viruses. 'Megavirales' were within the most genetically consistent viral orders, providing an additional argument for the introduction of this order in the ICTV classification.

| | Mean CCC | Mean DRC | Exclusive CC (%) | maxCore % | Mean %ID | Mean CCC | Mean DRC | Exclusive CC (%) | maxCore % | Mean %ID |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | PID 20 | | | | | PID 50 | | |
| **Baltimore** | | | | | | | | | | |
| Min | 0.88 | 1.49 | 66.57 | 20.16 | 39.37 | 0.91 | 1.73 | 83.87 | 9.59 | 61.45 |
| Mean | 0.92 | 7.47 | 88.38 | 41.17 | 46.63 | 0.94 | 5.61 | 90.92 | 21.79 | 68.97 |
| Median | 0.93 | 7.14 | 92.90 | 46.15 | 44.65 | 0.94 | 5.03 | 89.91 | 18.00 | 68.58 |
| Max | 0.95 | 22.97 | 99.28 | 73.68 | 59.94 | 0.94 | 5.61 | 90.92 | 21.79 | 68.97 |
| **Phylogenetic** | | | | | | | | | | |
| Min | 0.89 | 1.34 | 64.86 | 28.43 | 43.88 | 0.87 | 2.26 | 82.20 | 12.62 | 60.78 |
| Mean | 0.92 | 3.99 | 88.19 | 44.36 | 49.73 | 0.92 | 4.11 | 89.75 | 26.55 | 70.56 |
| Median | 0.90 | 2.43 | 96.35 | 48.15 | 47.01 | 0.92 | 3.42 | 89.57 | 15.46 | 71.00 |
| Max | 0.95 | 7.67 | 98.75 | 55.05 | 56.87 | 0.97 | 6.67 | 99.00 | 45.83 | 80.79 |
| **ICTV-Orders** | | | | | | | | | | |
| Min | 0.81 | 1.50 | 76.39 | 36.72 | 35.05 | 0.80 | 2.69 | 83.27 | 14.75 | 56.96 |
| Mean | 0.92 | 7.58 | 89.53 | 76.94 | 45.60 | 0.93 | 55.40 | 92.27 | 45.25 | 70.10 |
| Median | 0.94 | 6.88 | 94.13 | 85.11 | 42.94 | 0.94 | 7.91 | 91.02 | 42.07 | 71.09 |
| Max | 0.96 | 13.61 | 96.36 | 100.00 | 62.04 | 0.98 | 394.96 | 100.00 | 80.00 | 80.79 |
| **ICTV-Families** | | | | | | | | | | |
| Min | 0.61 | 1.58 | 17.65 | 29.03 | 30.39 | 0.66 | 0.34 | 25.00 | 15.38 | 55.20 |
| Mean | 0.91 | 27.31 | 80.63 | 89.32 | 49.42 | 0.92 | 23.36 | 90.69 | 60.48 | 71.65 |
| Median | 0.93 | 7.70 | 87.50 | 93.92 | 47.77 | 0.95 | 7.02 | 98.60 | 57.22 | 69.84 |
| Max | 1.00 | 331.98 | 100.00 | 100.00 | 84.29 | 1.00 | 331.98 | 100.00 | 100.00 | 99.99 |

**Table 2: Summary of statistics for 4 types of classifications:** SSN were constructed at the stringency levels (%ID) indicated above the table. Only classes with more than 2 viruses and 5 sequences were retained for the analysis. Details can be found in tables SI 1-4. CCC, average proportion of genetic conservation between sequences from the same cluster and from the same taxonomic class; DRC, deviation to random conductance; Excl.. %( Excl. #), percentage of exclusive CC (corresponding number); PID, average eedge weights (i.e. % identity) between two sequences from genomes of the same class.

Consequently, network analyses show that virus classifications face a pragmatic issue: overall genetic distinctiveness allows relatively safe assignments of viral sequences to existing

classes, however genetic diversity of viral taxa of similar ranks differs among the tested classifications. Therefore, virus classifications (especially ICTV classification at the family level) should be used carefully to avoid inaccurate estimates in metagenomic diversity surveys. Classes with broader genetic diversity will tend to be more easily detected in the environment than classes with reduced genetic diversity, since the former will necessarily be associated with more OTUs than the latter. Some alpha- and beta- diversity analyses of environmental data, which rely on counts and on contrasts of the abundance of taxonomic classes in different samples, will also be biased.

This conclusion suggests that there is a need for novel classifications of viruses, informed from a genomic perspective, and suited for diversity surveys. As a possible step in this direction, the elaboration of a special classification of viruses that would maximize the amount of genetic consistency across classes could be valuable (in agreement with*(113)*). Such a systematics could provide more comparable proxies of viral genetic diversity in the genomic and post-genomic era. Recent attempts to classify viruses by *(38)* may effectively come closer to this result. A similar approach could be applied on different types of classified lineages, i.e. to identify what groups of bacteria, archaea or eukaryotes with comparable taxonomical ranks are the most genetically heterogeneous, and what ranks of their classification are the least genetically consistent.

# Genome networks

Genome networks are often called “gene sharing networks” as they are best suited for summarising what genes are shared between different genomes, highlighting routes of gene sharing. The ability to explore gene sharing between all genomes in a network in a simple graph can have useful properties for reflecting microbial social life, inherently inclusive of gene sharing both as a consequence of vertical inheritance and lateral gene transfer (LGT). Bacteriophage and plasmid genomes are typically highly mosaic in nature due to a high level of horizontal gene transfer, making it difficult to classify their genomes *(37, 136)*. Lima-Mendez *et al.* proposed the use of genome networks as a new classification method that tackles this problem of mosaicism by classifying viruses based on their genome’s content *(37)*. Constructing genome networks using subsets of genes from different functional categories of genes can also be useful in exploring what kinds of genes are being shared by different genomes.

In a genome network, each genome is represented by a node, and two nodes are connected by an edge when the two corresponding genomes share homologous genes or gene families (Figure 9). These gene families can be identified from SSNs (of as CCs of LCs) or by alternative methods. In genome networks, edges can be weighted by the number of genes or gene families shared between the genomes. In this way genome networks enable the study of microbial social life, quantitatively displaying the gene families shared between genomes both as a result of vertical transmission and lateral gene transfer.

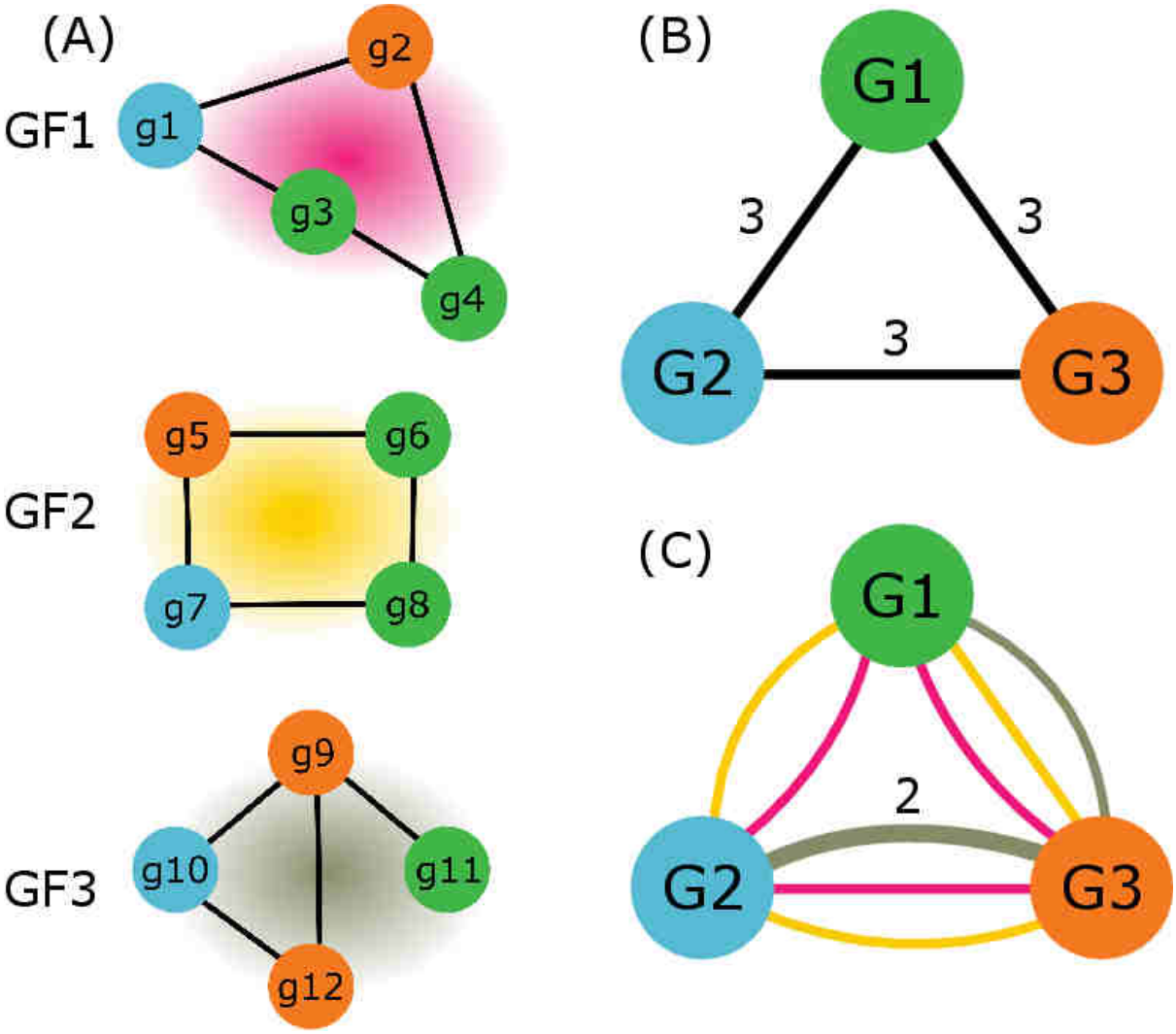



**Figure 9: Translating gene networks to genome networks.** A) Gene network for three gene families. Gene nodes are coloured based on their genome of origin. The background colour corresponds to the gene family colour in part C. B) The genome network corresponding to the gene network in A. Edges are weighted on the number of gene families shared by the genomes. C) Multiplex-genome network corresponding to the gene network in A. Genomes are connected by multiple edges with colours corresponding to different gene families. These edges are weighted based on the number of genes shared between two genomes for each family.

Genome networks are useful tools for exploring overall patterns of gene sharing between genomes. Recently Lord et al. developed BRIDES, a software package that specifically identifies different kinds of patterns in evolving genome networks after the addition of new genome nodes *(137)*. However, in genome networks the kind of gene families that are being shared is generally overlooked. To explore how functions are shared between different

genomes, genome networks can be built from genes using different subsets of functions (Figure 10) *(29)*. An alternative form of the genome network is the multiplex network. In this network nodes can be linked by edges of different types, for example, each edge representing a different gene family or different functional groups of gene families, thus retaining additional information compared to a simpler genome network (Figure 9) *(23)*. Multiplex networks can be useful for small scale analyses, however with large datasets they can rapidly become difficult to interpret and analyse. Importantly, multiplex networks are unimodal projections of bipartite graphs (discussed in the section "Bipartite Graphs") which can provide greater clarity and have a number of attractive properties for the analysis of larger datasets.

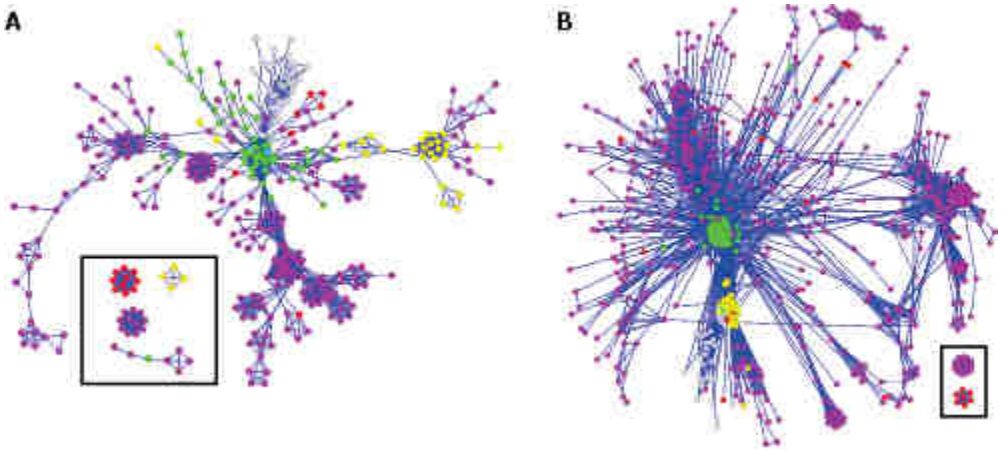



**Figure 10: Functional genome network reflecting the chimeric nature of eukaryotes.** These genome networks describing how genes in different functional categories are shared between bacteria (green), archaea (yellow), eukaryotes (grey), plasmids (purple) and viruses (red) from a published dataset *(29)*. In both cases a giant connected component is shown alongside examples of smaller connected components A) Genome network for COG category D: Cell division control. In this network, sequences of eukaryote origin (grey) cluster with bacterial sequences, reflecting their origin in the alpha-proteobacterial endosymbiont that would become the mitochondrion. B) Genome network for COG category K: Transcription machinery. In this network eukaryote sequence (gray) cluster with archaeal sequences; reflecting the origin of these genes in the archaeal-host for the eukaryotic endosymbiont.

## Classification of entities using genome networks

The possibility of summarising gene sharing between sets of entities with complex evolutionary histories means that genome networks can be useful for classifying organisms based on their gene content. Lima-Mendez *et al.* analysed bacteriophage genomes to generate two different phage genome networks that reflect their reticulate evolutionary history *(37)*. In the first genome network phage genomes (nodes) were connected by edges when shared significant similarity at the sequence level. This genome network was clustered using the previously discussed MCL algorithm *(138)*, identifying distinct groups of phages with sequence similarity. Following clustering, membership to a particular cluster was reassessed based on shared similarity with viruses in other clusters, reflecting their reticulate evolutionary history, allowing the generation of a matrix assigning a score describing the relative membership of any given viral genome to a particular classification group. In the

second approach, Lima-Mendez *et al*. generated a "module" based genome network, where edges are drawn between two phage genomes if they share a "module", in this case defined as a group of genes with similar phylogenetic profiles, enabling the exploration of what kinds of genes are shared between different groups of phages or are "signatures" for a particular group of phage genomes *(37)*.

## Exploring routes of gene sharing in genome networks

Two network metrics, also useful in the analysis of gene networks, can be used to attempt to identify "hubs" of gene sharing in the context of genome networks: node "degree" and "betweenness". Both metrics aim to determine the centrality of a node in a network. The degree of a node is simply the number of edges that it is connected to. The betweenness of a node is the frequency at which it is found in all the possible shortest paths between any two nodes in the network. Halary *et al.* used a combination of gene and genome networks based on DNA sequence similarity to explore gene sharing between prokaryotes and mobile genetic elements *(30)*. Plasmids were identified as hubs of gene sharing within this pool of genomes, suggesting that they are key vectors for genetic exchange between cellular genome and a potential DNA reservoir shared by genomes. Phages were more peripheral in the network, and mostly linked prokaryotes from the same lineage. Thus, genome provided insights on the evolutionary processes that shape the gene content of prokaryote genomes

The importance of plasmids in genetic worlds was further highlighted by exploring plasmid genome networks without inclusion of prokaryote genomes *(14, 36)*. Connecting 2,343 plasmid genomes based on shared gene content in a single graph demonstrated that plasmids tended to cluster based on the phylogenetic class of their corresponding host prokaryote rather than habitat, but that more mobile plasmids tended to be more "central" in the graph, indicating that these were hubs of gene sharing. Specifically, routes of gene sharing for gene families including antibiotic resistance markers were identified between actinobacterial plasmids and gammaproteobacterial plasmids, suggesting that actinobacteria may act as a reservoir for antibiotic resistance genes for gammaproteobacteria *(14)*.

The finding that plasmids are hubs of gene sharing for prokaryote genomes was supported by analysis of gene sharing in a Proteobacterial phylogenomic network including 329 Proteobacterial genomes *(32)*. A phylogenomic network is an extension of a genome network in which genome nodes are linked by edges if they share genes, however the genome nodes themselves are mapped to the base phylogeny for the set of genomes analysed *(34)*. This

study identified extensive evidence for lateral gene transfer among Proteobacteria, with at least 1 LGT event inferred in 75% of all gene families. Of these putative LGTs, more were related to plasmid related genes than phage related genes, suggesting plasmid conjugation was a more frequent source of gene transfer *(32)*. Directed graphs exploring directionality of LGT events between 657 prokaryote genomes allowed the polarisation of 32,028 putative LGT events finding that frequency of recent events correlates with genome sequence similarity, and most LGTs occurring between donor-recipient pairs with <5% difference in GC content, suggesting that there are some barriers to lateral gene transfer between prokaryotes, but that these are not insurmountable *(31)*. Later reconstruction of transduction events linking phage donors and recipients in a phylogenomic network demonstrated that LGT by transduction was generally highest in similar genomes and between clusters of closely related species, but that this constraint was occasionally broken, resulting in LGTs over long evolutionary distances *(35)*.

# Bipartite graphs

Bipartite graphs are excellent at summarising what genes are shared between sets of genomes, and as such are ideal for comparative genomics, including for the comparison of genomes reconstructed in metagenomic analyses. The potential to extend this approach to multi-level graphs, adding additional layers of information such as the environment in ecological studies, could provide a powerful summary of gene sharing in relatively complex datasets.

A multi-level network is a network in which edges exclusively connect nodes of different types, i.e. representing different levels of biological organisation. Thus, a bipartite graph is a graph with two types of nodes (top and bottom nodes), where edges exclusively connect nodes of different types (Figure 11) *(139)*. The types of nodes used can vary widely depending on the biological question, from linking diseases (top nodes) to their associated genes (bottom nodes) in order to explore the association between related disease phenotypes and their genetic causes *(140, 141)*, to exploring the concept of flavour pairings in food based on a graph of ingredients (top nodes) and the flavour compounds they contain (bottom nodes) *(142)*. For applications in molecular biology, a typical example of a bipartite graph may describe the relationships between genomes (top nodes) and gene families (bottom nodes), with edges between nodes indicating that a genome encodes at least one member of the corresponding gene family (Figure 11) *(23, 33, 38, 143)*. This kind of genome to gene family graph is particularly suited for the comparative analysis of the gene content of genomes in

microbial communities and for exploring patterns of gene sharing, for example between distantly related cellular genomes *(33)*, or between cellular genomes and their mobile genetic elements (Corel *et al.* forthcoming). It is possible to represent all genes shared between a given set of genomes, as a result of both vertical inheritance and horizontal gene transfer, in a single bipartite graph *(23)*. This feature was utilised by Iranzo *et al.* to explore gene sharing amongst the entire dsDNA virosphere, a group of entities typified by high rates of molecular evolution and gene transfer *(38)*.

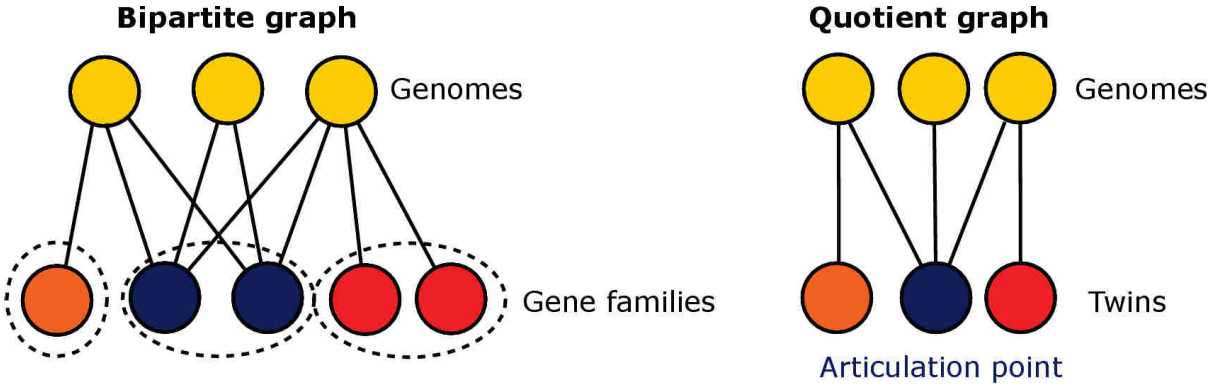



**Figure 11: A bipartite graph and its reduction to a quotient graph:** A) An example of a bipartite graph displaying how five gene families are shared between three genomes. B) A reduced form of the bipartite graph in which gene families are combined to "twin" nodes if they share identical taxonomic distributions. A single "articulation point" connects all three genomes.

Two topological features of bipartite graphs can be used to facilitate studies of gene sharing by an exact decomposition of the bipartite graph: twins and articulation points *(23, 144)*. A bipartite graph can be reduced to a quotient graph, a reduced variant of the bipartite graph where nodes from the bipartite graph have been combined based on sharing similar properties without the loss of information. For twin nodes ("twins"), this reduction is based on the combination of bottom nodes that have identical neighbours into a single "twin" supernode in the quotient graph (Figure 11). This is as a useful way of reducing the size of large graphs without losing information, but twin nodes also have useful properties for graph interpretation. The genomes supporting a twin node (its neighbours) define a club of genomes that share genes, through common ancestry and/or horizontal transfer. For example, in any given dataset any "core" set of gene families encoded by all species in the analysis will be represented by a single twin node. The gene families combined in twin supernodes can be viewed as gene families that are likely to be transmitted together *(23)*. An articulation point is a node that, when removed, will split the graph into two or more connected components. Within a gene family- genome bipartite graph, articulation points are expected to help to identify "public genetic goods", gene families that are shared by distantly related entities that may confer an advantage independent of genealogy *(23, 145)*, as well as selfish genetic elements such as transposases that also spread across multiple genomes. Two recently

developed tools, AcCNET *(143)* and MultiTwin (forthcoming), have simplified the process of constructing and analysing multi-level graphs without the need for custom programming (Box 3).

## Box 3: Considerations for the construction and analysis of bipartite graphs using AcCNET and MultiTwin

The default workflow for both ACcNet and MultTwin takes protein sequence data in fasta format as input, and generates a bipartite graph alongside a number of graph summary statistics and outputs for visualisation in standard tools (such as Gephi and Cytoscape), but with a number of important differences, including:

- **Graph levels:** Both AcCNET and MultiTwin can generate a bipartite graph using their default workflow; however MultiTwin can also be used to explore additional graph levels by adding additional node types (e.g. a tripartite graph). Multi-partite graphs mean that gene family level annotations can be associated with additional levels of biological information beyond which genomes they are found in. This may be particularly useful for the comparison of samples in metagenomics studies or time course experiments, allowing gene families to be associated directly with features such as environmental origin or time point.
- **Gene family identification**: AcCNET uses kClust *(146)* to assemble gene families, a kmer based method for rapid assembly of clusters of homologous proteins from sequence data. By default, MultiTwin identifies gene families using an all versus all BLAST search, followed by identification of connected components at a given threshold, as previously discussed for gene family detection from SSNs. MultiTwin can also be used in a modular way allowing for additional customisation, including the use of any custom gene family input in the form of a "community file": a tab delimited file linking every gene/protein id to a community identifier, with gene families defined using a clustering method of choice.
- **Edge weighting:** In AcCNET the edge weight is proportional to the inverse of the phylogenetic distance between proteins in a cluster from a given genome to other proteins within the same cluster. In MultiTwin the default edge weight is based on the number of genes present in a gene family from any given genome.
- **Graph compression:** While both methods can be used to identify "twin" nodes, only MultiTwin generates a quotient graph from these twin nodes and identifies articulation points.

AcCNET is available at: https://sourceforge.net/projects/accnet

MultiTwin is available at: http://www.evol-net.fr/index.php/en/downloads

### Using bipartite graphs to explore patterns of gene sharing between diverse entities

The simplest application of a bipartite graph is the summary of all genes shared between genomes in a single parsable graph, and this feature has been used to explore gene sharing in the dsDNA virome *(38)*, a range of *Escherichia coli* genomes to investigate *the E. coli* pangenome *(143)*, and between a broad range of prokaryotes that include newly discovered organisms *(33)*. In their analysis of prokaryote genomes, Jaffe *et al.* used the notion of "twins" to explore patterns of gene sharing between prokaryotes, including Archaea, and the recently discovered ultrasmall "Candidate Phyla Radiation" and TM6 bacteria with extremely unusual and reduced genomes. They found evidence for lateral gene transfer between ultrasmall bacteria and other prokaryotes, consistent with the suggestion they may be symbionts *(33)*. In their exploration of the dsDNA virome, Iranzo *et al.,* used graph module detection, algorithms designed to identify groups of densely connected nodes in a graph, to identify sets of densely connected viral genes and genomes that included a range of viruses with broad host ranges, as well as 14 hallmark viral genes that account for most of the gene sharing between all different viral modules *(38)*.

## Conclusions

This chapter has offered a brief introduction to the generation of commonly used sequence similarity networks in molecular biology, and a guide to how they can be generated and applied to a broad range of studies (Figure 12). Networks provide a highly scalable framework for the study of an increasingly broad range of applications in molecular biology and evolution and have already contributed to a number of important discoveries in the field. These include exploring patterns of introgression and horizontal transfer across all domains of life and mobile elements, the origin of eukaryotes, the contribution of new genes including novel fusion genes to major evolutionary transitions, shedding light on the "microbial dark matter" in metagenome sequencing datasets, and in testing ecological hypotheses about organism and gene distribution and environmental selection. New methods and tools for network analysis are becoming increasingly user-friendly and accessible to biologists without extensive programming experience, and enabling network analysis to become a more common parts of a biologists toolkit in the analysis of molecular sequence data.

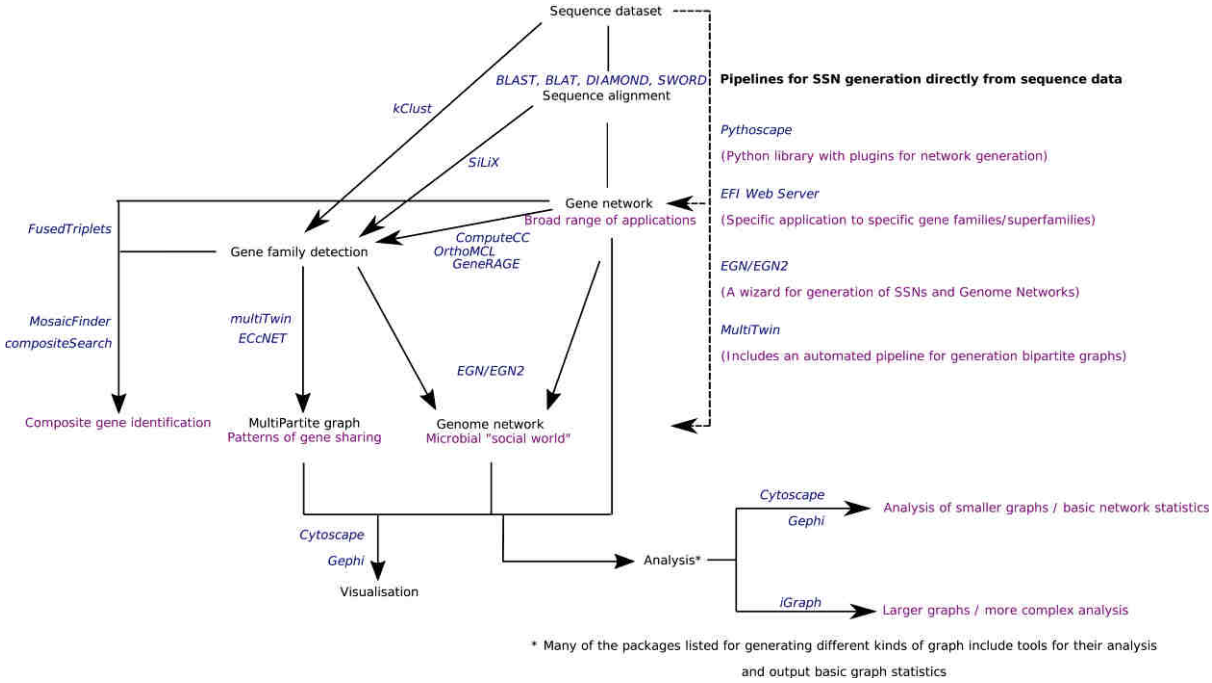



**Figure 12: A workflow highlighting some of the available routes for generation and analysis of SSNs, genome networks and bipartite graphs.** This workflow highlights just some of the many tools and routes for network construction and analysis.

# Exercises

The exercises use EGN*(49)* and require access to a local installation of BLAST+ *(55)* and Perl. The fasta sequence file "example.faa" provided with EGN and includes a dataset protein sequences from Archaea, Bacteria, Eukaryotes and mobile genetic elements, available at http://www.evol-net.fr/index.php/fr/downloads:

1) Perform a manual all *vs* all BLAST using search for a given protein sequence file from the unix terminal (requires local installation of BLAST). The output can be filtered to generate a network:
   a) Make the blast database using the "*makeblastdb*".
      i) Command: "*makeblastdb -dbtype prot -in example.faa –out example*"
   b) Performing the BLAST search using "blastp", remembering to output data in a tabular format for easy processing.
      i) Command: "*blastp -query example.faa -db example -evalue 1e-5 -seg yes -soft_masking true - max_target_seqs 5000 -outfmt "6 qseqid sseqid evalue pident bitscore qstart qend qlen sstart send slen" -out protein.blastpout*"
2) Generate a SSN using EGN from example.faa (requires local installation of BLAST and download of EGN from http://www.evol-net.fr/index.php/fr/downloads):
   a) Run EGN from the terminal using "perl egn.1.0.plus.pl" from the programs home directory.

b) Follow onscreen prompts sequentially to generate an alignment, filter the output, and generate a gene network with outputs compatible with both Cytoscape and Gephii.

3) Visualise SSN networks:

a) In Cytoscape: Import files named "*cc.*.txt*" as a network to visualise that set of connected components.

i) To associate nodes with their annotations, import "*cc*.atr*" as a table.

b) In Gephi: Open "*cc*.gxf*" files to import individual connected components from the network in to gephi. Use the "layout" menu to explore different kinds of layouts for the network.

# Glossary of terms

- Articulation point: A node in a graph whose removal increases the number of connected components of the resulting graph.
- Adjacency matrix: A numerical square matrix with row and columns labelled by network nodes, with 1 or 0 in the matrix indicating whether are they connected by an edge in the network.
- Assortativity: A measure of the preference for labelled nodes in a network to attach to other nodes with identical labels. This is the Pearsons correlation' coefficient of the degrees of pairs of linked nodes. See main text for full equation.
- Betweenness: A centrality measure for a node in a graph. Precisely, this is the proportion of shortest paths between all possible pairs of nodes in a connected component that pass through this node. A betweenness close to 1 is indicative of a highly central gene, whereas close to 0 is more peripheral.
- Bipartite graph: A graph with two types of nodes (top and bottom nodes), in which an edge only connects nodes of different types.
- Club of genomes: A group of entities that replicated separately but exploit common genetic material that may not trace back to the last common ancestor.
- Communities (also called modules): In graph terminology, a community is defined as a group of nodes that are more connected between themselves than to nodes in the rest of the graph.
- Composite gene: A gene that is made up of at least two component parts.
- Component genes: Genetic fragments sharing partial similarity to a composite gene.
- Conductance: A measure that quantifies whether a given group of nodes shares more edges between them than with the rest of the nodes. A conductance of zero implies that the graph is isolated, while a conductance close to one implies more connections are shared between that group of nodes and other nodes than are shared within the group of nodes.
- Connected component: A subgraph in which any pair of nodes is connected, either directly or indirectly, and that is not connected to the rest of the graph.
- Degree: The number of edges connected to a given node.
- Endosymbiont: An organism that lives inside another to the mutual benefit of both organisms.
- Edge: The link between two nodes in a network.
- E-value: The number of alignments in a sequence similarity search expected to be seen by chance searching against a database of a certain size.
- Introgression: Descent process through which the genetic material of an entity propagates into different host structures and is replicated within these new host structures.
- Lateral gene transfer (LGT; Or Horizontal gene transfer, HGT): Movement of genetic material between entities not mediated by vertical descent.
- Louvain community: A graph community identified using the Louvain algorithm.

- Network (or graph): A system of objects (nodes), some pairs of which are linked (edge).
- Multipartite graph: Similar to a bipartite graph, but with any number of types of nodes exclusively connected to nodes of other types
- Multiplex graph: A graph where nodes can be connected by edges of different types
- Modularity: The fraction of edges falling within given groups (e.g. communities or functional categories) in a network, minus the fraction of edges that would be expected with a random distribution of edges.
- Phylogenomic network: A phylogenetic network constructed from whole genome sequences where genomes are connected based on pairwise relationships including vertical and lateral gene transfer (LGT) events.
- Public genetic goods: Common genetic materials shared by clubs of phylogenetically distinct genomes
- Quotient graph: A simplified graph whose nodes represent disjoint subsets of nodes of the original graph; an edge in this new graph connects two such new nodes whenever an edge in the original graph connects at least one element of a new node with at least one from the other.
- Supporting genomes: The common set of neighbours that support a “twin” class in a multipartite graph.
- Twins: Nodes in a multipartite graph that share identical sets of neighbours.

# Bibliography


1. Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet 5:123–135. doi: 10.1038/nrg1271

2. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–30. doi: 10.1038/nature04546

3. Williams TA, Foster PG, Cox CJ, Embley TM (2013) An archaeal origin of eukaryotes supports only two primary domains of life. Nature 504:231–6. doi: 10.1038/nature12779

4. Alsmark C, Foster PG, Sicheritz-Ponten T, et al (2013) Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. Genome Biol 14:R19. doi: 10.1186/gb-2013-14-2-r19

5. Hirt RP, Alsmark C, Embley TM (2015) Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. Curr Opin Microbiol 23:155–162. doi: 10.1016/j.mib.2014.11.018

6. Nowack ECM, Price DC, Bhattacharya D, et al (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. Proc Natl Acad Sci U S A 113:12214–12219. doi: 10.1073/pnas.1608016113

7. McCoy JM, Mi S, Lee X, et al (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403:785–789. doi: 10.1038/35001608

8. Kondo N, Nikoh N, Ijichi N, et al (2002) Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. Proc Natl Acad Sci 99:14280–14285. doi: 10.1073/pnas.222228199

9. McInerney JO (2017) Horizontal gene transfer is less frequent in eukaryotes than prokaryotes but can be important (retrospective on DOI 10.1002/bies.201300095). BioEssays 39:1700002. doi: 10.1002/bies.201700002

10. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–38.

11. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A 104:870–5. doi: 10.1073/pnas.0606318104

12. Hooper SD, Mavromatis K, Kyrpides NC (2009) Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biol 10:R45. doi: 10.1186/gb-2009-10-4-r45

13. Nelson-Sathi S, Sousa FL, Roettger M, et al (2014) Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80. doi: 10.1038/nature13805

14. Tamminen M, Virta M, Fani R, Fondi M (2012) Large-scale analysis of plasmid relationships through gene-sharing networks. Mol Biol Evol 29:1225–40. doi: 10.1093/molbev/msr292

15. Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet 25:107–110. doi: 10.1016/j.tig.2008.12.004

16. Vos M, Hesselman MC, te Beek TA, et al (2015) Rates of Lateral Gene Transfer in Prokaryotes: High but Why? Trends Microbiol 23:598–605. doi: 10.1016/j.tim.2015.07.006

17. McInerney JO, McNally A, O'Connell MJ (2017) Why prokaryotes have pangenomes. Nat Microbiol 2:17040. doi: 10.1038/nmicrobiol.2017.40

18. Niehus R, Mitri S, Fletcher AG, Foster KR (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. Nat Commun 6:8924. doi: 10.1038/ncomms9924

19. Hotopp JCD, Clark ME, Oliveira DCSG, et al (2007) Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. Science (80- ) 317:1753–1756. doi: 10.1126/science.1142490

20. Wolf YI, Kondrashov AS, Koonin E V (2000) Interkingdom gene fusions. Genome Biol 1:research0013.1. doi: 10.1186/gb-2000-1-6-research0013

21. Becking LB (1934) Geobiologie of inleiding tot de milieukunde. W.P. Van Stockum & Zoon, Den Haag

22. Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC (2015) Remote homology and the functions of metagenomic dark matter. Front Genet 6:234. doi: 10.3389/fgene.2015.00234

23. Corel E, Lopez P, Méheust R, et al (2016) Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol 24:224–237. doi: 10.1016/j.tim.2015.12.003

24. Lopez P, Halary S, Bapteste E (2015) Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol Direct 10:64. doi: 10.1186/s13062-015-0092-3

25. Forster D, Bittner L, Karkar S, et al (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol 13:16. doi: 10.1186/s12915-015-0125-5

26. Fondi M, Karkman A, Tamminen M V, et al (2016) "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis.

Genome Biol Evol 8:1388–400. doi: 10.1093/gbe/evw077

27. Cheng S, Karkar S, Bapteste E, et al (2014) Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front Ecol Evol 2:72. doi: 10.3389/fevo.2014.00072

28. Thiergart T, Landan G, Schenk M, et al (2012) An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin. Genome Biol Evol 4:466–485. doi: 10.1093/gbe/evs018

29. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A 110:E1594-603. doi: 10.1073/pnas.1211371110

30. Halary S, Leigh JW, Cheaib B, et al (2010) Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A 107:127–32. doi: 10.1073/pnas.0908978107

31. Popa O, Hazkani-Covo E, Landan G, et al (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21:599–609. doi: 10.1101/gr.115592.110

32. Kloesges T, Popa O, Martin W, Dagan T (2011) Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. Mol Biol Evol 28:1057–1074. doi: 10.1093/molbev/msq297

33. Jaffe AL, Corel E, Sylvestre Pathmanathan J, et al (2016) Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. Environ Microbiol 18:5072–5081. doi: 10.1111/1462-2920.13477

34. Dagan T (2011) Phylogenomic networks. Trends Microbiol 19:483–491. doi: 10.1016/j.tim.2011.07.001

35. Popa O, Landan G, Dagan T (2017) Phylogenomic networks reveal limited phylogenetic range of lateral gene transfer by transduction. ISME J 11:543–554. doi: 10.1038/ismej.2016.116

36. Fondi M, Fani R (2010) The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. Environ Microbiol 12:3228–3242. doi: 10.1111/j.1462-2920.2010.02295.x

37. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Reticulate representation of evolutionary and functional relationships between phage genomes. Mol Biol Evol 25:762–777. doi: 10.1093/molbev/msn023

38. Iranzo J, Krupovic M, Koonin E V. (2016) The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. MBio 7:e00978-16. doi: 10.1128/mBio.00978-16

39. Tatusov RL, Koonin E V, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–7.

40. Tatusov RL, Galperin MY, Natale DA, Koonin E V (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36. doi: 10.1093/nar/28.1.33

41. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90. doi: 10.1038/47056

42. Pasternak G, Hochhaus a, Schultheis B, Hehlmann R (1998) Chronic myelogenous leukemia: molecular and cellular aspects. J Cancer Res Clin Oncol 124:643–60.

43. Watanabe H, Otsuka J (1995) A comprehensive representation of extensive similarity linkage between large numbers of proteins. Bioinformatics 11:159–166. doi: 10.1093/bioinformatics/11.2.159

44. Park J, Teichmann SA, Hubbard T, Chothia C (1997) Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273:349–54. doi: 10.1006/jmbi.1997.1288

45. Bolten E, Schliep A, Schneckener S, et al (2001) Clustering protein sequences--structure prediction by transitive homology. Bioinformatics 17:935–941. doi: 10.1093/bioinformatics/17.10.935

46. Bapteste E, Lopez P, Bouchard F, et al (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci 109:18266–18272. doi: 10.1073/pnas.1206541109

47. Jachiet P-A, Pogorelcnik R, Berry A, et al (2013) MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29:837–844. doi: 10.1093/bioinformatics/btt049

48. Méheust R, Zelzion E, Bhattacharya D, et al (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A 113:3579–84. doi: 10.1073/pnas.1517551113

49. Halary S, McInerney JO, Lopez P, Bapteste E (2013) EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol 13:146. doi: 10.1186/1471-2148-13-146

50. Martin AJM, Walsh I, Domenico T Di, et al (2013) PANADA: Protein Association Network Annotation, Determination and Analysis. PLoS One 8:e78383. doi: 10.1371/journal.pone.0078383

51. Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504. doi: 10.1101/gr.1239303

52. Bastian M, Heymann S, Jacomy M (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. Third Int AAAI Conf Weblogs Soc Media 361–362. doi: 10.1136/qshc.2004.010033

53. Csárdi G, Nepusz T The igraph software package for complex network research.

54. Hagberg AA, Schult DA, Swart PJ (2008) Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds) Proc. 7th Python Sci. Conf. Pasadena, CA USA, pp 11–15

55. Camacho C, Coulouris G, Avagyan V, et al (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi: 10.1186/1471-2105-10-421

56. Altschul SF, Gish W, Miller W, et al (1990) Altschul et al.. 1990. Basic Local Alignment Search Tool.pdf. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

57. Kent WJ (2002) BLAT--the BLAST-like alignment tool. Genome Res 12:656–64. doi: 10.1101/gr.229202. Article published online before March 2002

58. Vaser R, Pavlović D, Šikić M (2016) SWORD—a highly efficient protein database search. Bioinformatics 32:i680–i684. doi: 10.1093/bioinformatics/btw445

59. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. doi: 10.1093/bioinformatics/btq461

60. Buchfink B, Xie C, Huson DH (2014) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. doi: 10.1038/nmeth.3176

61. Dayhoff MO (1976) The origin and evolution of protein superfamilies. Fed Proc 35:2132–8.

62. Heger A, Holm L (2000) Towards a covering set of protein family profiles. Prog Biophys Mol Biol 73:321–337. doi: 10.1016/S0079-6107(00)00013-4

63. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci U S A 99:7821–6. doi: 10.1073/pnas.122653799

64. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp. doi: 10.1088/1742-5468/2008/10/P10008

65. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–84. doi: 10.1093/nar/30.7.1575

66. Altenhoff AM, kunca N, Glover N, et al (2015) The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res 43:D240–D249. doi: 10.1093/nar/gku1158

67. Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. doi: 10.1101/gr.1224503

68. Dessimoz C, Cannarozzi G, Gil M, et al (2005) OMA, A Comprehensive, Automated Project for the Identification of Orthologs from Complete Genome Data: Introduction and First Achievements. Springer, Berlin, Heidelberg, pp 61–72

69. Dessimoz C, Boeckmann B, Roth ACJ, Gonnet GH (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res 34:3309–3316. doi: 10.1093/nar/gkl433

70. Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinformatics 9:518. doi: 10.1186/1471-2105-9-518

71. Altenhoff AM, Gil M, Gonnet GH, et al (2013) Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. PLoS One 8:e53786. doi: 10.1371/journal.pone.0053786

72. Schneider A, Dessimoz C, Gonnet GH (2007) OMA Browser Exploring orthologous relations across 352 complete genomes. Bioinformatics 23:2180–2182. doi: 10.1093/bioinformatics/btm295

73. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116. doi: 10.1186/1471-2105-12-116

74. Penel S, Arigon A-M, Dufayard J-F, et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10:S3. doi: 10.1186/1471-2105-10-S6-S3

75. Dagan T, Roettger M, Bryant D, Martin W (2010) Genome Networks Root the Tree of Life between Prokaryotic Domains. Genome Biol Evol 2:379–392. doi: 10.1093/gbe/evq025

76. Jacob F (1977) Evolution and tinkering. Science (80-. ). 196:

77. Marcotte EM, Pellegrini M, Ng HL, et al (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285:751–3.

78. Kawai H, Kanegae T, Christensen S, et al (2003) Responses of ferns to red light are mediated by an unconventional photoreceptor. Nature 421:287–290. doi: 10.1038/nature01310

79. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313–26. doi: 10.1101/gr.101386.109

80. Marsh JA, Teichmann SA (2010) How do proteins gain new domains? Genome Biol 11:126. doi: 10.1186/gb-2010-

11-7-126

81. Promponas VJ, Ouzounis CA, Iliopoulos I (2014) Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey. Brief Bioinform 15:443–454. doi: 10.1093/bib/bbs072

82. McLysaght A, Guerzoni D (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc B Biol Sci 370:20140332. doi: 10.1098/rstb.2014.0332

83. Enright AJ, Ouzounis CA (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16:451–457. doi: 10.1093/bioinformatics/16.5.451

84. Snel B, Bork P, Huynen M (2000) Genome evolution. Gene fusion versus gene fission. Trends Genet 16:9–11.

85. Enright AJ, Ouzounis CA (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2:RESEARCH0034.

86. Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–31.

87. Nakamura Y, Itoh T, Martin W (2007) Rate and polarity of gene fusion and fission in Oryza sativa and Arabidopsis thaliana. Mol Biol Evol 24:110–121. doi: 10.1093/molbev/msl138

88. Ekman D, Björklund ÅK, Elofsson A (2007) Quantification of the Elevated Rate of Domain Rearrangements in Metazoa. J Mol Biol 372:1337–1348. doi: 10.1016/j.jmb.2007.06.022

89. Jachiet P-AA, Colson P, Lopez P, Bapteste E (2014) Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. Genome Biol Evol 6:2195–2205. doi: 10.1093/gbe/evu168

90. Song N, Joseph JM, Davis GB, et al (2008) Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins. PLoS Comput Biol 4:e1000063. doi: 10.1371/journal.pcbi.1000063

91. Yanai I, Derti A, DeLisi C (2001) Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. Proc Natl Acad Sci. doi: 10.1073/pnas.141236298

92. Dorrell RG, Gile G, McCallum G, et al (2017) Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. Elife. doi: 10.7554/eLife.23717

93. Martin W, Stoebe B, Goremykin V, et al (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165. doi: 10.1038/30234

94. Martin W, Rujan T, Richly E, et al (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A 99:12246–51. doi: 10.1073/pnas.182432999

95. Reyes-Prieto A, Hackett JD, Soares MB, et al (2006) Cyanobacterial Contribution to Algal Nuclear Genomes Is Primarily Limited to Plastid Functions. Curr Biol. doi: 10.1016/j.cub.2006.09.063

96. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Proceeding 17th Int. Conf. World Wide Web - WWW '08. ACM Press, New York, New York, USA, p 695

97. Newman MEJ (2003) Mixing patterns in networks. Phys Rev E 67:26126. doi: 10.1103/PhysRevE.67.026126

98. Newman M (2010) Networks. An introduction. Oxford Univ Press. doi: 10.1093/acprof:oso/9780199206650.001.0001

99. Rappé MS, Giovannoni SJ (2003) The Uncultured Microbial Majority. Annu Rev Microbiol 57:369–394. doi: 10.1146/annurev.micro.57.030502.090759

100. Williams TA, Embley TM (2014) Archaeal ?Dark Matter? and the Origin of Eukaryotes. Genome Biol Evol 6:474–481. doi: 10.1093/gbe/evu031

101. Castelle CJJ, Wrighton KCC, Thomas BCC, et al (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. Curr Biol 25:690–701. doi: 10.1016/j.cub.2015.01.014

102. Brown CT, Hug LA, Thomas BC, et al (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. doi: 10.1038/nature14486

103. Spang A, Saw JH, Jørgensen SL, et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. doi: 10.1038/nature14447

104. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. doi: 10.1038/nature21031

105. Prakash T, Taylor TD (2012) Functional assignment of metagenomic data: challenges and applications. Brief Bioinform 13:711–727. doi: 10.1093/bib/bbs033

106. Hingamp P, Grimsley N, Acinas SG, et al (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678–1695. doi: 10.1038/ismej.2013.59

107. de Vargas C, Audic S, Henry N, et al (2015) Eukaryotic plankton diversity in the sunlit ocean. Science (80- )

348:1261605–1261605. doi: 10.1126/science.1261605

108. Sunagawa S, Coelho LP, Chaffron S, et al (2015) Structure and function of the global ocean microbiome. Science (80- ) 348:1261359–1261359. doi: 10.1126/science.1261359

109. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, et al (2016) Uncovering Earth's virome. Nature 536:425–430. doi: 10.1038/nature19094

110. Shi M, Lin XD, Tian JH, et al (2016) Redefining the invertebrate RNA virosphere. Nature. doi: 10.1038/nature20167

111. van Regenmortel MH, Mayo MA, Fauquet CM, Maniloff J (2000) Virus nomenclature: consensus versus chaos. Arch Virol 145:2227–2232.

112. Gibbs AJ (2000) Virus nomenclature descending into chaos. Arch Virol 145:1505–1507.

113. Lawrence JG, Hatfull GF, Hendrix RW (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol 184:4891–4905.

114. Franklin LR (2007) Bacteria, sex, and systematics. Philos Sci 74:69–95. doi: Doi 10.1086/519476

115. Bapteste E, Boucher Y (2008) Lateral gene transfer challenges principles of microbial systematics. Trends Microbiol 16:200–207. doi: 10.1016/j.tim.2008.02.005

116. Bapteste E, O'Malley MA, Beiko RG, et al (2009) Prokaryotic evolution and the tree of life are two different things. Biol Direct 4:34. doi: 10.1186/1745-6150-4-34

117. Andam CP, Williams D, Gogarten JP (2010) Natural taxonomy in light of horizontal gene transfer. Biol Philos 25:589–602. doi: DOI 10.1007/s10539-010-9212-8

118. Koonin E V, Dolja V V (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78:278–303. doi: 10.1128/MMBR.00049-13

119. Lederberg J, Tatum EL (1946) Gene recombination in Escherichia coli. Nature 158:558.

120. Zinder ND, Lederberg J (1952) Genetic exchange in Salmonella. J Bacteriol 64:679–699.

121. Levin BR (1988) Frequency-dependent selection in bacterial populations. Philos Trans R Soc L B Biol Sci 319:459–472.

122. Rodriguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiol Lett 231:153–158.

123. Chen I, Christie PJ, Dubnau D (2005) The ins and outs of DNA transfer in bacteria. Science (80- ) 310:1456–1460. doi: 10.1126/science.1114021

124. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3:504–510. doi: 10.1038/nrmicro1163

125. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3:722–732. doi: 10.1038/nrmicro1235

126. Dagan T, Martin W (2009) Getting a better picture of microbial evolution en route to a network of genomes. Philos Trans R Soc L B Biol Sci 364:2187–2196. doi: 10.1098/rstb.2009.0040

127. Kulp A, Kuehn MJ (2010) Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu Rev Microbiol 64:163–184. doi: 10.1146/annurev.micro.091208.073413

128. McDaniel LD, Young E, Delaney J, et al (2010) High frequency of horizontal gene transfer in the oceans. Science (80- ) 330:50. doi: 10.1126/science.1192243

129. Dubey GP, Ben-Yehuda S (2011) Intercellular nanotubes mediate bacterial communication. Cell 144:590–600. doi: 10.1016/j.cell.2011.01.015

130. Desnues C, La Scola B, Yutin N, et al (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci U S A 109:18078–18083. doi: 10.1073/pnas.1208835109

131. Kutschera VE, Bidon T, Hailer F, et al (2014) Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. Mol Biol Evol 31:2004–2017. doi: 10.1093/molbev/msu186

132. Simmonds P (2014) Methods for virus classification and the challenge of incorporating metagenomic sequence data. J Gen Virol. doi: 10.1099/jgv.0.000016

133. International Committee on Taxonomy of Viruses (2014) ICTV Documents. In: http://talk.ictvonline.org/files/ictv_documents/default.aspx.

134. Baltimore D (1971) Expression of animal virus genomes. Bacteriol Rev 35:235–241.

135. Koonin E V, Senkevich TG, Dolja V V (2006) The ancient Virus World and evolution of cells. Biol Direct 1:29. doi: 10.1186/1745-6150-1-29

136. Iranzo J, Koonin E V., Prangishvili D, Krupovic M (2016) Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements. J Virol 90:11043–11055. doi: 10.1128/JVI.01622-16

137. Lord E, Le Cam M, Bapteste É, et al (2016) BRIDES: A New Fast Algorithm and Software for Characterizing Evolving Similarity Networks Using Breakthroughs, Roadblocks, Impasses, Detours, Equals and Shortcuts. PLoS One 11:e0161474. doi: 10.1371/journal.pone.0161474

138. Dongen SM van (2001) Graph clustering by flow simulation.

139. Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. Soc Networks 19:243–269. doi: 10.1016/S0378-8733(96)00301-2

140. Goh K-I, Cusick ME, Valle D, et al (2007) The human disease network. Proc Natl Acad Sci U S A 104:8685–90. doi: 10.1073/pnas.0701361104

141. Himmelstein DS, Baranzini SE, Rand V, et al (2015) Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLOS Comput Biol 11:e1004259. doi: 10.1371/journal.pcbi.1004259

142. Ahn Y-Y, Ahnert SE, Bagrow JP, et al (2011) Flavor network and the principles of food pairing. Sci Rep 1:196. doi: 10.1038/srep00196

143. Lanza VF, Baquero F, de la Cruz F, Coque TM (2017) AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics 33:283–285. doi: 10.1093/bioinformatics/btw601

144. Diestel R (2010) Graph theory. Springer

145. McInerney JO, Pisani D, Bapteste E, O'Connell MJ (2011) The public goods hypothesis for the evolution of life on Earth. Biol Direct 6:41. doi: 10.1186/1745-6150-6-41

146. Hauser M, Mayer CE, Söding J (2013) kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics 14:248. doi: 10.1186/1471-2105-14-248

# II.2 MULTI-LEVEL NETWORKS TO STUDY EVOLUTION

To explore how genes are shared between different genomes, genome networks can be built from gene similarity (Figure 8B). An alternative form of the genome network is the multiplex network. In this network, nodes can be linked by edges of different types, for example, each edge representing a different shared gene family or different functional groups of gene families, thus retaining additional information compared to a simpler genome network (Corel et al. 2016). Multiplex networks can be useful for small scale analyses, however with large datasets they can rapidly become difficult to interpret and analyze (Figure 8C).

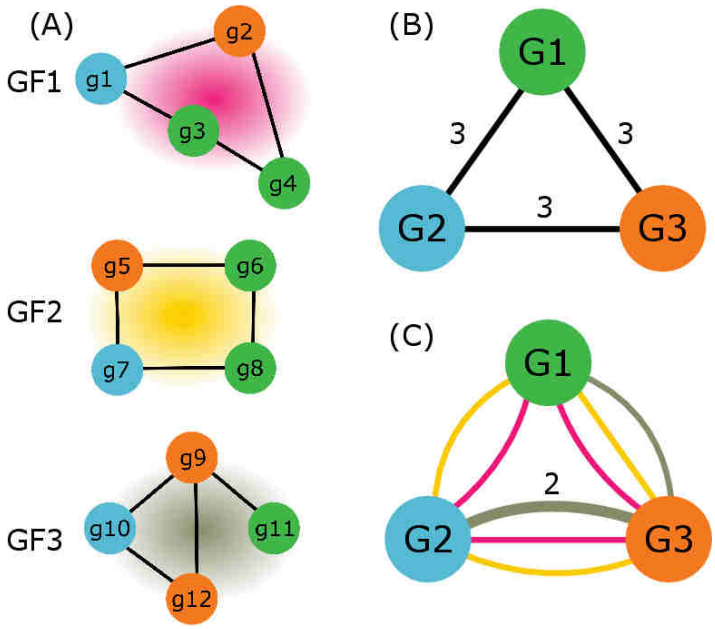



**Figure 8: Translating gene networks to genome networks.**
A) Gene network for three gene families. Gene nodes are coloured based on their genome of origin. The background colour corresponds to the gene family colour in part C. B) The genome network corresponding to the gene network in A. Edges are weighted on the number of gene families shared by the genomes. C) Multiplex-genome network corresponding to the gene network in A. Genomes are connected by multiple edges with colours corresponding to different gene families. These edges are weighted based on the number of genes shared between two genomes for each family.

Importantly, multiplex networks are unimodal projections of bipartite graphs which can provide greater clarity and have a number of attractive properties for the analysis of larger datasets. Bipartite graphs can be used to analyze the transfer or exchange of genetic material among organisms from same or different domains of life (e.g. transmission of antibiotic resistance in bacteria (Lanza et al. 2015) or diversification of Archaea and Bacteria by LGT (Jaffe et al. 2016)). They are ideal for comparative genomics, including the comparison of genomes reconstructed in metagenomic analyses as shown in the next section (I.3). A bipartite graph is a graph having two sets of nodes U and V, so that the edges only connect nodes of the set U to nodes of the set V (Figure 9).

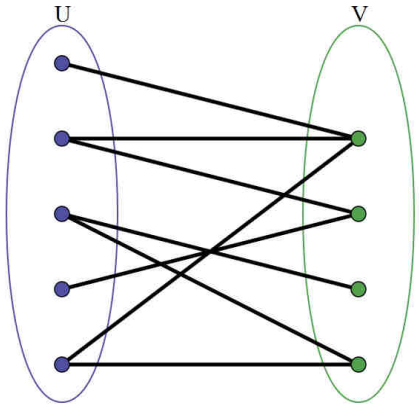



**Figure 9:** Example of a bipartite graph.

These types of graphs are useful to explore evolutionary processes at different level like gene family-genome bipartite graphs. Two topological features of bipartite graphs can be used to study gene sharing: twins and articulation points (Diestel 2006) (Figure 10). Twin nodes are useful since they describe entities having similar distributions. Articulation points, in contrast, is a bridge linking almost completely different entities, and are therefore indices of the graph's modularity. Extending bipartite graph approach to multi-level graphs, adding additional layers of information such as the environment in ecological studies (tripartite gene-genome-environment graphs), could provide a powerful summary of gene sharing in relatively complex datasets.

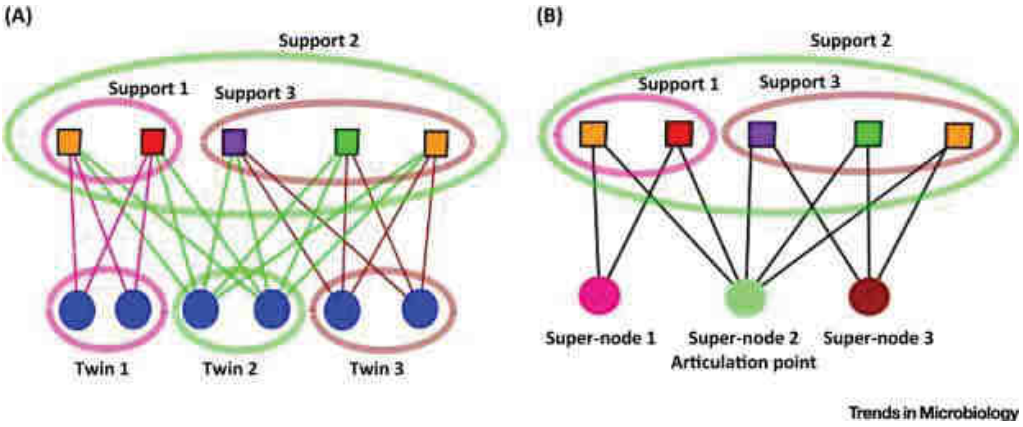



**Figure 10: Twins and articulation points in a bipartite graph.**
(A) Top nodes in this bipartite graph are genomes and bottom nodes gene families. Nodes in each colored ellipse at the bottom form a twin class, since their sets of neighbors (supports encircled by similarly colored ellipses on the top level) are identical (as highlighted by the coloring of their incident edges). (B) Collapsing twin nodes into super-nodes yields a reduced graph, without further bottom twin nodes. The supported groups of host genomes are unchanged, and are now defined as the neighbors of a single super-node. Due to the graph reduction, the green super-node is now an articulation point, since its removal disconnects the nodes in the pink and brown supports. (Corel et al. 2016)

In the article n°2, we present an integrated suite of software tools named MultiTwin, aimed at the construction, structuring and analysis of multipartite graphs for evolutionary biology. We illustrate the use of this tool with an application of the bipartite approach (using gene family-genome graphs) for the analysis of pathogenicity traits in prokaryotes. This article has been submitted to the journal "Molecular Biology and Evolution" and is under review.

# *MultiTwin*: a software suite to analyse multipartite graphs

Eduardo Corel[1*‡], Jananan S. Pathmanathan[1‡], Andrew K. Watson[1], Slim Karkar[2], Philippe Lopez[1] and Eric Bapteste[1]

[1]**Sorbonne Universités,** Université Pierre et Marie Curie**, Institut de Biologie Paris-Seine,** Centre National de la Recherche Scientifique, Unité Mixte de Recherche 7138 Evolution Paris-Seine, 7 quai St Bernard, 75005 Paris, France.

[2] **Rutgers University**, Department of Ecology, Evolution, and Natural Resources, 14 College Farm Rd, New Brunswick, NJ 08901- 8551, USA.

*To whom correspondence should be addressed. Eduardo Corel: eduardo.corel@upmc.fr

‡ Both authors contributed equally to this work.

**Abstract**

The inclusion of introgressive processes in evolutionary studies induces a less constrained view of evolution. Network-based methods (like large-scale similarity networks) allow to include in comparative genomics all extra-genomic carriers (like viruses, the most abundant biological entities on the planet) with their cellular hosts. The integration of several levels of biological organisation (genes, genomes, communities, environments) enables more comprehensive analyses of gene sharing and improved sequence-based classifications. However, the algorithmic tools for the analysis of such networks are usually restricted to people with good programming skills. We present an integrated suite of software tools named *MultiTwin*, aimed at the construction, structuring and analysis of multipartite graphs for evolutionary biology. We illustrate the use of this tool with an application of the bipartite approach (using gene family-genome graphs) for the analysis of pathogenicity traits in prokaryotes.

**Availability:** Source code freely available for download at **http://www.evol-net.fr/index.php/fr/downloads**, implemented in Python 2.7 and C++ and supported on Linux.

## Introduction

The network paradigm is increasingly used as a complement for phylogenetic tree reconstruction for biological evolutionary studies (Halary et al. 2010; Kloesges et al. 2011; Leigh et al. 2011; Tamminen et al. 2012; Corel et al. 2016; Iranzo, Krupovic, et al. 2016). We present here *MultiTwin*, an exploratory tool for multipartite graph analysis. Such graphs encompass several levels of biological organisation. Bipartite graphs have been up to now most commonly used (Ahn et al. 2011; Himmelstein et al. 2015; Lanza et al. 2017), and particularly gene family-genome bipartite graphs have already demonstrated their usefulness, like uncovering membrane-related genes shared between recently discovered ultra-small bacteria (CPR) and archae (Jaffe et al. 2016), proposing finer classifications of archaeal or ds-DNA viruses (Iranzo, Koonin, et al. 2016; Iranzo, Krupovic, et al. 2016), or analyzing the transmission of antibiotic resistance through Firmicute plasmids (Lanza et al. 2015). Higher-level application would also be of considerable interest, like the study of environmental adaptive traits with tripartite gene-genome-environment graphs, and are starting to gain attention both from the computational (Murata and Tsuyoshi 2010) and applied point of view (Alaimo et al. 2014). Our contribution consists in a general framework and dedicated tools developed in Python for the construction, structuring and analysis of multipartite graphs. Detecting regularities and singularities in genome-based graphs informs on the degree of redundancy of genomic data, and gives a summarization, both in terms of *compressibility* and *modularity* of the genomes under study, with possible applications to the detection of functional modules (*bio-bricks*).

Our tool implements the search of one type of regularities (twin nodes), and one type of singularities (articulation points). Twin nodes are useful since they describe entities having similar distributions, and achieve therefore a type of *lossless compression*. Articulation points, in contrast, represent the unique link between otherwise completely unrelated communities, and are therefore indices of the graph's *modularity*. Moreover, the *MultiTwin* suite can generate a bipartite gene family-genome graph from genomic data (*i.e.* either from sequences themselves or the output file of a BLAST all-against-all run on the set of sequences).

## New approach

### Graph Model

A graph $G = (V, E)$ is *k-partite* if there exists a partition of the set of nodes $V = V_1 \cup \ldots \cup V_k$ such that an edge only connects nodes from two *different* subsets of the partition. For example, a gene family-genome graph has two types of nodes, and an edge connects only a gene family and a genome where one member of the family is found.

Our model considers an initial graph structure (named the *root graph*), and distinguishes between *iterable* and *terminal* operations. This feature allows for a flexible and multi-level analysis of graphs: graph modifications can be nested one into another, and all intermediate steps can be considered in the analysis, thanks to a consistent trailing scheme for the intermediate graphs (*cf.* Figure 1). Some graph operations involve a node renaming. A key feature of our model is the use of a *trail file*, which maintains the correspondence between the original node identifiers (in the root graph) and those of the current (possibly terminal) graph. The rationale behind this choice is that some biological annotations are most likely available for the entities forming the nodes of the root graph. In our example, the root graph is the gene-genome bipartite graph. Functional and taxonomic annotations are available for individual genes and for genomes respectively (but usually not for gene families or clusters of genomes).

We implement the following basic operations on graphs: subgraphs, factoring and (overlapping) clustering.

- A *subgraph* is defined by a subset $E' \subset E$ of the graph's edges. It is an iterable operation, but involves no renaming.
- *Factoring* is defined by a surjective mapping (a non-overlapping node clustering) $f: V \to W$, resulting in a *factor graph* $G/f = (W, f(E))$ where $f(E) = \{(f(u), f(v)) | (u, v) \in E\}$. It is also iterable, and implies node renaming.
- *Overlapping clustering* is a terminal graph operation, and no node renaming is required.

*Iterable* and *terminal* refer to whether the operations preserve the graph's k-partite structure. Factoring through a non-overlapping clustering that groups nodes of *different types* or through an *overlapping* clustering, destroys the k-partite structure of the graph. Note, however, that it is possible to model any clustering of a k-partite graph as a (k+1)-partite graph, and hence to analyse it further with our tool.

Our suite includes a program named DetectTwins.py for the detection of *twin nodes*, that is, nodes whose neighbourhoods in *G* coincide, and their support (*i.e.* their common neighbourhood), as well as a dedicated tool to construct gene families (FamilyDetector). Finally, we also implemented a module (Description.py) to annotate the content of the graph's clusters and its possible intermediate levels. For instance, in Figure 1, intermediate levels are gene families (level 1) and twins (level 2). Clusters can be any kind of node subsets: groups of gene families or genomes, nodes forming a connected component, communities returned by an external clustering algorithm, and so on.

**Overview of functions**

The *MultiTwin* suite contains the following main scripts:

- CleanBlastp
- FamilyDetector
- InducedSubgraph.py
- FactorGraph.py
- DetectTwins.py
- Description.py

as well as a standalone program BiTwin.py which performs the pipeline described in Figure 1, and a few utilities used in the main scripts.

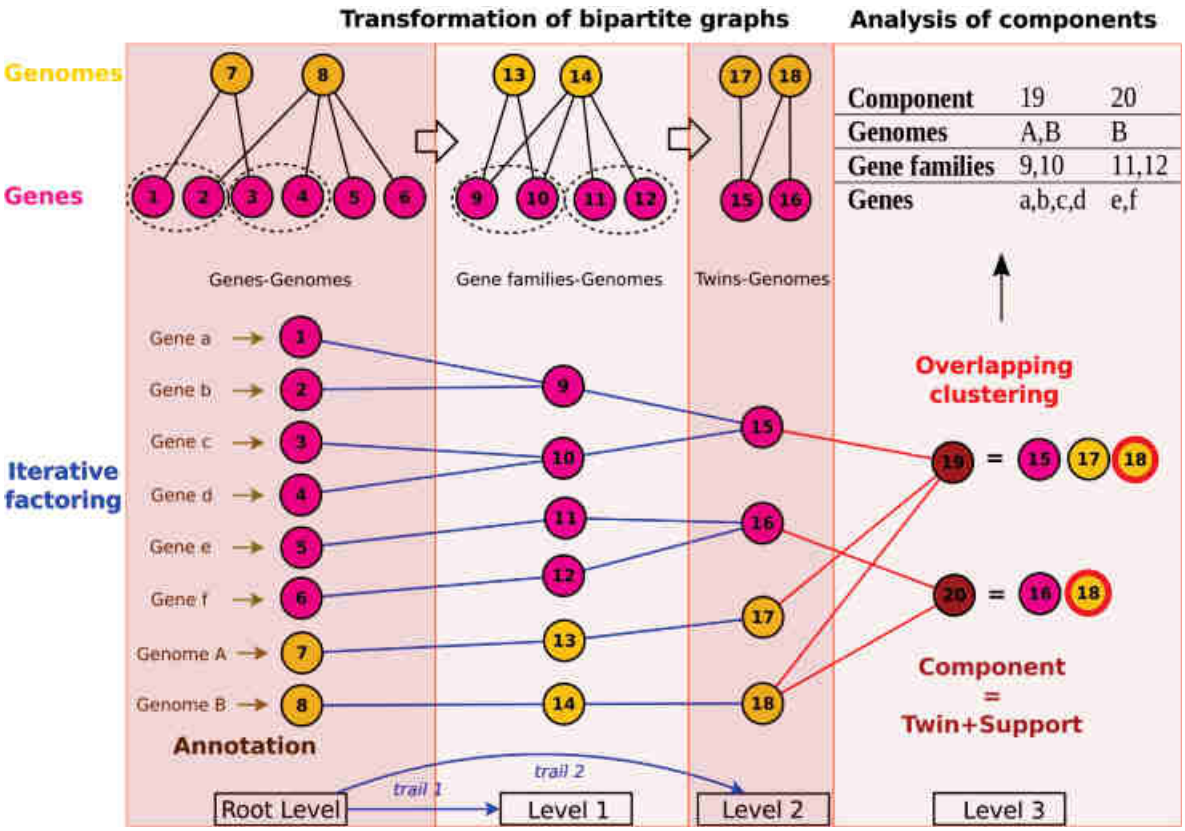



**Figure 1: Outline of the bipartite graph generation and analysis.** At the root level, the bipartite graph only consists in disjoint star graphs. Level 1 and level 2 are constructed by two successive runs of `FactorGraph.py` using the factoring maps described in blue. The first factoring is based is the gene family clustering produced by our script `FamilyDetector`. Different similarity thresholds can be used, resulting in differently structured graph (assuming a molecular clock, these graphs can be seen as time slices of evolution). The second corresponds to the identification of twins by `DetectTwins.py`. The change of identifiers in the graph is recorded in the trail files as indicated on the bottom line. At level 3, the operation is a terminal one, since it produces overlapping clusters. The analysis of the resulting components is performed by the `Description.py` script, and is based on the annotations (at the root level) and the specified trail files.

**File formats and types.**

All files generically follow the same syntax X TAB Y.

A graph is described by its *edge file*, where X and Y denote the head and the tail of an edge, and a *node type file*, where X is the node ID and Y is the node type (*e.g.* type 1 corresponding to genes and type 2 to genomes). Node type files can be omitted for unipartite graphs, and also for bipartite graphs, provided that the first column of the edge file only contains type 1 nodes, and the second column type 2 nodes. *Community files*, where X is the node ID and Y is a community ID, can be used both for overlapping and non-overlapping clusterings (depending on whether node IDs are repeated or not). For instance, the decomposition of the graph into *connected components* can be encoded as a community file.

A distinctive feature of our model is to track consistently the successive modifications of our multipartite graphs, by the use of a special community file with an additional two-line header, called a *trail file*. In this file, X refers to the node ID in the *root graph*, Y to the node ID in the current graph, and the header recalls which operation on which graph has produced the current graph.

Only the *annotation file* (containing the biological information) has a different format. It consists of a tabbed file with a compulsory header containing the attribute names, and whose rows begin with the identifier used in the root graph.

**Implementation**

The implementation of the framework was carried out in Python (version 2.7) with some additional original code in C++. The Python code includes efficient implementations of graph algorithms from the igraph package (Csárdi and Nepusz 2006), that can moreover be accessed through the python-igraph wrapper.

**Features of the multipartite analysis functions.**

The code available at the URL http://www.evol-net.fr/ index.php/fr/downloads accepts different kinds of inputs, depending on the user's objectives. A detailed file with installation and usage information is provided. The data used in the application is also available with a dedicated guide file that allows to replicate our analysis.

**Standalone generation of the bipartite gene family-genome graph.**

The standalone program BiTwin.py consists of four mainly independent modules:

- Construction of the sequence families: by default, we assume that the sequences have been subjected to an all-against-all BLAST run (that can optionally be performed if raw sequences are supplied). The sequence similarity graph resulting from keeping the reciprocal best hit is filtered above similarity, coverage and E-value thresholds (≥30% identity, ≥80% mutual coverage and E-value ≤$10^{-5}$ by default). The sequences are then grouped into families, either as connected components (option 1) or as "Louvain communities" (Blondel et al. 2008) (option 2) of this graph.
- Construction of the bipartite graph: this step consists in factoring the user-supplied genome-sequence file by the sequence families file resulting from the previous step, seen as a community file.
- Twin and articulation point detection: we implemented two algorithms to compute the twin nodes (and their supports), and as well as the articulation points of the bipartite graph.
- Formatting and analysis output: this step uses the tabbed annotation file, with a compulsory header with the attribute names.

All the resulting bipartite graphs produced by the pipeline are stored in a hierarchy of directories below the current working directory.

**Custom usage.**

The *MultiTwin* code can also be used as a framework for the analysis of user-supplied multipartite graphs (Figure 2). In this usage, the graphs can be modified iteratively, either by subgraph induction or by factoring according to a node clustering, either iterable or terminal. Any node clustering algorithm can be used, provided that the result is supplied to FactorGraph.py as a community file. The Cluster.py script of our suite produces a community file for several algorithms that are available in igraph. Finally, the obtained graph can be analysed on the basis of the resulting intermediate levels of factoring (see the README file for the Description.py script).

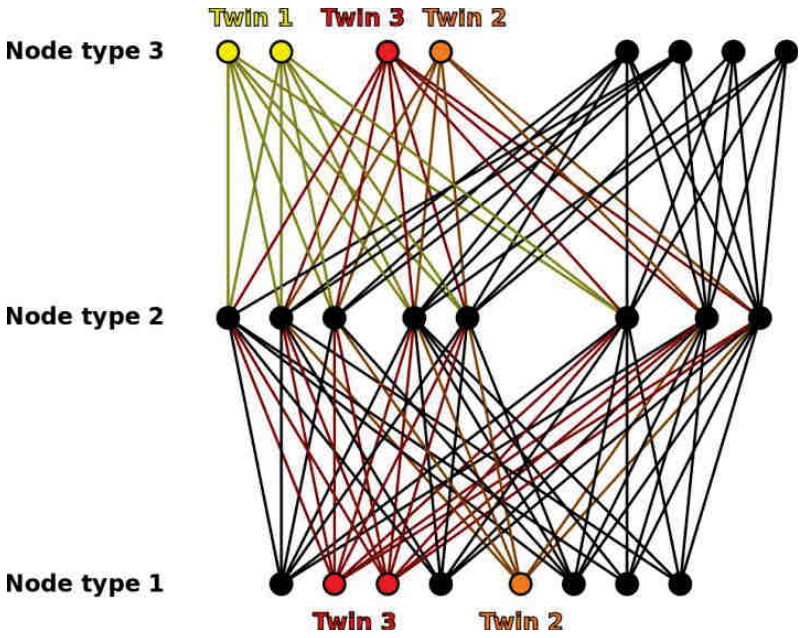



**Figure 2: Twin nodes in a toy example of tripartite graph.** Twin classes are formed by all the nodes having exactly the same neighbourhood. In this example, we highlighted in the same colour the nodes forming the graph's three non-trivial twin classes. All nodes in black have a different set of neighbours (and form thus each their own trivial twin class). In a multipartite graph, twins can be *homogeneous*, like twin 1 (in yellow) or *heterogeneous*, like twins 2 and 3. `DetectTwins.py` implements an option to detect only homogeneous twins (possibly even of a given type). In a tripartite graph where nodes of respective types 1, 2 and 3 are gene families, genomes and environments, it may be interesting to detect patterns like twin 2, where a gene family is found in the strict subset of those genomes that thrive in the same environment. Twin 3 is likely less informative, since the environment is non-discriminating (core genes are nevertheless detected on the lower layer).

## Results

We assembled a dataset of 20 pairs of genomes with comparable sizes, coming from phylogenetically closely related pathogen and non-pathogen organisms (Supp. Table 1). Organisms were assigned as “pathogens” or “non-pathogens” based on metadata from the GOLD (Mukherjee et al. 2017) and PATRIC (Wattam et al. 2017) databases. Protein sequences from these genomes were used in an all-against-all BLAST search with parameters as described in (Bittner et al. 2010), and the bipartite network was generated using BiTwin.py, with a minimum of 30% identity and 80% mutual coverage between sequences and gene family direction set to assemble connected components. COG annotations were assigned to gene families using

RPS-BLAST (Marchler-Bauer et al. 2002). The DetectTwins.py function was used to identify trivial and non-trivial twins. The twins were used as a community file for FactorGraph.py, collapsing gene-families that have an identical species distribution into a single node in the factored bipartite graph. Should the gene content of prokaryotic genomes have evolved largely in a tree-like fashion, one would expect to find mostly twins whose support have the same taxonomy. However, our study uncovered many twins with polyphyletic support.

In total, 26,228 gene families were compressed to 3,982 twin nodes. 3,197 were trivial twins (80.29%) - that is, they were single gene families with a unique distribution. 785 twins were non-trivial (19.71%), composed of multiple gene families with identical taxonomic distributions (Figure 3). Non-trivial twins include a “core” bacterial twin, composed of 50 gene families, and 4,371 genes that are universally conserved in all 40 genomes included in the analysis (Figure 3).

Additionally, 119 pathogen specific twins (plus 20 species specific twins) were identified that included sequences from more than one pathogen species. 58 twins were trivial and 61 were non-trivial (Supp. Table 2). The strongest cases for pathogen specific traits identified by bipartite analysis are the pathogen specific twins that are most broadly distributed within the group. The majority of pathogen specific twins (84) only included sequences from two different pathogen genomes. No single twin included sequences exclusive to all pathogen genomes, meaning that there is no “core” pool of gene families exclusively shared by pathogens. Additionally, two strategies were used to screen for twins enriched in pathogens but also present in non-pathogens - either a coarse cutoff value of >80% of genes within a given twin being pathogen-derived (42 twins total) or a hypergeometric test followed by FDR correction to identify twins significantly enriched in pathogen-derived genes (5 twins) (Supp. Table 2).

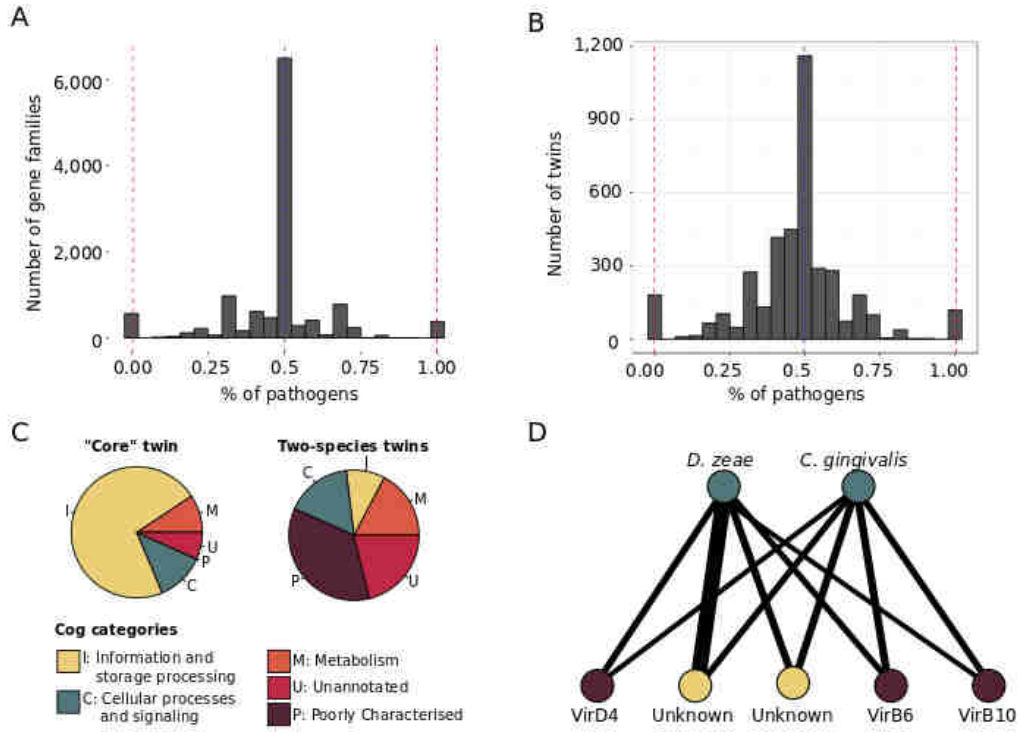



**Figure 3: Summary of the bipartite graph analysis of forty prokaryotic genomes.** A) The majority of gene families contained an equal proportion of pathogen and non-pathogen genes. Comparatively few are enriched in either pathogens or non-pathogens, with an extreme drop off from the peak at 0.5. A subset of gene families are exclusive to pathogens or to non-pathogens, indicated by peaks at 0 and 1, however the majority of these are only found in one genome. B) Most twins also contain an equal proportion of pathogens and non-pathogens, however the peak at 0.5 is less extreme in comparison to the surrounding distribution. There is a more gradual decline in number of twins from this peak towards the extremities at 0 and 1 than in the distribution at the gene family level. C) Functional analysis revealed that the twin containing all "core" gene families was predominantly composed of gene families involved in information and storage processing. This contrasts the twins containing gene families found in only two species, where informational genes are the least represented COG. Gene families found in two species are predominantly either associated with poorly characterised COGs or unannotated. (D) An example non-trivial bipartite twin of four gene families (bottom nodes) co-distributing in two relatively distantly related pathogen genomes (top nodes) from *Dickeya zeae* (Gamma-proteobacteria) and *Capnocytophaga gingivalis* (Flavobacteria). Two gene families (purple) contain components of the type IV secretion system, while two (yellow) have no known COG annotations. Their co-distribution with components of the type IV secretion system in distantly related taxa suggests that these may play a role in pathogenicity.

Both pathogen specific and pathogen enriched twins identified in this analysis include gene families with known roles in pathogenicity. One of the most broadly distributed pathogen specific twins, ADP-heptose:LPS heptosyltransferase (COG0859), is a part of the core machinery for LPS biosynthesis which, as an endotoxin, is a characterised factor in the pathogenesis of a broad range of gram-negative bacteria (Raetz and Whitfield 2002). Another pathogenicity factor, haemolysin co-regulated protein 1, was enriched in pathogens (based on the >80% cutoff). This is part of the type 6 secretion machinery, and has been proposed as a chaperone for effector protein secretion (Silverman et al. 2013). In addition to pathogenicity factors, a chloramphenicol-O-acetyl transferase and a beta-lactamase class D were enriched in pathogen genomes, enzymes conferring antibiotic resistance (Schwarz et al. 2004).

The identification of these known pathogen gene families within our set of twins can be seen as a proof of concept. It demonstrates the effectiveness of the bipartite graph approach for gene rediscovery. Moreover, this approach could be applied to identify novel genes associated with a particular feature. For example, in this dataset many twins unique to pathogens and enriched in pathogen genomes compared to non-pathogens are either annotated by COG as conserved proteins of unknown function, or unannotated in COG. Their enrichment in pathogen genomes compared to other twins suggests a potential role in pathogenicity for these thus far uncharacterised genes.

Likewise, 181 non-pathogen specific twins (plus 20 single species twins) were identified in this analysis, including 98 trivial and 83 non-trivial twins. No twins included genes from all non-pathogen genomes. Additionally, we identified 96 twins in which >80% of genes were from non-pathogens, and 29 twins enriched in non-pathogens using the hypergeometric test. The non-pathogen specific and enriched twins are more abundant and generally have broader distribution than those found in pathogens. This greater abundance is consistent with the findings of a broader analysis on 317 genomes (Merhej et al. 2009), which suggested that gene loss, in opposed to acquisition of virulence factors, has driven the evolution of parasites in their adaptation to their host cell. This included the loss of rRNA genes and transcriptional regulators, a result which is mirrored in our analysis. Another 5 broadly distributed non-pathogen enriched twins (two non-trivial) are associated with

loss of transcriptional regulation, supporting the idea that the evolution of pathogenesis could be related to the loss of regulation. Our approach independently found a correlation between nutrient acquisition (Merhej et al. 2009), and specifically a nitrogen fixation ability and a non-pathogenic lifestyle. Two large and broadly-distributed non-trivial twins enriched in non-pathogens are made up entirely of ABC-transport protein gene families, with predicted substrates including sugars and amino acids and oxoions. Four different twins also each include different components of the TRAP-type C4-dicarboxylate transport system, with substrates including succinate, malate and fumarate. This transport system is required for nitrogen fixation (Finan et al. 1983). Another more broadly distributed non-trivial twin exclusive to non-pathogens includes two gene families, a predicted Fe-S oxidoreductase and nitrogenase molybdenum-iron protein (alpha and beta chains). These are central components of the pathway for nitrogen fixation (Dixon and Kahn 2004). Moreover, a trivial twin unique to non-pathogens is annotated as a Sec-independent protein secretion pathway component. This secretion system has a broad range of functions, one of which is its requirement for nitrogen oxide reduction in the nitrogen cycle (Natale et al. 2008). Finally, two twins containing >80% non-pathogen genes include additional parts of the pathway for nitrogen fixation: nitrogenase subunit NifH and Nitrate/Nitrite transport proteins. While elements of the nitrogen fixation pathway are shared between pathogens and non-pathogens (Carvalho et al. 2010), our bipartite graph analysis reinforces the argument that nitrogen fixation is a predominantly a feature of non-pathogenic bacteria.

This relatively small scale bipartite graph analysis identified known signatures of pathogenesis and antibiotic resistance that were exclusive to or enriched in pathogen genomes, as well as genes of thus far unknown function which may play similar roles in pathogen biology, highlighting the potential of the approach for gene discovery. A larger number of twins were associated with non-pathogen genomes, consistent with the idea that pathogens undergo reductive evolution during their adaptation to the host environment including deregulation of gene expression (Merhej et al. 2009; Georgiades and Raoult 2011). Non-pathogen enriched twins associated were also associated with nitrogen fixation. Nitrogen fixation within a community can be viewed as an example of the production of a “public good” - it is a pathway that produces an important commodity that can be shared by an entire

community, but its phylogenetic distribution within that community is patchy. Though some pathogens are known to encode genes involved in the production of public goods, it would be interesting to explore whether there is a broad trend towards production public goods by non-pathogens. *MultiTwin* would allow to test this hypothesis on a larger scale dataset.

## Acknowledgements and funding

We thank J.O. McInerney, M. Habib, F. de Montgolfier and T. Hujsa for critical discussions, and R. Lannes for help with the Python code.

This work has been supported by the ERC (grant FP7/2007-2013 Grant Agreement #615274 to JSP, AKW, EC and EB); and a grant from Région Ile-de-France (DIM Malinf 2011-2013) to SK.

Conflict of Interest: none declared.

## References

Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L, Roque AC. 2011. Flavor network and the principles of food pairing. Sci. Rep. 1:196. doi:10.1038/srep00196. [accessed 2017 May 28]. http://www.nature.com/articles/srep00196.

Alaimo S, Giugno R, Pulvirenti A. 2014. ncPred: ncRNA-Disease Association Prediction through Tripartite Network-Based Inference. Front. Bioeng. Biotechnol. 2:71. doi:10.3389/fbioe.2014.00071.

Bittner L, Halary S, Payri C, Cruaud C, de Reviers B, Lopez P, Bapteste E, Casjens S, Luft B, Boucher Y. 2010. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biol. Direct 5:47. doi:10.1186/1745-6150-5-47. [accessed 2017 Apr 26]. http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-5-47.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. doi:10.1088/1742-5468/2008/10/P10008.

Carvalho FM, Souza RC, Barcellos FG, Hungria M, Vasconcelos ATR. 2010. Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of

the order Rhizobiales. BMC Microbiol. 10:37. doi:10.1186/1471-2180-10-37.

Corel E, Lopez P, Méheust R, Bapteste E, Darwin C, O'Hara RJ, Doolittle WF, Bapteste E, Bapteste E, Al. E, et al. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol. 24:224–237. doi:10.1016/j.tim.2015.12.003.

Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Syst. 1695:1695.

Dixon R, Kahn D. 2004. Genetic regulation of biological nitrogen fixation. Nat. Rev. Microbiol. 2:621–631. doi:10.1038/nrmicro954.

Finan TM, Wood JM, Jordan DC. 1983. Symbiotic properties of C4-dicarboxylic acid transport mutants of Rhizobium leguminosarum. J. Bacteriol. 154:1403–1413.

Georgiades K, Raoult D. 2011. Defining Pathogenic Bacterial Species in the Genomic Era. Front. Microbiol. 1:151. doi:10.3389/fmicb.2010.00151.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. Proc. Natl. Acad. Sci. U. S. A. 107:127–32. doi:10.1073/pnas.0908978107.

Himmelstein DS, Baranzini SE, Rand V, Lovering R, Bruford E, Khodiyar V. 2015. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. Tang H, editor. PLOS Comput. Biol. 11:e1004259. doi:10.1371/journal.pcbi.1004259. [accessed 2017 May 28]. http://dx.plos.org/10.1371/journal.pcbi.1004259.

Iranzo J, Koonin E V., Prangishvili D, Krupovic M. 2016. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements. Sandri-Goldin RM, editor. J. Virol. 90:11043–11055. doi:10.1128/JVI.01622-16.

Iranzo J, Krupovic M, Koonin E V. 2016. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. MBio 7:e00978-16. doi:10.1128/mBio.00978-16.

Jaffe AL, Corel E, Sylvestre Pathmanathan J, Lopez P, Bapteste E. 2016. Bipartite

graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. Environ. Microbiol. 18:5072–5081. doi:10.1111/1462-2920.13477.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. Mol. Biol. Evol. 28:1057–1074. doi:10.1093/molbev/msq297. [accessed 2017 Jul 5]. https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msq297.

Lanza VF, Baquero F, de la Cruz F, Coque TM. 2017. AcCNET (Accessory Genome Constellation Network): comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics 33:283–285. doi:10.1093/bioinformatics/btw601. [accessed 2017 May 28]. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw601.

Lanza VF, Tedim AP, Martínez JL, Baquero F, Coque TM, Lanza VF, Tedim AP, Martínez JL, Baquero F, Coque TM. 2015. The Plasmidome of Firmicutes: Impact on the Emergence and the Spread of Resistance to Antimicrobials. Microbiol. Spectr. 3:PLAS-0039-2014. doi:0.1128/microbiolspec.PLAS-0039-2014.

Leigh JW, Schliep K, Lopez P, Bapteste E. 2011. Let them fall where they may: Congruence analysis in massive phylogenetically messy data sets. Mol. Biol. Evol. 28:2773–2785. doi:10.1093/molbev/msr110.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30:281–3.

Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. 2009. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol. Direct 4:13. doi:10.1186/1745-6150-4-13.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, et al. 2017. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res. 45:D446–

D456. doi:10.1093/nar/gkw992.

Murata T, Tsuyoshi. 2010. Detecting communities from tripartite networks. In: Proceedings of the 19th international conference on World wide web - WWW ’10. New York, New York, USA: ACM Press. p. 1159.

Natale P, Brüser T, Driessen AJM. 2008. Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. Biochim. Biophys. Acta - Biomembr. 1778:1735–1756. doi:10.1016/j.bbamem.2007.07.015.

Raetz CRH, Whitfield C. 2002. Lipopolysaccharide endotoxins. Annu. Rev. Biochem. 71:635–700. doi:10.1146/annurev.biochem.71.110601.135414.

Schwarz S, Kehrenberg C, Doublet B, Cloeckaert A. 2004. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. FEMS Microbiol. Rev. 28:519–542. doi:10.1016/j.femsre.2004.04.001.

Silverman JM, Agnello DM, Zheng H, Andrews BT, Li M, Catalano CE, Gonen T, Mougous JD. 2013. Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates. Mol. Cell 51:584–93. doi:10.1016/j.molcel.2013.07.025.

Tamminen M, Virta M, Fani R, Fondi M. 2012. Large-scale analysis of plasmid relationships through gene-sharing networks. Mol. Biol. Evol. 29:1225–40. doi:10.1093/molbev/msr292. [accessed 2017 Jun 25]. https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr292.

Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, et al. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45:D535–D542. doi:10.1093/nar/gkw1017.

# II.3 APPLICATION OF BIPARTITE GRAPHS FOR THE ANALYSIS OF INTERDOMAIN LGT BETWEEN ULTRASMALL AND LARGER PROKARYOTES.

In recent years, environmental metagenomes studies have shed light on a number of new organisms revealing an unsuspected biodiversity of microbial ecosystems (Sunagawa et al. 2015). Among the newly discovered organisms are very small prokaryotes known as CPR (Candidate Phyla Radiation), encompassing 15% of the known bacterial phyla and thus arousing a great interest (Brown et al. 2015). It has been shown that these organisms are deprived of certain metabolic pathways and that they have a cell envelope sharing common characteristics with both Gram (-) bacteria and archaea (Brown et al. 2015; Luef et al. 2015). Because of their genomic properties, these groups of ultrasmall prokaryotes must be dependent on other Bacteria or Archaea, making them as potential candidates for endosymbiotic lifestyle.

In the article n°3, we used a bipartite approach to verify whether exchanges of genetic material occurred by symbiosis or endosymbiosis between CPR and other prokaryotes, which might have been potential hosts of CPR. For this, we compared the protein sequences of the CPRs with those of the complete prokaryotic genomes from NCBI. The similarities between the sequences were detected with BLASTP and only the hits passing a certain number of thresholds (percentage of identity, mutual coverage, and E-value,) were kept for analysis. Subsequently, this filtered SSN was treated as a bipartite graph. Analysis of these bipartite graphs suggest numerous horizontal gene transfer (LGT) between CPR and other prokaryotes, including Archaea. Functional analyses of the sequences involved in these LGT showed that they were, in the majority of cases, related to membrane proteins. This article has been accepted and published in the journal "Environmental Microbiology".

environmental microbiology

Environmental Microbiology (2016) **18**(12), 5072–5081 doi:10.1111/1462-2920.13477

# Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins

**Alexander L. Jaffe,† Eduardo Corel,† Jananan Sylvestre Pathmanathan, Philippe Lopez and Eric Bapteste***

*Equipe AIRE, UMR 7138, Laboratoire Evolution Paris-Seine, Université Pierre et Marie Curie, 7 Quai St. Bernard 75005, Paris, France.*

## Summary

**Based on their small size and genomic properties, ultrasmall prokaryotic groups like the Candidate Phyla Radiation have been proposed as possible symbionts dependent on other bacteria or archaea. In this study, we use a bipartite graph analysis to examine patterns of sequence similarity between draft and complete genomes from ultrasmall bacteria and other complete prokaryotic genomes, assessing whether the former group might engage in significant gene transfer (or even endosymbioses) with other community members. Our results provide preliminary evidence for many lateral gene transfers with other prokaryotes, including members of the archaea, and report the presence of divergent, membrane-associated proteins among these ultrasmall taxa. In particular, these divergent genes were found in TM6 relatives of the intracellular parasite *Babela massiliensis*.**

## Introduction

Recent metagenomic analyses are revealing a wealth of new, unusual microbes that challenge current knowledge about prokaryotic diversity and microbial symbiosis. Among these groups is a cosmopolitan clade termed the Candidate Phyla Radiation (CPR; Brown *et al.*, 2015; Luef *et al.*, 2015; Hug *et al.*, 2016), mostly ultrasmall cells nearing the lower theoretical size limit for viability predicted by physical models (Velimirov, 2001). This clade comprises 15% of the described bacterial phyla and shares cell envelope characteristics with both Gram-positive bacteria and archaea (Brown *et al.*, 2015; Luef *et al.*, 2015). Based on these unusual membranes, small cellular size/genomes, and lack of certain biosynthetic pathways, it has been suggested that these bacteria are obligate fermenters dependent on other microbial community members (Brown *et al.*, 2015). This makes them prime candidates for an endosymbiotic lifestyle.

However, larger novel microbes like the hydrothermal vent organism Lokiarchaeum have also been recently described. This surprising archaeal group harbors membrane-remodeling systems compatible with rudimentary phagocytic capability, and displays a composite proteome possibly acquired by LGT (Spang *et al.*, 2015). Thus, the lineage to which Lokiarchaeum belongs has been proposed as a prime candidate host for prokaryotic endosymbionts, with a possible contribution to eukaryogenesis (Koonin, 2015; Spang *et al.*, 2015; but see Nasir *et al.*, 2015). In principle, the discovery of these novel, candidate hosts and symbionts in the environment adds to debated theoretical suggestions that (i) massive gene transfers between archaea and bacteria (Nelson-Sathi *et al.*, 2015) and (ii) prokaryote-in-prokaryote endosymbiosis (Lake, 2009; Swithers *et al.*, 2011) might have facilitated major evolutionary transitions like the origin of eukaryotes and the emergence of Gram-negative bacteria. However, prokaryote-in-prokaryote symbioses remain extremely rare, with only one described example in the mealybug (Husnik *et al.*, 2013).

New environmental datasets provide a first opportunity to examine the genomic relationships among the CPR, Lokiarchaeota, and other prokaryotic groups. Given the particular characteristics described above, we tested whether members of the CPR might have been endosymbiotic or partners in gene exchange with other bacteria or archaea. More precisely, we looked for signs of endosymbiotic gene transfer—a process by which a symbiont transfers genetic material to the host (Timmis *et al.*, 2004: Martin *et al.*, 2015)—and LGT involving organisms from the ultrasmall size fraction published by Brown *et al.* (2015). To this end, we performed a large-scale BLAST comparison of protein sequences from both draft and complete genomes of CPR (and TM6, a related phylum)

Received 8 July, 2016; accepted 27 July, 2016. *For correspondence. *E-mail: eric.bapteste@upmc.fr; Tel. +330144272164. †These authors contributed equally to this work.

dataset against all complete bacterial and archaeal genomes on NCBI (February 1, 2016). Subsequently, we used a bipartite graph analysis (BGA) to examine the resulting patterns of gene sharing across these genomes. This approach to environmental sequence data allowed us to identify specific processes of transfer or diversity, like those involving membrane-related proteins.

## Results/discussion

We began by BLASTing a large dataset of predicted proteins from a set of binned and curated CPR/TM6 genomes ($n = 637,155$) against proteins from all complete bacterial and archaeal genomes on NCBI (4,600 genomes, $n = 15,373,158$ proteins). We dereplicated the CPR/TM6 sequences and partitioned them into four categories—those that showed an above-threshold BLAST hit with only archaeal genomes ($n = 2,236$), only bacterial genomes (BAC, $n = 124,022$), both (PROK, $n = 81,634$), or neither (CPR/TM6, $n = 158,245$). We first performed a BGA on the BAC subset, delineating groups of CPR/TM6 sequences with shared, exclusive similarity to a given set of prokaryotic genomes. These groupings of proteins exclusively associated with a given set of genomes are known as "twins," the detection of which is an efficient way to represent complex patterns of gene sharing among organisms (Corel *et al.*, 2016). For example, a "twin" associating a group of sequences to one or more complete CPR genomes indicates that these sequences are likely from a CPR organism. However, twins connecting a set of CPR/TM6 sequences with one or more bacteria or archaea distantly related to CPR/TM6 suggests a case of gene transfer, endosymbiotic or otherwise, between CPR/TM6 organisms and other distantly related prokaryotes (Fig. 1).

The BGA resulted in 82,953 "twins" that were sorted by decreasing number of CPR/TM6 proteins they contained. The twins containing the largest number of CPR/TM6 sequences (between 268 and 3,540) involved the complete CPR genomes from the Brown *et al.* (2015) dataset, as well as 5 strains of *Peribacter riflensis*, another phylum in CPR (Anantharaman *et al.*, 2016). This result is expected, and offers a good proof of concept for our methodology: given that most proteins in the Brown *et al.* (2015) dataset are already classified as CPR, the BGA approach should associate those proteins with complete CPR genomes. The next strongest signal (i.e., CPR/TM6 proteins exclusively associated with a particular prokaryotic genome) revealed 456 proteins showing distant (~39% mean sequence identity, Fig. 2) but exclusive similarity to *Babela massiliensis*, a gram-negative, intracellular amoeboid parasite in the candidate phylum TM6 (Pagnier *et al.*, 2015). These 456 genes were contained by 16 different bins (Supporting Information Table S1), all but one of which were taxonomically annotated as TM6.

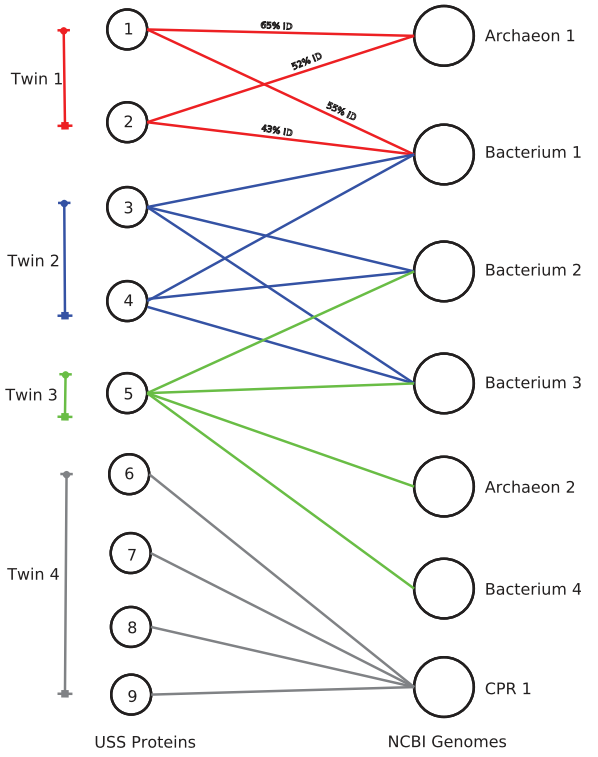



**Fig. 1.** The process of defining 'twins' in a BGA delineates groups of sequences with shared, exclusive similarity to a given set of genomes. For example, sequences 1 and 2 belong to a twin because they exclusively associate with the same set of genomes (Archaeon 1 and Bacterium 1). Note that genomes can be included in two or more different twins—Twin 2 also contains Bacterium 1 but involves a different set of proteins (3 and 4). In this case, sequences 2 and 3 show similarity to different genes within Bacterium 1 (i.e., they are not homologous). Twin 3 is an example of a twin where one or several CPR/TM6 sequences associate with many different genomes. Twin 4, in which multiple CPR/TM6 sequences associate exclusively with one genome, is an example of an interesting case that can allow attribution of CPR/TM6 sequences to a particular species (when the contained genome is a CPR/TM6 bacterium) or can hint at patterns of gene transfer or novel diversity (when contained genome is not a CPR/TM6 bacterium). Each edge between sequences and genomes has a corresponding weight, or percent identity (see Twin 1 for example), which were calculated from the BLAST results.

### *Patterns of gene similarity in the TM6*

Our results indicate that as many as 16 ultrasmall organisms in the Brown *et al.* dataset have genes with exclusive similarity to those in *Babela massiliensis*. This is interesting for two reasons: First, these relationships may help to begin constructing more detailed phylogeny among the TM6, which to date remains mostly unstudied. Specifically, that a set of genes among novel TM6 representatives resembles *Babela* (or one of its relatives) adds to existing

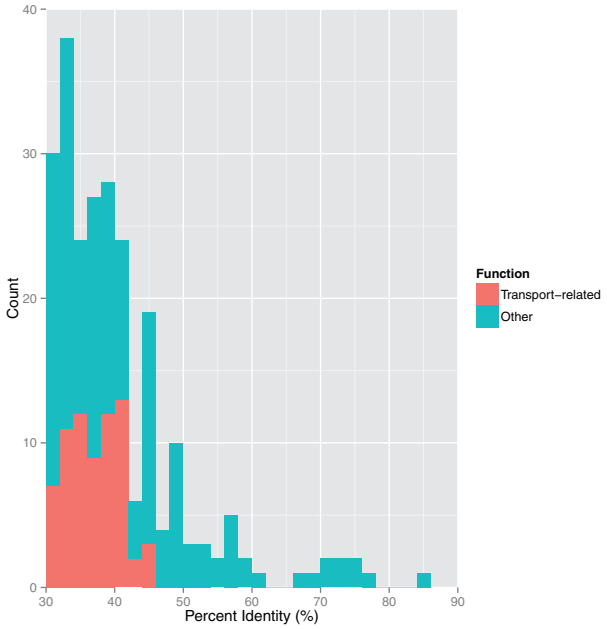



**Fig. 2.** The distribution of percent identity between CPR/TM6 proteins and their counterparts in *Babela massiliensis*, with highly divergent membrane, transporter, pump, and translocase-related ones highlighted in red. While the BGA twin relating the CPR/TM6 to this TM6 bacterium contained 456 proteins, only 236 could be annotated by RPS Blast.

knowledge of gene ancestry in this group. However, because only several complete TM6 genomes were included in our analysis, this relationship remains relative and may change as additional genomes in this phylum are discovered, described, and analyzed. Second, our results could help to shed light on the genomic consequences of ultrasmall size in the TM6. These organisms, based on the size fraction from which they were collected, would be 6–10 times smaller than their relative *Babela* (Pagnier *et al.*, 2015). Despite this, our twin analysis clearly confirms that small and ultrasmall TM6 have many related genes, some of which appear divergent.

Interestingly, *Babela* exhibits many of the genomic characteristics typical of intracellular symbionts, including reduction of genome size through loss of biosynthetic pathways (Pagnier *et al.*, 2015). In addition, *Babela* contains many genes related to transport, including ATP/ADP translocases, porins, an ABC-family permease, and other transporters (Pagnier *et al.*, 2015). These membrane-associated proteins could be important in the integration of metabolisms at the host-endosymbiont interface—in eukaryotes, specialized transporters currently play a role in moving small molecules across the inner envelope membrane of chloroplasts, connecting cytosolic and organellar pathways (Weber and Fisher, 2007). Interestingly, we recovered highly divergent versions of some of these same genes, among others, in the BGA twin associated with *Babela*—in particular, 17 amino acid transporters, 20 ATP/ADP and preprotein translocases, 18 multidrug pump/transporters, and several other related genes in the CPR/TM6 (Table 1). These membrane-related genes were related to that of *Babela* but with a low identity (~37% mean % ID, Fig. 2), and were contained in 14 bins also belonging to uncharacterized TM6 organisms. At any rate, further work should address the possibility that the highly divergent transport proteins recovered among the environmental TM6 play a role in adapting to a lifestyle in the ultrasmall size fraction. This lifestyle may not necessarily be parasitic, although recent work has indicated that this mode is likely both common and ancestral among the TM6 clade (Gong *et al.*, 2014; Yeoh *et al.*, 2015).

### *Lateral gene transfer between CPR/TM6 and other prokaryotes*

We repeated the BGA for the ARC data partition, again sorting the resulting twins by decreasing number of CPR/TM6 proteins they contained. This yielded three top twins, each of which linked sequences to a single archaeal genome—Woesearchaeota AR20 and Diapherotrites AR10, two ultrasmall size-fraction archaea from the superphylum DPANN (Rinke *et al.*, 2013; Castelle *et al.*, 2015), and Lokiarchaeum (Spang *et al.*, 2015)—with which there were 230, 131, and 53 exclusively associated proteins, respectively. We subsequently created a heatmap showing the distribution of sequence similarity between CPR/TM6 sequences in the ARC subset and genes within the complete archaeal genomes from NCBI (Fig. 3). This revealed further regions of interest.

First, we did not observe a pattern of high similarity (>70% ID) between CPR/TM6 proteins in the ARC subset and any archaeal genes, indicating that recent interdomain gene transfer is an unlikely explanation for the presence of numerous CPR/TM6 homologs in Archaea. However, the heatmap did reveal a “core group” of 62 CPR genes that showed distant homology (mean ~39% ID) to a large distribution of the complete archaeal genomes (Region A in Fig. 3). Region C and D generally corresponded to the two twins identified as top results in the BGA, involving the novel archaeal genomes Diapherotrites AR10 and Woesearchaeota AR20 (39–40% ID). Interestingly, these genomes were assembled from the same sample site and size fraction as the CPR dataset as part of a larger study identifying new members of the DPANN (Brown *et al.*, 2015; Castelle *et al.*, 2015). Additionally, the two groups of CPR/TM6 sequences associated with these genomes showed similar functional profiles, containing many divergent, membrane-related proteins (Table 2).

To determine whether these archaea-exclusive signals stemmed from inaccurate binning (and may therefore reflect that some contigs belong to archaea rather than

**Table 1.** Functional description of 236 CPR/TM6 proteins associating exclusively with *Babela massiliensis*, as annotated by RPS Blast. Highly divergent membrane, transporter, pump, and translocase-related proteins are in marked in bold face.

| Count | COG | Annotation |
|---|---|---|
| **17** | **COG0531** | **Amino acid transporters** |
| 14 | COG0612 | Predicted Zn-dependent peptidases |
| **13** | **COG3202** | **ATP/ADP translocase** |
| **12** | **COG0534** | **Na+-driven multidrug efflux pump** |
| 11 | COG0265 | Trypsin-like serine proteases, typ. perip. contain C-term PDZ dom. |
| 10 | COG0285 | Folylpolyglutamate synthase |
| 10 | COG2932 | Predicted transcriptional regulator |
| 9 | COG0544 | FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor) |
| 9 | COG0592 | DNA polymerase sliding clamp subunit (PCNA homolog) |
| **9** | **COG4775** | **Outer membrane protein/protective antigen OMA87** |
| 8 | COG2812 | DNA polymerase III, gamma/tau subunits |
| 6 | COG0681 | Signal peptidase I |
| **6** | **COG0706** | **Preprotein translocase subunit YidC** |
| **6** | **COG1132** | **ABC-type multidrug transport system, ATPase permease comps.** |
| 6 | COG1524 | Uncharacterized proteins of the AP superfamily |
| 5 | COG0712 | F0F1-type ATP synthase, delta sub. (mito. oligomycin sens. prot.) |
| 4 | COG3264 | Small-conductance mechanosensitive channel |
| 3 | COG0333 | Ribosomal protein L32 |
| 3 | COG0456 | Acetyltransferases |
| 3 | COG0596 | Pred. hydrolases/acyltransferases (alpha/beta hydrolase superf.) |
| 3 | COG0607 | Rhodanese-related sulfurtransferase |
| 3 | COG0636 | F0F1-type ATP synth. sub. c/Arch./vacuolar-type H+-ATPase, sub K |
| 3 | COG1011 | Predicted hydrolase (HAD superfamily) |
| 3 | COG1214 | Inactive homolog of metal-dep. proteases, putative mol. Chaperone |
| 3 | COG2165 | Type II secretory pathway, pseudopilin PulG |
| 3 | COG3031 | Type II secretory pathway, component PulC |
| 3 | COG3283 | Transcriptional regulator of aromatic amino acids metabolism |
| 3 | COG4972 | Tfp pilus assembly protein, ATPase PilM |
| 2 | COG0037 | Pred. ATPase of the PP-loop superf. implicated in cell cycle control |
| 2 | COG0200 | Ribosomal protein L15 |
| 2 | COG0224 | F0F1-type ATP synthase, gamma subunit |
| 2 | COG0355 | F0F1-type ATP synthase, epsilon sub. (mitochondrial delta subunit) |
| 2 | COG0360 | Ribosomal protein S6 |
| 2 | COG1974 | SOS-response trans. repressors (RecA-mediated autopeptidases) |
| 2 | COG2204 | Response reg. w/CheY-like receiver, ATPase, & DNA-bind. Doms |
| 2 | COG2267 | Lysophospholipase |
| 2 | COG3688 | Predicted RNA-binding protein containing a PIN domain |
| 2 | COG4564 | Signal transduction histidine kinase |
| **2** | **COG4591** | **ABC-type transport sys., inv. in lipop. release, permease comp.** |
| 1 | COG0006 | Xaa-Pro aminopeptidase |
| 1 | COG0204 | 1-acyl-sn-glycerol-3-phosphate acyltransferase |
| 1 | COG0269 | 3-hexulose-6-phosphate synthase and related proteins |
| 1 | COG0331 | (acyl-carrier-protein) S-malonyltransferase |
| 1 | COG0356 | F0F1-type ATP synthase, subunit a |
| 1 | COG0419 | ATPase involved in DNA repair |
| 1 | COG0545 | FKBP-type peptidyl-prolyl cis-trans isomerases 1 |
| 1 | COG0666 | FOG: Ankyrin repeat |
| 1 | COG0707 | UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase |
| 1 | COG0758 | Pred. Rossmann fold nucl.-binding protein involved in DNA uptake |
| 1 | COG0793 | Periplasmic protease |
| 1 | COG0858 | Ribosome-binding factor A |
| 1 | COG1221 | Trans. Regs. w/AAA-type ATPase domain & DNA-binding dom |
| 1 | COG1222 | ATP-dependent 26S proteasome regulatory subunit |
| **1** | **COG1297** | **Predicted membrane protein** |
| **1** | **COG1314** | **Preprotein translocase subunit SecG** |
| 1 | COG1450 | Type II secretory pathway, component PulD |
| **1** | **COG1463** | **ABC-type tranp. sys. Inv. in resisting org. solvents, peripl. comp.** |
| 1 | COG1544 | Ribosome-associated protein Y (PSrp-1) |
| 1 | COG1579 | Zn-ribbon protein, possibly nucleic acid-binding |
| 1 | COG1723 | Uncharacterized conserved protein |
| 1 | COG3027 | Uncharacterized protein conserved in bacteria |
| 1 | COG3829 | Trans. regulator w/PAS, AAA-type ATPase, & DNA-binding dom. |
| 1 | COG4232 | Thiol:disulfide interchange protein |
| **1** | **COG4907** | **Predicted membrane protein** |
| 1 | COG4970 | Tfp pilus assembly protein FimT |

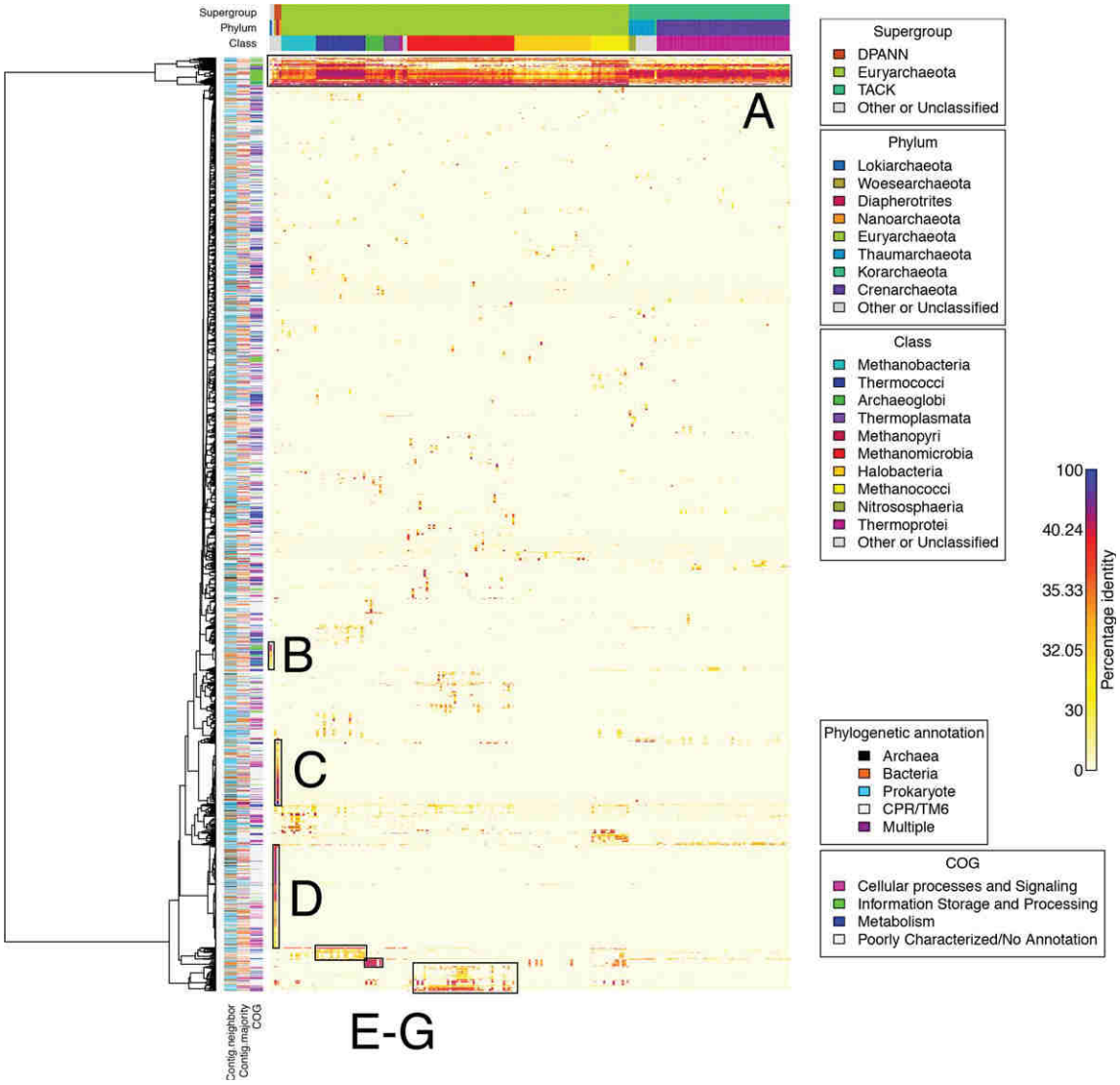



**Fig. 3.** A heatmap showing patterns of similarity between the CPR/TM6 proteins contained in the ARC subset and the archaeal genomes retrieved from NCBI. The percent identities shown were calculated from the BLAST hits between CPR/TM6 proteins and their corresponding proteins in the NCBI archaeal genomes. Taxonomic information for these genomes and genomic context/COG info for the CPR/TM6 proteins are shown in the heatmap sidebars (see Procedures and Supporting Information Methods).

CPR/TM6 organisms) or from interdomain gene transfer, we retrieved the number and taxonomy of genomic bins containing the genes in each twin. For the twin corresponding to Woesearchaeota AR20, 230 sequences were contained in 156 bins; for that corresponding to Diapherotrites AR10, 131 sequences were contained in 112 bins; for that corresponding to Lokiarchaeota, 53 sequences were contained in 52 bins; for that corresponding to the "core group" (Region A, Fig. 3), 62 sequences were contained in 52 bins; 90% or more of the genes in these twins were identified as belonging to the CPR phyla *Microgenomates* or *Parcubacteria*; a small number were of *Berkelbacteria*, *Peregrinibacteria*, or other origin (Supporting Information Table S1). Most were unannotated at the class level. We also examined the sequences at a contig level, retrieving the most frequent BLAST-assigned taxonomic annotations on each of the contigs containing twins with exclusive similarity to archaeal genes. These results showed that very few of these genes (<5% of each twin) were from contigs that met the majority rule for ARC placement. 72% or more of these contigs (containing genes exclusively similar to archaea) met the majority rule for bacterial (BAC) or CPR (no BLAST match to other bacteria, CPR, or archaea) origin (Supporting Information Table S1). Thus, while semiautomatic taxonomic assignments are limited, and contamination by low abundance,

**Table 2.** Functional description of CPR/TM6 proteins associating exclusively with singular archaeal genomes –Diapherotrites AR10, Woesearchaeota AR20, and Lokiarchaeum, respectively, as annotated by RPS Blast. COG annotations shared across genome groups ("twins") are marked in bold face.

| Count | COG | Annotation |
|---|---|---|
| Diapherotrites AR10 | | |
| 8 | COG4095 | Uncharacterized conserved protein |
| 5 | COG0785 | Cytochrome c biogenesis protein |
| **3** | **COG1651** | **Protein-disulfide isomerase** |
| 2 | COG0526 | Thiol-disulfide isomerase and thioredoxins |
| 2 | COG1378 | Predicted transcriptional regulators |
| 2 | COG1522 | Transcriptional regulators |
| 2 | COG2267 | Lysophospholipase |
| 1 | COG0089 | Ribosomal protein L23 |
| 1 | COG0189 | Glutathione synth/Ribo. Prot. S6 mod enzyme |
| **1** | **COG0438** | **Glycosyltransferase** |
| 1 | COG0451 | Nucleoside-diphosphate-sugar epimerases |
| 1 | COG0500 | SAM-dependent methyltransferases |
| 1 | COG1032 | Fe-S oxidoreductase |
| 1 | COG1305 | Transglutaminase-like enzymes, putative cysteine proteases |
| 1 | COG1418 | Predicted HD superfamily hydrolase |
| 1 | COG1577 | Mevalonate kinase |
| **1** | **COG2226** | **Methylase involved in ubiquinone/menaquinone biosynthesis** |
| 1 | COG2230 | Cyclopropane fatty acid synthase and related methyltransferases |
| 1 | COG2717 | Predicted membrane protein |
| 1 | COG3118 | Thioredoxin domain-containing protein |
| 1 | COG4106 | Transaconitate methyltransferase |
| 1 | COG5542 | Predicted integral membrane protein |
| 1 | COG5650 | Predicted integral membrane protein |
| Woesearchaeota AR20 | | |
| 22 | COG1215 | Glycosyltransferases, probably involved in cell wall biogenesis |
| **9** | **COG0438** | **Glycosyltransferase** |
| **6** | **COG2226** | **Methylase involved in ubiquinone/menaquinone biosynthesis** |
| 5 | COG4243 | Predicted membrane protein |
| 3 | COG0463 | Glycosyltransferases involved in cell wall biogenesis |
| 3 | COG2244 | Membrane protein involved in export of O-antigen/teichoic acid |
| 3 | COG2510 | Predicted membrane protein |
| 2 | COG2511 | Archaeal Glu-tRNAGln amidotrans. Sub. E (contains GAD domain) |
| 2 | COG3177 | Uncharacterized conserved protein |
| 1 | COG0262 | Dihydrofolate reductase |
| 1 | COG0719 | ABC-type trans. sys. inv. in Fe-S cluster assem., permease comp. |
| 1 | COG1216 | Predicted glycosyltransferases |
| 1 | COG1414 | Transcriptional regulator |
| 1 | COG1437 | Adenylate cyclase, class 2 (thermophilic) |
| **1** | **COG1651** | **Protein-disulfide isomerase** |
| 1 | COG1669 | Predicted nucleotidyltransferases |
| 1 | COG1814 | Uncharacterized membrane protein |
| 1 | COG1898 | dTDP-4-dehydrorhamnose 3,5-epimerase and related enzymes |
| 1 | COG2129 | Predicted phosphoesterases, related to the Icc protein |
| 1 | COG2259 | Predicted membrane protein |
| 1 | COG2887 | RecB family exonuclease |
| 1 | COG3882 | Predicted enzyme involved in methoxymalonyl-ACP biosynthesis |
| Lokiarchaeum | | |
| 14 | COG1449 | Alpha-amylase/alpha-mannosidase |
| 12 | COG0064 | Asp-tRNAAsn/Glu-tRNAGln amidotrans. B sub. (PET112 hom.) |
| 6 | COG0178 | Excinuclease ATPase subunit |
| 4 | COG0642 | Signal transduction histidine kinase |
| 3 | COG3259 | Coenzyme F420-reducing hydrogenase, alpha subunit |
| 2 | COG0084 | Mg-dependent DNase |
| 1 | COG0125 | Thymidylate kinase |
| 1 | COG0171 | NAD synthase |
| 1 | COG0183 | Acetyl-CoA acetyltransferase |
| 1 | COG0322 | Nuclease subunit of the excinuclease complex |
| 1 | COG0334 | Glutamate dehydrogenase/leucine dehydrogenase |
| 1 | COG0674 | Pyruvate:ferredoxin oxidoreductase, related oxidored., alpha sub. |
| 1 | COG0714 | MoxR-like ATPases |
| 1 | COG1042 | Acyl-CoA synthetase (NDP forming) |
| 1 | COG1690 | Uncharacterized conserved protein |

highly fragmented genomes is possible, wide-spread sequence misbinning in the CPR dataset seems unlikely. Furthermore, contigs containing genes with exclusive similarity to archaeal ones appeared to be of generally high quality—only several had a coverage below 5, with most higher (median coverage between 9 and 11, depending on the twin). Finally, only three genes in the examined twins were labeled as "Possibly archaeal contamination" in the original study. Overall, our contig majority analyses revealed that sequence binning was likely accurate.

Thus, we observed multiple twins exclusively associating CPR sequences to genes in a variety of archaea, including those from both major taxonomic groups as well as novel, ultrasmall DPANN. In other words, numerous ultrasmall bacteria presented genes exclusively similar to those of archaeal groups in their genomes. Ultimately, the results of the BGA, when combined with the binning and contig analyses, suggest that this similarity between CPR and known prokaryotic genomes may be the result of multiple interdomain LGT between these organisms. For one, an ancestor of Woesearchaeota AR20 and an ancestor of Diapherotrites AR10 could have exchanged genes with members of the CPR, helping in part to explain the observed patterns of similarity in Region C and D (Fig. 3). An RPSBlast of genes in these regions revealed that many coded for proteins relating to the membrane—integral proteins, cytochrome biogenesis, methylase involved in ubiquinone/menaquinone biosynthesis, and glycosyltransferases involved in cell wall biogenesis (Table 2). Common in certain ultrasmall archaeal genomes these glycosyltransferases are predicted to play a key role in synthesizing structural and signaling saccharides (Castelle *et al.*, 2015).

Under the hypothesis of LGT between CPR and archaea, the large "core" band of similarity seen across all groups in the ARC heatmap (Region A, Fig. 3) is surprising, as it includes more conserved archaeal genes like those in information storage and processing. Indeed, an RPS Blast analysis indicated that Region A included many genes that coded for ribosomal proteins, DNA polymerase, and tRNA synthethases (green on COG sidebar, also see Supporting Information Table S2). Several of these synthethases appeared to show higher homology with thermophilic archaeal classes, whereas some ribosomal genes were restricted to members of the phylum Euryarchaeota. This pattern may indicate ancient gene exchange involving CPR and some broad distributions of relatively large, varied archaea. However, it may also reflect an ancient phylogenetic relationship between CPR and archaea, if CPR are indeed relatively basal in the prokaryotic tree of life as suggested by a recent concatenation of 16 ribosomal proteins (Hug *et al.*, 2016). Nonetheless, there are several other regions apparent on the heatmap with exclusive above-threshold homology of CPR/TM6 genes with particular classes of Archaea, for example, Region E (Fig. 3) with varied similarity to members of Thermococci, Region F (Fig. 3), with ~40% similarity to the members of Archaeglobi, or Region G (Fig. 3), with varied similarity to members of Methanomicrobia. These other patterns of similarity strengthen the suggestion that ancient gene transfer may have occurred among members of the ultrasmall size fraction.

Lastly, we also recovered a large group of CPR/TM6 proteins ($n = 53$) that showed distant homology exclusively with Lokiarchaeum, which is already known to have a proteome nearly 30% homologous with bacteria (Region B, Fig. 3; Spang *et al.*, 2015). As above, these genes were placed in quality bins of diverse bacterial origin, and so interdomain gene transfer with a relative of Lokiarchaeota is a possible explanation for the observed pattern of similarity. However, the functional profile of these CPR/TM6 genes was largely different from that of the genes matching with AR10 and AR20. Genes shared exclusively between CPR/TM6 and Lokiarchaeota were composed mostly of amidotransferases involved in tRNA biosynthesis and a family of enzymes involved in carbohydrate metabolism, but lacking membrane-related genes (Table 2). We can only speculate that these different functional patterns may hint at different gene-capture mechanisms among archaea. Lokiarchaeota, if phagotrophic, could prey on a diversity of ultrasmall bacteria, while AR20/AR10 may be involved in symbiotic relationships with CPR. This could then lead to the convergent sharing of membrane-related genes compatible with such a lifestyle.

The observed results for the ARC subset are consistent with literature suggesting that ancient gene transfer from bacteria to archaea can play a major role in evolution of specific lineages (Nelson *et al.*, 1999; Lopez-García *et al.*, 2015; Nelson-Sathi *et al.*, 2015). In the striking case of the Haloarchaea, as many as 157 gene families coding for transporters were imported from Eubacteria (Nelson *et al.*, 1999). These transfers can facilitate colonization of new niche space, for example, Lopez-García *et al.* (2015) details the convergent acquisition of metabolism, transport, and membrane genes allowing adaptation to mesophilic conditions among three distant archaeal lineages. Ancient transfer of metabolic genes from bacteria to archaea has also been implicated in the origin of several major archaeal groups (Nelson-Sathi *et al.*, 2015). While polarity of any CPR/TM6-Archaea gene transfers in this dataset would be difficult to determine, transfer events among these domains are generally believed to be skewed towards those in which bacteria act as donors (Lopez-García *et al.*, 2015; Nelson-Sathi *et al.*, 2015). This may be due to adaptive gains made by use of new metabolic strategies and a lower fitness cost to archaea of incorporating foreign genetic material (Lopez-García *et al.*, 2015). Transfer of membrane-related genes could also be achieved endosymbiotically, where the symbiont (by lysis or another process) donates genes to the host. In fact, this scenario

has been suggested in the literature—as a possible step in the retention of the mitochondrial progenitor during early eukaryogenesis (Martin *et al*., 2015), or as a mechanism to regulate the cell wall of an intracellular bacterium (Husnik *et al*., 2013). Although LGT of membrane transporters was observed primarily between ultra-small donors and recipients (CPR, DPANN), we speculate that, in the case of a hypothetical CPR/TM6-large archaeon symbiosis, transfer and subsequent expression of the symbiont's transporters or other membrane-related genes could be critical.

Finally, graph analyses of the BAC and PROK subsets provided several other examples of more recent transfer between the CPR/TM6 and larger prokaryotes. We ran BGA analyses maintaining an 80% cover requirement between CPR/TM6 sequences and their homologs, but for the PROK subset generated gene families with more stringent percent identity (>=50, >=60, >=70, >=80, >=90% ID, see Supporting Information Methods). These gene families would later be detected as twin members. At 80% similarity, we observed that CPR/TM6 mannose isomerase genes paired with both genomes of methanogenic archaea like *Methanosarcina* and with complete CPR genomes (Supporting Information Table S3). We also observed several sets of CPR/TM6 genes associated with single deltaproteobacterial genomes, for example, a set of recombinases with *Hippea maritima* and a set of GTP-binding protein TypA/BipA with *Desulfomonile tiedjei*. Likewise, at 90% ID, a set of transcriptional regulators from the CPR/TM6 showed homology to a wide array of *Bacillus* genomes (these genes also showed similarity to archaea, just at a lower threshold). These patterns may indicate that CPR/TM6 have also exchanged genes with other bacteria, and expand upon Brown *et al*. (2015) proposal of a possible ribosomal protein transfer among members of the CPR.

## Conclusion

Recent phylogenetic analyses have underscored the importance of studying ultrasmall microbial groups like CPR in expanding our knowledge of the tree of life (Hug *et al*., 2016). The patterns of genetic diversity and gene transfer reported in the present study contribute to this body of knowledge and bring forward a reticulate aspect of their evolution. Methods complementary to environmental metagenomics, like single cell genomics, could help to better elucidate relationships among organisms and their gene content (Stepanauskas, 2012) and ultimately shed additional light on patterns of transfer among these organisms. Furthermore, as we report the unusual membranes in a second domain of life (Castelle *et al*., 2015), we propose that these characteristics may be the result of a convergent evolutionary pressure. The ultrasmall niche may require underappreciated membrane adaptations, and further work should address the role of these proteins in adapting to or managing this lifestyle. Future analysis of massive environmental datasets from this size fraction, like that of TARA Oceans (Karsenti *et al*., 2011), could help to shed more light on gene transfer and phylogeny in these organisms and ultimately further our understanding of any drivers underlying their evolution.

## Procedures

We downloaded the full dataset of CPR/TM6 proteins from the online repository (ggkbase.berkeley.edu/CPR-complete-draft/organisms) listed in Brown *et al*. (2015). We then removed sequences with mid-protein stop codons, leaving a final dataset of 637,155 proteins. We also downloaded all proteins from all complete archaeal and bacterial genomes on NCBI (4,600 genomes, 15,373,158 sequences, February 1, 2016). This NCBI dataset included the eight complete CPR genomes from the Brown *et al*. (2015) dataset but not the ~800 other draft genomes also reported in that study. Full taxonomy information for the complete genomes was retrieved from the NCBI taxonomy database (ncbi.nlm.nih.gov/taxonomy). We performed a BLAST analysis of all CPR/TM6 proteins against all proteins from the complete genomes on a distributed cluster (version 2.3.0+, with the following options: -seg yes, -soft_masking true, and -max_target_seqs 5000). We filtered these results for sequence hits $\geq$ 30% identity, $\geq$ 80% mutual cover, and e-value $\leq$ 1e-5 to retain only full sized homologs of CPR/TM6 proteins in complete prokaryotic genomes. We partitioned the CPR/TM6 proteins into ARC, BAC, PROK, and CPR/TM6 groups as explained above, de-replicating each set using cd-hit (version 4.6, -c 1 –s 1; Li and Godzik, 2006) to yield only unique CPR/TM6 sequences. PROK CPR/TM6 genes were further clustered into gene families (Supporting Information Methods).

We performed a BGA on the BLAST results for each subset, delineating groups of CPR/TM6 proteins with shared, exclusive similarity to a given set of prokaryotic genomes (Corel *et al*., 2016). This procedure defines "twins" composed of the CPR/TM6 sequences and the NCBI genomes hosting homologs of these sequences (Fig. 1). Twins were sorted on the number of included CPR/TM6 proteins and were filtered to retain those with low numbers of included NCBI genomes, as these allowed us to look more easily for candidate gene transfers. Recent gene transfer among the PROK and BAC subsets was detected using a BGA with higher identity thresholds (i.e., to be included in a twin, a link between a CPR/TM6 protein and a gene in an NCBI genome must be of >=50, >=60, >=70, >=80, or >=90 percent identity). For each CPR/TM6 protein in the ARC subset, we retrieved the identity of its home contig from the original sequence metadata and used this to retrieve all other sequences, regardless of

annotation, for those contigs. From these data, we created the "contig majority" parameter, or the highest frequency annotation on that contig among ARC, PROK, BAC, or CPR/TM6, and "contig neighbor," the annotations of the genes flanking the ARC gene on that contig (Supporting Information Methods). For the ARC, PROK, and *Babela*-associated CPR/TM6 genes, we performed an RPS BLAST (version 2.3.0+, with the following options: -seg yes, -soft_masking true and -max_target_seqs 5) with the NCBI COG database (ncbi.nlm.nih.gov/COG/) to obtain full gene annotations and COG categories. Finally, bin analyses were performed for the relevant gene subsets by retrieving the original sequence headers from Brown *et al*. (2015) and extracting bin/taxonomy information.

### Author contributions

All authors designed the study. E.C., A.L.J., and J.S.P. performed the bioinformatic analyses; A.L.J. and E.B. wrote and revised the manuscript. All authors discussed results and commented on the manuscript.

### Acknowledgements

We thank two anonymous reviewers for their insightful and constructive comments. We also thank Raphaël Méheust and Adrien Danzon for their aid in designing bioinformatic analyses. E.C., J.S.P., and E.P. were funded by the European Research Council (FP7/2007-2013 Grant Agreement 615274) and A.L.J. by the Alex G. Booth Traveling Scholarship and the Benjamin Franklin Travel Grant.

### Data availability

Results of the bipartite graph analyses will be made available at http://www.evol-net.fr/index.php/en/downloads. Sequence files available upon request.

### References

Anantharaman, K., Brown, C.T., Burstein, D., Castelle, C.J., Probst, A.J., Thomas, B.C., *et al*. (2016) Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ* **4:** e1607.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., *et al*. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523:** 208–211.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., *et al*. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* **25:** 690–701.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol* **24:** 224–227.

Gong, J., Qing, Y., Guo, X., and Warren, A. (2014) "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). *Syst Appl Microbiol* **37:** 35–41.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., *et al*. (2016) A new view of the tree of life. *Nat Microbiol* 16048.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., *et al*. (2013) Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell* **153:** 1567–1578.

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., *et al*. (2011) A holistic approach to marine ecosystems biology. *PLoS Biol* **9:** e1001177.

Koonin, E.V. (2015) Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biol* **13:** 84.

Lake, J.A. (2009) Evidence for an early prokaryotic endosymbiosis. *Nature* **460:** 967–971.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658–1659.

López-García, P., Zivanovic, Y., Deschamps, P., and Moreira, D. (2015) Bacterial gene import and mesophilic adaptation in archaea. *Nat Rev Microbiol* **13:** 447–456.

Luef, B., Frischkorn, K.R., Wrighton, K.C., Holman, H.Y.N., Birarda, G., Thomas, B.C., *et al*. (2015) Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat Commun* **6:** 6372.

Martin, W.F., Garg, S., and Zimorski, V. (2015) Endosymbiotic theories for eukaryote origin. *Philos Trans R Soc Lond B Biol Sci* **370:** 20140330.

Nasir, A., Kim, K.M., and Caetano-Anollés, G. (2015) Lokiarchaeota: eukaryote-like missing links from microbial dark matter? *Trends Microbiol* **23:** 448–450.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* **399:** 323–329.

Nelson-Sathi, S., Sousa, F.L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., *et al*. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517:** 77–80.

Pagnier, I., Yutin, N., Croce, O., Makarova, K.S., Wolf, Y.I., Benamar, S., *et al*. (2015) Babela massiliensis, a representative of a widespread bacterial phylum with unusual adaptations to parasitism in amoebae. *Biol Direct* **10:** 1–17.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431–437.

Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., *et al*. (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521:** 173–179.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15:** 613–620.

Swithers, K.S., Fournier, G.P., Green, A.G., Gogarten, J.P., and Lapierre, P. (2011) Reassessment of the lineage fusion hypothesis for the origin of double membrane bacteria. *PloS One* **6:** e23774.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., and Martin, W. (2004) Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* **5:** 123–135.

Velimirov, B. (2001) Nanobacteria, Ultramicrobacteria and Starvation Forms: A Search for the Smallest Metabolizing Bacterium. *Microbes Environ* **16:** 67–77.

Weber, A.P. and Fischer, K. (2007) Making the connections–the crucial role of metabolite transporters at the interface between chloroplast and cytosol. *FEBS Lett* **581:** 2215–2222.

Yeoh, Y.K., Sekiguchi, Y., Parks, D.H., and Hugenholtz, P. (2015) Comparative genomics of candidate phylum TM6 suggests that parasitism is widespread and ancestral in this lineage. *Mol Biol Evol.* **33:** 915–927.

## Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site.

**Fig. S1.** Using multi-step BLAST to create 'gene families'.

**Table S1.** Characteristics of each 'twin' or gene grouping described in the text, including the number of genes in each twin/grouping, the number of unique genomic bins associated with the twin/grouping and their taxonomic composition, and the number of unique contigs associated with the twin/grouping and their BLAST-assigned phylogenetic annotations.

**Table S2.** Functional description and contig majority/neighbor information for 62 CPR/TM6 proteins in Region A of Fig. 3, matching exclusively with archaeal genomes. A contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent'.

**Table S3.** An exemplar 'twin' from the PROK BGA analysis at $\geq$ 80% identity. Each CPR/TM6 protein listed in the first table shows exclusive similarity at this threshold with a gene in all of the genomes listed in the second table, which includes both archaea and a member of the CPR. This pattern suggests possible inter-domain gene transfer. Also noted is the BLAST-assigned phylogenetic annotation for the neighbor of each gene, and the most frequent annotation on the contig that contains it. As above, a contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent'.

SUPPORTING INFORMATION

## 1. METHODS

### 1.1 Gene Family Construction

To create 'gene families,' we performed an all-against-all BLAST analysis of the PROK dataset (81,634 proteins) (version 2.3.0+, with the following options : -seg yes, -soft_masking true and -max_target_seqs 5000), in effect creating a sequence similarity network (Corel *et al.,* 2016**).** 'Gene families' are isolated groups of input sequences that are connected by BLAST hits of ≥ 30% sequence identity, ≥ 80% mutual coverage, and an e-value ≤ 1e-5. In our analysis of the PROK dataset, gene families were also constructed at >=40, >=50, >=60, >=70, >=80, and >=90 percent identity[1]. In each iteration of that analysis, the threshold used to create gene families matched that used to create 'twins' with genes from the NCBI genomes.

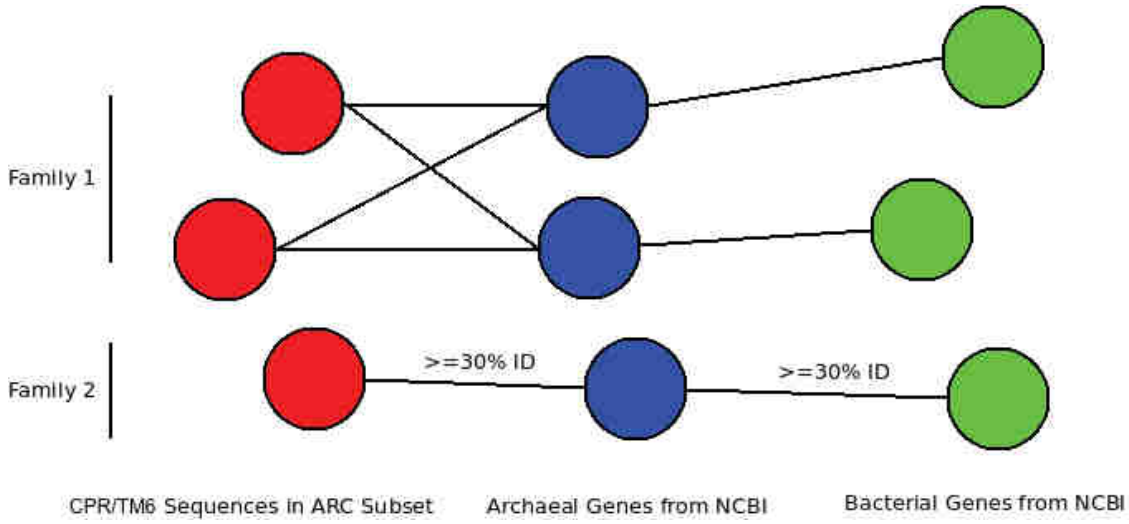
### 1.2 'Contig Neighbors'

After retrieving all genes on the home contig of each ARC gene, we could then assign an overall annotation to the set of genes directly flanking it. ARC genes with one or more 'PROK' neighbors or with an 'ARC' and 'BAC' neighbor were assigned the contig_neighbor designation 'PROK', those without 'PROK' but with one or more 'BAC' were assigned the contig_neighbor designation 'BAC', those without 'PROK' or 'BAC' but with one or more 'ARC' were assigned the contig_neighbor designation 'ARC'. Those with only 'CPR/TM6' kept that label (shortened to CPR in the tables below). This metric was used to provide genomic context for possible gene transfers among the CPR/TM6.

### 1.3 Tree Construction

We combined the CPR/TM6 sequences in the ARC subset with their homologous archaeal genes. We also performed a BLAST analysis of these archaeal homologs against the bacterial genomes previously taken from NCBI, allowing us to retrieve distant bacterial homologs of the CPR/TM6 sequences. Addition of these genes to the dataset resulted in 160,284 total sequences - the CPR/TM6 genes in the ARC grouping, their homologs in archaea, and any more distant bacterial homologs (see diagram below). We then computed 'gene families' (see above) and selected those that contained sequences from Region A of Fig. 3 (n=21 families). Many of these proteins were assigned to the "Information Storage and Processing" category (see Supp. Table 1). For each of these selected families, we aligned the contained sequences with MAFFT (v7.273, linsi setting), refined the alignments with Gblocks (v0.91b, -b4=6, -b5h), and generated trees with FastTree (v2.1.8 SSE3, default parameters).

[1] Méheust, Raphaël, et al. (2016) Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proceedings of the National Academy of Sciences: 201517551.




**Supp. Fig. 1:** Using multi-step BLAST to create 'gene families.'

2. TABLES

**Supp. Table 1.** Characteristics of each 'twin' or gene grouping described in the text, including the number of genes in each twin/grouping, the number of unique genomic bins associated with the twin/grouping and their taxonomic composition, and the number of unique contigs associated with the twin/grouping and their BLAST-assigned phylogenetic annotations.

| Gene Grouping/Twin | Contained Genes | Unique Genomic Bins | Bin Taxonomic Composition | Unique Contigs | Contig Majority Annotation |
|---|---|---|---|---|---|
| **Core' (Region A)** | 62 | 52 | 48% Microgenomates, 44% Parcubacteria, 8% Other | 56 | 73% CPR, 18% BAC, 2% ARC, 2% PROK, 5% Multiple |
| **Loki.-associated (Region B)** | 53 | 52 | 30% Microgenomates, 60% Parcubacteria, 10% Other | 53 | 40% CPR, 32% BAC, 17% PROK, 11% Multiple |
| **AR10-associated (Region C)** | 131 | 112 | 27% Microgenomates, 64% Parcubacteria, 9% Other | 123 | 52% CPR, 22% BAC, 15% PROK, 1% ARC, 10% Multiple |
| **AR20-associated (Region D)** | 230 | 156 | 27% Microgenomates, 66% Parcubacteria, 7% Other | 193 | 52% CPR, 30% BAC, 10% PROK, 1% ARC, 7% Multiple |
| ***Babela*-associated** | 456 | 16 | 99.8% TM6, 0.2% Other | 145 | n/a |

**Supp. Table 2.** Functional description and contig majority/neighbor information for 62 CPR/TM6 proteins in Region A of Fig. 3, matching exclusively with archaeal genomes. A contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent.'

| Gene | COG | Annotation | Contig Neigh. | Contig Majority |
|---|---|---|---|---|
| 5618 | COG1958 | Small nuc. Ribonucleo. (snRNP) homolog | PROK | CPR |
| 5621 | COG2007 | Ribosomal protein S8E | CPR | CPR |
| 11123 | COG0162 | Tyrosyl-tRNA synthetase | CPR | CPR |
| 38916 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 44794 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 71328 | COG1899 | Deoxyhypusine synthase | CPR | CPR |
| 71337 | COG3620 | Pred. trans. Reg. with C-terminal CBS domains | PROK | CPR |
| 71828 | COG1601 | Trans. Init. factor 2, beta sub. (eIF-2beta)/eIF-5 N-term. Dom. | CPR | CPR |
| 86106 | COG0162 | Tyrosyl-tRNA synthetase | PROK | BAC |
| 93815 | COG0162 | Tyrosyl-tRNA synthetase | CPR | BAC/CPR |
| 96811 | COG0008 | Glutamyl- and glutaminyl-tRNA synthetases | PROK | PROK |
| 100907 | COG0162 | Tyrosyl-tRNA synthetase | CPR | CPR |
| 102461 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 102977 | COG1608 | Predicted archaeal kinase | PROK | BAC |
| 109588 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 114255 | COG2019 | Archaeal adenylate kinase | BAC | CPR |
| 115832 | COG1056 | Nicotinamide mononuc. adenylyltransferase | CPR | CPR |
| 116018 | COG0162 | Tyrosyl-tRNA synthetase | CPR | CPR |
| 138094 | COG0162 | Tyrosyl-tRNA synthetase | BAC | BAC/CPR |
| 143243 | COG0162 | Tyrosyl-tRNA synthetase | CPR | BAC |
| 144048 | COG0162 | Tyrosyl-tRNA synthetase | BAC | CPR |
| 144459 | COG1632 | Ribosomal protein L15E | PROK | CPR |
| 156000 | COG0358 | DNA primase (bacterial type) | CPR | CPR |
| 157714 | COG0162 | Tyrosyl-tRNA synthetase | ARC | CPR |
| 165156 | COG1632 | Ribosomal protein L15E | PROK | CPR |
| 169053 | COG0162 | Tyrosyl-tRNA synthetase | BAC | CPR |
| 175411 | COG0162 | Tyrosyl-tRNA synthetase | CPR | CPR |
| 177135 | COG0468 | RecA/RadA recombinase | CPR | CPR |
| 197611 | COG0162 | Tyrosyl-tRNA synthetase | BAC | BAC |
| 207505 | COG1471 | Ribosomal protein S4E | CPR | CPR |
| 237384 | COG1056 | Nicotinamide mononuc. adenylyltransferase | CPR | CPR |
| 259717 | COG0052 | Ribosomal protein S2 | BAC | BAC |
| 260515 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 269612 | COG0162 | Tyrosyl-tRNA synthetase | CPR | BAC |
| 292544 | COG2023 | RNase P subunit RPR2 | CPR | CPR |
| 293093 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 334126 | COG1056 | Nicotinamide mononuc. adenylyltransferase | PROK | CPR |
| 335241 | COG0180 | Tryptophanyl-tRNA synthetase | PROK | CPR |

| | | | | |
|---|---|---|---|---|
| 345689 | COG0162 | Tyrosyl-tRNA synthetase | BAC | CPR |
| 350374 | COG0162 | Tyrosyl-tRNA synthetase | BAC | CPR |
| 358822 | COG0361 | Translation initiation factor 1 (IF-1) | ARC | CPR |
| 358823 | COG2412 | Uncharacterized conserved protein | PROK | CPR |
| 358825 | COG1601 | Trans. Init. factor 2, beta sub. (eIF-2beta)/eIF-5 N-term. dom. | BAC | CPR |
| 369776 | COG1019 | Predicted nucleotidyltransferase | CPR | CPR |
| 406402 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 406638 | COG0162 | Tyrosyl-tRNA synthetase | PROK | CPR |
| 459632 | COG0162 | Tyrosyl-tRNA synthetase | CPR | BAC |
| 462534 | COG1632 | Ribosomal protein L15E | CPR | BAC/CPR |
| 476790 | COG0162 | Tyrosyl-tRNA synthetase | BAC | CPR |
| 486443 | COG1571 | Pred. DNA-binding prot. w/ a Zn-ribbon dom. | PROK | BAC |
| 494502 | COG0638 | 20S proteasome, alpha and beta subunits | ARC | ARC |
| 494503 | COG1500 | Predicted exosome subunit | ARC | ARC |
| 497340 | COG2520 | Predicted methyltransferase | CPR | BAC |
| 527889 | COG2123 | RNase PH-related exoribonuclease | PROK | CPR |
| 585776 | COG1632 | Ribosomal protein L15E | CPR | CPR |
| 613765 | COG0592 | DNA poly. sliding clamp sub. (PCNA homolog) | CPR | CPR |
| 617704 | COG1899 | Deoxyhypusine synthase | BAC | CPR |
| 619736 | COG1056 | Nicotinamide mononuc. adenylyltransferase | PROK | BAC |
| 620955 | COG1056 | Nicotinamide mononuc. adenylyltransferase | PROK | CPR |
| 621607 | COG2125 | Ribosomal protein S6E (S10) | PROK | CPR |
| 623664 | COG1594 | DNA-dir. RNA poly. Sub. M/Trans. Elong. Fact. TFIIS | CPR | CPR |
| 625354 | COG1746 | tRNA nucleotidyltransferase (CCA-adding enzyme) | PROK | CPR |

**Supp. Table 3.** An exemplar 'twin' from the PROK BGA analysis at ≥ 80% identity. Each CPR/TM6 protein listed in the first table shows exclusive similarity at this threshold with a gene in all of the genomes listed in the second table, which includes both archaea and a member of the CPR. This pattern suggests possible inter-domain gene transfer. Also noted is the BLAST-assigned phylogenetic annotation for the neighbor of each gene, and the most frequent annotation on the contig that contains it. As above, a contig majority marker including a '/' indicates a contig where multiple gene types are 'most frequent.'

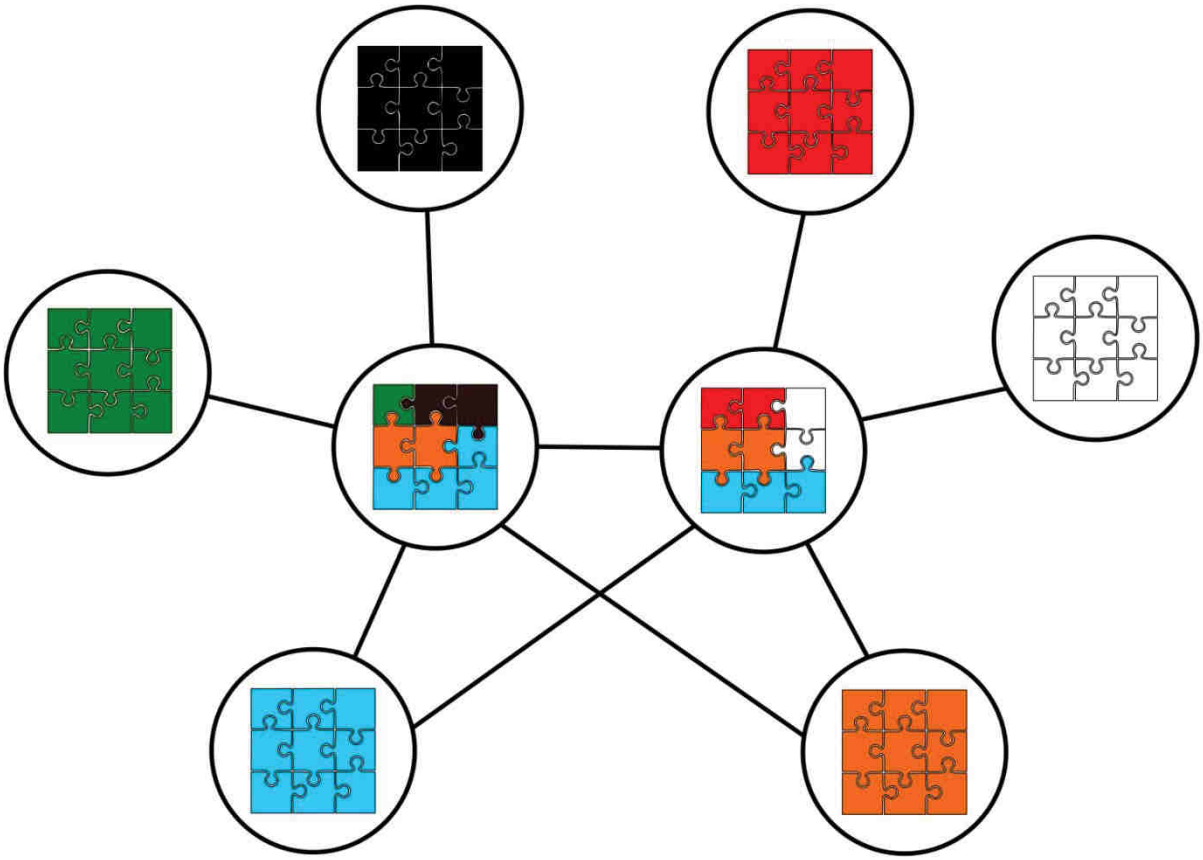
| Gene No. | COG | Annotation | Contig Neighbor | Contig Majority |
|---|---|---|---|---|
| 1535 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |
| 10266 | COG0662 | Mannose-6-phosphate isomerase | CPR | CPR |
| 35626 | COG0662 | Mannose-6-phosphate isomerase | BAC | CPR |
| 43945 | COG0662 | Mannose-6-phosphate isomerase | PROK | PROK |
| 44626 | COG0662 | Mannose-6-phosphate isomerase | BAC | CPR |
| 56961 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |
| 64503 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |
| 64889 | COG0662 | Mannose-6-phosphate isomerase | PROK | CPR |
| 65506 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |

| | | | | |
|---|---|---|---|---|
| 66119 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC/CPR |
| 92294 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |
| 164764 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC/PROK |
| 193855 | COG0662 | Mannose-6-phosphate isomerase | CPR | BAC |
| 194601 | COG0662 | Mannose-6-phosphate isomerase | CPR | CPR |
| 218473 | COG0662 | Mannose-6-phosphate isomerase | PROK | CPR |
| 223543 | COG0662 | Mannose-6-phosphate isomerase | BAC | BAC |
| 250887 | COG0662 | Mannose-6-phosphate isomerase | BAC | BAC |
| 266479 | COG0662 | Mannose-6-phosphate isomerase | BAC | BAC |
| 314490 | COG0662 | Mannose-6-phosphate isomerase | PROK | PROK |
| 338550 | COG0662 | Mannose-6-phosphate isomerase | CPR | BAC |
| 349979 | COG0662 | Mannose-6-phosphate isomerase | BAC | BAC/CPR |
| 370791 | COG0662 | Mannose-6-phosphate isomerase | ARC | PROK |
| 395094 | COG0662 | Mannose-6-phosphate isomerase | CPR | CPR |
| 399012 | COG0662 | Mannose-6-phosphate isomerase | PROK | BAC |
| 424197 | COG0662 | Mannose-6-phosphate isomerase | CPR | BAC |
| 470542 | COG0662 | Mannose-6-phosphate isomerase | CPR | CPR |
| 477729 | COG0662 | Mannose-6-phosphate isomerase | CPR | CPR |
| 489194 | COG0662 | Mannose-6-phosphate isomerase | CPR | BAC |
| 514095 | COG0662 | Mannose-6-phosphate isomerase | BAC | CPR |
| 517959 | COG0662 | Mannose-6-phosphate isomerase | PROK | PROK |
| 537423 | COG0662 | Mannose-6-phosphate isomerase | BAC | PROK |

Genome

*Methanosarcina mazei Go1*
*Methanosarcina acetivorans C2A*
*Methanolobus psychrophilus R15*
*Methanomethylovorans hollandica DSM 15978*
*Methanosarcina mazei Tuc01*
*Methanosarcina thermophila TM-1*
*Methanosarcina vacuolata Z-761*
*Methanosarcina thermophila CHTI-55*
*Methanosarcina sp. Kolksee*
*Methanosarcina sp. WWM596*
*Methanosarcina barkeri str. Wiesmoor*
*Methanosarcina sp. WH1*
*Methanosarcina barkeri MS*
*Methanosarcina sp. MTP4*
*Methanosarcina barkeri 227*
*Methanosarcina siciliae HI350*
*Methanosarcina siciliae C2J*
*Methanosarcina mazei WWM610*
*Methanosarcina mazei SarPi*
*Methanosarcina mazei S-6*
*Methanosarcina mazei LYC*
*Methanosarcina mazei C16*
*Methanosarcina barkeri 3*
*Methanosarcina barkeri CM1*
*Parcubacteria (Wolfebacteria) bacterium GW2011_GWB1_47_1*

# III. MODULARITY IN EVOLUTION



*Network of puzzles showing the relation between each entities. Nodes represent evolving modular objects, an edge is drawn between two nodes if the corresponding objects share components.*

# III.1 MOLECULAR EVOLUTION

The paradigm of evolution of modular proteins could be expressed as follows: there exists a limited repertoire of domains from which the set of current proteins have been formed (Gilbert 1978; Patthy 1999; Apic et al. 2001; Bashton and Chothia 2002; Chothia et al. 2003). New genes arise via molecular mechanisms (Figure 11) such as duplication of pre-existing genes, shuffling of genetic fragments, fusion and fission, as well as *de novo* DNA synthesis (Kawai et al. 2003; Marsh and Teichmann 2010; Wu et al. 2012; Promponas et al. 2014; McLysaght and Guerzoni 2015; Meheust et al. 2016).

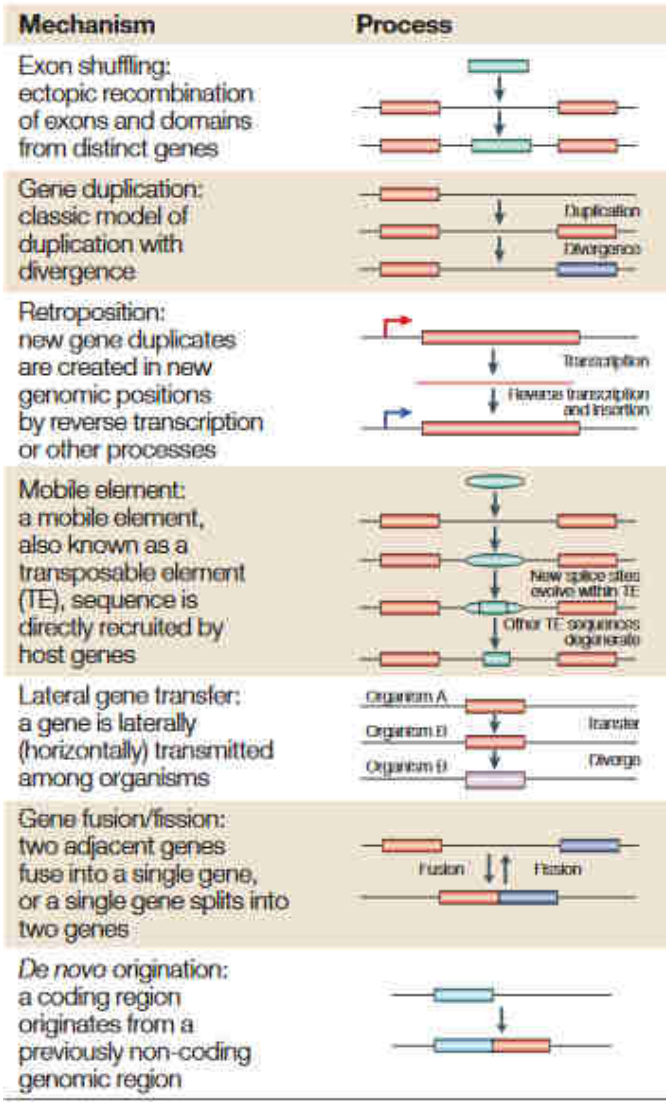



**Figure 11:** Molecular mechanisms for creating new gene structures. (Long et al. 2003)

Gene duplication, producing gene copies that can show different expression patterns and undergo neofunctionalisation, is a general process for evolutionary change. Gene duplication results in multiple related gene copies (paralogs) in the genomes. The analysis of the gene structure in many eukaryotic organisms showed a fragmented structure where the exons, the coding regions, are separated by the introns, non-coding intragenic regions. The intron-exon structure of eukaryotic genes promotes non-homologous recombination (Gilbert 1978). Exon shuffling, when it associates genetic fragments and domains in original combinations, also produces genetic novelty (Orgel and Crick 1980; Patthy 1999; Liu and Grigoriev 2004). It creates new genes, coding for new proteins, involved in novel protein-protein interactions and functions (Marcotte et al. 1999). Therefore, exon shuffling can be characterized via the identification of novel domains associations (i.e. the physical association between domains).

In this part of the thesis, we will focus on genes formed via combinatorial evolution processes, such as the fusion and recombination of genetic fragments from different gene families or the loss of a stop codon between two unrelated ORFs (Open Reading Frames) (Jones and Begun 2005). These genes are known as chimeric genes, fusion genes or composite genes (Enright et al. 1999). These composite genes are traditionally defined based on their detectable modularity: they are composed of segments (i.e. components) that can evolve separately in distinct gene families (Figure 12). Under this definition, composite genes can be the result of fusion of components, or involved as progenitors in fission events, after which associations of components are split in separate gene families.

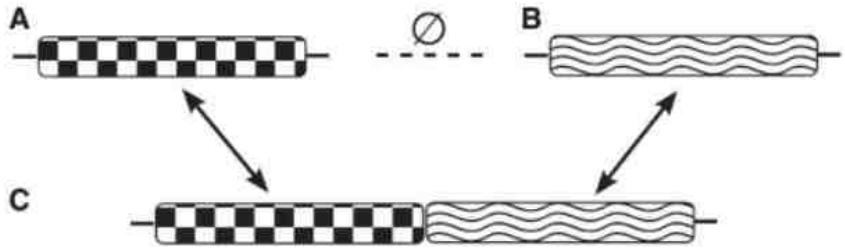



**Figure 12: Composite gene.**
Composite (fused) gene C and its two components A and B. A and B are similar to disjoint parts of C. A and B are dissimilar. (Jachiet et al. 2013)

Composite genes are produced by saltational processes. Unlike the gradual processes that involve slow and progressive evolutionary changes within a lineage (here a gene lineage), saltational processes will create macromutations involving large scale evolutionary jumps that can occur in a single generation, frequently involving several genes lineages (Suetsugu et al.

2005). These complex genetic changes producing novel combinations of existing modular elements have been described as a potential source of novelty upon which selection can act (Rogers and Hartl 2012). Genes from these unusual genetic combinations have been reported in the three domains of life (Jones et al. 2005; Rodrigues et al. 2007; Nie et al. 2011; Salim et al. 2011) but they appear to be more ommon in multicellular organisms' genomes, including humans (Courseaux and Nahon 2001; Brennan et al. 2008; Wilson et al. 2008; Kaessmann 2010; Avelar et al. 2014). Well-understood and well-characterized examples of remodeled genes include the Drosophila gene named *Jingwei*, from a fusion of a retrotransposed copy of an *Adh* locus and the 5' end of the *yande* gene (Wang et al. 2000) and *Kua-UEV* fusion gene from two adjacent genes (*Kua* and *UBE2V1*) in human (Thomson et al. 2000). As a matter of fact, it has been estimated that two-fifths of the prokaryotic genes and more than two-thirds of the eukaryotic genes are composed of several domains (Han et al. 2007). A recent study, conducted by Jachiet et *al*., allowed to extend the estimation of composite genes in viruses, showing that 8-15 percent of the viral sequences were composite (Jachiet et al. 2014). Although many of these combinations are likely to be dysfunctional or neutral, some appear to be advantageous like fusion of genes coding for proteins that interact in PPI networks (Enright et al. 1999; Marcotte et al. 1999; Enright and Ouzounis 2001; Marsh et al. 2013) or functionally biased genes encoding for proteins involved in the same metabolic pathways (Tsoka and Ouzounis 2000; von Mering et al. 2003; Hagel and Facchini 2017). For example, *Adiantum* ferns' adaptation to low light environment relies upon a composite photoreceptor, joining phytochrome and phototropin genes (Figure 13), which enables these ferns to use red light effectively (Nozue et al. 1998; Suetsugu et al. 2005)

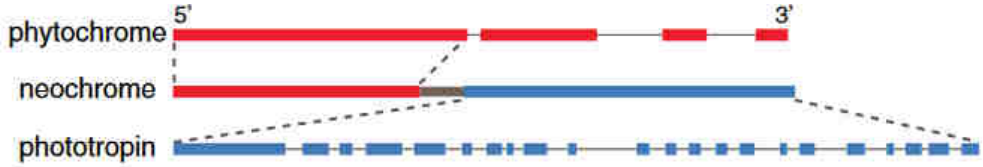



**Figure 13: The origin of fern neochrome.**
Neochrome is a chimeric photoreceptor in which the N terminus consists of a phytochrome sensory module fused to an almost complete phototropin sequence at the C terminus. Thick and thin lines represent exons and introns, respectively (length not to scale). (Li et al. 2014)

Another interesting example is the discovery of two fused genes in *Tetrathymena thermophila* free-living ciliates (Salim et al. 2009). Its fused genes *mtnAK* and *mtnBD* are each the result of the fusion of two different genes. Genes involved in these fusions catalyze different single steps of methionine salvage pathway in other organisms. Moreover in the case of *mtnBD*, the single fusion of *mtnB* and *mtnD* created a multifunctional enzyme replacing three independent enzymes in the salvage pathway. As stated by François Jacob, "Nature is a tinkerer and not an inventor". These unconventional genes from evolutionary 'bricolage' are important factors in molecular evolution, as well as contributors to genomic content (Jacob 1977; Duboule and Wilkins 1998).

The processes leading to detectable composite genes have been well studied in eukaryotic genomes but little is known about their impact on soil, marine, gut microbial communities or mobile genetic elements like plasmids (MGE) (Alvarez-Ponce et al. 2013; Nasir et al. 2014). However, the evolutionary processes shaping composite genes, have not been systematically studied, because relatively few composite genes have been identified and sufficiently characterized. Where and how composite genes are created in the environment is poorly understood. An increasing amount of molecular data with a considerable genetic diversity is now available from metagenomics projects, allowing addressing these fundamental issues beyond eukaryote genomes:

- Where are composite genes created ?

In terms of taxonomical lineage, eukaryotic model genomes are particularly concerned by these gene remodeling mechanisms. But the global distribution of composite genes in prokaryotes and MGE remains unknown, as well as the environments in which this complex molecular evolution occurs.

- How are composite genes created ?

Composite genes are not randomly assembled. The rules for association and dissociation of their components, e.g. the conditions structuring molecular evolution, are also poorly understood, particularly in the environment.

Systematic studies of composite genes are well formulated within the framework of sequence similarity networks. As in the introduction, SSN could be represented as an undirected graph where each node represents a unique sequence and each edge represents the

similarity between connected sequences. This kind of networks appear to be well suited to quantify and study this genetic remodeling (Bapteste et al. 2013). This approach enables the application of efficient graph theory concepts and tools to mine similarity information (Tordai et al. 2005; Song et al. 2008; Atkinson et al. 2009; Halary et al. 2010). Typically, a sequence similarity network can be reconstructed for a large dataset by connecting genes that are related in a BLAST search, with an E-value score better than a user-defined threshold. The structure of this network captures much of the history of gene evolution: not only divergence by point mutations but also recombinations, fusions and fission events (Adai et al. 2004). At the beginning of this thesis, bioinformatic tools, such as FusedTriplets 2.0 and MosaicFinder (Jachiet et al. 2013), were available to detect composite genes (by triplet analysis) and composite gene families (by clique analysis), respectively, in sequence similarity networks (Figure 14). However, these tools are neither optimal nor adapted for the study of composite genes in very large data sets, comprising several million proteins.

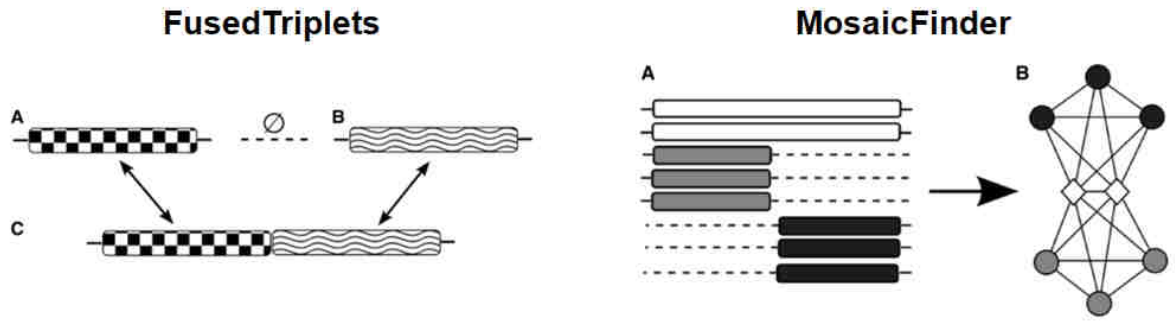



**Figure 14: Composite genes detection by FusedTriplets and MosaicFinder.**
FusedTriplets: detect composite (fused) gene C and its two components A and B. A and B are similar to disjoint parts of C. A and B are dissimilar.
MosaicFinder: (A) Multiple alignment of composite genes (white) and component genes (grey and black).(B) Similarity network of those genes. The white nodes correspond to a composite gene family. They are a clique minimal separator of the network. The black and grey nodes correspond to two separate component families. The evolution of genes families does not always result in cliques when some homologous sequences evolved beyond recognition by BLAST . Thus, a quasi-clique approach need to be developed.
(Jachiet et al. 2013)

During my thesis, I developed new fast and memory-efficient software called CompositeSearch, to improve the detection of composite genes and families of composite genes, using (quasi-cliques) (Article n°4). Afterwards, I investigated the biological properties of component and composite genes to infer what functions, genomes and environments were affected by such genetic reorganizations. I used CompositeSearch to study the impact of gene remodeling in plasmids (Article n°5) and in eukaryotes during transition from unicellularity to multicellularity (Article n°6).

### III.1.1 CompositeSearch: A new tool for studying modularity in molecular evolution

In the article n°4, I present CompositeSearch, a memory-efficient, fast and scalable method to detect composite gene families in large datasets (typically in the range of several million sequences). The method generalizes the use of similarity networks to detect composite and component gene families with a greater recall, accuracy, and precision than FusedTriplets and MosaicFinder. Moreover, CompositeSearch provides user-friendly quality descriptions regarding the distribution and primary sequence conservation of these gene families allowing critical biological analyses of these data. CompositeSearch was applied to a microbial environmental dataset of 3,906,323 sequences from 3 increasingly polluted sites (Sangwan et al. 2012) to test whether the evolutionary processes affecting gene remodeling in polluted environments followed some detectable rules. We report that increasingly polluted samplings sites present increasing percentages of composite genes, whereas the rules of functional associations of their components remain identical between sites. This article has been submitted to the journal "Molecular Biology and Evolution" and is under major revision.

# Introducing CompositeSearch, a generalized network approach for composite gene families detection: an application to the analysis of environmental gene remodeling

Jananan Sylvestre Pathmanathan,[1] Philippe Lopez,[1] François-Joseph Lapointe [2] and Eric Bapteste [*,1]

[1]Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), Paris F-75005, France.

[2]Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (QC) H3C 3J7 Canada.

[*]Corresponding author: E-mail: eric.bapteste@upmc.fr

## Abstract

Genetic sequences evolve through multiple processes beyond point mutations. In particular, the remodeling of genes by shuffling, fusion and fission of genetic fragments, as well as *de novo* DNA synthesis, contribute to the creation and diversification of gene families. Therefore, genetic sequences show similarity with one another for diverse reasons, i.e. common ancestry producing homology, and/or partial sharing of component fragments. These processes must be disentangled to understand the rules and constraints on gene evolution. This task is especially challenging in large molecular datasets, since computational analyses remain a bottleneck. In this article, we present CompositeSearch, a memory-efficient, fast and scalable method to detect composite gene families in large datasets (typically in the range of several million sequences). CompositeSearch generalizes the use of similarity networks to detect composite and component gene families with a greater recall, accuracy, and precision than recent programs (FusedTriplets and MosaicFinder). Moreover, CompositeSearch provides user-friendly quality descriptions regarding the distribution and primary sequence conservation of these gene families allowing critical biological analyses of these data. We applied CompositeSearch on a microbial environmental dataset of 3,906,323 protein sequences from 3 increasingly polluted sites, to test whether the evolutionary processes affecting gene remodeling in polluted environments may obey some detectable rules. Our

results suggest a possible correlation between sites level of pollution and proportion of composite genes, while the rules of functional associations of their components tends to remain identical between sites.

# Introduction

Genetic sequences evolve through multiple processes beyond point mutations. In particular, the remodeling of genes by shuffling of genetic fragments, fusion and fission, as well as *de novo* DNA synthesis, contributes to the creation and diversification of gene families (Kawai et al. 2003; Kaessmann 2010; Marsh and Teichmann 2010; Wu et al. 2012; Promponas et al. 2014; McLysaght and Guerzoni 2015; Meheust et al. 2016). Therefore, genetic sequences show similarity with one another for diverse reasons, i.e. common ancestry producing homology, and/or partial sharing of component fragments (Song et al. 2008; Haggerty et al. 2014). These processes must be disentangled to understand the rules and constraints on genes evolution. This task is especially challenging in large molecular datasets since computational analyses remain a bottleneck (Salim et al. 2011; Jachiet et al. 2013). While gene remodeling has been especially studied in eukaryotes (Kawai et al. 2003; Patthy 2003; Ekman et al. 2007; Nakamura et al. 2007; Meheust et al. 2016) and in cultured prokaryotes (Enright et al. 1999; Marcotte et al. 1999; Enright and Ouzounis 2000; Snel et al. 2000; Enright and Ouzounis 2001; Jachiet et al. 2013), it is particularly exciting to consider microbial environmental data to test whether the evolutionary processes affecting gene evolution in nature are similar to those described for cultured micro-organisms, and whether these processes obey rules. In the prokaryotic world (i.e. archaea, bacteria), cultured organisms are estimated to represent less than 1% of species diversity (Rappe and Giovannoni 2003), suggesting that evolutionary inferences based on cultivation studies need to be tested and complemented by analyses on actual environmental taxa, genes and processes (sometimes referred to as “the microbiological dark matter”) (Philippe et al. 2013; Cordero and Polz 2014; Brown et al. 2015; Lopez et al. 2015). In particular, environmental genetic diversity and its causes largely remains to be explored and explained (Koonin 2007; Lopez et al. 2015; Culligan and Sleator 2016; Fondi et al. 2016; Solden et al. 2016). For example, there are few studies on the effect of pollution on the evolution of environmental genetic diversity in microbes, although these studies certainly suggest that genetic diversity can be impacted (Zhang et al. 2009; Kriwy and Uthicke 2011; Sangwan et al. 2012; Chemerys et al. 2014; Staehlin et al. 2016).

In this article, we present CompositeSearch, a memory-efficient, fast and scalable method to detect composite gene families in large datasets (typically in the range of several million sequences). Composite genes are detected as a result of the fusion of partial or complete non-homologous DNA fragment, called component, or as a result of fission from a larger gene into dissociated persistent fragment. CompositeSearch generalizes the use of similarity networks to detect composite and component gene families with a greater recall, accuracy, and precision than recent programs (FusedTriplets and MosaicFinder (Jachiet et al. 2013)). Moreover, it provides user-friendly quality descriptions regarding the distribution and primary sequence conservation of these gene families allowing critical biological analyses of these data, and it is used as an input for the reconstruction of mutirooted gene networks (Haggerty et al. 2014).

We applied this new method to a published dataset of 3,906,323 environmental sequences from soil microbial communities across three hexachlorocyclohexane (HCH) contamination levels (Sangwan et al. 2012) to quantify the proportion of composite genes in each of these samples, and to describe the functional rules of gene components associations. We observed that components associated with a given composite gene family tend to belong to similar functional categories (consistent with the findings that genes coding for proteins in functional interactions can fuse or fission (Tsoka and Ouzounis 2000; Yanai et al. 2001)). We also report that increasingly polluted samplings sites present an observable increase of the proportion of composite genes. These observations are raising the hypothesis that “environmental hotspots of gene remodeling” might exist, and larger datasets could test in the future whether these hotspots may be associated with specific forms of pollution.

## New approach

Here, we present CompositeSearch, a memory-efficient, fast and scalable method, implemented in C++, which detects composite gene families in large datasets (typically in the range of several million sequences). Composite genes are traditionally defined based on their apparent modularity: they are composed of segments (i.e. components) that can evolve separately in distinct gene families. Under this definition, composite genes can be the result of fusion of components, or involved as progenitors in fission events, after which associations of components are split in separate gene families. CompositeSearch generalizes the use of

similarity networks (SSN) to detect composite and component gene families. SSN are undirected graphs, where each node represents a unique sequence and each edge represents the similarity between connected sequences (given similarity criteria, such as a minimum percentage identity, BLAST E-value (Altschul et al. 1990) and minimum mutual coverage, i.e., the minimal length covered by the matching parts with respect to the total length of each compared sequence)(Jachiet et al. 2013; Corel et al. 2016). For a given comparison between two sequences, the alignment, score and E-value are not symmetric. They can vary depending on which sequence is used as the query. Thus, the network is first symmetrized by considering the best match of each pairwise comparison. As the greatest asymmetry is found in the better-scoring comparisons [i.e. at a much more stringent threshold than the ones used for network reconstruction (Atkinson et al. 2009)], this procedure does not impact the topology.

This network's structure captures much of the history of gene evolution: not only divergence by point mutations but also recombinations, fusions and fission events (Adai et al. 2004; Jachiet et al. 2013). Typically, gene families form sub-graphs with high connectivity, in which connected sequences display significant BLAST E-values $\leq 1E^{-5}$, mutual covers $\geq 80\%$, %ID $\geq 30\%$. By contrast, superfamilies (Atkinson et al. 2009) and composite gene families (Song et al. 2008; Jachiet et al. 2013; Haggerty et al. 2014; Jachiet et al. 2014; Meheust et al. 2016) introduce more complex informative patterns in SSNs.

Using these graphs to identify composite genes and gene families, CompositeSearch shows a greater recall, accuracy, and precision than recent programs (FusedTriplets and MosaicFinder). In short, these two programs are helpful but limited in scope. FusedTriplets cannot handle large datasets and does not define composite gene families. MosaicFinder is also unable to analyze large datasets (due to memory and speed limitations). While it identifies composite and component gene families, MosaicFinder is only meant to find highly conserved composite gene families that form minimal clique separators in sequence similarity network. The ‘clique’ condition implies that MosaicFinder misses divergent (e.g. ancient or fast evolving) composite gene families (whose members do not necessarily connect all together in sequence similarity networks) (Fig.1). The ‘separator’ condition implies that composite genes will remain undetected for datasets with highly remodeled genes by MosaicFinder. Indeed the repeated use of gene components introduces cyclic paths in sequence similarity networks, which turns composite families into local, but not global separators. Beyond its larger scope and better performance, CompositeSearch can also provide quality descriptions (absent from MosaicFinder and FusedTriplets) regarding the size

and primary sequence conservation of composite and component gene families, easing critical biological analyses of these data. CompositeSearch is available at http://www.evol-net.fr/downloads/.

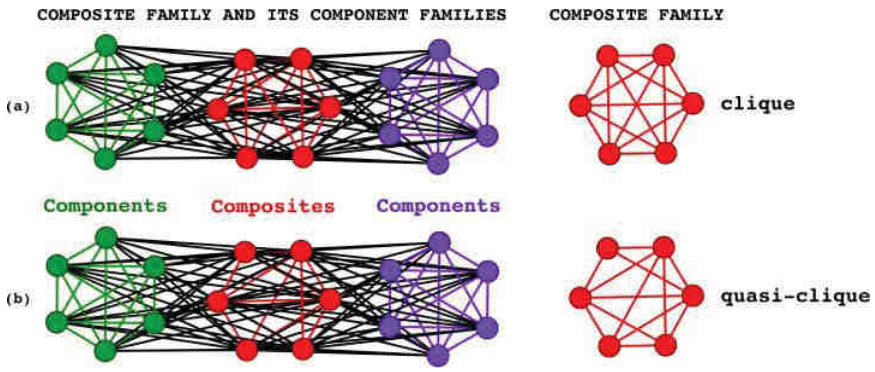



**Fig. 1.** Similarity network of a composite gene family and its components gene family. (a) represents a composite gene family (red) forming a clique and (b) represents a composite gene family forming a quasi-clique. The component gene families are represented in green and purple. MosaicFinder will detect only the case (a) instead of CompositeSearch which is able to detect composite gene family forming clique (a) and quasi-clique (b).

# Results

## Benchmarking

## Simulated data

We tested and compared CompositeSearch with FusedTriplets and MosaicFinder (Jachiet et al. 2013) using data simulated with Seq-Gen (Rambaut and Grassly 1997), as there is no large manually curated database of composite genes to use as a test bed to our knowledge. We explored the effect of gene family divergence and multiple domain reassortments on composite gene detection under the hypothesis that the more divergent gene families are, the harder they are to detect (Supplementary Figure S4). We produced gene families with different degrees of divergence as follows. We scaled ultrametric phylogenetic trees with Seq-Gen (option -d) so that the total length of a tree can be measured as the distance from the root to any of the leaves in units of mean number of substitutions per site (MNSS). We explored 3 evolutionary rates (0.1, 0.5 and 1.0) to cover the range from highly conserved to highly divergent gene families (parameter 1).

Three component families (A, B and C) have been evolved under the Whelan and Goldman model of amino acid substitution and a site-specific rate heterogeneity following a continuous gamma distribution (alpha=1). Ancestral sequences of 300 amino acids were generated randomly for each component family. These sequences were then evolved along perfect (complete) binary trees with five levels at the same evolution rate, i.e. symmetric and balanced trees with 32 leaves at the fifth level, resulting in component families with 32 genes. These component families will be used for fusion events leading to composite genes with two and three domains.

First a pair of sequences *sA* and *sB* is selected from component family A and B at the same distance *k* from the tree root (from 0 to 5 (parameter 2)) to create a composite sequence with 2 domains. Second, a sequence *sC* is selected from component family C at a distance $p >= k$ from the tree root (from *k* to 5 (parameter 3)) to create composite sequences with 3 domains and reassortments.

We used *sA* and *sB* to create a novel 300 amino acids composite sequence *sAB* made of 30-50% of the first sequence fused with 70-50% of the second sequence (parameter 4). This ancestral composite sequence *sAB* was then evolved along a perfect binary tree with *q* levels ($q=p-k$). This composite family was evolved at the same 3 evolutionary rates (parameter 5) that were used for the component families, thereby producing highly conserved to highly divergent composite families. Finally we used an evolved sequence of *sAB* to create three new composite sequences with domain reassortments (*sABC*, *sCAB* and *sACB*) made of 30-50% of *sC* (parameter 6). These three composite families were then evolved along a perfect binary tree with *z* levels ($z=5-p$) at the same 3 evolutionary rates explained previously (parameter 7). For recent fusion events (fusion level = 0), composite sequences were left unmodified. This protocol was repeated 100 times for each combination of the 7 parameters.

The sensitivity and specificity of each program were summarized using in Supplementary Table S1. In terms of detection of composite genes, CompositeSearch performs equally well with FusedTriplets (identical TPR and FPR), but, unlike FusedTriplets, CompositeSearch returns composite gene families. In terms of detection of gene families, CompositeSearch outperforms MosaicFinder. CompositeSearch has higher TPR than MosaicFinder, especially for divergent composite sequences, without enhancing its FPR. Therefore, CompositeSearch will find additional composite genes with respect to MosaicFinder (thanks to the detection of composite genes forming quasi-cliques).

As CompositeSearch is able to detect the number of domains (or components) for each composites, we created a more detailed table (Supplementary Table S2) showing the sensitivity and specificity of CompositeSearch to detect the exact number of domains.

## Computational performances

Because its algorithm uses a dichotomous search to browse the network and because CompositeSearch is multithreaded, CompositeSearch outperforms both FusedTriplets and MosaicFinder in terms of speed and memory use, when these parameters are contrasted on a Linux machine with Intel Xeon CPU E5-2630 v2 2.60 GHz processors and 256 GB RAM, even on one CPU. This is especially noticeable for large metagenomic data sets (Table 1). By contrast, construction the SSN composite genes and composite gene families detection run in a few second to few minutes depending on the network's size.

**Table 1.** CompositeSearch, FusedTriplets and MosaicFinder performances comparison.

| Data | Nodes | Edges | Software | #Cpu | Runtime | Memory |
|---|---|---|---|---|---|---|
| 1 | 338,868 | 71,946,457 | MosaicFinder | 1 | 548 h 27 min | 82 GB |
| | | | FusedTriplets | 1 | 70 h 47 min | 18 GB |
| | | | CompositeSearch | 1 | 00 h 12 min | 2.5 GB |
| | | | CompositeSearch | 10 | 00 h 06 min | 2.5 GB |
| 2 | 3,166,706 | 282,789,792 | MosaicFinder | 1 | - | - |
| | | | FusedTriplets | 1 | - | - |
| | | | CompositeSearch | 10 | 08 h 48 min | 32 GB |

*We compared the performance of CompositeSearch, FusedTriplets and MosaicFinder on the same Linux machine with Intel Xeon CPU E5-2630 v2 2.60 GHz processors and 256 GB RAM. The data (1) is a SSN from plasmids complete genomes (NCBI December 2014) and (2) HCH metagenomes (Sangwan et al. 2012). CompositeSearch outperform FusedTriplets and MosaicFinder even with one CPU as shown for data (1). On the data (2) FusedTriplets and MosaicFinder stop by running out of memory, which was not the case for CompositeSearch.*

## Application to metagenomic data

As an application illustrating the features of CompositeSearch, we detected composite genes in the metagenomes of 3 distinct, increasingly polluted sites, gathered from the MG-RAST server (Meyer et al. 2008) as indicated in Sangwan *et al* (2012). The contamination was caused by a pesticide used for agriculture crops, hexachlorocyclohexane (HCH). Site 1 was considered pristine since it presented a concentration of 0.03 mg HCH/g soil. By contrast, site 2 presented a concentration of 0.7 mg HCH/g soil, and site 3 presented a concentration of 450 mg HCH/g soil (Sangwan et al. 2012). Here, we used CompositeSearch to retain all composite genes with component gene families having at least 2 genes. Interestingly, the proportion of such composite genes per metagenome weakly yet but statistically significantly increased with pollution levels. There were 36% of composite genes (594,395 sequences out of 1,613,523) at site 1; 40% of composite genes (444,495 sequences out of 1,102,372) at site 2; and 42% of composite genes (499,532 sequences out of 1,190,337) at site 3. We tested the significance of these results with a pairwise Fisher exact test and p-values were corrected with the false discovery rate (FDR) method ($P < 2.2e^{-16}$). Significance was assessed using a jackknife procedure by 500 independent resamplings of 500 000 sequences from each site, followed by composite genes detection and a pairwise Fisher exact test.

There are many possible, non-exclusive, interpretations for this correlation between an elevated proportion of composite genes in the environments and their increasing contamination by HCH. A first hypothesis however is simply that the difference in proportions of composite genes across these 3 sites is due to spatial variation. Second, HCH pollution may select for taxa whose genomes are intrinsically richer in composite genes. Third, HCH pollution may select for specific composite genes. For example, 2.5% of the composite families that were detected had a much higher abundance in the most polluted sites than in the pristine one. In Figure 2, we show one of these particular composites. After annotating with the Kegg Orthology database (KOD), the functional analysis showed that this composite gene family is involved in thiamine biosynthesis pathway (TBP) (ID:K03149), a pathway of importance for microbial metabolism. These composite genes are formed by the association of two components (Supplementary Figure S5): a C-terminal component, annotated as *ThiS* (COG2104) domain, and a N-terminal component, annotated as *ThiG* (COG2022) domain. Both *ThiS* and *ThiG* have been reported to form gene clusters (Rodionov et al. 2002). Moreover, this finding is consistent with the literature on gene fusions of genes involved in TBP, e.g. fusion of *ThiE-ThiD*; *ThiE-ThiM* or *ThiO-ThiG* were described

(Rodionov et al. 2002). The detection of environmental *ThiS-ThiG* genes, a novel combination thus extends the description of fusion events of TBP genes. In the TBP, *ThiS*, which is a sulfur carrier, interacts with *ThiG* for thiazole formation in the *ThiS-COSH* chemical form. The fusion of *ThiS* with *ThiG* couples this later protein with its sulfur donor, which ensures the proximate presence of a thiol donor next to the *thiG* sequence. Interestingly, this original association may provide an additional selective advantage for the composite *ThiS-ThiG* gene in environments polluted by HCH. *ThiG* proteins are notoriously sensitive to reactive oxygen species (ROS), and to chlorine (RCS). These ROS and RCS cause post-translational thiol-modifications leading to the super-oxidation of the thiol residues of *ThiG*, which critically alter *ThiG* activities (Loi et al. 2015). Model organisms protect *ThiG* against thiol-modification by various processes of thiolations, which rely upon the presence of low-molecular-weight thiol-redox buffers. In the environment, the coupling of a thiol donor such as *ThiS* with *ThiG* might therefore interfere with thiolations, and provide an emergent mechanism to protect *ThiG* activity. Testing this hypothesis will of course require experimental evidence. Fourth, HCH pollution may enhance the formation of composite genes in microbial genomes (possibly by introducing stop codons that split complete genes, or by enhancing the rate of compensatory mutations between genes coding for interacting proteins, as postulated by the theory of constructive neutralism (Gray et al. 2010)). Fifth, the different proportions of composite genes across sites may be related to other factors than HCH pollution. At any rate, all these interpretations suggest that different environments have different proportions of composite genes, hinting at the existence of environmental hot spots of gene remodeling.

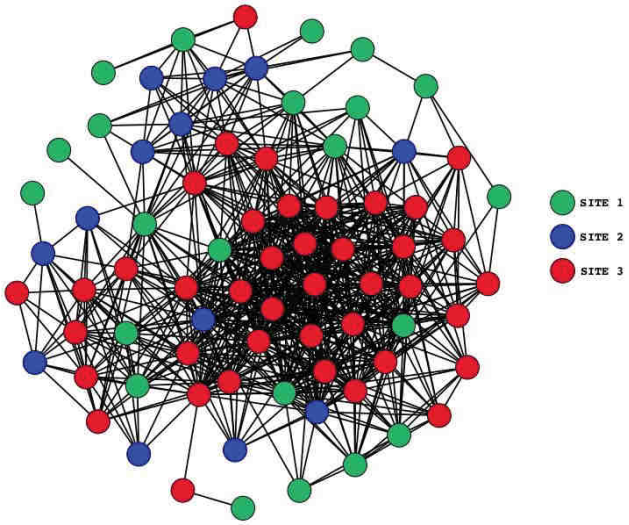



**Fig. 2.** Network of a composite gene family detected by CompositeSearch. This family is composed of genes belonging to site 1 (green), site 2 (blue) and site 3 (red). Composite genes are more abundant in the most polluted site (red). These genes have been annotated with the Kegg Orthology ID: K03149 and are involved in thiamin biosynthesis metabolic pathways.

Interestingly CompositeSearch can also be used to investigate the rules of component association. We split each composite gene into its constitutive protein domains, as detected by CompositeSearch. For each domain the functional categories was assigned using eggNOG-mapper (Huerta-Cepas et al. 2016; Huerta-Cepas et al. 2017). For each environment, we summarized the information about the functional assignation of pairs of protein domains present along a composite gene. We reported the proportion of all combinations of functional categories realized by pairs of domains in a matrix. Thus, this matrix provides a functional profile of protein domains associated in the composite genes for each environment (Fig. 3 shows the matrix for site 1, 2 and 3). If gene remodeling strongly depends on the functions of protein domains, and if similar constraints apply for the functional association/dissociation of genetic components, we expect similar functional profiles for the pairs of associated protein domains across all environments. We used the Mantel test to compare the 3 matrices and verify if the profiles were similar or different. We did a pairwise comparison of these 3 matrices using the "CADM.post" function of the Mantel test from the *ape* (v. 3.5) library (Paradis et al. 2004) of the R statistical package (v. 3.2.5) (R Core Team 2016). For our purpose we modified the "CADM.post" function to account for values on the diagonal and fixed the number of permutation to 999. We observed a quasi perfect correlation ($r^2 = 0.99$) for all pairs of matrices (Supplementary Table S3). Therefore, we can reject the null

hypothesis that the 3 profiles of association of domains forming composite genes were different. This suggests that associations/dissociations of protein domains are strongly constrained by functions and that the same rules regarding the functions of protein domains subjected to gene remodeling apply across environments (here with different pollution levels). Typically, the higher frequency of composite genes we report in the most polluted site does not involve different rules of functional associations between protein domains than those observed at the other 2 sites.

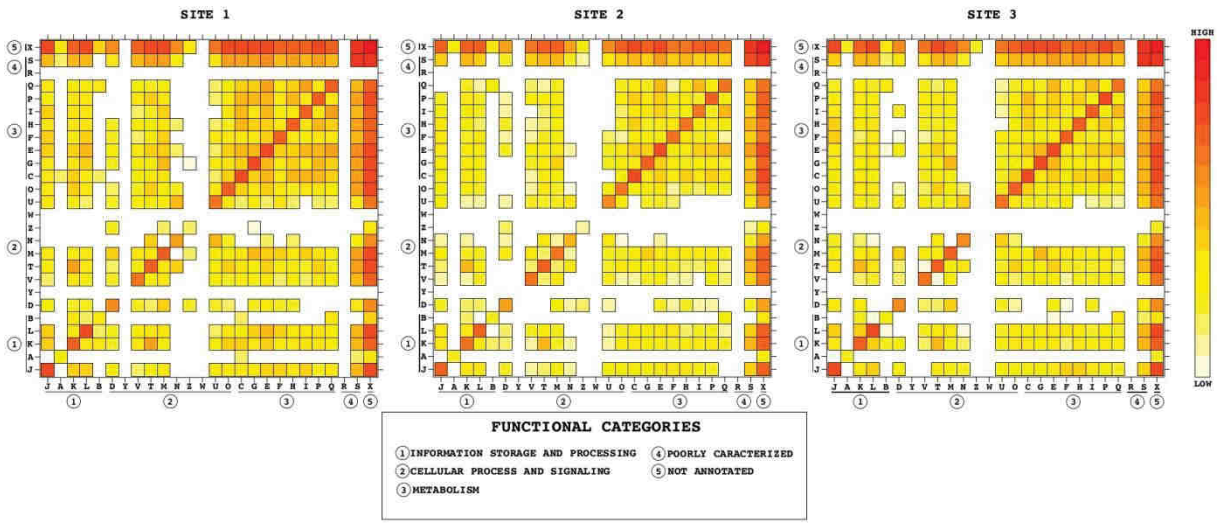



**Fig. 3.** Matrix showing the proportion of all combinations of functional categories realized by pairs of domains for each site.

## Discussion

CompositeSearch is an efficient tool that detects composite genes and composite gene families. It allows investigating the process of gene remodeling in large datasets, for example metagenomes and/or thousands of complete genomes. While CompositeSearch is faster than currently available software, like FusedTriplets and MosaicFinder, it still can be improved. We observed that in CompositeSearch, the most time consuming step is the detection of gene families, using a DFS algorithm than runs on a single CPU. Parallelized algorithms that detect connected components are available (Kang et al. 2009; Iverson et al. 2015), but they usually require high computational resources. As CompositeSearch was developed with maximum portability in mind, these algorithms are not implemented yet could be in a future version.

This software provides new opportunities to better understand how gene remodeling has shaped the evolution of organisms, i.e to detect whether gene remodeling obeys some rules, and whether these rules change across different environments and lineages. In particular, investigating additional polluted environments and larger datasets could allow, in the future, to test whether functional associations of protein domains remain constant at larger geographical scales and for different types of pollution, and to better understand the causes of environmental genetic diversity.

## Materials and methods

CompositeSearch is a multithreaded tool, which detects both composite genes and their families. Composite genes are traditionally defined based on their apparent modularity: they are composed of segments (i.e. components) that can evolve separately in distinct gene families. Under this definition, composite genes can be the result of fusion of components, or involved as progenitors in fission events, after which associations of components are split in separate gene families.

### STEP 1: Construction of the SSN

The SSN is constructed by CompositeSearch, based on the cleaned result of an all-against-all BLAST sequence comparison. This preliminary step relies on a C++ program called *cleanBlastp*, provided along with CompositeSearch. *cleanBlastp* uniquely numbers each sequence in the BLAST output, and removes all self-hits, keeping the best hit (i.e. lowest E-value) amongst multiple hits between pairs of sequences. At the end of this preliminary step, the input file used by CompositeSearch contains BLAST information about matches between pairs of sequences (qstart, qend, sstart, send), sequence length (qlen, slen) and their symmetrized similarity scores (E-value, pident). The selection of unique pairs of hits avoids simultaneous memory access issues and allows to parallelize the SSN construction, by splitting the cleaned BLASTP results file into a user defined number of CPUs. CompositeSearch utilizes user-defined similarity scores (default E-value $\leq 10$, default Pident $\geq 30\%$) to construct the SSN. The results are then represented as an undirected network

G=(V,E), where V is the set of sequences, and edge is (u,v) ϵ E if the similarity score $S_{uv}$ or $S_{vu}$ is higher than a user-defined threshold.

**STEP 2: Definition of gene families**

CompositeSearch clusters sequences into gene families in two steps. First, it uses a modified Depth First Search (DFS) algorithm on a thresholded SSN (default: mutual coverage between two sequences ≥ 80%) that defines connected components (CCs). Each CC is considered as a putative gene family, when the minimum mutual sequence coverage criterion is high (≥ 80 %), but gene family definition is then further refined in a second step as follows. Each time a CC is detected a mutual coverage score ($S_{mc}$) is calculated. If $S_{mc} < 1$, this CC is subjected to the Louvain community detection algorithm (Blondel et al. 2008), using C++ *igraph* 0.7.1 library (Csardi and Nepusz 2006). Indeed, BLAST matches can be over-extended (Mills and Pearson 2013), with the consequence that non-homologous sequences may be introduced in a CC in pathological cases (Supplementary Figure S1). This second step of community detection allows to define, at a finer granularity, the groups of sequences forming communities (e.g. cliques and/or quasi-cliques) within the CC, which are finally considered as a gene family. Thus each sequence from the original dataset is assigned to a given gene family and a connectivity score is computed for each family. If gene families are pre-computed, a tab delimited file with the gene ID and its family ID can be given as an input and for each of these gene families a connectivity score will be also computed.

This step returns 3 files:

- family.nodes: a file where the nodes for each family is listed which will be useful for post-analysis of the gene families;
- family.edges: a file where the edges for each family is listed;
- family.info: a file storing the number of nodes and edges for each family and their connectivity, which can be used for a sized-based or connectivity-based selection of composite gene families by the user. Connectivity is measured as :

$$C_{family} = \frac{(2 * N_{edges})}{N_{nodes} * (N_{nodes} - 1)}$$

**STEP 3: Detection of composite genes**

Unlike MosaicFinder and FusedTriplets, CompositeSearch starts from the assumption that each node could be a composite gene. This decision allows to parallelize detection of composite genes by distributing a list of nodes to visit for each CPU, which takes into account node degree to produce computationally balanced lists of nodes to be distributed among the CPUs. CompositeSearch checks whether a node's neighbors belong to different gene families and their size is higher or equal to the minimum number of genes to be used as component gene families. If all neighbors of a node belong to only one gene family, this node is not a composite gene. If at least two neighbors of this node belong to distinct gene families, CompositeSearch takes the sequence corresponding to the node as a reference and maps the matches from all different families along that sequence. Each region with matches from different families along a composite sequence is called a "protein domain" hereafter. For each "protein domain", CompositeSearch computes an average position for the start of the domain and an average position for the end of the domain (Supplementary Figure S2). If there is no overlap between at least two "protein domains" along the reference sequence, then the reference sequence is considered as composite, since it is composed of at least two non-overlapping regions with homology to different gene families. In practice, a maximum overlap can be allowed (by default ≤ 20 AA, in order not to discard *bona fide* composite genes despite possible BLAST short overextensions introducing artefactual overlaps between protein domains).

During this step, CompositeSearch produces 2 files:

- file.composites : a file in fasta format with the number of the composite sequences and the position and identity of the component families matching along this composite;
- file.compositesinfo: a file containing the number of protein domains along a composite sequence, and a non-overlapping score ($S_{no}$) between all of these domain. The $S_{no}$ score is measured as: $N_i/N_T$, where $N_i$ is the number of non-overlapping pairs of "protein domains", and $N_T$ is the number of all possible pairs of domains ($N_T$). This measure allows the user to sort composite gene families based on the neat separation of all their protein domains ($S_{no}$ close to 1) or the separation of some of their protein domains only (lower $S_{no}$) (Supplementary Figure S3).

**STEP 4: Detection of composite gene families**

CompositeSearch goes through all gene families to check whether a family is composite or not. Any gene family containing at least one composite gene and with a size higher or equal to the minimum number of genes fixed for composite gene family detection is considered as composite family. This process can be parallelized by distributing a list of gene families to analyze for each CPU.

During this step, CompositeSearch produces 2 files:

- file.compositefamilies: a file in fasta format with the number of the composite gene family and the position and identity of the component families matching along this composite;
- file.compositefamiliesinfo: a file containing the connectivity, percentage of composite genes, mean number of "protein domains" of the composite gene families.


## Acknowledgements

We thank Raphaël Méheust for his helpful and constructive discussion during the development of CompositeSearch. We also thank our beta testers James O. McInerney, Mary J. O'Connell, Raymond Moran and Rob Leigh.

## Funding

J.S.P. and E.B. are funded by the European Research Council (FP7/2017-2013 Grant Agreement #615274).

*Conflict of Interest:* none declared.

# References


Adai AT, Date SV, Wieland S, Marcotte EM. 2004. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. J Mol Biol 340(1):179-90.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One 4(2):e4345.

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

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523(7559):208-11.

Chemerys A, Pelletier E, Cruaud C, Martin F, Violet F, Jouanneau Y. 2014. Characterization of novel polycyclic aromatic hydrocarbon dioxygenases from the bacterial metagenomic DNA of a contaminated soil. Appl Environ Microbiol 80(21):6591-600.

Cordero OX, Polz MF. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol 12(4):263-73.

Corel E, Lopez P, Meheust R, Bapteste E. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol 24(3):224-37.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems.

Culligan EP, Sleator RD. 2016. Editorial: From Genes to Species: Novel Insights from Metagenomics. Front Microbiol 7:1181.

Ekman D, Bjorklund AK, Elofsson A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. J Mol Biol 372(5):1337-48.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402(6757):86-90.

Enright AJ, Ouzounis CA. 2000. GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16(5):451-7.

Enright AJ, Ouzounis CA. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2(9):RESEARCH0034.

Fondi M, Karkman A, Tamminen MV, Bosi E, Virta M, Fani R, Alm E, McInerney JO. 2016. "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis. Genome Biol Evol 8(5):1388-400.

Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. 2010. Cell biology. Irremediable complexity? Science 330(6006):920-1.

Haggerty LS, Jachiet PA, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, Pisani D, Wilkinson M, Bapteste E, McInerney JO. 2014. A pluralistic account of homology: adapting the models to the data. Mol Biol Evol 31(3):501-16.

Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M et al. . 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44(D1):D286-93.

Iverson J, Kamath C, Karypis G. 2015. Evaluation of connected-component labeling algorithms for distributed-memory systems. Parallel Computing 44:53-68.

Jachiet PA, Colson P, Lopez P, Bapteste E. 2014. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. Genome Biol Evol 6(9):2195-205.

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29(7):837-44.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res 20(10):1313-26.

Kang U, Tsourakakis CE, Faloutsos C. 2009. PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations. 2009 9th Ieee International Conference on Data Mining:229-238.

Kawai H, Kanegae T, Christensen S, Kiyosue T, Sato Y, Imaizumi T, Kadota A, Wada M. 2003. Responses of ferns to red light are mediated by an unconventional photoreceptor. Nature 421(6920):287-90.

Koonin EV. 2007. Metagenomic sorcery and the expanding protein universe. Nat Biotechnol 25(5):540-2.

Kriwy P, Uthicke S. 2011. Microbial diversity in marine biofilms along a water quality gradient on the Great Barrier Reef. Syst Appl Microbiol 34(2):116-26.

Loi VV, Rossius M, Antelmann H. 2015. Redox regulation by reversible protein S-thiolation in bacteria. Front Microbiol 6:187.

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

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428):751-3.

Marsh JA, Teichmann SA. 2010. How do proteins gain new domains? Genome Biol 11(7):126.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci 370(1678):20140332.

Meheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A 113(13):3579-84.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A et al. . 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

Mills LJ, Pearson WR. 2013. Adjusting scoring matrices to correct overextended alignments. Bioinformatics 29(23):3007-13.

Nakamura Y, Itoh T, Martin W. 2007. Rate and polarity of gene fusion and fission in Oryza sativa and Arabidopsis thaliana. Mol Biol Evol 24(1):110-21.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20(2):289-90.

Patthy L. 2003. Modular assembly of genes and the evolution of new functions. Genetica 118(2-3):217-31.

Philippe N, Legendre M, Doutre G, Coute Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C et al. . 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341(6143):281-6.

Promponas VJ, Ouzounis CA, Iliopoulos I. 2014. Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey. Brief Bioinform 15(3):443-54.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Staistical Computing, Vienna, Austria.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci 13(3):235-8.

Rappe MS, Giovannoni SJ. 2003. The uncultured microbial majority. Annu Rev Microbiol 57:369-94.

Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. 2002. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J Biol Chem 277(50):48949-59.

Salim HM, Koire AM, Stover NA, Cavalcanti AR. 2011. Detection of fused genes in eukaryotic genomes using gene deFuser: analysis of the Tetrahymena thermophila genome. BMC Bioinformatics 12:279.

Sangwan N, Lata P, Dwivedi V, Singh A, Niharika N, Kaur J, Anand S, Malhotra J, Jindal S, Nigam A et al. . 2012. Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels. PLoS One 7(9):e46219.

Snel B, Bork P, Huynen M. 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet 16(1):9-11.

Solden L, Lloyd K, Wrighton K. 2016. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol 31:217-26.

Song N, Joseph JM, Davis GB, Durand D. 2008. Sequence similarity network reveals common ancestry of multidomain proteins. PLoS Comput Biol 4(4):e1000063.

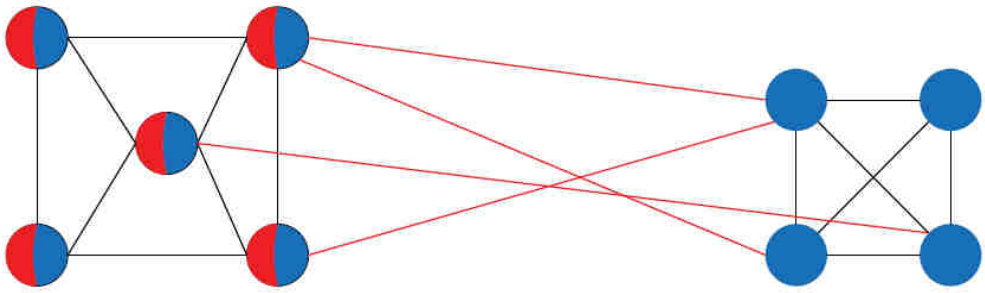
Staehlin BM, Gibbons JG, Rokas A, O'Halloran TV, Slot JC. 2016. Evolution of a Heavy Metal Homeostasis/Resistance Island Reflects Increasing Copper Stress in Enterobacteria. Genome Biol Evol 8(3):811-26.

Tsoka S, Ouzounis CA. 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26(2):141-2.

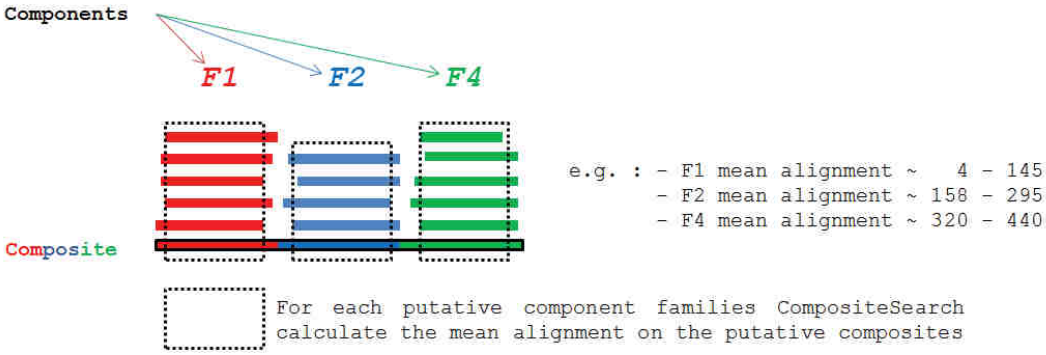
Wu YC, Rasmussen MD, Kellis M. 2012. Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny. Mol Biol Evol 29(2):689-705.

Yanai I, Derti A, DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A 98(14):7940-5.

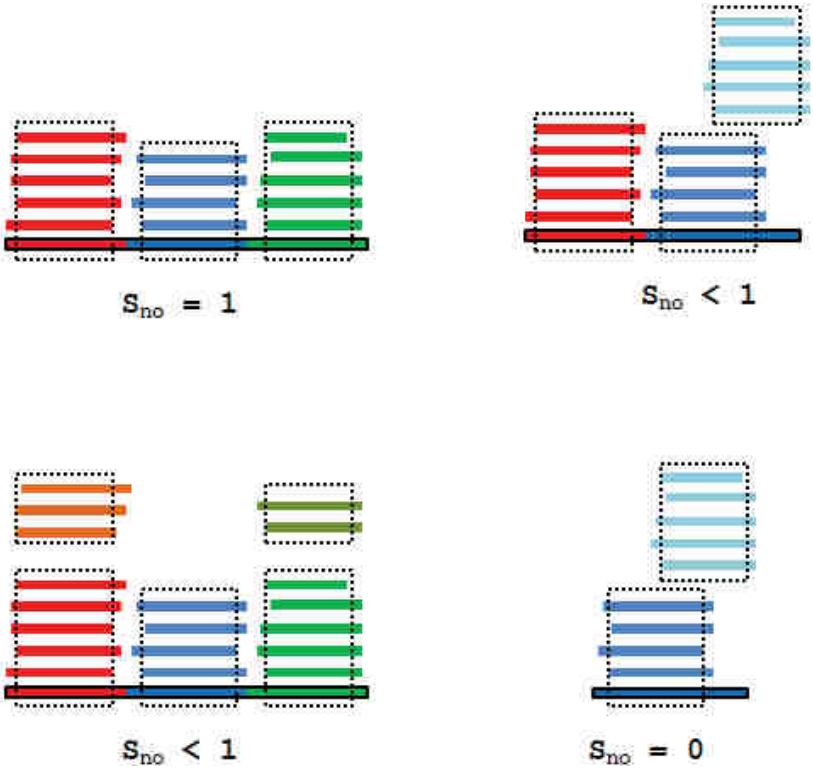
Zhang Y, Zhang X, Zhang H, He Q, Zhou Q, Su Z, Zhang C. 2009. Responses of soil bacteria to long-term and short-term cadmium stress as revealed by microbial community analysis. Bull Environ Contam Toxicol 82(3):367-72.



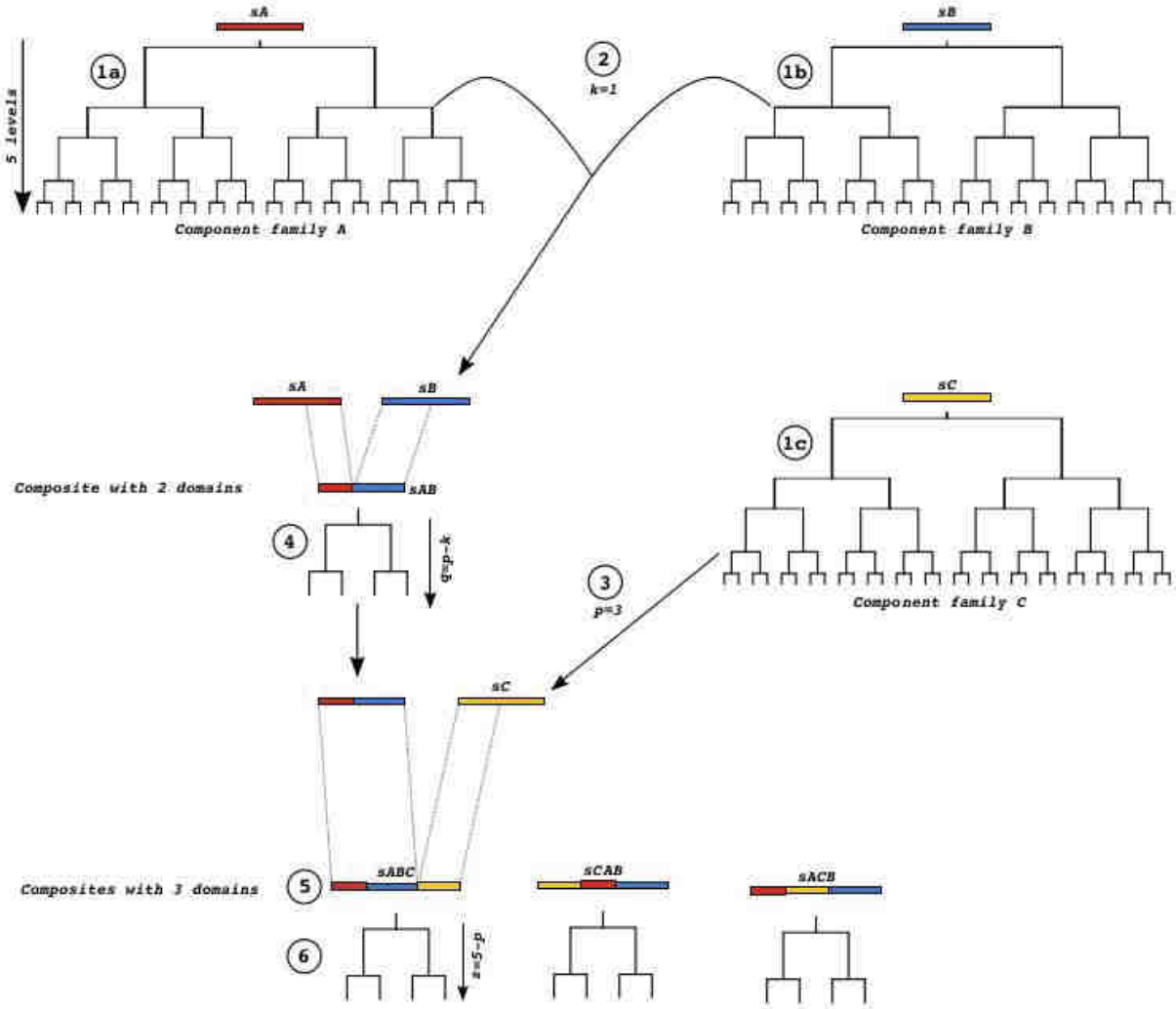
**Supplementary Figure S1.** Example of a composite gene family (red/blue nodes) and one of its component gene family (blue nodes) detect as one gene family due to BLAST overextensions (red edges). Applying the Louvain clustering will correct the gene families detection by splitting them.




**Supplementary Figure S2.** The average position of domains' start and end computed by CompositeSearch.




**Supplementary Figure S3.** Example of non-overlapping score ($S_{no}$) between all of these domain.




1 - Generate 3 random sequences with a length of 300 amino acids
- Evolve these sequences along a five levels perfect binary tree (1a, 1b and 1c)

2 - Select a pair of sequences (***sA***, ***sB***) along trees 1a and 1b at the same level ***k*** [from 0 to 5]
- Fuse fragments of ***sA*** [30%-50%] and ***sB*** [70%-50%] sequences given a percentage [30%-50%] to create a composite sequence (***sAB***) of 300 amino acids

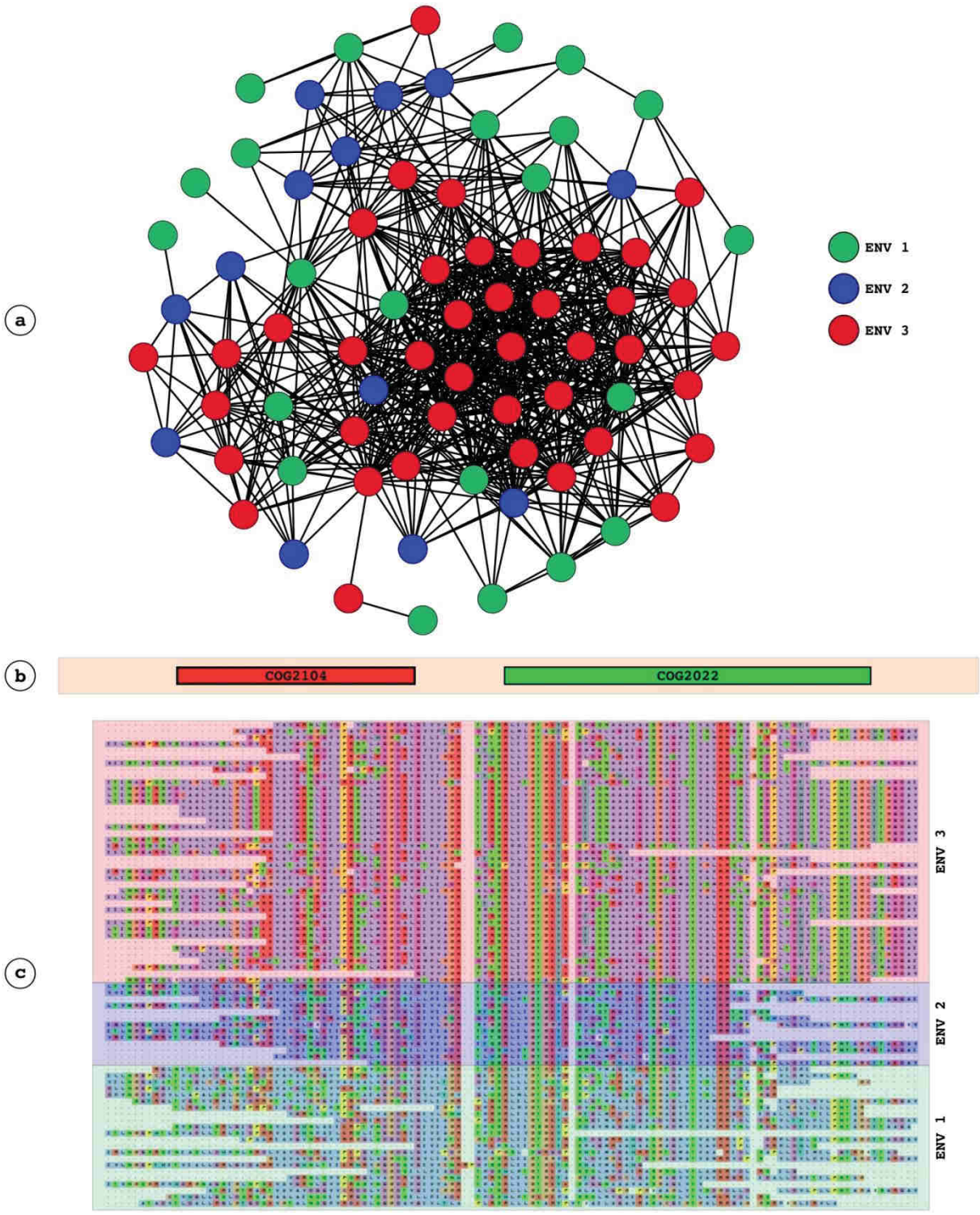
3 - Select a sequence ***sC*** along the tree 1c at distance ***p*** >= ***k***

4 - Evolve ***sAB*** along a perfect binary tree of ***q*** level (***q*** = ***p*** - ***k***)

5 - Select an evolved sequence ***sAB***
- Fuse a fragment of sequence ***sC*** [30%-50%] at different position with ***sAB*** to create composite sequences (***sABC, sCBA and sACB***)

6 - Evolve ***sABC, sCAB*** and ***sACB*** along a perfect binary tree of ***z*** level (***z*** = 5 - ***p***)

**Supplementary Figure S4.** Simulation and evolution of composite genes.




**Supplementary Figure S5.** (a) Network of a composite gene family detected by CompositeSearch. This family is composed of genes belonging to site 1 (green), site 2 (blue) and site 3 (red). Composite genes are more abundant in the most polluted site (red). These genes have been annotated with the Kegg Orthology ID: K03149 and are involved in thiamin biosynthesis metabolic pathways. (b) COG annotation of each component, COG2104 (ThiS) and COG2022 (ThiG). (c) Multiple alignment of composite gene sequences of the family shown in (a).

| Evolution Rate | | | METHODS | | | |
|---|---|---|---|---|---|---|
| | | | Composite detection | | Composite family detection | |
| Component | Composite 1 | Composite 2 | CompositeSearch | FusedTriplets | CompositeSearch | MosaicFinder |
| 0.1 | 0.1 | 0.1 | 99.50 | 99.50 | 98.92 | 95.47 |
| 0.1 | 0.1 | 0.5 | 99.48 | 99.48 | 98.93 | 95.43 |
| 0.1 | 0.1 | 1.0 | 99.48 | 99.48 | 99.03 | 41.68 |
| 0.1 | 0.5 | 0.1 | 99.52 | 99.52 | 99.58 | 95.60 |
| 0.1 | 0.5 | 0.5 | 99.54 | 99.54 | 99.56 | 95.59 |
| 0.1 | 0.5 | 1.0 | 99.58 | 99.58 | 99.58 | 39.27 |
| 0.1 | 1.0 | 0.1 | 99.56 | 99.56 | 99.68 | 95.02 |
| 0.1 | 1.0 | 0.5 | 99.62 | 99.62 | 99.65 | 85.06 |
| 0.1 | 1.0 | 1.0 | 99.62 | 99.62 | 99.67 | 35.82 |
| 0.5 | 0.1 | 0.1 | 99.94 | 99.94 | 99.57 | 98.24 |
| 0.5 | 0.1 | 0.5 | 99.94 | 99.94 | 99.54 | 98.34 |
| 0.5 | 0.1 | 1.0 | 99.95 | 99.95 | 99.61 | 44.80 |
| 0.5 | 0.5 | 0.1 | 99.95 | 99.95 | 99.81 | 98.47 |
| 0.5 | 0.5 | 0.5 | 99.95 | 99.95 | 99.81 | 98.33 |
| 0.5 | 0.5 | 1.0 | 99.96 | 99.96 | 99.81 | 40.68 |
| 0.5 | 1.0 | 0.1 | 99.95 | 99.95 | 99.85 | 97.54 |
| 0.5 | 1.0 | 0.5 | 99.95 | 99.95 | 99.85 | 88.03 |
| 0.5 | 1.0 | 1.0 | 99.95 | 99.95 | 99.87 | 38.01 |
| 1.0 | 0.1 | 0.1 | 99.94 | 99.94 | 99.82 | 99.36 |
| 1.0 | 0.1 | 0.5 | 99.96 | 99.96 | 99.84 | 99.37 |
| 1.0 | 0.1 | 1.0 | 99.96 | 99.96 | 99.86 | 42.68 |
| 1.0 | 0.5 | 0.1 | 99.96 | 99.96 | 99.92 | 99.49 |
| 1.0 | 0.5 | 0.5 | 99.96 | 99.96 | 99.92 | 95.94 |
| 1.0 | 0.5 | 1.0 | 99.96 | 99.96 | 99.93 | 38.60 |
| 1.0 | 1.0 | 0.1 | 99.87 | 99.87 | 99.93 | 91.30 |
| 1.0 | 1.0 | 0.5 | 99.89 | 99.89 | 99.94 | 83.99 |
| 1.0 | 1.0 | 1.0 | 99.87 | 99.87 | 99.93 | 34.46 |

**Supplementary Table S1.** Detection of composite genes and composite gene families on simulated data. This table shows the true positive rate (TRP) for the detection of composite genes (CompositeSearch and FusedTriplets) and the detection of composite gene families (CompositeSearch and MosaicFinder). Depending on the algorithm, CompositeSearch can detect composite genes and composite gene families. For composite detection, TRP is defined as the percentage of genes identified as composite that are indeed composite in the simulation. For composite family detection, TRP is defined as the percentage of genes in the detected composite families that are indeed composite in the simulation. These percentages are computed on 189 possible combinations of parameters explained in the Methods section (component lengths and composite tree levels variation) replicated 100 times averaged over two-component and three-component composites for each of these 27 combinations of evolutionary rates (see Supplementary Figure 4 and Methods section).

| Evolution Rate | | | TPR | | | | FPR |
|---|---|---|---|---|---|---|---|
| | | | Composite 1 | | Composite 2 | | |
| Component | Composite 1 | Composite 2 | EXACT | NON EXACT | EXACT | NON EXACT | |
| 0.1 | 0.1 | 0.1 | 97.83 | 0.07 | 90.33 | 9.60 | 0.92 |
| 0.1 | 0.1 | 0.5 | 97.77 | 0.14 | 94.63 | 5.31 | 0.87 |
| 0.1 | 0.1 | 1.0 | 98.03 | 0.07 | 96.32 | 3.63 | 0.85 |
| 0.1 | 0.5 | 0.1 | 99.11 | 0.10 | 88.44 | 11.51 | 0.85 |
| 0.1 | 0.5 | 0.5 | 99.05 | 0.12 | 92.31 | 7.64 | 0.86 |
| 0.1 | 0.5 | 1.0 | 99.12 | 0.08 | 93.73 | 6.21 | 0.87 |
| 0.1 | 1.0 | 0.1 | 99.35 | 0.07 | 86.66 | 13.28 | 0.84 |
| 0.1 | 1.0 | 0.5 | 99.29 | 0.07 | 89.55 | 10.40 | 0.83 |
| 0.1 | 1.0 | 1.0 | 99.34 | 0.06 | 88.38 | 11.54 | 0.79 |
| 0.5 | 0.1 | 0.1 | 99.12 | 0.08 | 92.67 | 7.26 | 1.02 |
| 0.5 | 0.1 | 0.5 | 98.98 | 0.16 | 93.68 | 6.26 | 1.12 |
| 0.5 | 0.1 | 1.0 | 99.09 | 0.18 | 94.17 | 5.77 | 1.09 |
| 0.5 | 0.5 | 0.1 | 99.54 | 0.14 | 90.20 | 9.74 | 1.03 |
| 0.5 | 0.5 | 0.5 | 99.55 | 0.12 | 91.18 | 8.76 | 1.10 |
| 0.5 | 0.5 | 1.0 | 99.57 | 0.09 | 89.96 | 9.99 | 1.05 |
| 0.5 | 1.0 | 0.1 | 99.65 | 0.09 | 88.24 | 11.70 | 0.85 |
| 0.5 | 1.0 | 0.5 | 99.66 | 0.09 | 87.97 | 11.96 | 0.91 |
| 0.5 | 1.0 | 1.0 | 99.71 | 0.08 | 83.28 | 16.66 | 0.90 |
| 1.0 | 0.1 | 0.1 | 99.63 | 0.07 | 88.92 | 11.00 | 0.96 |
| 1.0 | 0.1 | 0.5 | 99.62 | 0.11 | 88.00 | 11.95 | 1.05 |
| 1.0 | 0.1 | 1.0 | 99.60 | 0.16 | 87.00 | 12.94 | 0.99 |
| 1.0 | 0.5 | 0.1 | 99.77 | 0.11 | 86.73 | 13.21 | 0.85 |
| 1.0 | 0.5 | 0.5 | 99.76 | 0.13 | 85.13 | 14.82 | 0.88 |
| 1.0 | 0.5 | 1.0 | 99.82 | 0.08 | 80.62 | 19.33 | 0.82 |
| 1.0 | 1.0 | 0.1 | 99.85 | 0.07 | 84.55 | 15.38 | 0.56 |
| 1.0 | 1.0 | 0.5 | 99.84 | 0.08 | 79.74 | 20.20 | 0.58 |
| 1.0 | 1.0 | 1.0 | 99.81 | 0.10 | 71.37 | 28.58 | 0.53 |

**Supplementary Table S2.** More detailed performance of CompositeSearch. This table shows the true positive rate (TPR) and the false positive rate (FPR) of CompositeSearch when applied on two-components composites (Composite1) and three-components composites (Composite2). Identification is described as EXACT when the correct number of components is found and NON EXACT otherwise. The number of replicates is the same as in Supplementary Table S1. The FPR values represent occurrences of component sequences detected as composite.

| | Site 1 | Site 2 | Site 3 |
|---|---|---|---|
| Site 1 | 1.0000000 | 0.9998911 | 0.9996207 |
| Site 2 | 0.9998911 | 1.0000000 | 0.9998151 |
| Site 3 | 0.9996207 | 0.9998151 | 1.0000000 |

**Supplementary Table S3.** $R^2$ values (correlation coefficients) of sites pairwise comparisons with Mantel test.

## III.1.2 Impact of genomic structure and mobility on gene remodeling in plasmids, allowing the evolution of new dependency systems.

Plasmids are essential in the transfer of genetic information between bacteria (Sorensen et al. 2005). They play a key role in the evolution of the prokaryotic world. The genomic organization of many plasmids has been described as modular, involving important functional and evolutionary consequences. For example, conjugative plasmids, involved in lateral transfers of genes between cells, are made up of well-characterized genetic modules, ie sets of genes encoding a common process. On the other hand, pairs of genes, particularly those involved in Toxin-Antitoxin (STA) systems, have been studied in plasmids (Van Melderen 2010). STAs are also referred to as dependency systems since the survival of their hosts depends on the presence of TA genes (Van Melderen and Saavedra De Bast 2009) (Figure 15).

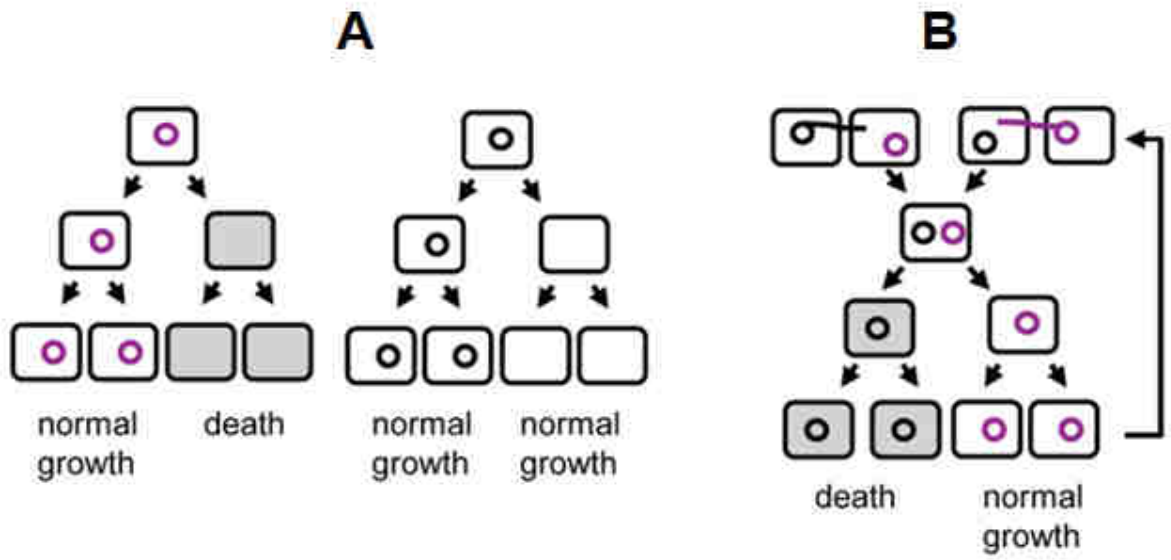



**Figure 15: Advantage conferred by plasmid-encoded TA systems.**
(A) Vertical transmission. TA systems increase plasmid prevalence in growing bacterial populations by post-segregational killing (PSK). PSK+ plasmid is shown in purple, left panel. Daughterbacteria that inherit a plasmid copy at cell division grow normally. If daughter bacteria do not inherit a plasmid copy, degradation of the labile antitoxin proteins by the host ATP-dependent proteases will liberate the stable toxin. This will lead to the selective killing of the plasmid-free bacteria (in gray). B) Horizontal transmission. Plasmid–plasmid competition. The PSK+ plasmid (in purple) and the PSK plasmid (in black) belong to the same incompatibility group and are conjugative. Under conditions in which conjugation occurs, conjugants containing both plasmids are generated. Because the two plasmids are incompatible, they can not be maintained in the same bacteria. (Van Melderen and Saavedra De Bast 2009)

These genes work together but are expressed separately, which is not the case for composite genes. Composite genes are formed by the fusion of at least 2 distinct genes (or fragment of genes). The potential impact of host cells, mobility, and plasmid genome structure on gene remodeling events leading to the evolution of these composite genes in plasmids has never been studied. Similarly, the functions of these composite genes and the rules for combining their components are very little known.

In the article n°5, we studied composite genes in plasmids in order to answer the following questions:

(1) In which type of plasmids are most of the composite genes found?

(2) What are the functions of these composite genes?

We used 4,393 complete genomes of plasmids (NCBI December 2014) to quantify the proportion of composite genes and to analyze the gene remodeling in these plasmids and some of its functional consequences. We have observed that plasmids with different proportions of composite genes are present in the same host, lineages and / or eventually same cell, indicating that selection for composite genes is only weakly bounded by the host. On the other hand, our results show that the different properties of plasmids, such as mobility and their genomic structure, have an impact on the remodeling of their genes. The functional analysis of these composite genes revealed the presence of composite genes combining components for which at least one component was not assigned to any of the functional categories COG, e.g. "X". We can use these "COG-X" composite genes as molecular Rosetta stones to decipher the hypothetical functionality of these associations. We report composite genes probably involved in the evolution of novel dependence modules, which possess at least one toxin or antitoxin component, as well as the evolution of plasmid encoded composite genes involved in cell cycle control and cell division. Therefore, genome remodeling on plasmids, although more constrained at the plasmid level than at the host level, can have important effects on the dynamics and evolution of the host cell population. This article is in preparation and will be submitted to the journal "Proceedings of the National Academy of Sciences".

# Plasmid's structure affects the distribution of remodelled genes within microbial populations

**JS. Pathmanathan, P. Lopez and E. Bapteste**

***In Preparation***

**ABSTRACT**

Plasmids are extrachromosomal genetic elements, which play important roles in their host cells, and affect the stability of microbial communities. Conjugative plasmids have been shown to contribute to the acquisition and exchange of genes by lateral gene transfer, introducing genetic variations in microbial communities, unraveling the plasticity of plasmid genomes in terms of their gene content. Moreover, plasmids of obligate intracellular parasites, which are less frequently mobilized, such as the plasmids of Borrelia, have been proposed to contribute to the generation of genetic variation from within their host cells via gene remodeling. Together, these observations support the hypothesis that plasmids are hosts to a diversity of novel, potentially adaptive, genes, arising via a diversity of introgressive processes. Here, we realized a systematic survey of the remodeled genes encoded on plasmids, followed by a functional analyses of these genes. We report that the proportion of remodeled genes (23% in average per genome) seems more affected by the nature of individual plasmids than by the phylogeny of the plasmid host cell. Furthermore, since some remodeled genes affect toxin-antitoxin systems and host cell division genes, we postulate that some remodeled genes on plasmids may affect the evolutionary dynamics of plasmids and their hosts.

## INTRODUCTION

Plasmids are central evolutionary players, which carry and/or mobilize genes across the prokaryotic world (Heinemann 1991; Sprague 1991; van Elsas and Bailey 2002; Smillie et al. 2010) (Conjugative plasmids: vessels of the communal gene pool). The genomic organisation of numerous plasmids has been described as modular, which has significant functional and evolutionary consequences (Bosi et al. 2011). For example, conjugative plasmids, involved in lateral gene transfer events between host cells, are comprised of well characterized genetic modules, i.e. sets of genes coding for a common process (Springael and Top 2004; Frost et al. 2005; Frost and Koraimann 2010; Smillie et al. 2010; Guglielmini et al. 2011). While the products of these genes functionally interact, genes from the *par* modules however are loosely connected because their encoded proteins are not directly physically linked after translation (Casart et al. 2008; Okibe et al. 2013; Li et al.

2015). Likewise, tighter couplings of genes on plasmids have also been well studied, in particular toxin-antitoxin addiction modules (Jaffe et al. 1985; Gerdes et al. 1986; Pandey and Gerdes 2005; Leplae et al. 2011; Unterholzner et al. 2013; Mruk and Kobayashi 2014; Fasani and Savageau 2015; Rocker and Meinhart 2016). These modules are famously involved in the maintenance and distribution of plasmids in the microbial world by a process of post-segregational killing. Microbial hosts, which after cellular division have lost plasmids harbouring such toxin-antitoxin modules are thus sentenced to death. When associated in operons as they have been repeatedly described (Zielenkiewicz and Ceglowski 2005; Van Melderen and Saavedra De Bast 2009; Yamaguchi et al. 2011), bork operon) genes forming the addiction modules are transcribed and translated together, which means that the toxin and antitoxin they produced interact but are present on physically separated molecules. By contrast, composite genes constitute a stronger form of genetic association(Huynen et al. 2000; Meheust et al. 2016). Composite genes unite genes (or genes fragments) from different gene families in a single open reading frame. Composite genes result from gene remodelling, which either involves a fusion process (when components encoding separate gene products in some genomes combine into a common open reading frame in other genomes), or involves a fission process (when a modular gene present in some genomes split into distinct components, which subsequently encode separate gene products in other genomes). In the former case, genes and proteins become physically coupled, in the latter case, they become physically decoupled. Gene remodelling consequently has predictable functional effects, such as easing domain-domain interactions between associated components (Meheust et al. 2016), enhancing the co-location of interacting proteins in the host cell since they will be present together at the same place and time (Tsoka and Ouzounis 2000; Yanai et al. 2001; Fani et al. 2007; Henry et al. 2016), or allowing for finer processual regulation (Snel et al. 2000). Significant proportions of plasmid-encoded composite genes have been reported in a general analysis including genomes from the three domains of life and mobile elements, including plasmids (~ 20%,(Jachiet et al. 2013)). However, the functions of these plasmid encoded composite genes, the rules of combinations of their components, and the potential impact of plasmids and host cells on the gene remodelling events, as well as the potential impact of these events on microbial populations have never been studied.

Here, we used 4,393 complete genomes of plasmids to quantify the proportion of composite genes per plasmid genome, and to analyze gene remodelling in plasmids, and some of its functional consequences. The nature of the host cell, the mobility and the genomic structure of the plasmids could affect the distribution of plasmid-encoded remodelled genes. Typically, the microbial host could impact the gene content of all its plasmids, as a result of this host lifestyle (free-living or

intracellular) and of the effective size of its populations, especially when plasmids impose a genetic load to their host cells. However, assuming selection on genes occurs more generally at multiple levels (Campos et al. 2015), one could also expect that the make-up of the plasmids themselves could influence their gene content. We tested whether hosts cells and/or plasmids had detectable effects on the presence of remodelled genes in plasmids. We observed that plasmids with different proportions of composite genes are present within the same lineage and/or cell, indicating that the selection for composite genes on plasmids is only weakly constrained by the microbial host. By contrast, linear plasmids contain in average significantly more composite genes than circular plasmids. Moreover, mobilizable plasmids contain significantly more composite genes than non-mobilizable and conjugative plasmids. These results indicate that plasmids properties impact the distribution of remodelled genes. Moreover, detailed analyses of the functions of components that were coupled or decoupled in plasmids show that gene remodelling in different genomic and mobility classes operated with different rules. Interestingly, these analyses unravelled composites genes combining components for which at least one component was not assigned to any functional COG category. Using these composite genes as Rosetta stones (Adai et al. 2004) to decipher the hypothetical functions of these associations, we propose that 1886 composite genes (from 244 clusters of homologous genes) relate to the evolution of unknown addiction modules, since they involve at least one toxin or antitoxin component, and 964 composite genes (from 28 clusters of homologous genes) relate to the evolution of genes involved in cell cycle control, cell division and chromosome partitioning. Therefore, gene remodelling on plasmids, while apparently more constrained at the plasmid level than at the host level, can have substantial effects on the dynamics and evolution of microbial populations.

# MATERIALS & METHODS

## Data

We downloaded 4,393 complete genomes of plasmids from NCBI (December 2014) which is composed by 3,951 circular and 442 linear plasmids. This led to a dataset of 338,930 protein sequences. We used CONJscan-T4SSscan (Guglielmini et al. 2013) to assign the mobility of each plasmid (Table.1).

| | Plasmid Shape | | |
|---|---|---|---|
| | Circular | Linear | *Total* |
| **CONJ** | 749 | 17 | ***766*** |
| **MOB** | 984 | 27 | ***1,011*** |
| **NOMB** | 395 | 33 | ***428*** |
| **Unknown** | 1,823 | 365 | ***2,188*** |
| ***Total*** | ***3,951*** | ***442*** | ***4,393*** |

**Table 1:** Complete genomes of plasmids mobility and structural information.

## Composite genes detection

### *Construction of the SSN*

SSN were constructed based on the cleaned result of an all-against-all BLAST sequence comparison. This preliminary step relies on a C++ program called *cleanBlastp*. *cleanBlastp* uniquely numbers each sequence in the BLAST output, and removes all self-hits, keeping the best hit (i.e. lowest E-value) amongst multiple hits between pairs of sequences. This preliminary step produces an input file which contains BLAST information about matches between pairs of sequences (qstart, qend, sstart, send), sequence length (qlen, slen) and their symmetrized similarity scores (E-value, pident). The results are then represented as an undirected network G=(V,E), where V is the set of sequences, and edge is (u,v) $\epsilon$ E if the similarity score $S_{uv}$ or $S_{vu}$ is higher than a user-defined threshold (here E-value $\leq$ 10, default Pident $\geq$ 30%).

***Definition of gene families***

Sequences were next clustered into clusters of homologous genes (CHG) in two usual steps. First, we defined connected components (CCs) by thresholding the SSN, keeping only edges when the mutual coverage between two sequences in the BLAST search ≥ 80%. When the minimum mutual sequence coverage criterion is high (≥ 80 %), each CC is commonly considered as a putative CHG (Jachiet et al. 2013; Corel et al. 2016; Meheust et al. 2016). Here, we further refined this first definition of CHG by implementing a mutual coverage score ($S_{mc}$) for each CC. $S_{mc}$ is equal to 1 when all hits between the nodes of the tested CC have a mutual coverage > 80% in the BLAST search, i.e. when no weaker edges exists between the nodes of that CC. If $S_{mc} < 1$, this CC was subjected to the Louvain community detection algorithm (Blondel et al. 2008), using C++ *igraph* 0.7.1 library (Csardi and Nepusz 2006). This second step of community detection allows to define, at a finer granularity, the groups of sequences forming communities (e.g. cliques and/or quasicliques) within the CC, which are finally considered as a CHG. Thus, each sequence from the original dataset was assigned to a given CHG.

***Detection of composite genes***

Composite genes were detected by checking whether a node's neighbors in the SSN belong to different CHG. If all neighbors of a node belong to only one CHG, this node is not a composite gene. If at least two neighbors of this node belong to distinct CHG, we used the sequence corresponding to the node as a reference and mapped the matches from all different CHG along that sequence. Each region of the reference sequence with matches from different CHG along a composite sequence corresponds to a component. For each component associated with a given reference, we computed an average position for the start of the component and an average position for the end of the component along the reference sequence. If there is no overlap between at least two components along the reference sequence, then the reference sequence is considered as composite, since the reference sequence is composed of at least two non-overlapping regions with homology to different CHG.

***Detection of composite gene families***

All nodes, for all CHG, were tested to determine whether a CHG is composite or not. Any CHG containing at least one composite gene according to the protocol above was considered as composite family.

## Statistical analysis

We analyzed the composite gene proportions for plasmids different characteristics (shape, mobility, host kingdom and phylum). We performed a pair wise Mann-Whitney-Wilcoxon test (p value <= 0.05) to verify whether the observed differences were significant or not. P values were adjusted using Bonferroni method. In order to check that the obtained results were not biased, for each case we performed a Jackknife test with 10,000 resampling. Resampling size was fixed to the smallest sample size.

# RESULTS & DISSCUSSION

Our approach aimed at detecting composite genes and composite gene families (see M&M). 66,083 CHG with 31,438 singletons (CHG with only one gene) were detected. 5,448 CHG (~ 8%) were tagged as composite gene family with 2,184 singletons. These results indicate that 76,997 (~ 23%) of plasmid genes are composite, which is more than in viruses (8%) (Jachiet et al. 2014). This high proportion of composite genes detected in plasmids suggests that plasmids could play a major role in the distribution of composite genes among bacteria.

## Host cell lineage weakly constrains the proportion of composite genes in plasmids

We analyzed the proportions of composite genes and composite gene families in a diversity of host taxa (Figure.1). When plasmids were grouped according to the Domain to which their host cell belonged, we observed that bacterial plasmids have a significantly higher average percentage of composite gene families than the archaeal and eukaryotic plasmids (Mann-Whitney-Wilcoxon test,

p-value ≤ 0.05 see M&M). Moreover, eukaryotic plasmids have a significantly lower average percentage of composite than bacterial and archaeal plasmids (Mann-Whitney-Wilcoxon test, p-value ≤ 0.05 see M&M), although the difference is less pronounced between eukaryotic and archaeal hosts. These differences may reflect the differences between the biology of prokaryotes and eukaryotes, these latter preferentially encoding abundant composite genes on their chromosomes rather than on their plasmids. However, the larger sample of bacterial plasmids in our dataset may also explain this result.

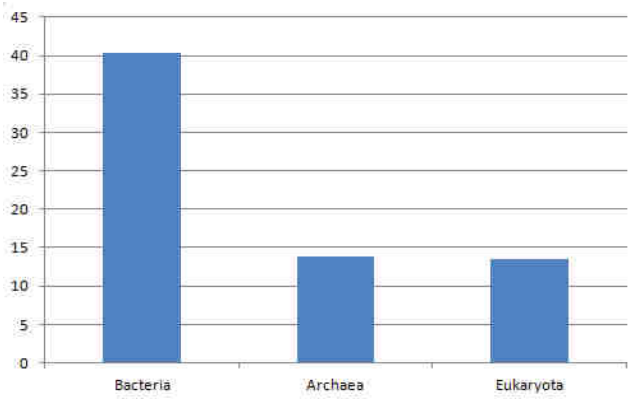



**Figure 1:** Average proportion of composite gene in the three Domains of life.

At the level of host phyla, we compared the percentage of remodeled genes between 26 groups. Phyla with low numbers of plasmids were likely to hosts too limited a number of composite genes to allow significant statistical tests, therefore we only retained host phyla with > 10 plasmids (Figure.2). The proportions of remodeled genes on plasmids vary widely between phyla (from around 60% to 5%). Plasmids of *Spirochaetes* and *Chlamydiae* show the highest proportions of composite gene families (i.e. greater than 50%), which is significantly higher than the other phyla (Mann-Whitney-Wilcoxon test, p-value ≤ 0.05, see M&M). Hosts of these plasmids, which are belonging to *Spirochaetes* and *Chlamydiae* phyla, are mostly obligate intracellular pathogens. Interestingly, it had been formerly suggested that hosts with such a reclusive lifestyle may benefit from introducing genetic variations from within, and that their extrachromosomal replicons could be used as 'organs enhancing gene evolution' (Halary et al. 2013)

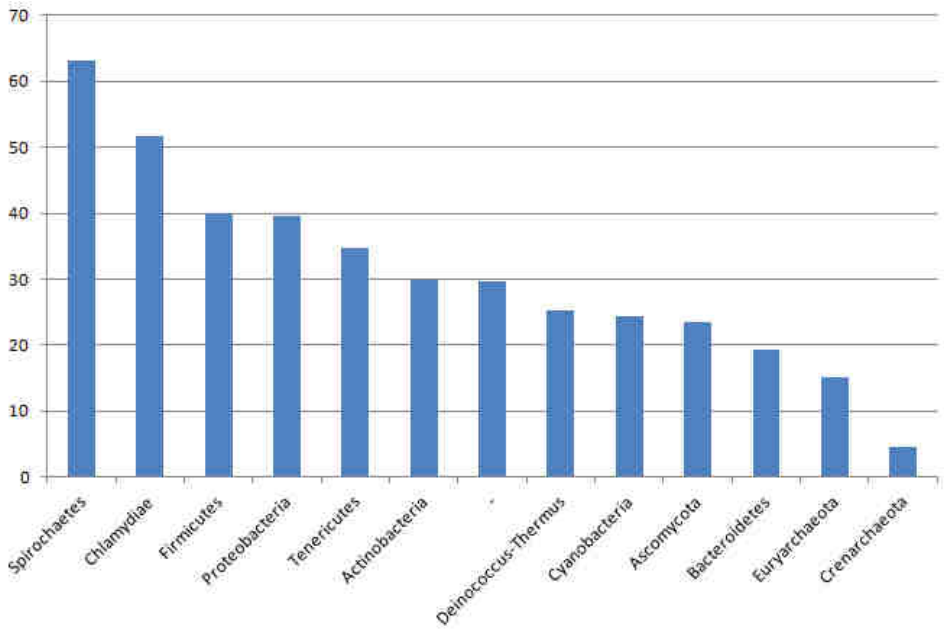



**Figure 2:** Average proportion of composite genes in various host phyla.

At the level of genera and species, we further noticed that the proportions of composite genes varied within a given host lineage. This was for example noticeable between the plasmids of *Borrelia* (Table.2). These differences within a genus, a species, and eventually a cell suggest that the proportion of plasmid encoded remodelled genes is not imposed by a general selective pressure exerted by the host on all its plasmids. Therefore, we searched for another possible cause for the differences between the proportions of composite genes in the biology of the plasmids themselves.

| PLASMID_ID | SIZE(bp) | SHAPE | MOBILITY | %COMPOSITES |
|---|---|---|---|---|
| NC_012261 | 26526 | circular | Unknown | 22.00 |
| NC_012262 | 17070 | circular | Unknown | 42.00 |
| NC_012257 | 30341 | circular | Unknown | 46.00 |
| NC_012268 | 30171 | circular | Unknown | 47.00 |
| NC_012251 | 30858 | circular | Unknown | 48.00 |
| NC_012253 | 60942 | circular | Unknown | 48.00 |
| NC_012264 | 30117 | circular | Unknown | 50.00 |
| NC_012266 | 30611 | circular | Unknown | 51.00 |
| NC_012256 | 8692 | circular | Unknown | 77.00 |
| NC_012229 | 17205 | linear | Unknown | 16.00 |
| NC_012241 | 54027 | linear | Unknown | 27.00 |
| NC_012245 | 27241 | linear | Unknown | 30.00 |
| NC_012232 | 27759 | linear | Unknown | 33.00 |
| NC_012233 | 29233 | linear | Unknown | 40.00 |
| NC_012236 | 28746 | linear | Unknown | 40.00 |
| NC_012238 | 24765 | linear | Unknown | 45.00 |
| NC_012243 | 18209 | linear | Unknown | 5.00 |
| NC_012248 | 27336 | linear | Unknown | 60.00 |
| NC_012240 | 29802 | linear | Unknown | 62.00 |

**Table 2 :** Proportion of composite genes in plasmids hosted by *Borrelia* species.

## Plasmid's biology impact the distribution of remodelled genes

We first considered the topology of the plasmid genome (Figure.3). Linear plasmids have a significantly higher average percentage of composite gene families than circular plasmids (Mann-Whitney-Wilcoxon test, p-value ≤ 0.05 see M&M). We do not favor the intuitive hypothesis that this difference could be explained by the relative simplicity of the linear plasmids. Whereas the addition of novel DNA at the termini of linear plasmids would provide a unique mechanisms for the evolution of remodelled genes to linear plasmids, because introducing DNA into a circular genomes requires the additional step of opening the genomes, we observed that remodeled genes were distributed along all the linear chromosomes, and not mainly at their termini.

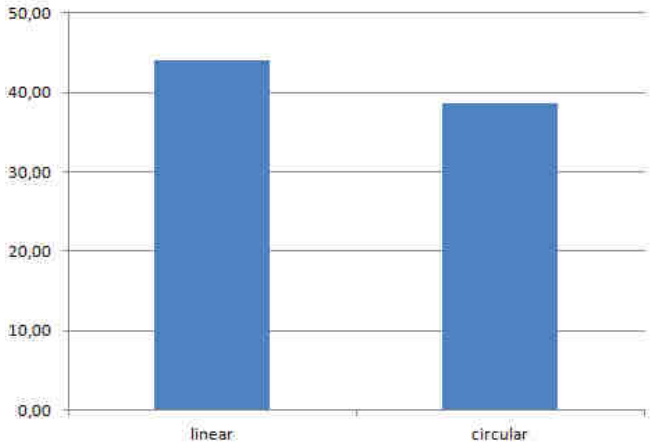



**Figure 3:** Average proportion of composite genes in linear and circular plasmids.

Second, we considered the plasmid mobility (Figure.4). The non-mobilizable (NOMOB) plasmids have a significantly lower average percentage of composite gene families than the mobilizable, conjugative and unassigned plasmids (Mann-Whitney-Wilcoxon test, p-value ≤ 0.05 see M&M). The mobilizable plasmids have a significantly higher average percentage of composite than conjugates and unassigned. Moreover, there was no significant difference between conjugative and unassigned plasmids. We verified that these differences were not trivially explained by size differences between these groups of plasmids. We detected no correlation between individual genome size and the proportion of plasmid encoded composite genes families, even though when plasmids are grouped into mobility classes genomes from non-mobile plasmids are (2-4x) larger than those of the mobilizable and conjugative plasmids, and that genomes of mobilizable plasmids are

smaller (3x) than those of conjugative plasmids. Rather than genome size, the frequency at which a given plasmid meets foreign DNA, seems a more natural explanation for these differences of remodeled gene families on different mobility classes. We speculate that mobile plasmids (be they mobilizable or conjugative) have a great opportunity to be in physical vicinity with a diversity of genomes than non-mobile DNA, which have a more restricted host distribution.

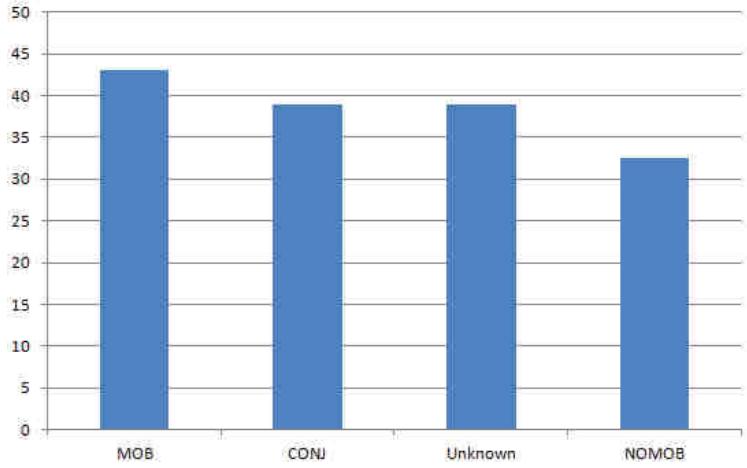



**Figure 4:** Average proportion of composite genes in mobilizable plasmids (MOB), non-mobilizable plasmids (NOMOB), conjugative plasmids (CONJ) and uncharacterized plasmids (Unknown).

## Functional analyses of coupled and decoupled components

We summarized the information about the functional assignation of pairs of components present along a composite gene for each genomic (circular, linear) (Figure.5) and mobility (mobile, non-mobile, conjugative and unassigned) (Figure.6) classes of plasmids. To this end, we split each composite gene into its constitutive component. For each component, its functional category was assigned using eggNOG-mapper (Huerta-Cepas et al. 2016; Huerta-Cepas et al. 2017). For each composite gene family, we computed the average proportion of components with a given COG function associated with all the composite genes belonging to that gene family. For each group of plasmids, we summarized the information about the functional assignation of pairs of components present along all composite gene families. We reported the proportion of all combinations of functional categories realized by pairs of components in a matrix. Thus, each matrix provides a functional profile of components associated in the composite genes for each group of plasmid (Figure 4 shows the matrix for sites 1, 2 and 3). If gene remodelling strongly depends on the functions of components, and if similar constraints apply for the functional association/dissociation

of genetic components across groups of plasmids, we expect similar functional profiles for the pairs of associated components across all groups of plasmids.

We used the Mantel test to compare the matrices and to verify if the profiles were similar or different, achieving a pairwise comparison of these matrices using the "CADM.post" function of the Mantel test from the ape (v. 3.5) library (Paradis et al. 2004) of the R statistical package (v. 3.2.5) (R Core Team 2016). This required modifying the "CADM.post" function to account for values on the diagonal and fixed the number of permutations to 999. This analysis revealed that remodelled genes present in different genomic and mobility classes had different functional profiles of components associations. In particular, circular plasmids explore a broader range of functional combinations than linear plasmids. By contrast, mobilizable plasmids realize a narrower range of functional combinations than non-mobile and conjugative plasmids, which harbour a broader diversity of metabolic remodelled genes. These distinct profiles can be explained either by the fact that gene remodelling follows different rules in these groups of plasmids, or that the composite genes associated with these groups of plasmids fulfil different functions, and therefore show different functional profiles. In both cases, this confirms the impact of plasmids on the association of components in remodelled genes.

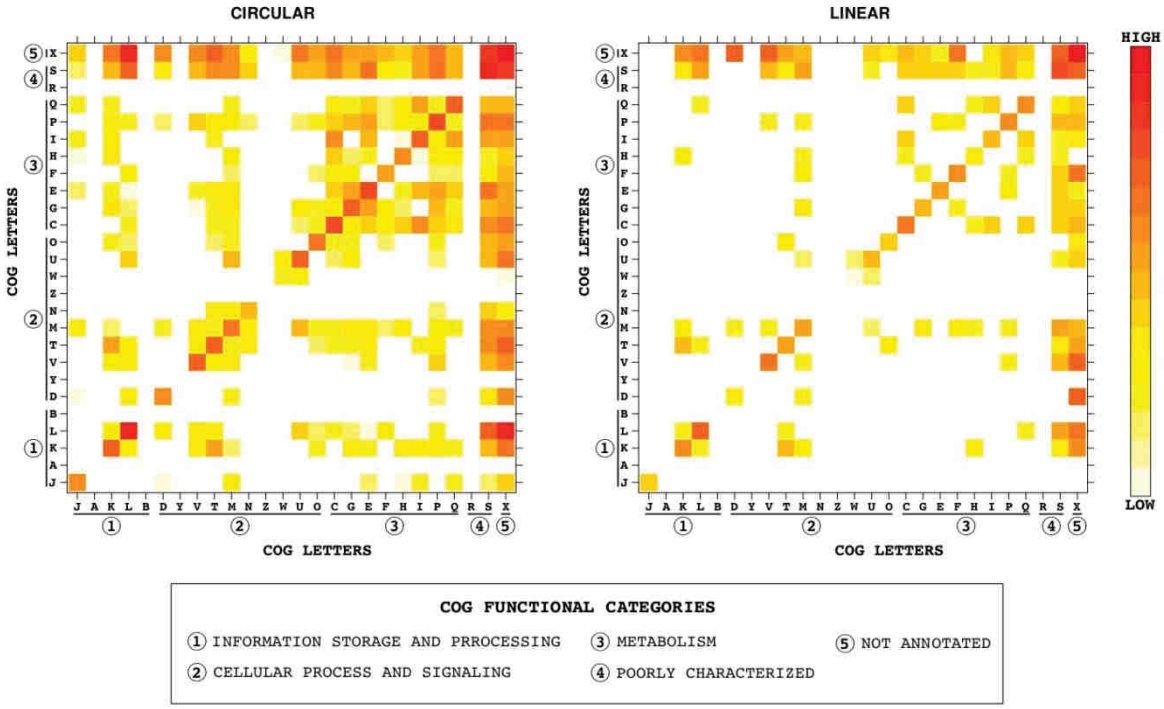



**Figure 5:** Relative abundance of two-component composite genes in linear and circular plasmids. COG categories of both components are given in abscissa and ordinate, and relative abundance is color coded from low (yellow) to high (red).

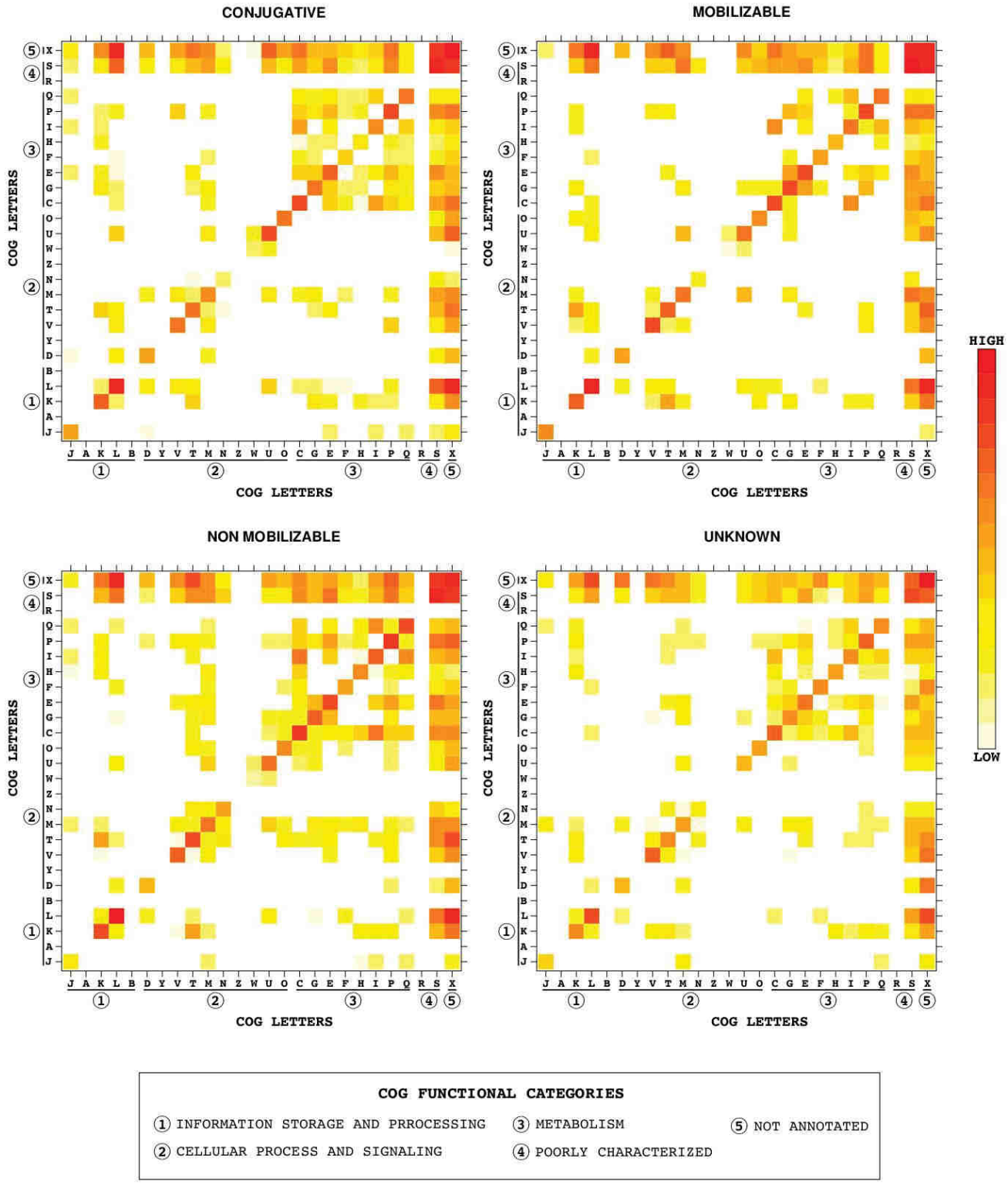



**Figure 6:** Relative abundance of two-component composite genes in plasmids, according to their mobility. Color code is the same as Fig. 5.

## Functional analysis of plasmid encoded composite genes

We annotated plasmids genes with using eggNOG-mapper. Roughly 40% of the plasmid genes were annotated, and the remaining sequences have been annotated as "Unknown". The proportion of composite genes in each COG functional category is represented in Figure 7. Remodelled genes were over-represented (Fisher test, p-value ≤ 0.05) in some critical functional categories: Transcription (K), Replication, recombination and repair (L), cell cycle control, cell division, chromosome partitioning (D), defence mechanisms (V), Transduction signal mechanisms (T), Energy production and conversion (C), Amino-acid transport and metabolism (E), Lipid transport and metabolism (I), inorganic ion transport and metabolism (P) and secondary metabolite biosynthesis, transport and metabolism (Q).

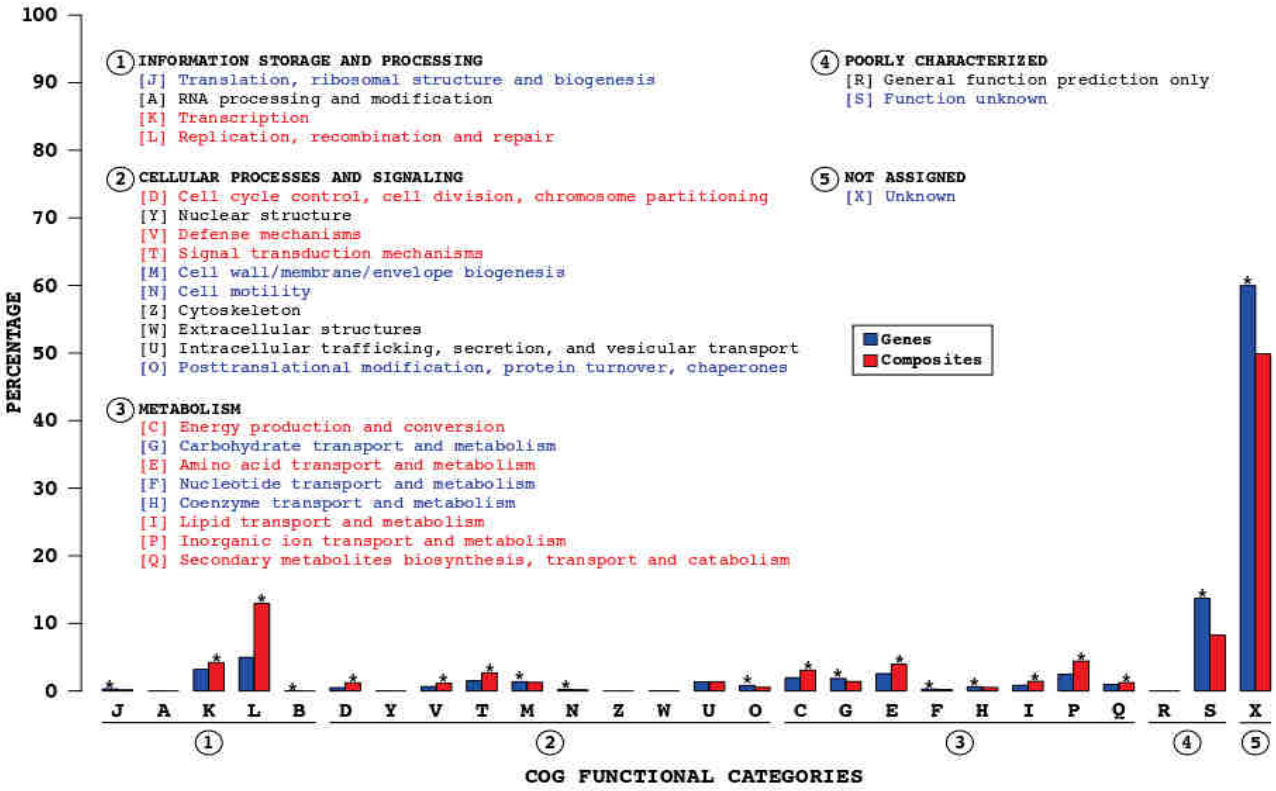



**Figure 7:** Relative abundance of COG functional categories in plasmid-encoded genes (blue) and composite genes (red). Categories that are significantly over-represented in composite genes (resp. in all genes) are highlighted in red (resp. in blue).

Consistently, some of the remodelled gene family had the potential to impact plasmids dynamics across the microbial populations. This was especially true for remodelled genes involved in Toxin-Antitoxin (STA) systems. We identified 1886 composite genes potentially related to STA, since these composite genes contained at least one toxin or one antitoxin component. 1855 of these

remodelled genes combined components of unknown functions with known toxins or antitoxins, and we predict they may constitute novel addiction modules. Other plasmid-encoded remodelled genes had the potential to directly affect host cell dynamics, since one of their components was either homologous to the cell division protein FtsK, or to the cell filamentation protein. If that prediction is correct, composite genes carried on plasmids could interfere with the host cell division.

## CONCLUSION

Plasmids host large proportions of remodelled genes. This high abundance confirms that plasmids are essential to introduce genetic variability in microbial populations, and that their fluid genomes are not only affected by lateral gene transfers of full-sized genes and selection of optimized genes; in fact plasmids genomes are also plastic at the sub-genic level. These remodeled genes are likely under some selection. We suggest that this selection possibly occurs at the gene level (typically for novel addiction modules) and at the plasmid level, since the proportions of plasmid encoded remodelled genes seems more affected by the biological properties of their host plasmids rather than by the evolutionary history of their host cells. Consequently, gene remodelling has the potential to alter the dynamics of microbial populations from within.

## FUNDING

J.S.P. and E.B. are funded by the European Research Council (FP7/2017-2013 Grant Agreement #615274).

*Conflict of Interest: none declared.*

## REFERENCES

Adai AT, Date SV, Wieland S, Marcotte EM. 2004. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. J Mol Biol 340(1):179-90.

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

Bosi E, Fani R, Fondi M. 2011. The mosaicism of plasmids revealed by atypical genes detection and analysis. BMC Genomics 12:403.

Campos M, Llorens C, Sempere JM, Futami R, Rodriguez I, Carrasco P, Capilla R, Latorre A, Coque TM, Moya A et al. . 2015. A membrane computing simulator of trans-hierarchical antibiotic resistance evolution dynamics in nested ecological compartments (ARES). Biol Direct 10:41.

Casart Y, Gamero E, Rivera-Gutierrez S, Gonzalez YMJA, Salazar L. 2008. par genes in Mycobacterium bovis and Mycobacterium smegmatis are arranged in an operon transcribed from "SigGC" promoters. BMC Microbiol 8:51.

Corel E, Lopez P, Meheust R, Bapteste E. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol 24(3):224-37.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems.

Fani R, Brilli M, Fondi M, Lio P. 2007. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. BMC Evol Biol 7 Suppl 2:S4.

Fasani RA, Savageau MA. 2015. Unrelated toxin-antitoxin systems cooperate to induce persistence. J R Soc Interface 12(108):20150130.

Frost LS, Koraimann G. 2010. Regulation of bacterial conjugation: balancing opportunity with adversity. Future Microbiol 5(7):1057-71.

Frost LS, Leplae R, Summers AO, Toussaint A. 2005. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 3(9):722-32.

Gerdes K, Rasmussen PB, Molin S. 1986. Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. Proc Natl Acad Sci U S A 83(10):3116-20.

Guglielmini J, de la Cruz F, Rocha EP. 2013. Evolution of conjugation and type IV secretion systems. Mol Biol Evol 30(2):315-31.

Guglielmini J, Quintais L, Garcillan-Barcia MP, de la Cruz F, Rocha EP. 2011. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. PLoS Genet 7(8):e1002222.

Halary S, McInerney JO, Lopez P, Bapteste E. 2013. EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol 13:146.

Heinemann JA. 1991. Genetics of gene transfer between species. Trends Genet 7(6):181-5.

Henry CS, Lerma-Ortiz C, Gerdes SY, Mullen JD, Colasanti R, Zhukov A, Frelin O, Thiaville JJ, Zallot R, Niehaus TD et al. . 2016. Systematic identification and analysis of frequent gene fusion events in metabolic pathways. BMC Genomics 17:473.

Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M et al. . 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44(D1):D286-93.

Huynen M, Snel B, Lathe W, 3rd, Bork P. 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10(8):1204-10.

Jachiet PA, Colson P, Lopez P, Bapteste E. 2014. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. Genome Biol Evol 6(9):2195-205.

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29(7):837-44.

Jaffe A, Ogura T, Hiraga S. 1985. Effects of the ccd function of the F plasmid on bacterial growth. J Bacteriol 163(3):841-9.

Leplae R, Geeraerts D, Hallez R, Guglielmini J, Dreze P, Van Melderen L. 2011. Diversity of bacterial type II toxin-antitoxin systems: a comprehensive search and functional analysis of novel families. Nucleic Acids Res 39(13):5513-25.

Li H, Angelov A, Pham VT, Leis B, Liebl W. 2015. Characterization of chromosomal and megaplasmid partitioning loci in Thermus thermophilus HB27. BMC Genomics 16:317.
Meheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A 113(13):3579-84.
Mruk I, Kobayashi I. 2014. To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. Nucleic Acids Res 42(1):70-86.
Okibe N, Suzuki N, Inui M, Yukawa H. 2013. pCGR2 copy number depends on the par locus that forms a ParC-ParB-DNA partition complex in Corynebacterium glutamicum. J Appl Microbiol 115(2):495-508.
Pandey DP, Gerdes K. 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. Nucleic Acids Res 33(3):966-76.
Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20(2):289-90.
R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Staistical Computing, Vienna, Austria.
Rocker A, Meinhart A. 2016. Type II toxin: antitoxin systems. More than small selfish entities? Curr Genet 62(2):287-90.
Smillie C, Garcillan-Barcia MP, Francia MV, Rocha EP, de la Cruz F. 2010. Mobility of plasmids. Microbiol Mol Biol Rev 74(3):434-52.
Snel B, Bork P, Huynen M. 2000. Genome evolution. Gene fusion versus gene fission. Trends Genet 16(1):9-11.
Sprague GF, Jr. 1991. Genetic exchange between kingdoms. Curr Opin Genet Dev 1(4):530-3.
Springael D, Top EM. 2004. Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. Trends Microbiol 12(2):53-8.
Tsoka S, Ouzounis CA. 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26(2):141-2.
Unterholzner SJ, Poppenberger B, Rozhon W. 2013. Toxin-antitoxin systems: Biology, identification, and application. Mob Genet Elements 3(5):e26219.
van Elsas JD, Bailey MJ. 2002. The ecology of transfer of mobile genetic elements. FEMS Microbiol Ecol 42(2):187-97.
Van Melderen L, Saavedra De Bast M. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet 5(3):e1000437.
Yamaguchi Y, Park JH, Inouye M. 2011. Toxin-antitoxin systems in bacteria and archaea. Annu Rev Genet 45:61-79.
Yanai I, Derti A, DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A 98(14):7940-5.
Zielenkiewicz U, Ceglowski P. 2005. The toxin-antitoxin system of the streptococcal plasmid pSM19035. J Bacteriol 187(17):6094-105.

## III.1.3 Evolution of genes and rules of gene remodeling during the transition and stabilization of animal multicellularity.

The transition from unicellular to multicellular organisms has occurred several times with independent origins in eukaryotes (Figure 16) (Ruiz-Trillo et al. 2007). The evolution of animals derived from protozoan lineages (*Ichthyosporea*, *Filasterea* and *Choanoflagellata*), which are closest unicellular eukaryotes relatives of metazoans, is one of the major transition in life's history (Torruella et al. 2012; Torruella et al. 2015). The mechanisms involved in this transition are not well known. Various theories stress on different causes of this event. Ecological considerations, considerations on cell-cell signalling and population effective size, as well as genetic changes have been discussed (King 2004). Regarding the genetic facet, gene duplication, exon shuffling and changes in genes regulatory networks have been underlined as major contributors in the origin of metazoans (King et al. 2008; Zmasek and Godzik 2011; Suga et al. 2012; Suga et al. 2013; Grau-Bove et al. 2017). These processes leave different clues in genomes. Additionally, novel gene families, invented within the metazoan lineage and/or subsequent molecular tinkering affecting preexisting sequences (e.g. insertion) may have contributed to the origin and maintenance of the multicellular lifestyle (Grau-Bove et al. 2017), and thus constitute important animal synapomorphies. These processes affecting the nature, number, length, and evolutionary rates of genes, are mutually non exclusive. They may have introduced substantial genetic variation in the sequences of metazoans that might be difficult to analyze comprehensively. For example, gene duplication, when associated with increased evolutionary rates, may have produced highly divergent and hardly detectable homologs. Likewise, within the genome, the association of genetic fragments belonging to unrelated gene family produces complex reticulate patterns. Network analyses provide a powerful broad-scale systematic comparative framework with the potential to unravel a diversity of genetic patterns, and therefore to investigate multiple aspects of molecular evolution and their potential connection to the evolution of multicellularity.

In the article n°6 in collaboration with Pr Iñaki Ruiz-Trillo, we performed a comparative approach using complete proteomes from 27 animals and 5 closely related unicellular relatives (representing the Holozoa clade: *Choanoflagellates*, *Filastereans* and *Ichthyosporeans*). We used sequence similarity networks to understand the evolution of genes and rules of gene remodeling during the transition from unicellular protists to animals, without relying on functional annotations for the definition of gene clusters and the identification of remodeled genes. This article is in preparation and will be submitted to the journal "Current Biology".

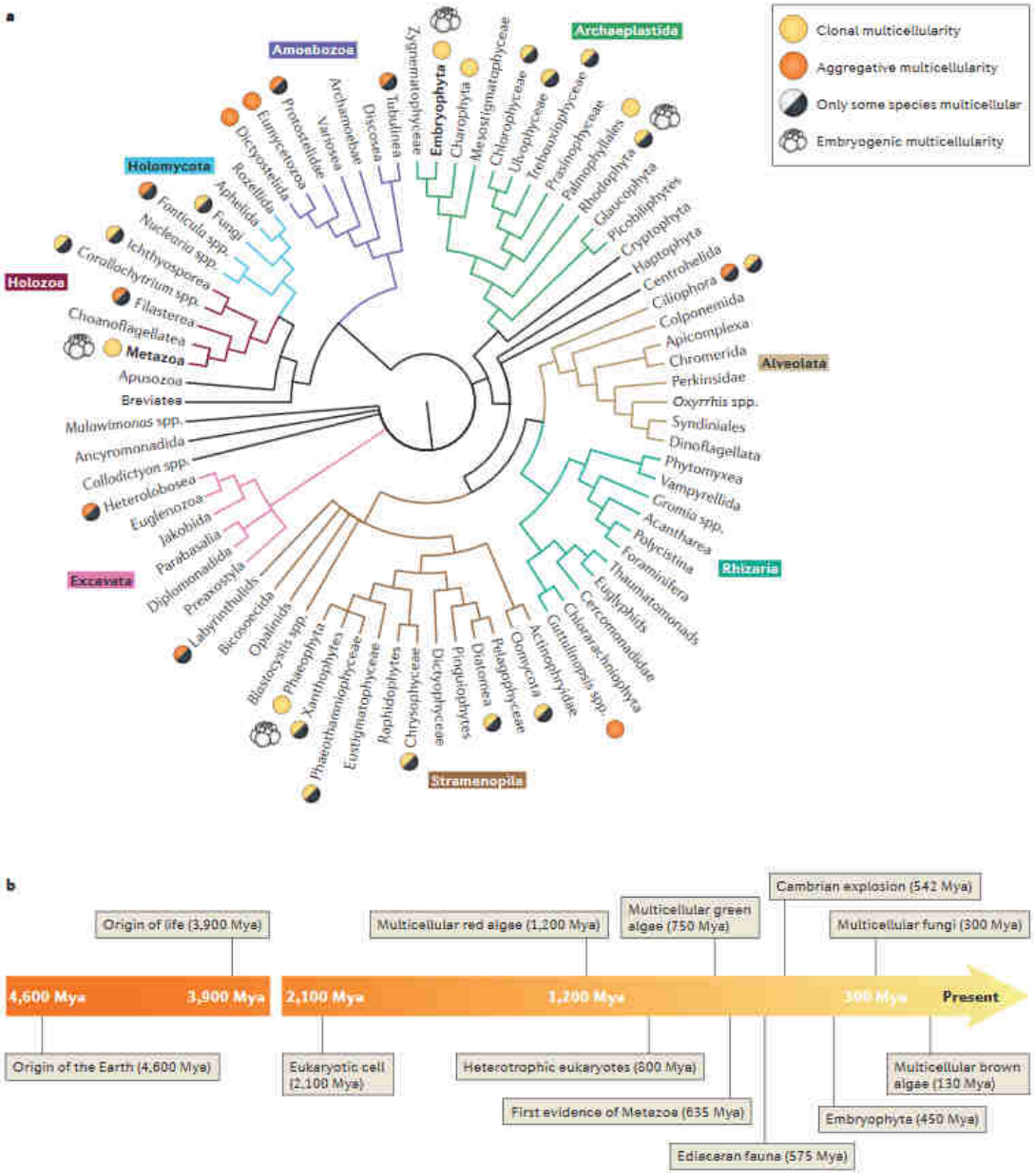



**Figure 16: The multiple origins of multicellularity.**
(a) The phylogenetic distribution of multicellularity among eukaryotes.
(b) A timeline of the origins of the major multicellular eukaryotic clades showing that transitions to multicellularity have occurred at very different times in the history of life. (Sebe-Pedros et al. 2017)

# Two bursts of specific and uniquely remodeled genes during animal evolution

***In preparation***

Jananan S. Pathmanathan[1] , Philippe Lopez[1], Iñaki Ruiz-Trillo[2,3,4]* and Eric Bapteste[1]*

1. Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), Paris F-75005, France
2. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Maritím de la Barceloneta 37-49, 08003 Barcelona, Spain
3. Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Catalonia, Spain.
4. ICREA, Pg. Lluís Companys 23, 08010 Barcelona.

## Summary

The emergence of animals from their unicellular ancestors is an important evolutionary event. Recent genomic data from unicellular relatives of animals had already shown that unicellular ancestors of animals were genetically much more complex than previously thought (Richter and King 2013; Suga et al. 2013; Sebe-Pedros et al. 2017). Analyses of gains and losses of domains and genes, based on Gene Ontology and Pfam, showed some gene innovations (i.e. novel gene families and protein domains) at the origin of animals (Zmasek and Godzik 2011; Fairclough et al. 2013; Suga et al. 2013; Grau-Bove et al. 2017). Moreover, it has been proposed that gene remodeling was an important mechanism for animal evolution (King et al. 2008; Suga et al. 2012; Suga et al. 2013; Grau-Bove et al. 2017). However, such analyses may be biased because most protein domains and Gene Ontology have been defined from animal taxa. To provide an unbiased analysis of genetic innovations during animal evolution, we used a complementary network-based approach, which does not rely on functional annotations for the definition of gene clusters and the identification of remodeled genes. We report

clusters of homologous genes and fusion genes, mapped onto the Holozoa (ie., animals and their closest relatives) tree. Our data confirmed the burst of genes associated with Metazoa and Eumetazoa (Grau-Bove et al. 2017). Moreover, we observed two successive trends of functional enrichment in animal evolution: first into Cytoskeleton and Extracellular Structures and later into Transcription. Additionally, we found that animals exploit a significantly different and broader range of functional combinations than their unicellular relatives. In particular, animals display unique combinations of genetic segments associated with extracellular matrix, signal transduction mechanisms and transcription. This suggest that metazoans and eumetazoans expanded their set of specific genes, while exploring a larger space of functional combinations in their remodeled genes. Overall, we report two major remodeling of the genetic landscape during animal evolution.

## Results and Discussion

To understand the evolution of genes during the transition from unicellular protists to animals, we performed a comparative approach using complete proteomes from both animals and their closest unicellular relatives: choanoflagellates, filastereans, and ichthyosporeans (the Holozoa clade (Torruella et al. 2012; Torruella et al. 2015)). We BLASTed all these proteins against each other to construct a protein similarity network, which allowed us to define clusters of homologous genes (CHG, i.e. connected sets of proteins, including proteins with a BLAST Evalue < $1E^{-5}$, ≥ 30%ID, and ≥ 80% mutual cover). This clustering approach provided us with 299,446 clusters (CHG). We assigned each CHG to a specific holozoan clade by searching for homologs of the CHG members in the entire NCBI database (January 2015). All homologs of the CHG were classified according to the NCBI taxonomy, and all CHG distributed in non-holozoan taxa (including prokaryotes) were considered as more ancient than the Holozoa. We also functionally annotated CHGs using KOG categories in order to further interpret our findings.

In order to understand the evolutionary origin of holozoan-specific CHGs, we

mapped them onto the holozoan tree using Dollo parsimony. This procedure allowed us to infer the CHG content of ancestral nodes within that tree, as well as their associated functional content. We observed a continuous evolution of CHGs along the tree of Holozoa, with two bursts: one at the Metazoa clade (N4, Figure 1) and the other at the Eumetazoa (Bilateria + Cnidaria) clade (N6, Figure 1). These observations are consistent with previous works (Zmasek and Godzik 2011; Grau-Bove et al. 2017), that were based on different and more standard methods such as Gene Ontology and/or Pfam protein domains assignations. Interestingly, these two bursts were associated with functions that are known to be relevant with multicellularity, and with two functional trends, spanning over multiple ancestral nodes of the tree (Supplementary Table 2). In particular, we observed an enrichment of CHGs associated to Cytoskeleton (Z) and Extracellular Structures (W) at the onset of Metazoa. This enrichment in Cytoskeleton CHG had started at the Choanozoa (Metazoa + Choanoflagelates) clade (Brooke and Holland 2003; Cavalier-Smith and Chao 2003). We also observed successive enrichments in CHGs associated to Transcription (K) in the Placozoa + Eumetazoa clade, and in the Eumetazoa. Thus, enrichments in genes involved in specific functions initiated in an ancestral lineage continued over long evolutionary periods. Such contingent trends (in this case first the complexification of the cytoskeleton and the extracellular structure, then the complexication of transcription) may have played a role in the emergence and further evolution of animals. It should be noted however, that CHGs without known functions (corresponding here to the S+X KOG categories) also significantly increased within animals. This observation suggests that many genes without known functions may have played important roles in animal origin and evolution.

Exon and domain shuffling have been proposed as an important mechanism involved in the evolution of multicellular lineages (King et al. 2008; Zmasek and Godzik 2011; Suga et al. 2012; Suga et al. 2013; Grau-Bove et al. 2017). These events lead to gene remodeling. To better understand some rules of gene remodeling at the onset of animals, we also used protein similarity networks to identify composite genes

in Metazoans and their closest unicellular relatives, without depending on domain annotations. A composite gene is formed through evolutionary combinatorial processes such as fusion and recombination of segments derived from different gene families or fission. Sequence similarity networks, where each node represents a unique sequence and each edge represents the similarity between connected sequences, appear to be well suited to identify and study this genetic mosaicism (Alvarez-Ponce et al. 2013; Bapteste et al. 2013; Jachiet et al. 2013). For each Holozoan clade, we distinguished three classes of composite CHGs: i) CHGs that evolved *via* the fusion of genetic material already present in the ancestor (novel fusion CHGs), ii) CHGs that appeared in the lineage, but underwent fission events in subsequent lineages (fission CHGs), and iii) CHGs for which the polarisation into fusion or fission CHGs was unclear, since these composite genes were comprised of components with complicated evolutionary histories. To compare the distribution of all these remodeled genes in unicellular and multicellular Holozoa, we mapped them onto the Holozoa tree using Dollo parsimony. This unraveled that the proportion of composite genes amongst the new CHGs of each clade is in general limited (within the range of 7-25%). In particular, there was a continuous evolution of novel fusion CHGs along the Holozoa tree, yet in limited and rather constant proportion (16,2% at the Metazoa, 15,6% at the Eumetazoa). Interestingly, enrichment of fusion CHGs in Metazoa and Eumetazoa only concerned CHG of unknown functions. However, in the (Placozoa + Eumetazoa) clade, we observed an enrichment of fusion CHGs associated with transcription and extracellular matrix.

We then focused on the components of the fusion CHGs to further investigate which specific functions were associated during the evolution of these composite CHGs, for each ancestral node of the Holozoa tree. We summarized the frequency of associations of components of fusion CHGs, based on the functional categories of these components, thus producing matrices of functional associations within fusion CHGs (Figure 1) for all animals and their closely related unicellular lineages, as well as for each ancestral node of the Holozoa tree (Figure 1). These matrices were then compared using a Mantel test to test whether they described significantly different functional

associations. Interestingly, we observed significant major changes in the rules of remodeling in animals compared to their unicellular ancestors, especially in the same two clades (Metazoa and Eumetazoa) for which we observed a burst of CHG evolution (Figure 1). These two types of genetic innovations, i.e. a burst of original combinations leading to fusion CHGs and a burst of CHG evolution, may not be directly causally related, because only a minority of novel CHGs are fusion CHGs. However, our finding suggests that these two types of expansions (i.e. one introducing novel specific CHG and another introducing new ways of combining genes) were going on simultaneously during animal history. Thus, the genomes of the ancestors during early animal evolution were remarkably dynamic. This genetic dynamism can be further witnessed with the detection of fission CHGs, since this class of composite CHGs suggests that a certain proportion of components forming new composite genes along the holozoan tree tend to get dissociated later during animal evolution (Supplementary Table 1).

We further compared the functional composition of fusion CHGs of animals and of their closely related protists to understand whether the diversity of these combinations could be related to key functions for animal evolution. We found that animals explored the space of functional combinations more extensively than their close unicellular relatives (Figure 2). For example, components involved in extracellular structures (W) were associated with components involved in 13 other functional categories in animals (Figure 2 C), something that is not happening in the unicellular holozoan taxa. Moreover, fusion CHGs of animals presented exclusive functional associations, which were not observed in their close unicellular relatives in our dataset. Interestingly, a very large fraction of these exclusive functional combinations in fusion CHGs concerned functions that were likely important for animal evolution. A first group of such unique combinations implicated components involved in signal transduction mechanisms (T), namely “T+T” and “T+J”, suggesting that signal transduction was remodeled during animal evolution. Coupling signal transduction with translation, ribosomal structure and biogenesis (the “T+J” fusion CHG) may in particular have affected the regulation of protein synthesis. A second group of animal-specific functional combinations

implicated components involved in transcription (K), giving rise to “K+K” and “X+K” fusion CHGs. This original remodeling of transcriptional functions fits well with findings indicating that transcription factors have likely played a major role in animal evolution. This functional remodeling of transcriptional regulation may have had a relevant role in the development of the fine tune and cell-type-specific transcriptional regulation observed in extant animals (Meyerowitz 2002; Levine and Tjian 2003; de Mendoza et al. 2013). Indeed most of those composite CHGs novel for eumetazoans are homeobox genes, which is known to have expanded in eumetazoans (de Mendoza et al. 2013). A third group of unique combinations implicated components involved in Extracellular structures (W), i.e. the “W+W” and “O+W” fusion CHG. Such remodeling are consistent with the fact that animal cells operate in a different environment than their close unicellular relatives, since animal cells must sense the environment of their tissue, as well as the signal coming from other tissues and organs. Here, we observed some composite CHGs involving syntrophin, laminins, integrins, and other extracellular components involved in adhesion and signaling. Unique combinations of the O category, i.e. the “O+O” fusion CHGs, were also exclusively observed in animal fusion CHGs. The repeated implication of the O category, coding posttranslational modification, protein turnover, and chaperones, into exclusive animal fusion CHGs, is also consistent with the observation that animals substantially remodel their proteins, as assessed, for example, by their increased abundance in ubiquitin (Grau-Bove et al. 2015). Likewise, unique combination of the Z category i.e. “Z+Z” fusion CHG in animals, combining components involved in the cytoskeleton, matches well with the discovery of new motor proteins in animals (Sebe-Pedros et al. 2014).

Overall, our data confirm, using a complementary approach, that animal genomes encode a significantly larger proportion of novel genetic regions compared to their ancestors, some of them resulting from fusion events. More importantly, our results show that animals use a larger genetic functional landscape than their unicellular relatives in composite genes, including novel combinations of genetic regions that are significantly enriched in “multicellular” functions such as extracellular matrix,

cytoskeleton, signal transduction and transcriptional regulation.

### Conclusion

We used an inclusive approach, largely complementary to analyses conducted in other studies, as our work did not *a priori* rely on standard Gene Ontology and Pfam definition, and allowed us to investigate 299,446 of CHGs in a single analysis, providing a broad picture on the genetic evolution associated with the transition to animal multicellularity. In particular, we observed both an increase in novel CHGs in animal lineages, and an increase in the diversity of the functional combinations giving rise to animal-specific fusion CHGs. Both of these bursts of genetic innovation at the Metazoa and Eumetazoa clades involved functions that were likely critical for the emergence of multicellular animals. Thus, our work provides a novel demonstration that genome evolution was particularly dynamic at the onset of animals, both at the genetic and sub-genetic levels.

## Materials &Methods

### Constitution of the dataset

We used the proteomes from 27 animal taxa and 5 closely related protist genomes representing the Holozoa clade: choanoflagellates, filastereans and ichthyosporeans. In total we had 855,506 protein sequences (Supplementary Table 1). We used eggNOG-mapper (Huerta-Cepas et al. 2016; Huerta-Cepas et al. 2017), in DIAMOND mode with the default parameters to annotate the protein sequences. Sequences without significant hits were annotated as X.

### Definition of CHG, and detection of composite CHG

We constructed a sequence similarity network (SSN) using the results of an all-against-all BLASTP (Altschul et al. 1990) of 855,506 sequences. The parameters

used for the BLASTP are: -seg yes -soft_masking true -max_seq_target 5000. In this undirected network, two proteins are connected based on their similarity scores (E-value <= 1e-5, Pident >= 30%). The SSN has been symmetrised by keeping only the best match of each pairwise comparison. A CHG is a cluster of homologous genes with high connectivity, in which connected sequences display significant BLAST E-values ≤ 1E-5, mutual covers ≥ 80%, Pident ≥ 30%.

We detected the composite genes and composite CHGs in this SSN using *CompositeSearch* (Pathmanathan JS et al, 2017). Composite genes are detected as a result of the fusion of partial or complete non-homologous DNA fragment, called component, or as a result of fission from a larger gene into dissociated persistent fragment. CompositeSearch generalizes the use of similarity networks to detect composite and component CHGs.

**Classification of CHGs**

Composite CHGs have been classified in 3 main categories (fusion, fission and non-polarisable) comparing their position and their components position in the tree (Supplementary Figure 1, for more details). This classification depend on the tree topology, e.g composite genes at basal node of a tree cannot be classified as a fusion.

## Acknowledgements

J.S.P. and E.B. are funded by the European Research Council (FP7/2017-2013 Grant Agreement #615274).

## REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A 110(17):E1594-603.

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L et al. . 2013. Networks: expanding evolutionary thinking. Trends Genet 29(8):439-41.

Brooke NM, Holland PW. 2003. The evolution of multicellularity and early animal genomes. Curr Opin Genet Dev 13(6):599-603.

Cavalier-Smith T, Chao EE. 2003. Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. J Mol Evol 56(5):540-63.

de Mendoza A, Sebe-Pedros A, Sestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci U S A 110(50):E4858-66.

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ et al. . 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome Biol 14(2):R15.

Grau-Bove X, Sebe-Pedros A, Ruiz-Trillo I. 2015. The eukaryotic ancestor had a complex ubiquitin signaling system of archaeal origin. Mol Biol Evol 32(3):726-39.

Grau-Bove X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. Elife 6.

Huerta-Cepas J, Forslund K, Pedro Coelho L, Szklarczyk D, Juhl Jensen L, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M et al. . 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44(D1):D286-93.

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29(7):837-44.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I et al. . 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451(7180):783-8.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature 424(6945):147-51.

Meyerowitz EM. 2002. Plants compared to animals: the broadest comparative study of development. Science 295(5559):1482-5.

Richter DJ, King N. 2013. The genomic and cellular foundations of animal origins. Annu Rev Genet 47:509-37.

Sebe-Pedros A, Degnan BM, Ruiz-Trillo I. 2017. The origin of Metazoa: a unicellular perspective. Nat Rev Genet 18(8):498-512.

Sebe-Pedros A, Grau-Bove X, Richards TA, Ruiz-Trillo I. 2014. Evolution and classification of myosins, a paneukaryotic whole-genome approach. Genome

Biol Evol 6(2):290-305.
Suga H, Chen Z, de Mendoza A, Sebe-Pedros A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sanchez-Pons N et al. . 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325.
Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. Sci Signal 5(222):ra35.
Torruella G, de Mendoza A, Grau-Bove X, Anto M, Chaplin MA, del Campo J, Eme L, Perez-Cordon G, Whipps CM, Nichols KM et al. . 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr Biol 25(18):2404-10.
Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol Biol Evol 29(2):531-44.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol 12(1):R4.

**Figure 1:** Evolution of the number of Clusters of Homologous Genes (CHG) along the Holozoan tree. The number of gains and losses of CHGs (used as a proxy for gene families) is given above each internal branch in blue and red respectively. The average proportion of composite genes in genomes (resulting from fusion of components in purple, from fission in green, and not polarisable in orange) is given as a pie charts for extant phyla (tips), as well as for ancestors, as reconstructed by Dollo parsimony. For each internal branch, composite genes that were assumed to be synapomorphies by Dollo parsimony had both their fragments functionally assigned and the relative abundance of pairs of functions is displayed as a matrix. Axes of the matrices are the 20 functional KOG categories.

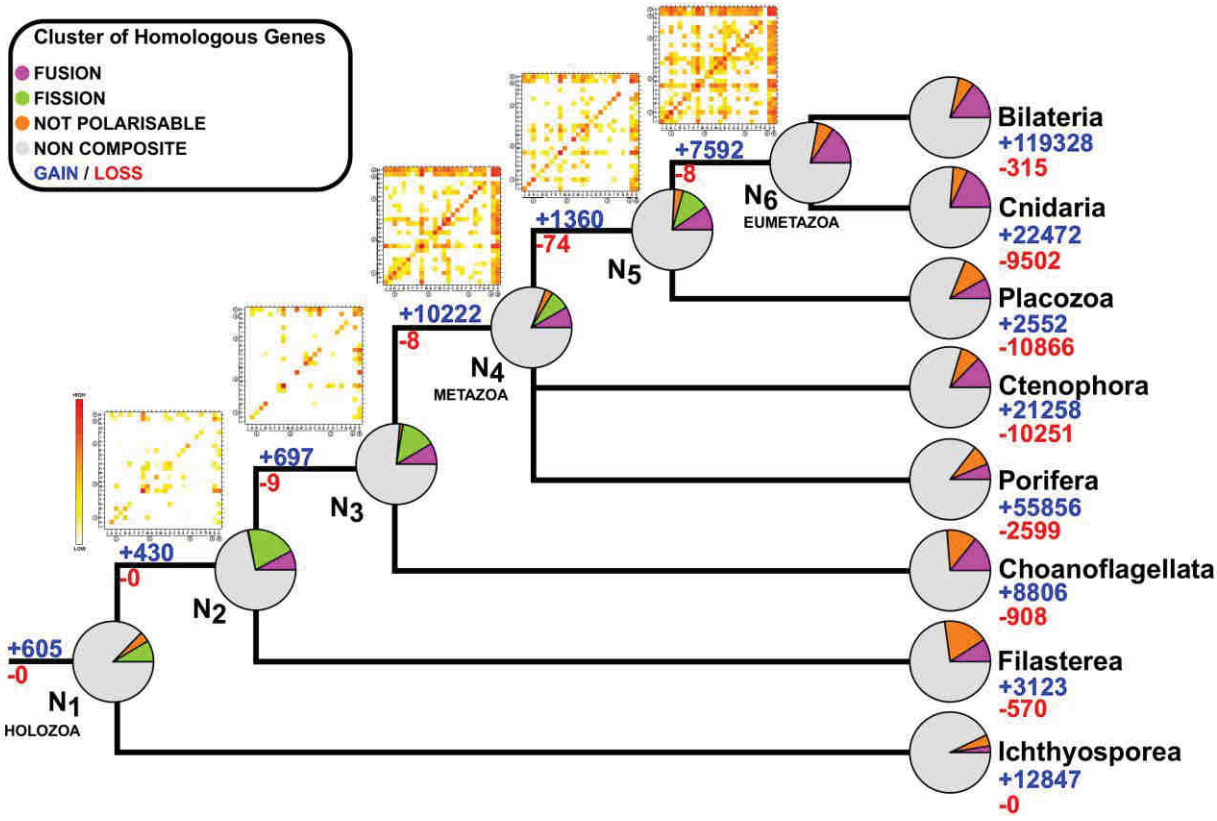
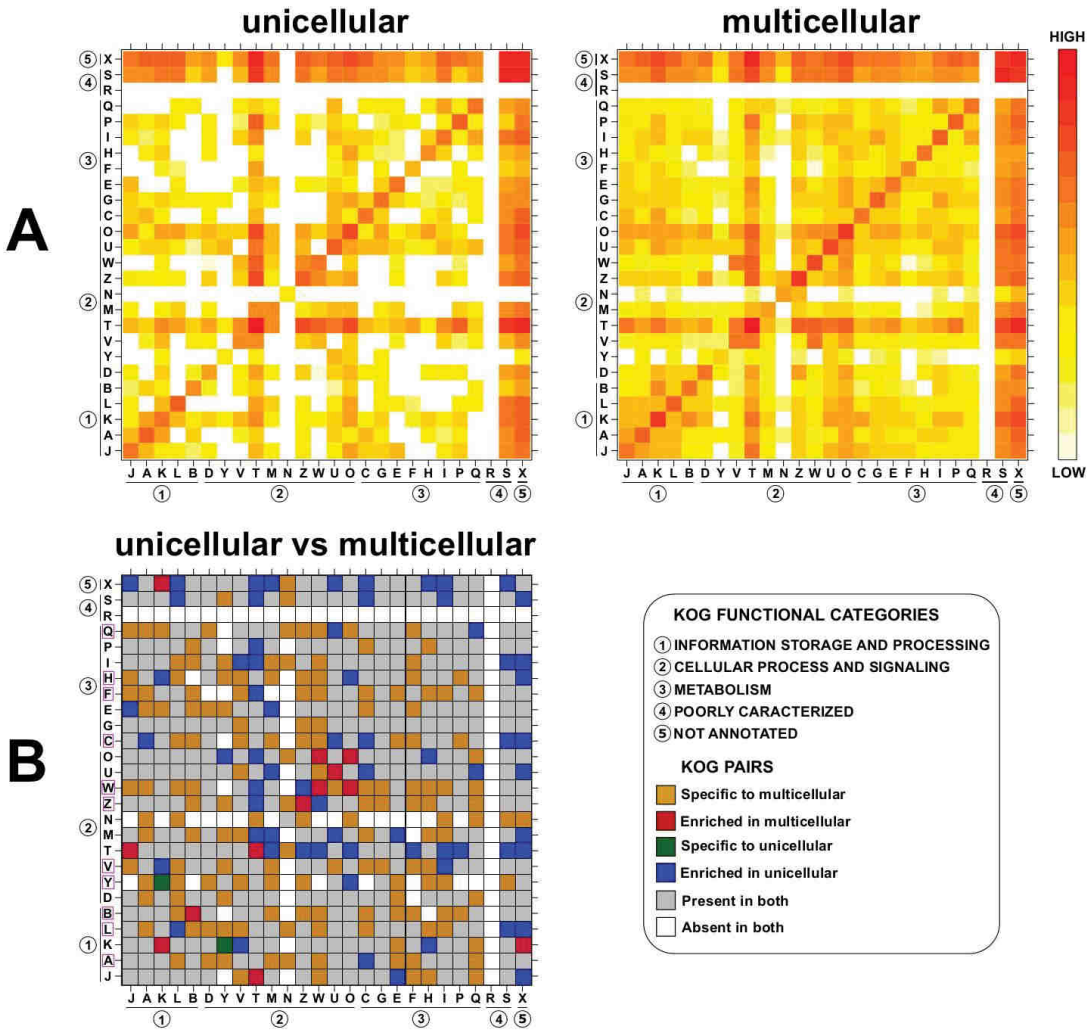


**Figure 2:** Functional associations in holozoan composite genes. (A) Following the same representations as in Figure 1, the relative abundance of pairs of functions found in composite genes is displayed for extant unicellulars of the dataset (top left) and for extant multicellular (top right). (B) Both matrices are contrasted in the bottom left matrix, where pairs of functions specific to or enriched in multicellulars are displayed in orange and red respectively, and pairs of functions specific to or enriched in unicellulars are displayed in green and blue respectively. KOG functions that appear mostly in multicellular composite are highlighted with a pink square (matrices are symmetrical).



**Supplementary Table 1 :** Information about the genome used for the analysis

| Genome | Muli- or Unicellular | Taxa |
|---|---|---|
| *Branchiostoma floridae* | Multicellular | Bilateria |
| *Caenorhabditis elegans* | Multicellular | Bilateria |
| *Callorhinchus milii* | Multicellular | Bilateria |
| *Capitella teleta* | Multicellular | Bilateria |
| *Ciona intestinalis* | Multicellular | Bilateria |
| *Crassostrea gigas* | Multicellular | Bilateria |
| *Danio rerio* | Multicellular | Bilateria |
| *Drosophila melanogaster* | Multicellular | Bilateria |
| *Echinococcus multilocularis* | Multicellular | Bilateria |
| *Homo sapiens* | Multicellular | Bilateria |
| *Ixodes scapularis* | Multicellular | Bilateria |
| *Lottia gigantea* | Multicellular | Bilateria |
| *Mus musculus* | Multicellular | Bilateria |
| *Saccoglossus kowalevskii* | Multicellular | Bilateria |
| *Strigamia maritima* | Multicellular | Bilateria |
| *Tribolium castaneum* | Multicellular | Bilateria |
| *Xenopus tropicalis* | Multicellular | Bilateria |
| *Acropora digitifera* | Multicellular | Cnidaria |
| *Hydra magnipapillata* | Multicellular | Cnidaria |
| *Nematostella ectensis* | Multicellular | Cnidaria |
| *Mnemiopsis leidyi* | Multicellular | Ctenophora |
| *Pleurobrachia bachei* | Multicellular | Ctenophora |
| *Trichoplax adhaerens* | Multicellular | Placozoa |
| *Amphimedon queenslandica* | Multicellular | Porifera |
| *Leucosolenia complicata* | Multicellular | Porifera |
| *Oscarella carmella* | Multicellular | Porifera |
| *Sycon ciliatum* | Multicellular | Porifera |
| *Monosiga brevicollis* | Unicellular | Choanoflagellata |
| *Salpingoeca rosetta* | Unicellular | Choanoflagellata |
| *Capsaspora owczarzaki* | Unicellular | Filasterea |
| *Creolimax fragrantissima* | Unicellular | Ichthyosporea |
| *Sphaeroforma arctica* | Unicellular | Ichthyosporea |

**Supplementary Table 2:** Functional enrichment of composite and non-composite genes at each ancestral node represented in the tree Figure 1.

| | ENRICHED FUNCTION | |
|---|---|---|
| **NODES** | **GENE** | **COMPOSITE** |
| **N1** | - | - |
| **N2** | [T] Signal transduction mechanisms<br>[S] Function unknown | - |
| **N3** | [Z] Cytoskeleton | - |
| **N4** | [Z] Cytoskeleton<br>[C] Energy production and conversion<br>[W] Extracellular structures | [S] Function unknown<br>[X] Not annotated<br>[W] Extracellular structres |
| **N5** | [S] Function unknown<br>[K] Transcription | - |
| **N6** | [S] Function unknown<br>[X] Not annotated<br>[K] Transcription | [X] Not annotated |

**Supplementary Table 3:** Proportion of composite genes (Fusion, Fission, non-polarisable) and non-composite genes at each internal node and each tip of the tree presented in Figure 1.

| POSITION | FUSION | FISSION | NON POLARISABLE | NON COMPOSITE |
|---|---|---|---|---|
| N1 | 0,00 | 8,60 | 4,20 | 87,20 |
| N2 | 7,80 | 20,10 | 0,30 | 71,80 |
| N3 | 8,50 | 13,90 | 1,30 | 76,30 |
| N4 | 8,40 | 7,80 | 3,10 | 80,70 |
| N5 | 9,60 | 11,00 | 3,60 | 75,80 |
| N6 | 15,60 | 0,00 | 6,50 | 77,90 |
| *Bilateria* | 15,30 | 0,00 | 6,30 | 78,40 |
| *Cnidaria* | 17,90 | 0,00 | 5,80 | 76,30 |
| *Placozoa* | 8,00 | 0,00 | 10,60 | 81,40 |
| *Ctenophora* | 12,60 | 0,00 | 7,90 | 79,50 |
| *Porifera* | 6,30 | 0,00 | 8,20 | 85,50 |
| *Choanoflagellata* | 14,20 | 0,00 | 12,10 | 73,70 |
| *Filasterea* | 9,10 | 0,00 | 18,10 | 72,80 |
| *Ichthyosporea* | 2,60 | 0,00 | 4,60 | 92,80 |

# III.2 MORPHOLOGICAL EVOLUTION

The study of molecular changes is not sufficient to understand the evolution of living organisms. It requires ecological, developmental, palaeontological and phylogenetic considerations. Palaeontology gives us invaluable information about anatomies, ecologies, physiologies, as well as spatial and temporal dynamics of past life (Jablonski and Shubin 2015). In the past 20 years, great technological improvements have been done not only in the molecular biology field but also in paleontology leading to the discovery of new early life fossils (Reisz and Sues 2015). The discovery and analysis of fossils from key intervals in the history of life can inform about the sequence, pattern, and phylogenetic dynamics underlying the origin of major functional and anatomical novelties (Jablonski and Shubin 2015; Parry et al. 2016).

In phylogeny, the use of character matrices from fossils is a widespread to analyze similarities and trace the evolutionary history of different traits, mostly in animals. The evolution of these traits does not necessarily follow that of the species; some traits may have appeared or disappeared several times independently in different lineages (Figure 17). Some of pre-existing traits can be dissociated, recycled and used to fulfill new functions. Study of this morphological modularity allows understanding the evolvability and plasticity of organismal form. Therefore, the analysis of the complex evolution of these morphological components requires the development of methods complementary to those used in classical phylogeny. Network approaches can be used to analyse the interdependency between characters in order to describe a broader range of changes and stases in organisms.

In the article n°7 in collaboration with Pr Pierre-Olivier Antoine, we propose to use network-based methods to study the co-occurrence of the traits in the panarthropods (Smith and Caron 2015) and rhinocerotid mammals, thanks to the fossil and current data that are available. We transformed the character matrices into traits matrices to focus on relationships between individual character states. We used these trait matrices to construct ‘trait networks’ to describe and to analyse patterns of co-occurrence between the character states that constitute the organisms. Trait networks provide a picture of character state combinations, but are not phylogenetic inferences. We have thus been able to analyze the co-occurrence relationships between character states during the evolution of panarthropods since the

Cambrian, and the evolution of rhinocerotid mammals during the last 50 million years. We observed a substantial general dissociability of traits during evolution for these two sets of organisms, and identified pivotal and relatively stable traits forming the structural backbone of the panarthropod and rhinocerotid morphological organisations. This article is in preparation and will be submitted to the journal "BMC Biology". The supplementary tables for this article can be downloaded from *http://www.evol-net.fr/downloads/*

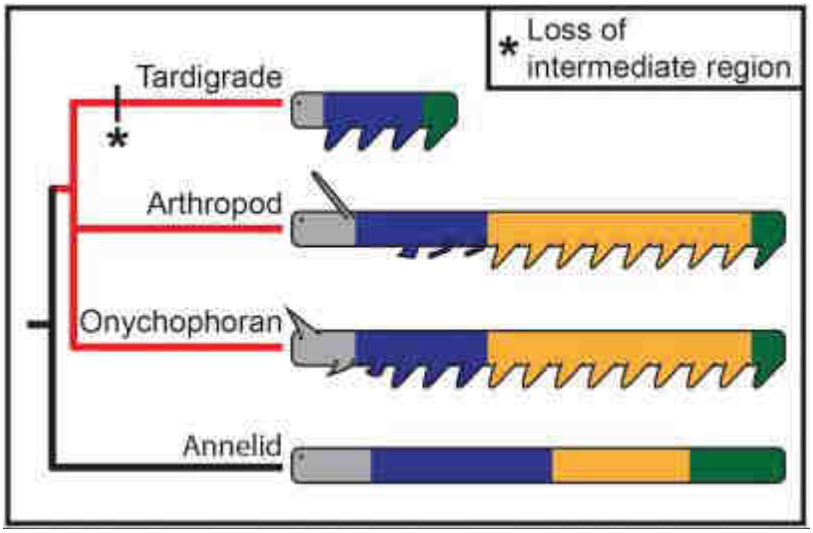



**Figure 17: The evolution of the tardigrade body plan.**
Hypothesis for the evolution of the tardigrade body plan by the loss of an intermediate trunk region (orange). Panarthropod branches are red in the phylogenetic tree. (Smith et al. 2016)

**Article**

# Palaeontological trait networks identify fluidity in organismal evolution

Etienne Lord[1,5,*], Jananan Sylvestre Pathmanathan[2,*], Eduardo Corel [2], Vladimir Makarenkov[1], Philippe Lopez[2], Frédéric Bouchard[3], Debashish Bhattacharya[4], Pierre-Olivier Antoine[6], Hervé Le Guyader[2], François-Joseph Lapointe[2,5] and Eric Bapteste[2,7,§]

[1]Département d'Informatique, Université du Québec à Montréal, CP 8888, Succursale Centre-Ville, Montréal (QC) H3C 3P8 Canada

[2]Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), Paris F-75005, France

[3]Département de Philosophie, Université du Québec à Montréal, CP 6128, Succursale Centre-Ville, Montréal (QC) H3C 3J7 Canada

[4]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA

[5]Département de Sciences Biologiques, Université du Québec à Montréal, CP 6128, Succursale Centre-Ville, Montréal (QC) H3C 3J7 Canada

[6]Institut des Sciences de l'Evolution, UMR 5554 CNRS, IRD, EPHE, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier cedex 5

[7]CNRS, UMR7138, Institut de Biologie Paris-Seine, Paris F-75005, France

[§]Corresponding author

*These authors contributed equally to this work

**Summary**

Explaining the evolution of animals requires ecological, developmental, paleontological and phylogenetic considerations, because organismal traits are affected by complex evolutionary processes. During evolution, traits can become tightly interdependent, dissociated, or used to fulfil novel functions. Therefore, how to better describe and analyse the evolution of taxa, not merely as lineages, but also as evolving organisations of traits, is becoming a key issue. Modelling a plurality of processes operating at distinct time-scales on potentially interdependent traits requires complementary treatments to phylogenetic analyses. We develop network approaches on paleontological and extant data to analyse the co-occurrence relationships between character states during the evolution of panarthropods since the Cambrian, and the evolution of rhinocerotid mammals during the last 50 million years. The pieces of the morphological toolkits of these taxa appear highly dissociable, hinting at repeated developmental changes during evolution, but some traits are significantly stable, unravelling backbones in these fluid body plans. Our evo-systemic framework supports a pluralistic modelling of organismal evolution, including trees and networks.

**Keywords:** *networks*, *tinkering, evolution, palaeontology, co-occurrence, Burgess fauna*

## Introduction

Organismal evolution is often investigated using phylogenetic approaches, which analyse 'characters X taxa' matrices to infer relationships between organismal lineages. The major focus of such, usually tree-based, analyses is generally to determine what groups of organisms derive from a last common ancestor, forming clades, and what their shared derived features (e.g. the synapomorphies of these clades) are. While invaluable, phylogenies can fruitfully be complemented by adopting a system-based perspective[1-3], using network approaches that explicitly analyse the interdependency between characters in order to describe a broader range of changes and stasis in organisms. This conception of organisms is deeply rooted in the biological field, as illustrated by the (idealistic) notion of correlation of parts[4], and its many critical refinements, as it became clear that correlations between animal traits can change in an irregular fashion[5]. Contra von Baer's laws of developments, Dollo, De Beer, and others[5-7] popularised the notion that individual organs can have independent phyletic histories, despite the obvious correlation of parts within any organisms, a clear challenge for the study of organismal evolution. Consistently, Evo-devo experiments characterised cases of co-options and tinkering of animal traits[8-13], and showed that structural biases built into genetic and developmental networks[2,10] can offer relevant explanations of convergences and parallelisms between organisms at the morphological level. Since these important aspects of organismal evolution resist traditional analyses[14-16], how to better describe and analyse the evolution of relationships between traits is becoming a pivotal question to enhance the understanding of organismal evolution[2].

Here, we propose to study organismal evolution, by using a different way of enumerating the signal of a given 'characters X taxa' matrix. More precisely, these matrices can be recoded into 'traits X taxa' matrices to focus on relationships between individual character states. Based on these recoded matrices, 'trait networks' can be used to describe and to analyse a rich body of patterns of co-occurrence between the character states making the organisms. Thus, trait networks provide a picture of character state combinations, but are not phylogenetic inferences. They are a way to organise information about various types of co-occurrence of morphological traits in organisms, and to effectively exploit the evolutionary signal associated with these patterns, while taking advantage of the tools of graph theory. The main focus of this strategy is to detect traits holding remarkable roles in trait networks, and to identify groups of traits with remarkable behaviours, in order to stimulate hypotheses about the processes affecting the morphological toolkits of organisms over the course of evolution. In particular, trait networks can be used to characterise the relative stability of the structural backbone of organisms and to lay out potential rules of associations for some of the pieces of their morphological toolkit. For example, strictly co-occurring morphological traits form 'complexes of character states', which may result from a common developmental regulation. Since (groups of) traits displaying evolutionary informative patterns in networks may or may not simply map onto an organismal phylogeny, patterns in trait networks can be used to detect and to highlight evolutionary events and processes that are neither naturally captured nor primarily brought forward in analyses of organismal trees or in character compatibility analyses[17]. Yet, trait networks do not aim at replacing phylogenetic approaches. Indeed, phylogenetic considerations can further illuminate the outcome of trait networks analyses. For example, traits complexes may be associated with clades,

and correspond to synapomorphies of these groups. But traits complexes can also be found in paraphyletic groups of taxa, requiring more complex explanations of their distributions.

Below, we introduce a new method and a user-friendly tool called ComponentGrapher for the construction and analyses of trait networks. We applied it to two distinct and well-established palaeontological–neontological datasets, featuring panarthropods from the Cambrian Burgess fauna[18] and focusing on Cenozoic rhinoceroses[19,20], respectively. These two phyla are very different, in particular in their body plans. Hence, they are likely subjected to different evolutionary constraints, i.e. metamerism is visible in arthropods, which harbors relatively independant metamers (except in their heads), whereas metamerism is less visible in mammals, that present very integrated metamers. We observed a substantial general dissociability of traits during evolution for these two sets of organisms, and identified pivotal and relatively stable traits forming the structural backbone of the panarthropod and rhinocerotid morphological organisations. Whilst the two datasets strongly contrast in terms of body plans and temporal/taxonomic scales, the general observation that many traits are used repeatedly in different combinations in different taxa, which usually do not form a clade, constitutes an additional incentive to further couple developmental and palaeontological studies.

## 1. **Trait networks, a novel comparative approach**

We developed a new method, which enumerates the signal present in 'characters X taxa' matrices to extract patterns of co-occurrence between the character states making the organisms, and therefore to generate and to test hypotheses about the evolution of traits relationships during organismal evolution. This method is implemented in a software called COMPONENT-GRAPHER (https://github.com/etiennelord/ComponentGrapher). It differs from clique / compatibility analysis in its approach, scope, and goals[21], and it produces a picture, instead of an inference, of character states relationships.

The main steps of our analyses are described below (see also Figure 1, Extended Data Figures 1 & 2). First, for each dataset of interest, COMPONENT-GRAPHER read a 'characters X taxa' matrix by columns. Second, each unique character state associated with a given character was extracted, i.e. if an original character had three states (0, 1, 2), this was now split into three characters states for which the presence/absence of each one was scored. Although the effect of recoding multistate characters as binary presence/absence data has been shown to be problematic for the reconstruction of phylogenetic trees[22-24], this coding is not a problem here as the number of nodes in trait networks is determined only by the number of character states and not the number of characters. Missing (*), ambiguous (?) and non-applicable (-) tokens were discarded at this stage of the analysis. Third, we selected all character states that do not signify absence; only those traits (e.g. character states corresponding to a present feature) were considered in subsequent analytical steps. Fourth, the nodes of the trait network were created: each node corresponds to a distinct character state. Thus, the nodes (and traits) of our networks are not equivalent to the characters used for phylogenetic analyses. Fifth, the type of co-occurrence between all pairs of character states from different characters was assessed to build the edges of the trait network (see Figures 1, 2 and Extended Data Figures 1, 2).

Four different types of relationships were characterised. In a type I relationship, two traits have identical taxa distributions. Since these traits are always found together, they form remarkable sets of features, which we call complexes. In a type II relationship, one trait shows a broader taxonomic distribution, entirely including that of the other trait. This is observed when, while two traits are simultaneously present in some taxa, a more broadly distributed trait has also evolved separately from the other trait in some taxa. By contrast, the least widespread trait is never observed without the most broadly distributed one. In a type III relationship, two traits have overlapping taxonomic distributions. While these traits are simultaneously present in some taxa, both have also evolved separately in distinct organisms. In a type IV relationship, two traits, never found in any common taxa, show mutually exclusive taxonomic distributions. Note that with our protocol only pairs of character states associated with distinct characters (not from the same character) were assigned a type IV relationship.

Sixth, based on these relationships, the network was constructed and stored as a list of nodes and a list of type I-IV edges. The network was then analyzed by COMPONENT-GRAPHER to identify the patterns described in Figure 2. This network construction is by definition robust: a given data matrix returns only one network of each type (and always the same, since it is an exact 'picture' of the relationships between character states). Seventh, we used COMPONENT-GRAPHER to compute two types of network measures: i) measures relative to the general topological properties of the trait network, and ii) specific topological properties of each of its node (i.e. centrality measures). For example, in-degree and out-degree of nodes were computed by counting the number of incoming/outgoing type II edges of each node. Since all network measures used in our analyses relied on exact graph metrics and not on heuristics, the values inferred from the network analyses are also robust. Finally, we used permutation tests implemented in COMPONENT-GRAPHER to assess the statistical significance of these network values. In short, COMPONENT-GRAPHER uses a null model of uncoordinated evolution, in which all characters states would be evolving independently. Thus this test amounts to a permutation of the states for each character, as already proposed to test for the presence of phylogenetic signal in character data[25,26]. COMPONENT-GRAPHER outputs all these results, as well as exportable networks (edgelist, and graphml formats, compatible with Cytoscape[27] and Gephi[28]).

## 2. Interpreting trait networks

Simple motifs with evolutionary significance can be exactly searched for in trait networks. We focused on several of them (Figure 2). Traits connected by type I edges are always associated in organisms. For example, in panarthropods, the three or four circumoral enlarged plates, the pre-ocular limb pair with arthrodial membranes, and the strengthening rays in lateral flaps are always found together, and exclusively so in *Anomalocaris*, *Peytoia*, and *Hurdia* (Supplementary Table 1B and Figure 3a). For such tight associations, it is therefore compelling to look for explanations, such as common developmental regulations affecting the genes coding for these traits, in particular when these pieces of the morphological toolkit were *a priori* assumed to evolve independently. Such

complexes may be synapomorphies of clades, but this is not necessary. By contrast, disjoint traits are simply never found in the same organisms, such as, in the rhino dataset, the separated metacone and hypocone on the fourth upper premolar (present in *Hyrachyus eximius*, *Trigonias osborni*, *Huaqingtherium lintungense* and *Aceratherium incisivum*), and the lingual bridge of the protocone and hypocone on the third and fourth upper premolar (present in *Ceratotherium simum*, *Diceratherium armatum*, *Teleoceras fossiger*, and *Lartetotherium sansaniense*), i.e. two aspects of upper premolar molarisation[19], whilst they could be considered intuitively as evolving interdependently. These traits may be encoded by genes undergoing antagonistic regulations, or simply by genes that appeared separately during evolution. Nested traits, such as the very convex base of the corpus mandibulae present in the rhinocerotids *Ceratotherium simum*, *Diceros bicornis* and *Coelodonta antiquitatis*, and the rugose frontal bone present in the former taxa plus *Dicerorhinus sumatrensis* (related to the emblematic diagnostic presence of a frontal horn), convey information regarding the relative stability of traits (Figure 4b). This asymmetric taxonomic distribution means that some traits are only present when the other trait is also present. Thus, we say that the latter, i.e. traits with larger in-degree (number of incoming type II edges), are more stable relative to other traits with which they co-existed. Such relatively stable traits are remarkable because they provide a structural backbone, around which the rest of the organismal traits changes. The detection of backbone traits suggests that past organisation constrains, and in effect biases, the future evolution of the traits that evolve in organisms. This is understandable from a systemic perspective, i.e. central or essential traits, for example those interacting with many others, have less flexibility to change than traits that are more peripheral in biological organisations. Nested traits can correspond to nested synapomorphies of clades, but it is not a logical obligation.

Finally, overlapping traits are distributed across non-nested sets of taxa. For example, in panarthropods, the sclerotized pharyngeal 'teeth', and the terminal mouth opening orientation only occur together in *Priapulus*, *Cricocosmia*, *Paucipodia*, *Hallucigenia sparsa*, and *Jianshanopodia*, while their evolution is dissociated in other organisms (*Aysheaia*, *Siberion*, *Onychodictyon ferox*, *Onychodictyon gracilis*, *Diania*, *Microdictyon*, *Cardiodictyon*, *Hallucigenia fortis*,*Halobiotus* (Eutardigrada), *Siberian 'Orsten' tardigrade*, *Kerygmachela*, *Actinarctus* (Heterotardigrada), *Halobiotus* (Eutardigrada), *Hurdia*, *Supella longipalpa)*, in which they do not occur together. Such a distribution is a sign of complex evolution of the traits: it may involve losses, reversions, convergences, and/or parallelisms. When three traits entertain a type III relationship with each other, they form a triangle in the type III trait network. A triangle means that the evolution of these traits is dissociated in at least some taxa, and suggests that the presence of these traits is not under a common developmental regulation over evolutionary time. A high proportion of triangles in the type III trait network means that a high proportion of traits can evolve in such a dissociated fashion, and therefore it measures a general dissociablity on traits in the studied organisms. We call organismal fluidity the fact the same traits (rather than different traits) can be found in distinct combinations. Organismal fluidity is higher when the proportion of triangles is higher, i.e. when the type III networks increasingly resemble a clique because the highest proportion of triangles obtains when all nodes are connected together by a type III edge in the graph. This fluidity should be not confused with the dissociations of genes produced by introgressive processes in prokaryotic taxa. The multiple traits of a single fluid metazoan are likely derived from a single common ancestor, however the genes coding for these traits, and thus the interactions between these traits, have not necessarily been subjected to

simultaneous regulation, activation, and inactivation during organismal evolution, which decouples their presence in organisms.

Finally, some traits (central in type D triplets, Figure 2) are alternatively found with traits that never occur together. We call these central traits 'pivotal', because they have taken part in distinct morphological organisations. This behaviour is an extreme form of versatility. The morphological organisations including a pivotal trait are all the more different (in terms of composition) than there are type D triplets centered on the pivotal trait. The detection of pivotal traits is a pre-condition to evaluate their role during organismal evolution. They may typically have been co-opted for novel functions, hinting at regulatory changes for their coding genes, or may have helped to recruit novel traits, before becoming superfluous.

## 3. Application to two palaeontological datasets

We investigated two datasets covering distinct geological intervals and phyla (Phanerozoic panarthropods and Cenozoic mammals). First, we recoded the data set in[18], describing 141 traits present in 40 taxa of panarthropods, including 35 fossils and five members of extant lineages (see Methods and Supplementary Table 1A). The detection of type I relationships between traits returned 14 complexes, which is significantly higher than expected by chance (Table 1). Finding complexes opens the intriguing possibility that maybe some character states that seemed to belong to different characters are in fact inseparable instances of a common developmental regulation, hence may constitute a single character that was not previously characterized as such. It is of course for the experts to determine whether they want to use the detection of unexpected complexes in this way, particularly for the 12 complexes, which associated traits from different regions of the body plan (such as cluster4 : Mouth + Head + Appendages, or cluster 2: Mouth + Head + Bodyplan, Figure 3a). Six complexes mapped perfectly with the organismal phylogeny, suggesting that each of these complexes was assembled once in a last common ancestor, and four complexes were on terminal branches. By contrast, cluster 14(the 3 neuromeres integrated into the dorsal condensed brain, and the deutocerebral innervation) shared by *S. longipalpa*, *Fuxianhuia, Alalcomenaeus*, or cluster 3 (the pre-oral chamber and sclerites comprise stacked elements) shared by *Hallucigenia* and *E. kanangrensis* are, for example, not merely explained by common ancestry since these taxonomic groups do not correspond to clades on the ecdysozoan phylogeny[18]. Secondary losses or convergent evolution likely occurred for these complexes. Remarkably, all of these 14 complexes are small. The largest complex merely associates four traits. Four other complexes associate three traits, and the nine remaining complexes associate two traits. Such small and rare complexes, encompassing a total of 34 traits, represent only limited portions of the morphological toolkit of organisms. Therefore, most traits of panarthropods present in this dataset do not form undissociable groups during evolution. Consistently, there are 1,766 type II edges, associated traits that are occasionally decoupled, which is significantly higher than expected by chance. These pairs of traits with nested taxonomical distribution are very rarely clades: only 78 (4.4%) of the type II edges correspond to nested clades; 381 (21.6%) correspond to a clade included in a non monophyletic group, and 1307

(74%) correspond to two nested non monophyletic groups. For example, the distribution of annulation on trunk and limbs convergently evolved with the presence of secondary structures on non-sclerotized (lobopodous) limbs, the latter never existing without the former (Figure 3b). Thus, nested traits cannot usually be simply explained by the evolution of synapomorphies. Detailed analysis of type II edges, contrasting in-degrees and out-degrees for all traits of the network, shows that the organisation of the pieces forming the "puzzle" of panarthropods is rather labile: no trait is especially stable in a large number of taxa. The vast majority of traits have similar and rather small in-degrees. However 42 traits, such as the paired appendages, the permanently inverted pharynx or distinct pre-ocular limb pair, were significantly more stable relatively to the other traits than expected by chance (Supplementary Table 1C, Figure 3a), introducing backbones, around which various combinations of morphological pieces have evolved in panarthropods. The majority of these significantly stable traits involves character states from different characters, but 16 of these traits were couplets, i.e. alternative states of the same features, such as the ventral and the posterior mouth opening orientation, indicating that a minority of the characters of panarthropods are structurally more stable.

Additionally, there were 2,937 type III edges in the trait network. Although significantly less abundant than by expected by chance, these relationships provide supplemental evidence of the general dissociability of traits during panarthropods evolution. The density of the type 3 graph reaches 0.38, and its diameter (defined as the longest shortest path that must be traversed to connect any pair of nodes in this graph) is 4. Altogether, these graph measures confirm that the evolution of panarthropods frequently involved similar traits albeit in different combinations in different organisms. Interestingly, 10 traits, such as the uniform distribution around the pharynx of pharyngeal teeth or *aciculae* appear significantly overrepresented at the center of type D triplets (Supplementary Table 1D). All these pivotal traits come from different characters, and suggest some transitionist[10] changes at the morphological level that occurred during the gradual evolution in panarthropods[29]. For example, after lobopodous organisations, lobopodous organisations with the trunk exites evolved, then distinct organisations with both trunk exites and appendages comprising fewer than 15 podomeres evolved, and finally a third organisation with appendages comprising fewer than 15 podomeres (Extended Data Figure 3).

Overall, mapping unstable, stable, significantly stable and pivotal traits on the body plans of panarthropods allowed us to analyze whether in different regions of the body plan the morphology is affected by different evolutionary processes. Most unstable traits (i.e. relatively to other traits that showed a broader taxonomic distribution) can be found in all body compartments, to the exception of the eyes, already well-structured in panarthropods (Figure 5). These unstable features occur mainly in the anterior parts. In general, Onychophora and Tactopoda display comparable proportions of unstable traits (Exact Fisher test, p-value 0.62). However, unstable traits are not evenly distributed in the same body regions for these two groups (Extended Data Figures 4, 5, 6). The "brain", "first post-ocular" regions (and to a lesser extent "mouth" and "appendages" of Onychophora appear to be evolutionarily more flexible than those of Tactopoda. This trend is even more pronounced when the stability of traits exclusive to Onychophora is compared with that of traits exclusive to Tactopoda. This difference in modes and regimes of evolution along the body plan is likely explained by the diverse feeding adaptations in marine fossils of Onychophora, and thus highlights the high

evolvability of this clade. By contrast, Tactopoda display a greater proportion of exclusive stable traits, likely correlated with the stability of the body plan of Euarthropoda and Tardigrada, even though their "heads" in general show proportionally more unstable features than Onychophora for this dataset. This observation is consistent with the evolutionary importance of this body part for Tardigrada, rightly described as walking heads. Overall, our analysis of trait stability provides complementary evidence that Onychophora and Tactopoda show distinct evolutionary profiles, compatible with the recent proposal of the monophyly of each clade.

Second, we recoded a data set primarily modified from[19,20], describing 120 traits present in 21 taxa of ceratomorph mammals, without missing data, primarily focused on rhinocerotids (rhinos), and including 15 fossil species and 6 members of extant lineages among rhinos and tapirs (see Methods and Supplementary Table 2A). We detected eight complexes, which does not differ from expectations by chance (Table 1). Six of them associated traits from different regions of the body plan (Figure 4a). More precisely, complexes occur at both terminal (1, 3, 4, 6, and 8) and internal nodes (2, 5, and 7). They are mainly documented in the subfamily of living rhinos, the Rhinocerotinae. Within the latter clade, Miocene Aceratheriini have two dental-based complexes (complexes 7 and 8) and the short-limbed and hippo-like teleoceratine *Brachypotherium brachypus* yields a jaw- and teeth-based complex (complex 4). Two-horned rhinos, either living (Sumatran, white and black rhinos) or recently extinct (woolly rhino), comprise more integrative complexes, containing skull and tooth characters (complexes 1 and 3), skull and forelimb characters (complex 2). The most inclusive complex (complex 5) encompasses jaw, tooth, and forelimb features, observed in the morphologically well-supported woolly, white, and black rhino clade[19]. Conversely, no complex characterises the sister group to Rhinocerotinae, i.e., Elasmotheriinae. At first sight, all complexes located at internal nodes involve closely related taxa: complexes 2 (two-horned rhinos), 5 (grazers among two-horned rhinos), and 7 (Aceratheriini). In other words, they may be good indicators of strongly supported morphological clusters. Moreover, one complex concerns the non-rhinocerotid taxa of the rhino dataset, i.e. the outgroups (the extant Brazilian tapir *Tapirus terrestris* and the early diverging hyrachyid *Hyrachyus eximius*) gathering tooth and hind limb characters.

As for panarthropods, all of these complexes are small, associating at most four traits. Collectively, complexes encompass a total of 22 traits, hence less than 18% of the morphological toolkit of rhinos. Most traits of rhinos happen to be dissociated during evolution. Consistently, there are 5,100 type II edges, which is significantly higher than expected by chance. Only eight (0.16%) of the type II edges correspond to nested clades while 492 (9.6%) correspond to a clade included in a non monophyletic group, and 4600 (90%) correspond to two nested non monophyletic groups. For example, a distal articulation strongly oblique with respect to the trochlea on the astragalus convergently evolved with the orientation of lower molar hypolophids, the latter never existing without the former (Figure 4b). Interestingly, there was no reason to consider these postcranial and dental features as being related *a priori*. Thus, like for panarthopods, nested traits of rhinos cannot usually be simply explained by the evolution of synapomorphies. Fifty traits were significantly more stable relatively to other traits than expected by chance (Supplementary Table 2C), constituting a detectable backbone in rhinos. The majority of these significantly stable traits involves character states from different characters, but 18 of these traits were couplets, such as the narrow and the very broad rostral ends of the nasal bones, indicating that a minority of the characters of rhinos are

structurally more stable. Among them, there is a certain predominance of "iconic" features (e.g., nasal and frontal horns, crown height, dental formula, shape of the last upper molar, and tridactylous hand), considered as diagnostic in pre-Hennigian/phylogenetic classifications, while phylogenetic analyses based on equivalent datasets have demonstrated that these traits were strongly tainted of convergence and/or parallelism[19,20,30,31]. In other words, these traits seem to be relevant for understanding the rhinocerotid body plan.

Additionally, there were 16,063 type III edges in the trait network (significantly less abundant than expected by chance). The density of the type 3 graph was much higher than for panarthropds (0.71), as well as the proportion of triangles in the type II graph (0.43), for a comparable diameter (of 2). These network metrics show that the evolution of rhinos also frequently involved similar traits albeit in different combinations in different organisms. For example, in some organisations a short metastyle on the first-second upper molars is present along with a low zygomatic width (with respect to frontal width), whilst in others such a low zygomatic width occurs with a crochet on upper molars. Conversely, neither a short metastyle and a crochet on upper molars, nor a short metastyle, a low zygomatic width, and a crochet on upper molars occurred simultaneously in any rhinocerotid (Extended Data Figure 7). Interestingly, 21 traits, such as the foramen mentale in front of p2 or at the level of p2-4, appear significantly overrepresented at the centre of type D triplets (Supplementary Table 2D), and they were in large majority couplets (16 out of 21; mainly on teeth, and to a lesser extent on jaw and limbs; e.g., tibia and fibula independent or fused). Be they plesiomorphic or derived states[19], these features have taken part in distinct morphological organisations among rhinocerotids.

Mapping the traits on the rhino body plan unravelled a significant regionalization of unstable traits (Fisher exact test, p-value 0.05) (Figure 6). These unstable traits were significantly more abundant in the cranio-dental region (c. 10% of cranio-mandibular and dental features) than in the postcranial region. Unstable traits consist of independent characteristics or singletons, instead of couplets. The total absence of unstable characters recognised for the body plan or the limb bones (0/66) was striking. The postcranial skeleton is remarkably stable within the controlled rhinocerotids with respect to cranio-mandibular region and teeth, pointing to an early implementation of postcranial Bauplan among rhinocerotids, without major changes since then. This is particularly contrasting with the results regarding the distribution of homoplasy in phylogenetic analyses focused on similar datasets[19,30,31], where all the considered body regions yield a similar amount of homoplastic characters, further showing that instability does not equal homoplasy and that both network- and phylogenetic-based approaches are thus complementary in depicting distinct aspects of trait versatility. Other differences were not statistically significant.

Overall, the panarthropod and rhino datasets show major discrepancies, in particular in terms of instability amount (28/141 vs. 19/229). These differences may be due to highly distinct scaling, both taxonomic and temporal: for comparable sizes (141 characters in 40 taxa vs. 229 characters in 21 taxa, respectively), these data sets embrace representatives of either a superphylum (Panarthropoda) throughout the Phanerozoic interval (540 million years) or of a suborder (Ceratomorpha) during the last 50 million years.

## 4. Discussion: "Fluid animals"

Our approach provides a new strategy allowing for complementary re-analyses of currently available data from a systemic perspective, in particular palaeontological data. Network analyses describing how associations of traits evolved should contribute greatly to a mechanistic explanation of evolution. They confirm that not all components of the anatomy of a given organism change at the same time, at the same rate, or in the same way, but probably as a result of various structural constraints, and that this heterogeneity of modes of evolution can probably not be captured by evolutionary models that treat characters as if they were evolving independently, since the uncoordinated model of traits evolution was rejected. Moreover, our method highlighted traits with remarkable behaviour during evolution, in terms of their relative stability, their pivotal distribution, and their contribution to complexes. It showed that panarthropods and rhinos instantiate different sorts of fluidity, since relatively less stable traits are observed everywhere (but in the eyes) of panarthropods, yet only in the heads of rhinos. Moreover, the general observation that many of these animals traits are used repeatedly, in different combinations, in different taxa, which usually do not form clades, suggests that the genes encoding these traits might be inherited without expression (or decimation by genetic drift) from a common ancestor, and might be recruited into novel gene regulation networks during the course of evolution, unless similar traits can be invented on multiple occasions and coded from different gene sets, or traits losses are massive during organismal evolution.

The former interpretations would agree with the description of the main developmental stages in terms of gene regulatory networks, proposed in the pioneering work by Britten and Davidson[32], now theoretically and experimentally validated[9,33-36]. As stated by[9], "*it is obvious that if there is indeed a finite repertoire of network sub-circuits used to effect development, the evolution of development has to be considered as the process of assembly, reassembly, and redeployment of these sub-circuits.*" A certain morphological fluidity echoes with this genomic fluidity. Thus, importantly, fossils could contribute to generate hypotheses about the role of important aspects of developmental evolution, namely regulation and heterochrony, in evolutionary changes, when the resulting network patterns suggest frequent parallelism, and convergence. Therefore, our analysis encourages an openly pluralistic modelling of organismal evolution, including trees and networks, and constitutes a major incentive to further couple developmental studies with palaeontological studies. Such consideration does not belittle the importance of phylogenetic reconstruction, but stresses the need for a further integration of network-thinking into evolutionary analyses[2], because it has the potential to enhance the retrodictive dimension of evolutionary biology. Precisely, we hope that our study opens an avenue for network analyses of palaeontological data. In that process, the current implementation of COMPONENT-GRAPHER could be critically improved. On the one hand, the use of variable, non-applicable, and missing tokens may be considered in the future. On the other hand, while we report here an apparent major signal of organismal fluidity, as with any comparative analysis, the conclusions still depend (to some extent) on the quality of the available matrix. A different treatment of missing data may affect the inference regarding the general dissociablity of traits, even though the rhinocerotids dataset was not affected by this possible bias.

Interestingly, because our graph-theoretical approach investigates types of distribution of traits (or more generally components) at higher levels, without the absolute need for an underlying

phylogeny, it could already be broadly applied to analyse organisations from the molecular level (i.e. by analysing the distributions of active sites across homologous genes) up to the ecosystemic level (i.e. by analysing the distributions of OTUs or species across environmental samples). In these -omics days, the types (and amount) of data to be compared between taxa are increasing faster than accurate evolutionary models to describe their rules of changes are implemented. In that sense, networks can contribute to further the integration of systems and evolutionary biology. We believe such an evo-systemic could be especially informative, since evolution from molecules to ecosystems depends on changes in organisations as well as on the divergence and merging of lineages.

## Methods

Constitution of the dataset.

For panarthropods, we retrieved the dataset in[18] describing 141 components present in 40 Phanerozoic taxa, including 35 fossils and five extant species, and removed character states describing absent features to focus only on the components making up organisations (Supplementary Table 1A). For rhinocerotids, we used a matrix derived from[19,20,30,31], including 120 morpho-anatomical characters scored in 15 extinct and six living ceratomorph mammal species (tapirs, rhinoceroses and their kin), ranging from the last 50 million years (Supplementary Table 2A). Non-applicable and missing characters were removed from the original matrix[19], as well as insufficiently characterised fossil taxa, so that the dataset is fully documented for a taxonomic sample gathering all suprageneric clades usually recognised within Rhinocerotidae[31].

Network construction and analyses.

We implemented COMPONENT-GRAPHER (https://github.com/etiennelord/ComponentGrapher), and provided it with the above matrix to construct and analyse the network. To assess whether the results could have been obtained by chance alone, a permutation test based on the null hypothesis that characters states are randomly distributed among taxa is performed. Namely, this test permutes character states in each column of the data matrix in order to break the phylogenetic structure[25]. New networks are then obtained from these permuted data sets, from which the corresponding graph statistics are computed. The test values obtained from the actual data matrix are declared significant when the vast majority of the values obtained under the null hypothesis are more extreme than the original values. For each data set, the number of permutations was set so as to make sure that the corresponding p-values could reach a predetermined significance level fixed at 0.05, following a Bonferroni correction for multiple tests.

Detection of stable components.

Degree analysis of the network of inclusion (type II) quantifies the relative stability of each trait. Type II in-degree quantifies how many direct neighbours of a given trait point toward a given node, hence how many traits have a more restricted taxonomic distribution than a focal trait. Type II out-degree quantifies toward how many direct neighbours each individual trait is pointing, indicating that a focal trait has a more restricted distribution than these neighbours. Very precarious traits have a null in-degree and a positive out-degree. By contrast, stable traits have a higher in-degree and a lower out-degree. To determine which traits are more stable than by chance alone, another permutation test was applied directly to the nodes of the networks, using the same protocol as described above. A trait was declared to be significantly stable when its type II in-degree was more extreme than the vast majority (95%) of in-degrees obtained under the null hypothesis.

Detection of organismal fluidity.

The extensiveness of trait dissociability was tested by investigating topological features of the type III graph. We computed (i) the density of the graph of type III, (ii) the number of triangles, and (iii) the diameter of the type III graph. The use of the same traits in multiple different morphological combinations, rather than their irremediable replacement in diverging lineages, produces dense type III graphs, with reduced diameters.

## References


1 Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. (2006).

2 Wilkins, A. Between "design" and "bricolage": genetic networks, levels of selection, and adaptive evolution. *Proc Natl Acad Sci U S A.* **104**, 8590-8596. (2007).

3 Yafremava, L. S. *et al.* A general framework of persistence strategies for biological systems helps explain domains of life. *Frontiers in genetics***4**, 16, doi:10.3389/fgene.2013.00016 (2013).

4 Cuvier G. in *Recherches sur les ossements fossiles où l'on rétablit les caractères de plusieurs animaux dont les révolutions du globe ont détruit les espèces.* (Déterville, Paris 1812).

5 Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History*. 352 (1989).

6 Brigandt, I. Homology and Heterochrony: The Evolutionary Embryologist Gavin Rylands de Beer (1890-1972). *Journal of Experimental Zoology (Molecular and Developmental Evolution)***306B**, 317-328 (2006).

7 Gould, S. J. *The Structure of Evolutionary Theory*. (Harvard University Press, 2002).

8 Carroll, S. B. *Endless Forms Most Beautiful*. 350 (Quercus, 2005).

9 Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature***468**, 911-920 (2010).

10 Duboule, D. & Wilkins, A. S. The evolution of 'bricolage'. *Trends in genetics : TIG***14**, 54-59 (1998).

11 Jacob, F. Evolution and Tinkering. *Science***196**, 1162 (1977).

12 Jacob, F. Complexity and tinkering. *Annals of the New York Academy of Sciences***929**, 71-73 (2001).

13 Shubin, N. *Your inner fish*. 237 (Vintage Books, 2009).

14 Bolker, J. A. Modularity in Development and Why It Matters to Evo-Devo. *American Zoologist***40**, 770-776 (2000).

15 Hall, B. K. Homoplasy and Homology: Dichotomy or continuum? *Journal of Human Evolution***52**, 473-479 (2007).

16 Young, R. L. & Wagner, G. P. Why ontogenetic homology criteria can be misleading: lessons from digit identity transformations. *J. Exp. Zool. (Mol. Dev. Evol)***316B**, 165-170 (2011).

17 Meacham, C. A. & Estabrook, G. F. Compatibility methods in systematics. *Annual Rev. Ecol. Syst.***16**, 431-446 (1985).

18 Smith, M. R. & Caron, J. B. Hallucigenia's head and the pharyngeal armature of early ecdysozoans. *Nature***523**, 75-78, doi:10.1038/nature14573 (2015).

19 Antoine, P.-O. *Phylogénie et évolution des Elasmotheriina (Mammalia, Rhinocerotidae)*, (2002).

20 Antoine, P.-O., Duranthon, F. & Welcomme, J.-L. Alicornops (Mammalia, Rhinocerotidae) dans le Miocène supérieur des Collines Bugti (Balouchistan, Pakistan) : implications phylogénétiques. *Geodiversitas***25.**, 575-603 (2003).

21 Salisbury, B. A. Strongest evidence in compatibility: Clique and tree evaluation using apparent phylogenetic signal. *Taxon***48**, 755-766 (1999).

22 Hawkins, J. A., Hughes, C. E. & Scotland, R. W. Primary Homology Assessment, Characters and Character States. *Cladistics***13**, 275–283 (1997).

23 Maddison, W. P. Missing Data versus Missing Characters in Phylogenetic Analysis. *Syst Biol***42**, 576-581 (1993).

24 Seitz, V., Ortiz Garcia, S. & Liston, A. Alternative Coding Strategies and the Inapplicable Data Coding Problem. *Taxon***49**, 47 (2000).

25 Archie, J. W. A randomization test for phylogenetic information in systematic data. *Syst. Zool.***38**, 239-252 (1989).

26 Faith, D. P. & Cranston, P. S. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics***7**, 1-28 (1991).
27 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research***13**, 2498-2504 (2003).
28 Bastian, M., Heymann, S. & Jacomy, M. in *International AAAI Conference on Weblogs and Social Media* (California: San Jose, 2009).
29 Smith, M. R. & Ortega-Hernandez, J. Hallucigenia's onychophoran-like claws and the case for Tactopoda. *Nature***514**, 363-366, doi:10.1038/nature13576 (2014).
30 Antoine, P.-O. *et al.* A revision of Aceratherium blanfordi Lydekker, 1884 (Mammalia: Rhinocerotidae) from the early Miocene of Pakistan: postcranials as a key. *Zoological Journal of the Linnean Society***160**, 139-194 (2010).
31 Becker, D., Antoine, P.-O. & Maridet, O. A new genus of Rhinocerotidae (Mammalia, Perissodactyla) from the Oligocene of Europe. *Journal of Systematic Palaeontology***11**, 947-972 (2013).
32 Britten, R. J. & Davidson, E. H. Gene Regulation for Higher Cells - a Theory. *Science***165**, 349-&, doi:DOI 10.1126/science.165.3891.349 (1969).
33 Erkenbrack, E. M. & Davidson, E. H. Evolutionary rewiring of gene regulatory network linkages at divergence of the echinoid subclasses. *Proceedings of the National Academy of Sciences of the United States of America***112**, E4075-E4084, doi:10.1073/pnas.1509845112 (2015).
34 Gao, F. & Davidson, E. H. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. *Proceedings of the National Academy of Sciences of the United States of America***105**, 6091-6096, doi:10.1073/pnas.0801201105 (2008).
35 Gillis, J. A. & Hall, B. K. A shared role for sonic hedgehog signalling in patterning chondrichthyan gill arch appendages and tetrapod limbs. *Development***143**, 1313-1317, doi:10.1242/dev.133884 (2016).
36 Peter, I. S. & Davidson, E. H. Evolution of Gene Regulatory Networks Controlling Body Plan Development. *Cell***144**, 970-985, doi:10.1016/j.cell.2011.02.017 (2011).

**Acknowledgements**

We thank P. Janvier, A de Ricqlès, and J. McInerney and 3 anonymous reviewers for critical reading of the MS. EL is supported by a Natural Sciences and Engineering Research Council (NSERC) scholarship. VM and FJL respectively hold NSERC discovery grants OGP0155251 and OGP249644. EB is funded by the European Research Council under the European Community's Seventh Framework Program FP7 (Grant Agreement n°615274).

**Author contributions**

EB, FJL, FB & DB designed the analysis, EL, VM & JSP implemented COMPONENT-GRAPHER, EL, EC, JSP & POA analysed the data, EB, FJL & PL interpreted the results, EB, FB, FJL, POA & HLG wrote the manuscript. All authors read and approved the final manuscript.

**Author information**

DB: Bhattacharya@AESOP.Rutgers.edu

EB[§]: eric.bapteste@upmc.fr

EC: eduardo.corel@upmc.fr

EL: lord.etienne@courrier.uqam.ca

FB: f.bouchard@umontreal.ca

FJL: francois-joseph.lapointe@umontreal.ca

HLG: herve.le_guyader@upmc.fr

JSP: jananan.pathmanathan@etu.upmc.fr

POA:pierre-olivier.antoine@umontpellier.fr

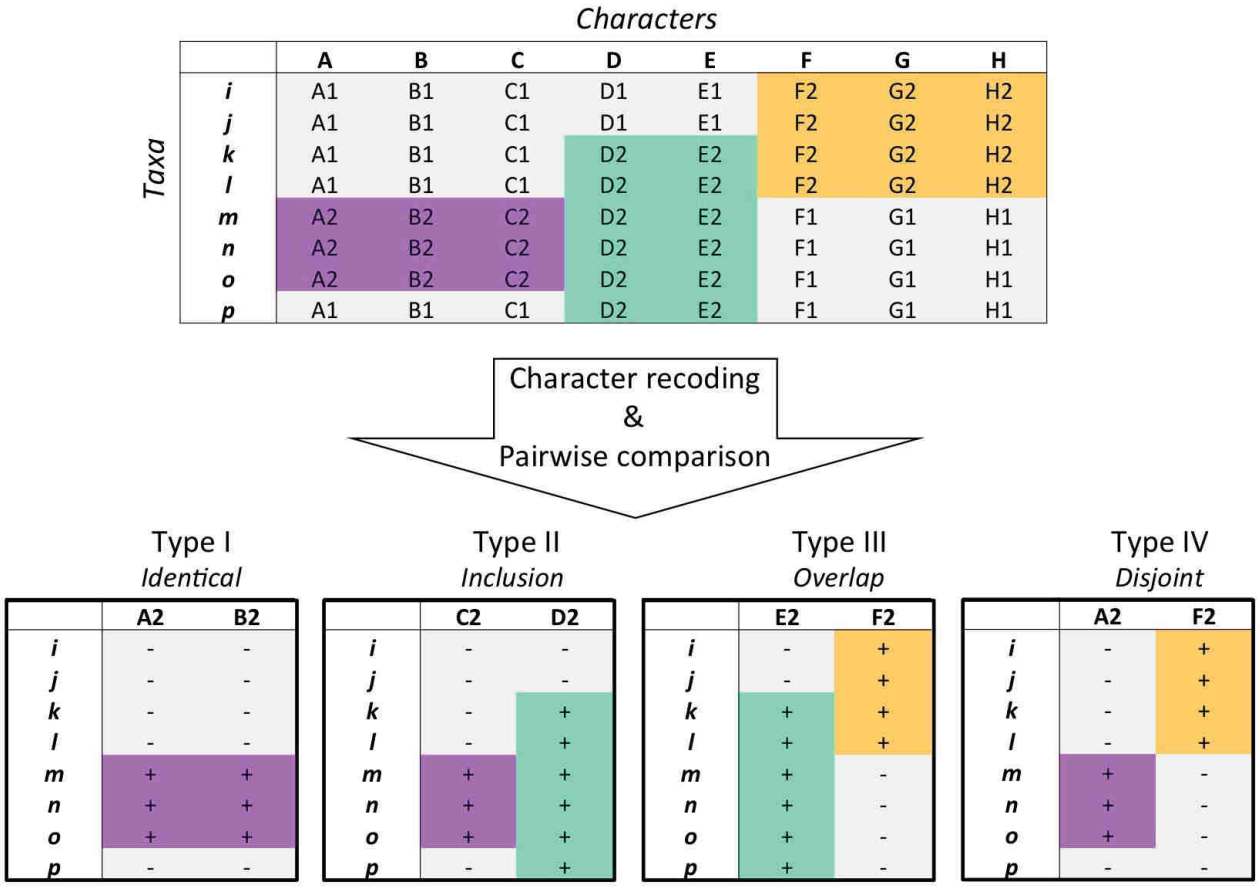
PL: philippe.lopez@upmc.fr

VM: makarenkov.vladimir@uqam.ca

**We declare no competing interest.**

**Legends:**

**Figure 1.** Principle of the matrix analysis. Our approach exploits existing phylogenetic data matrices featuring taxa as rows and homologous characters as columns. Each original column is replicated in as many new columns as there are character states (e.g. A2, B2), defining a new matrix of taxa by traits, where the presence of each trait is indicated by a '+' and its absence indicated by a '-'. All pairs of columns of this new matrix are then compared with one another, distinguishing four types of distribution of traits across taxa, therefore characterising four possible types of relationships between all pairs of traits.




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| *i* | A1 | B1 | C1 | D1 | E1 | F2 | G2 | H2 |
| *j* | A1 | B1 | C1 | D1 | E1 | F2 | G2 | H2 |
| *k* | A1 | B1 | C1 | D2 | E2 | F2 | G2 | H2 |
| *l* | A1 | B1 | C1 | D2 | E2 | F2 | G2 | H2 |
| *m* | A2 | B2 | C2 | D2 | E2 | F1 | G1 | H1 |
| *n* | A2 | B2 | C2 | D2 | E2 | F1 | G1 | H1 |
| *o* | A2 | B2 | C2 | D2 | E2 | F1 | G1 | H1 |
| *p* | A1 | B1 | C1 | D2 | E2 | F1 | G1 | H1 |

| | A2 | B2 |
|---|---|---|
| *i* | - | - |
| *j* | - | - |
| *k* | - | - |
| *l* | - | - |
| *m* | + | + |
| *n* | + | + |
| *o* | + | + |
| *p* | - | - |

| | C2 | D2 |
|---|---|---|
| *i* | - | - |
| *j* | - | - |
| *k* | - | + |
| *l* | - | + |
| *m* | + | + |
| *n* | + | + |
| *o* | + | + |
| *p* | - | + |

| | E2 | F2 |
|---|---|---|
| *i* | - | + |
| *j* | - | + |
| *k* | + | + |
| *l* | + | + |
| *m* | + | - |
| *n* | + | - |
| *o* | + | - |
| *p* | + | - |

| | A2 | F2 |
|---|---|---|
| *i* | - | + |
| *j* | - | + |
| *k* | - | + |
| *l* | - | + |
| *m* | + | - |
| *n* | + | - |
| *o* | + | - |
| *p* | - | - |

**Figure 2.** Some remarkable network patterns and their biological meaning. The first column displays the relationships between a pair of traits (here character states). The second column represents the corresponding network pattern. The third column introduces the terms specifically used to describe and analyse these patterns. The fourth column highlights some possible biological meanings of these patterns.

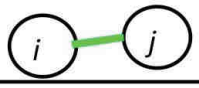
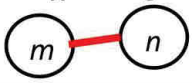
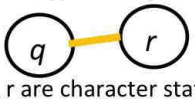
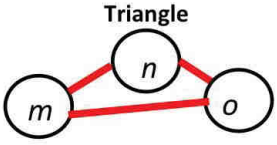
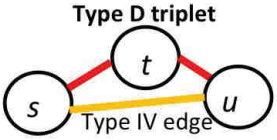
| Pattern in the recoded '*traits* X taxa' matrix | Network pattern | Associated terminology | Biological meaning |
|---|---|---|---|
| ***i*** + + + +<br>***j*** + + + + | **Type I edge**<br>  | **Complex** | Traits are **indissociable.**<br>**Common developmental regulation ?** |
| ***k*** + + + -<br>***l*** + + + + | **Type II edge** (directed)<br>  <br>*k out-degree* = 1 *l* in-degree = 1 | **Nested traits** | The higher the in-degree of a node,<br>the higher the **relative stability** of the trait.<br>**More stable traits are structural backbones** |
| ***m*** + + -<br>***n*** - + + | **Type III edge**<br>  | **Overlapping traits** | **Complex trait evolution** |
| ***q*** + + - -<br>***r*** - - + + | **Type IV edge**<br>  <br>(q, r are character states from different characters) | **Disjoint traits** | These traits never appeared in the same organism. |
| ***m*** + + - +<br>***n*** - + + -<br>***o*** + - + + | **Triangle**<br>  | ***m,n, o*** are **versatile traits** | ***m, n, o*** can be ***dissociated*** in organisms. **Traits** involved **in more triangles** are **more versatile**. **The higher the proportion of triangles** in type III graphs, **the higher the general dissociablity of traits**. |
| ***s*** + + - -<br>***t*** - + + -<br>***u*** - - + + | **Type D triplet**<br>  | ***t*** is a **pivotal trait** | ***t*** is shared between different morphological organisations. When **many traits are common across different organisations, many nodes are central in type D triplets**. |

**Figure 3.** Phylogeny of panarthropods, modified from [18] to depict a) 14 trait complexes. Each complex is represented by its corresponding motif (each node represents a trait, each green edge represents the type I relationships between 2 traits) along the phylogeny, based on its taxonomic distribution. Each complex is also identified by a circled number; blue circles representing complexes shared by a common ancestor and all its descendants (putative synapomorphy), non-blue circles representing complexes whose distribution does not map simply onto the phylogeny (homoplasy). The top left squared box identifies the distribution of complexes over the main regions of the panarthropod body plan (H: head; M: mouth; E: eye; B: brain; PO: 1st post-ocular; BP: body plan; TT: trunk and tail; A: appendages). Blue letters highlight complexes of traits from different regions. b) Phylogeny of panarthopods showing 2 examplars of traits with type II relationships. The distribution of trait 120 is nested in that of trait 27 (clade within clade). The distribution of trait 99 is nested in that of trait 62 (non clade within non clade). 120 : Appendages comprise 15 or more podomeres | Fewer than 15 podomeres; 27 : Type of eyes | multiple visual units (including compound eyes); 99 : Secondary structures on non-sclerotized (lobopodous) limbs | present; 62 : Annulation distribution | trunk and limbs.

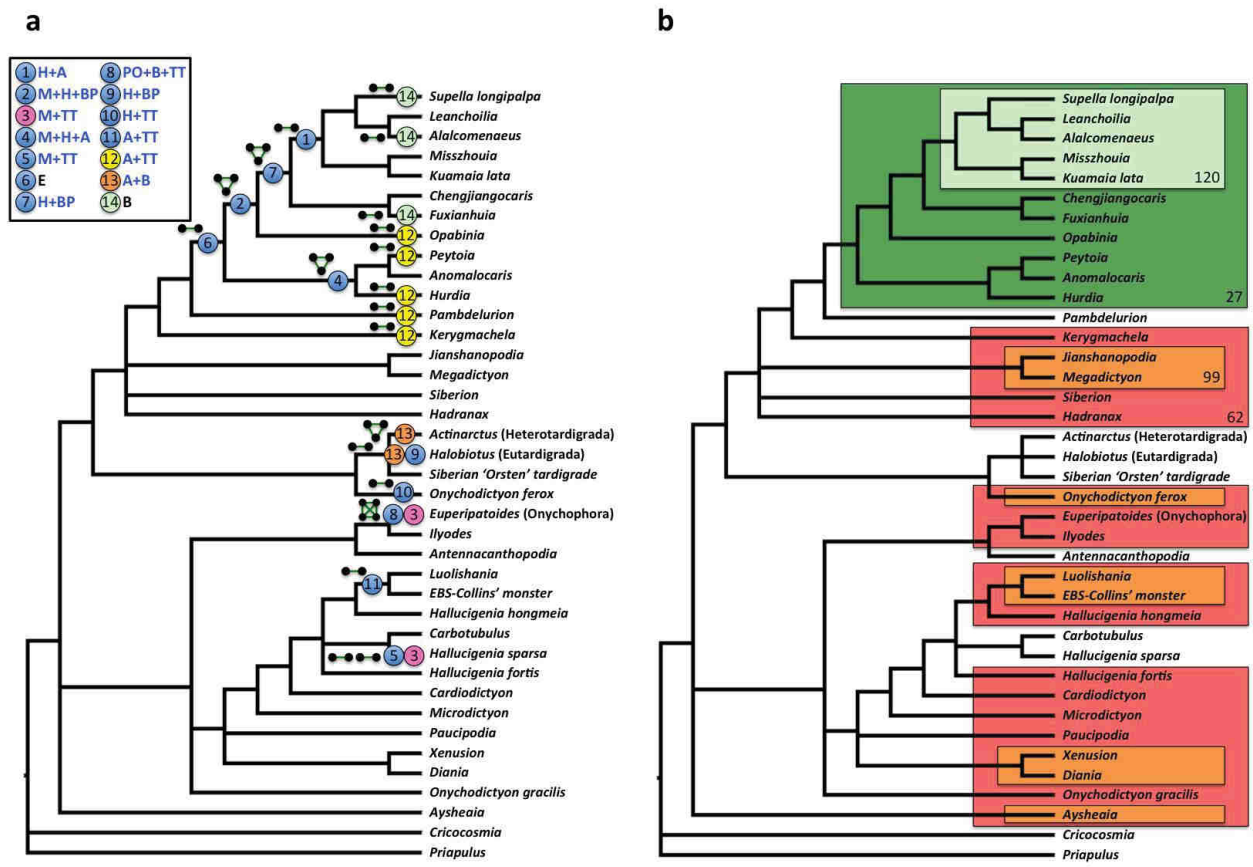


**Figure 4:** Composite phylogenetic tree of selected Rhinocerotidae, resulting from the parsimony analyses of[19,20,30], based on 282 cranio-mandibular, dental, and postcranial characters, to depict a) 8 trait complexes. Each complex is represented by its corresponding motif (each node represents a trait, each green edge represents the type I relationships between 2 traits) along the phylogeny, based on its taxonomic distribution. Each complex is also identified by a circled number; blue circles representing complexes shared by a common ancestor and all its descendants (putative synapomorphy), yellow circles representing a complex whose distribution does not map simply onto the phylogeny (homoplasy). The top left squared box identifies the distribution of complexes over the main regions of the rhinocerotid body plan (S: skull; T: teeth; J: jaw; BP: body plan; FL: Forelimb and HL: Hindlimb). Blue letters highlight complexes of traits from different regions. b) Phylogeny of Rhinocerotidae showing 2 examplars of traits with type II relationships. The distribution of trait 44 is nested in that of trait 23 (clade within clade). The distribution of trait 160 is nested in that of trait 217 (non clade within non clade). 23: Frontal bone: aspect|'rugose '; 44: Corpus mandibulae: base|'very convex '; 160: Lower molars: hypolophid|'transverse '; 217: Astragalus: orientation trochlea/distal articulation|'very oblique '.

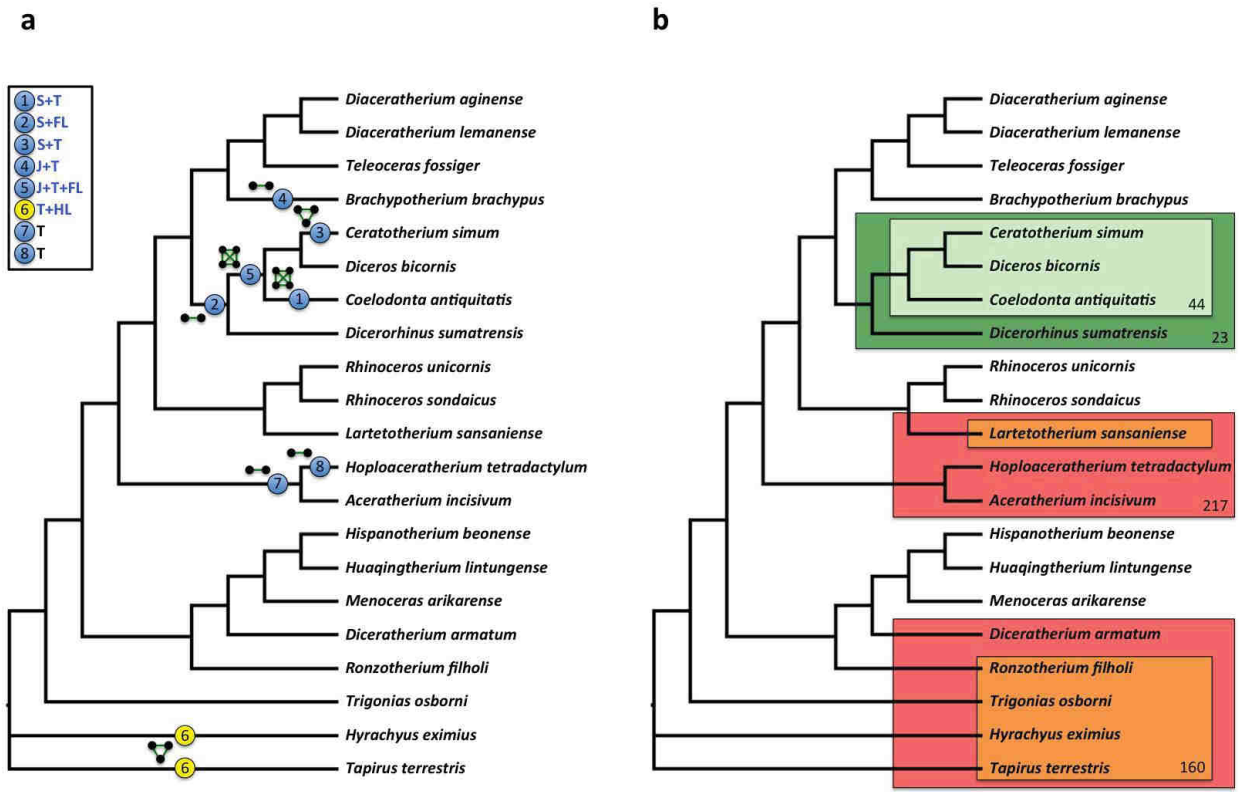


**Figure 5:** Schematic mapping of morphological traits on the panarthropod body plan. Main regions are indicated in boxes. Red squares are relatively unstable traits (i.e. type II in-degree is null); blue squares are relatively stable traits (i.e., type II in-degree is positive); yellow squares indicate traits with significant relatively stability (p-value < 0.05, permutation test). Numbers in squares correspond to NodeID. Black boxed squares correspond to traits that are significantly central in type D triplets (p-value < 0.05, permutation test). The barplot indicates the relative frequencies of traits in main regions of the panarthropod body plan, observed in all species. Areas in red/blue/yellow are versatile/relatively stable/significantly stable traits respectively. The main regions are H: head; M: mouth; E: eye; B: brain; PO: 1st post-ocular; BP: body plan; TT: trunk and tail; A: appendages.

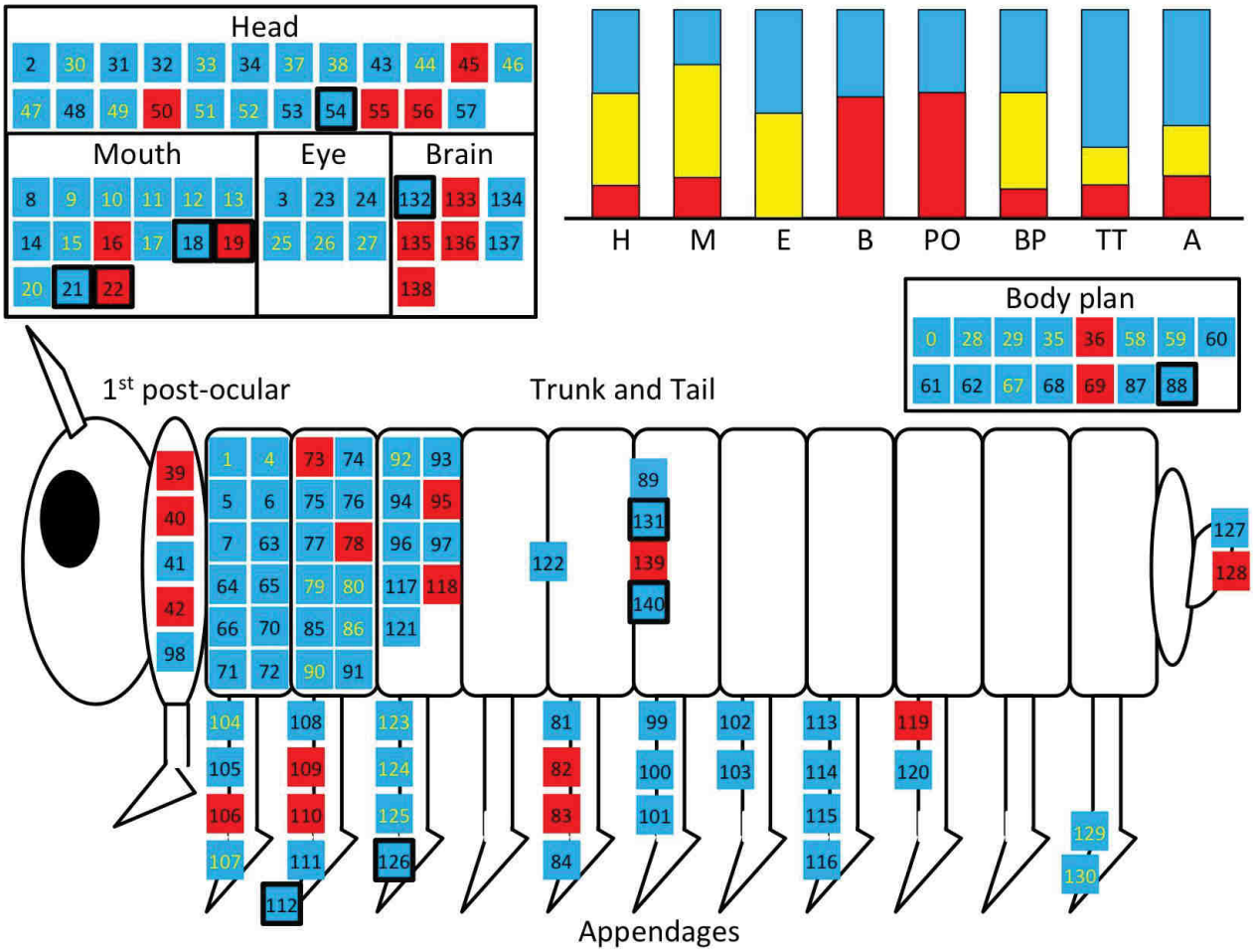


**Figure 6:** Schematic mapping of morphological traits on therhinocerotid body plan. Main regions are indicated in boxes. Red squares are relatively unstable traits (i.e. type II in-degree is null); blue squares are relatively stable traits (i.e., type II in-degree is positive); yellow squares indicate traits with significant relatively stability(p-value < 0.05, permutation test). Numbers in squares correspond to NodeID. Black boxed squares correspond to traits that are significantly central in type D triplets (p-value < 0.05, permutation test). The barplot indicates the relative frequencies of traits in main regions of the rhinocerotid body plan, observed in all species. Areas in red/blue/yellow are versatile/relatively stable/significantly stable traits respectively. The main regions are T: teeth; S: skull; J: jaw; BP: body plan; FL: Forelimb and HL: Hindlimb.

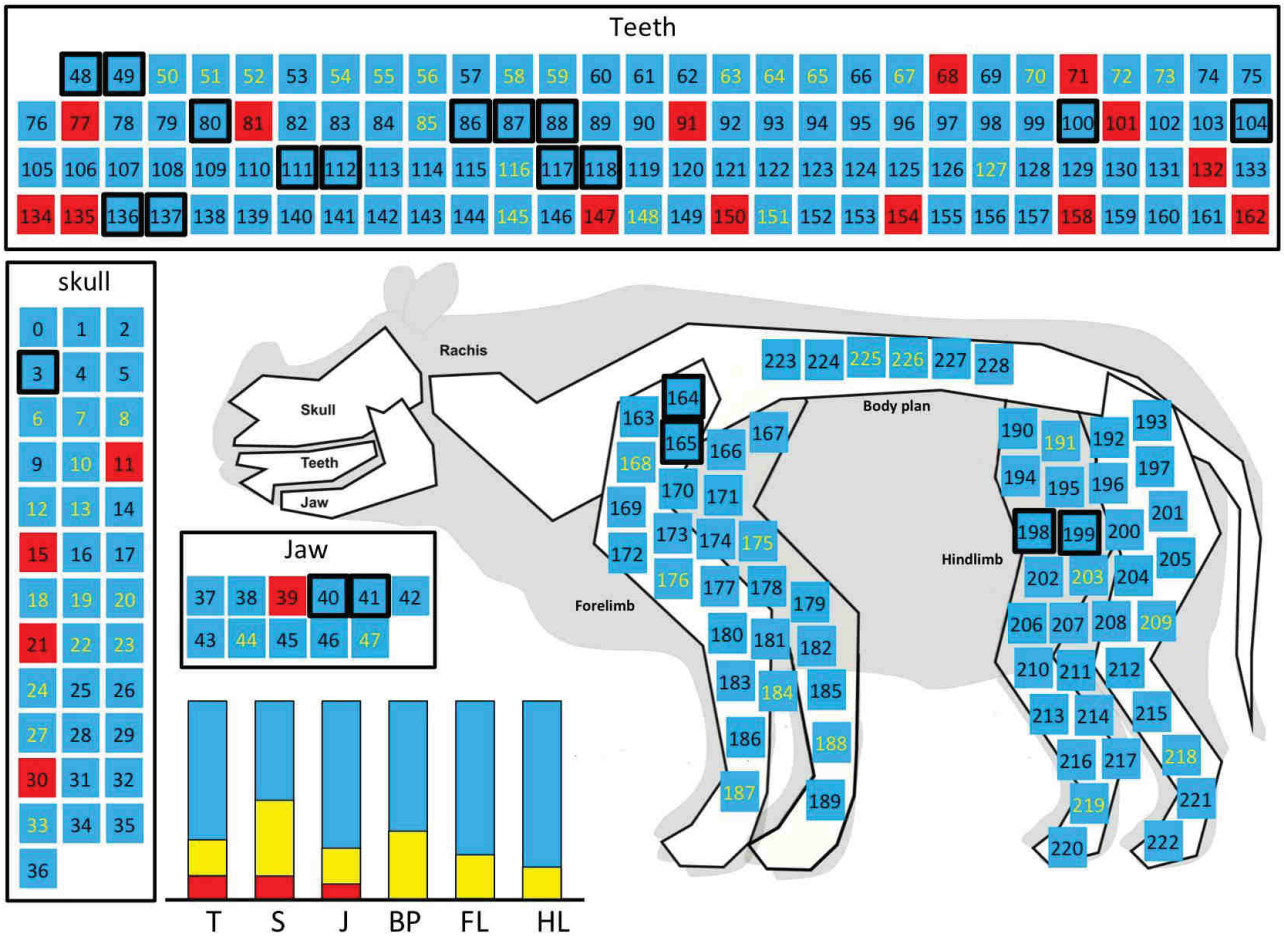


**Table 1.** Summary of network metrics with results of corresponding permutation test for a. Panarthropoda and b. rhinocerotidae.P-values were adjusted for multiple tests with a Bonferroni correction. Higher: significantly higher than expected by chance; lower:significantly lower than expected by chance; NS: non significant.

**a. PANARTHROPODA**

| Trait networks | Network metrics | Reference value | p-value | significance |
|---|---|---|---|---|
| Type I network | Number of complexes | 14 | 0,000333222 | higher |
| | Number of edges | 27 | 0,000333222 | higher |
| Type II network | Number of directed edges | 1766 | 0,000333222 | higher |
| | Number of significantly stable traits | 42 | | |
| Type III network | Number of edges | 2937 | 0,000333222 | lower |
| | Number of triangles | 36057 | 0,000333222 | lower |
| | Proportion of triangles | 0,1162664 | 0,000333222 | lower |
| | Density | 0,385129809 | 0,000333222 | lower |
| Type IV network | Number of edges | 1108 | 0,000333222 | higher |
| Type III+IV network | Number of type D triplets | 5671 | 0,000333222 | lower |
| | Number of significantly pivotal traits | 10 | | |

**b. RHINOCEROTIDAE**

| Trait networks | Network metrics | Reference value | p-value | significance |
|---|---|---|---|---|
| Type I network | Number of complexes | 8 | 0,040191962 | NS |
| | Number of edges | 22 | 0,00079984 | NS |
| Type II network | Number of directed edges | 5100 | 0,00019996 | higher |
| | Number of significantly stable traits | 50 | | |
| Type III network | Number of edges | 16063 | 0,00019996 | lower |
| | Number of triangles | 680642 | 0,00019996 | lower |
| | Proportion of triangles | 0,4286196 | 0,00019996 | lower |
| | Density | 0,711444795 | 0,00019996 | lower |
| Type IV network | Number of edges | 4774 | 0,00019996 | higher |
| Type III+IV network | Number of type D triplets | 186504 | 0,00179964 | NS |
| | Number of significantly pivotal traits | 21 | | |

**Extended Data Figure 1.** Pseudocode of the two algorithms in COMPONENT-GRAPHER

**Algorithm 1:** Algorithm to compute the co-occurrence network from a character matrix

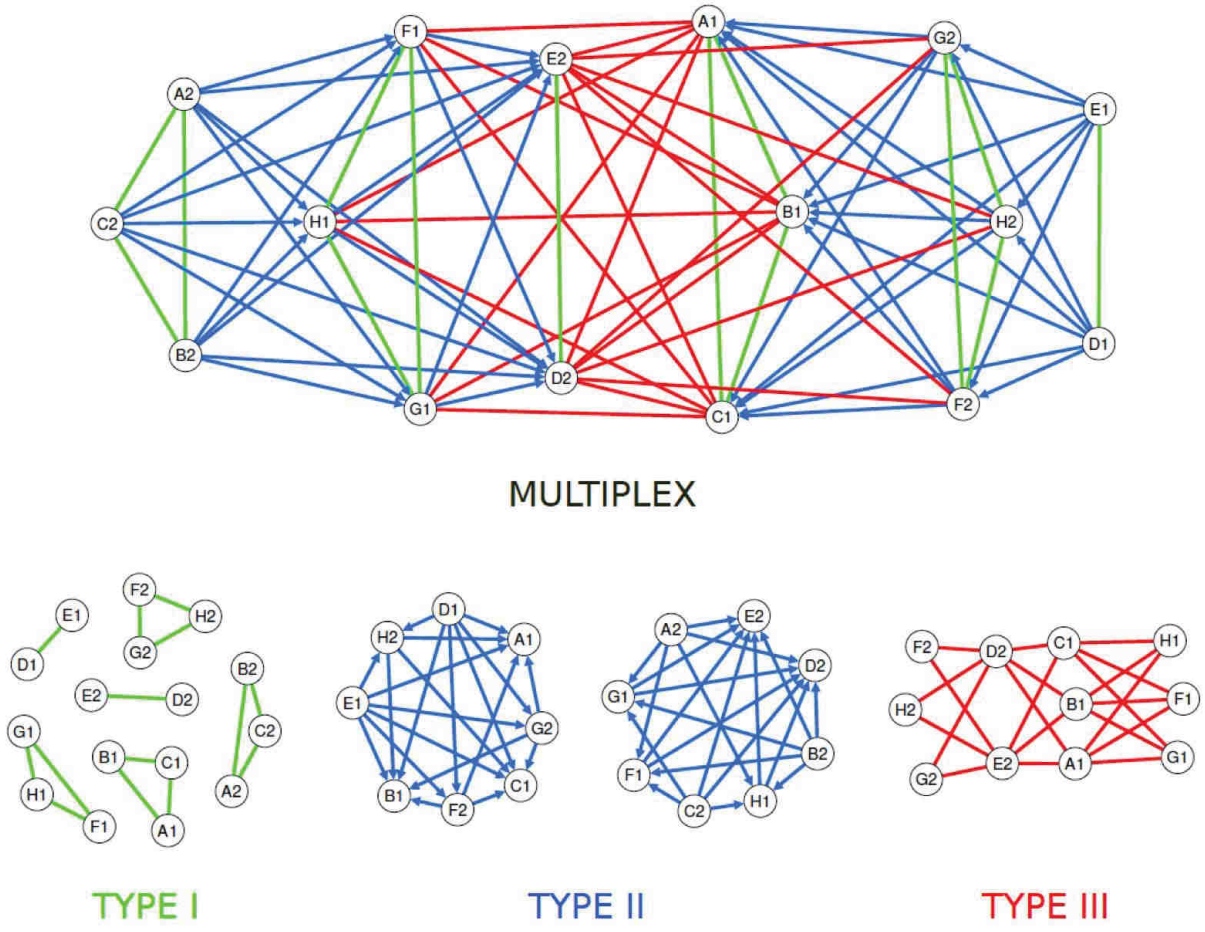
```
function ComputeNetwork (M)
  Input  : matrix M(T,C) of T taxa (rows) and C characters (columns)
  Output: a co-occurrence network N(n,e) with n nodes and e edges
Compute each node n and associated encoding
for i ← 1 to len(C) do
    foreach unique valid[1] character s of the column C[i] do
        new node n= Encode(C[i],s)
        n.column = i
        Add node n to network N
    end
end
Compute each edge e
foreach node n1 associated with column i do
    foreach node n2 associated with column j do
        /* edges are undirected unless otherwise stated */
        new edge e
        X=n1.states ∩ n2.states
        if X = ∅ then
            e.type = 4 (disjoint)
        else if n1.total=n2.total=|X| then
            e.type = 1 (identical)
        else if |X| = n1.total and |X| < n2.total then
            /* directed edge from n1 → n2 */
            e.type = 2 (Inclusion)
        else if |X| = n2.total and |X| < n1.total then
            /* directed edge from n2 → n1 */
            e.type = 2 (Inclusion)
        else
            e.type = 3 (Overlap)
        end
        Add edge e to network N
    end
end
return network N
```

[1]Valid character are {0..9, A..Z}. Unvalid characters {$, * and -} are ignored during the node creation while polymorphic character *e.g.* {0,1} at a unique position in matrix $M$ must either be selected or removed prior to running the algorithm.

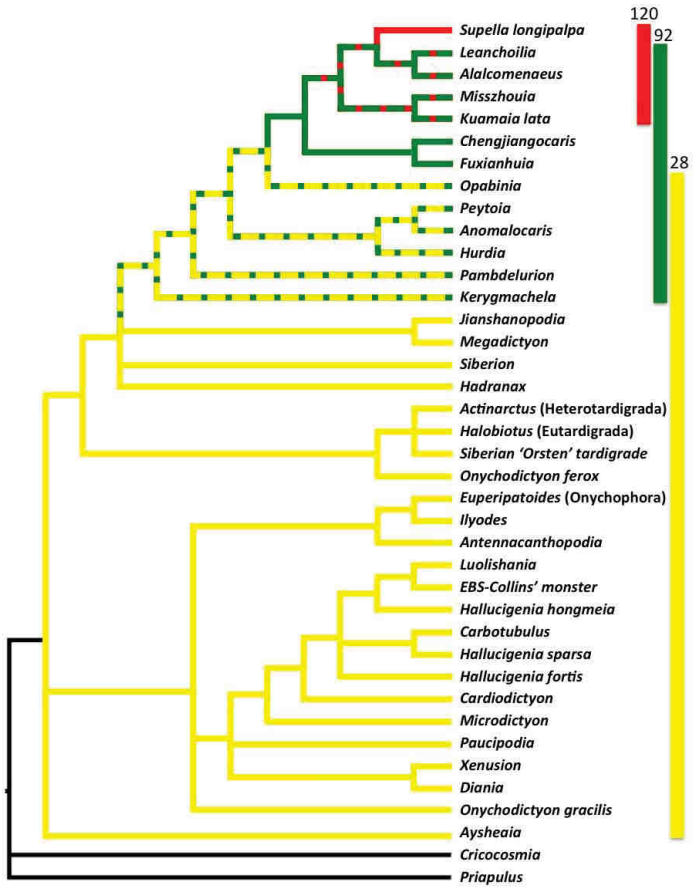
**Algorithm 2:** Algorithm to create a new node $n$ with associated binary states

```
function Encode (c,s)
  Input  : column c of matrix M, character s
  Output: new node
new node n
n.states ← {}
n.total = 0 /* total taxa associated with this node */
for i ← 1 to len(c) do
    if c[i] = s then
        n.states[i] =1
        n.total++
    else
        n.states[i] =0
    end
end
return node n
```

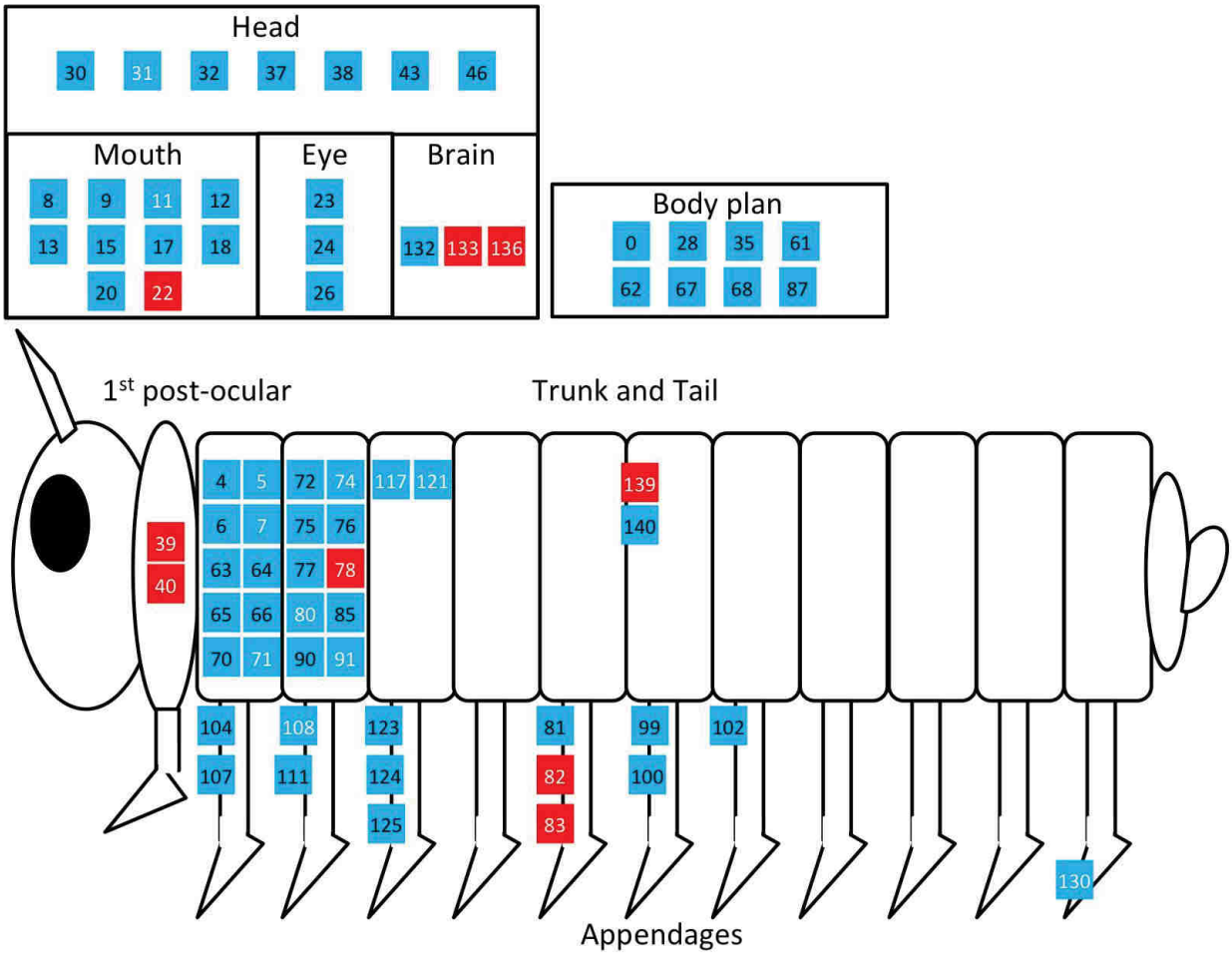
**Extended Data Figure 2.** Co-occurrence networks for palaeontological studies. Each trait is treated as an individual node. Two nodes are directly connected by an edge indicating their type of relationship (I, II or III), but are disconnected otherwise (Type IV). This inclusive co-occurrence graph can also be decomposed into three networks: a green network of identity featuring only nodes connected by type I edges; a blue network of inclusion featuring only nodes connected by oriented type II edges, an arrow pointing from the least stable toward the most stable trait; a red network of overlaps featuring only nodes connected by type III edges.



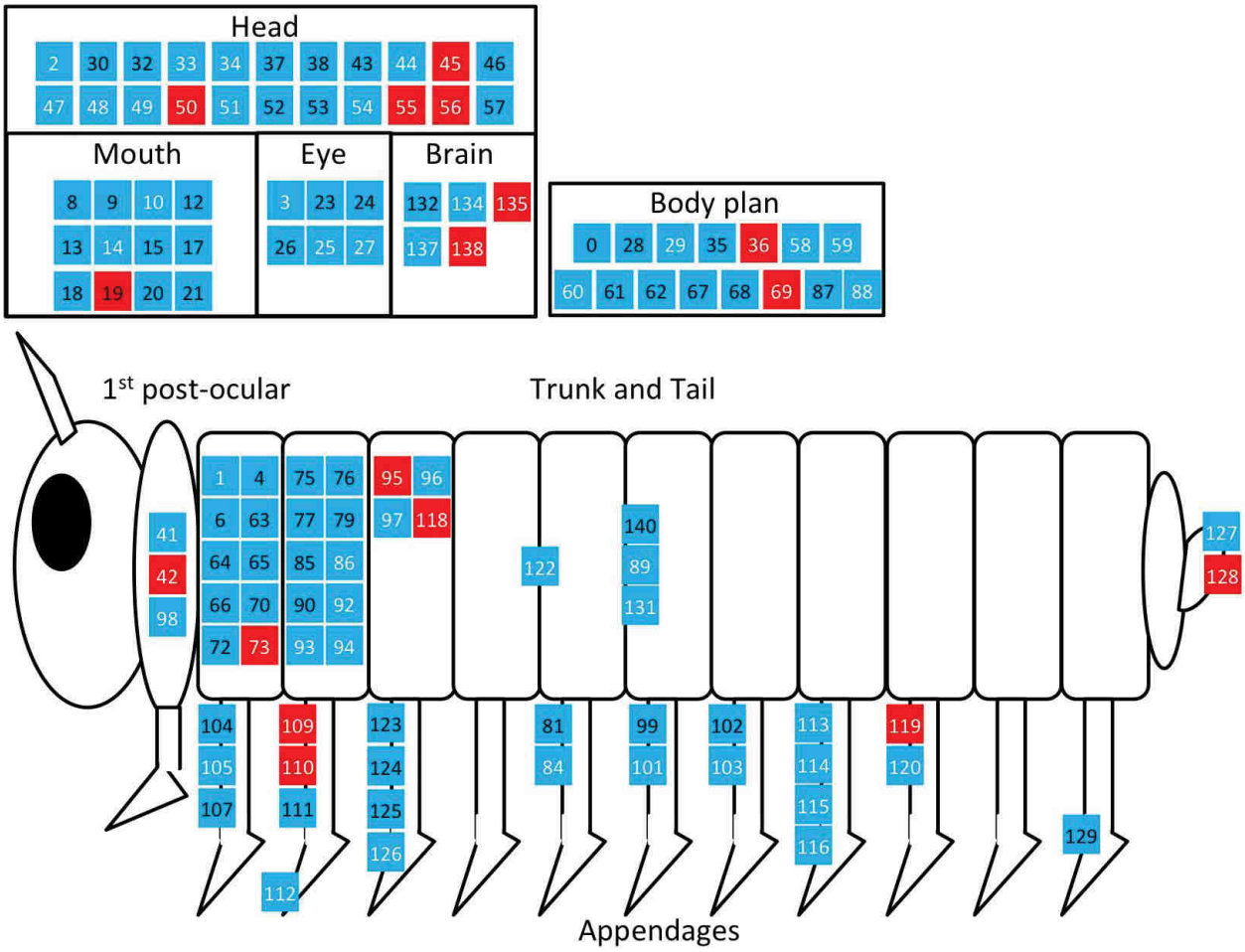
**Extended Data Figure 3.** Mapping of a type D triplet along the phylogeny of panarthropods. Each trait is represented by a different color. The distribution of trait 120 overlaps with that of trait 92; the distribution of trait 92 overlaps with that of trait 28; however the distributions of trait 120 and 28 are disjoint. 28: Sclerotized post-ocular (post-protocerebral) body appendages with arthrodial membranes | 'lobopodous' ; 92: Trunk exites | present; 120:Appendages comprise 15 or more podomeres | Fewer than 15 podomeres.



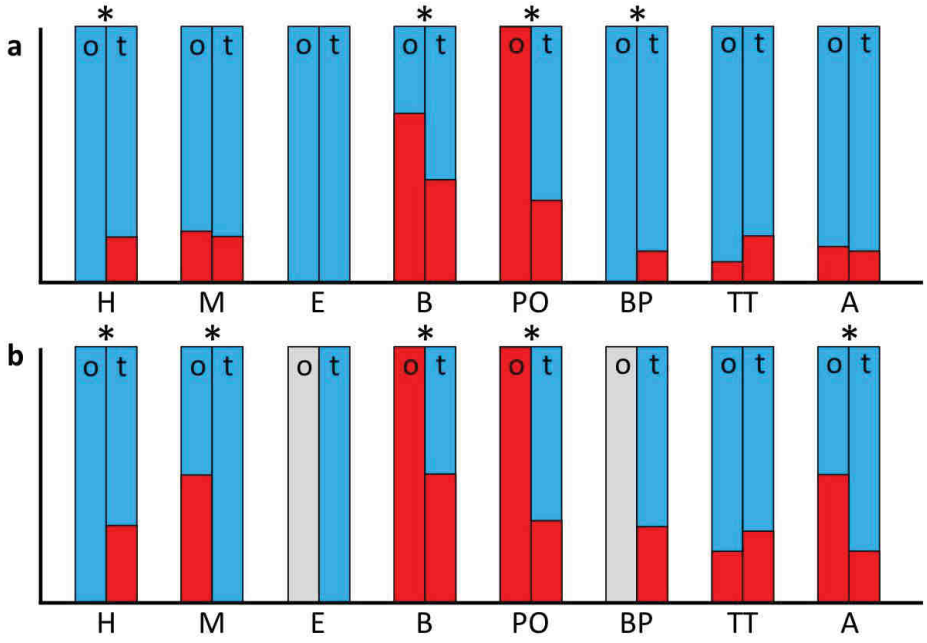
**Extended Data Figure 4.** Schematic mapping of morphological traits on the onychophoran body plan. Main regions are indicated in boxes. Red squares are relatively unstable traits (i.e. type II in-degree is null); blue squares are stable traits (i.e., type II in-degree is positive). Numbers in squares correspond to NodeID. Font colours represent the distribution of the traits: black, present in Onychophora and relatives; white, exclusively present in Onychophora.



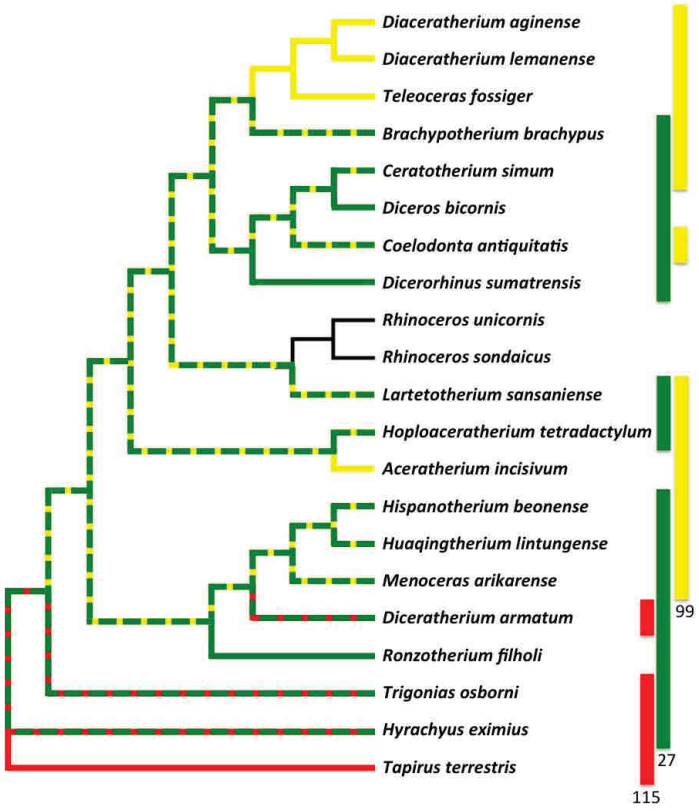
**Extended Data Figure 5.** Schematic mapping of morphological traits on the tactopodan body plan. Main regions are indicated in boxes. Red squares are relatively unstable components (i.e. type II in-degree is null); blue squares are stable traits (i.e., type II in-degree is positive). Numbers in squares correspond to NodeID. Font colours represent the distribution of the traits: black: present in Tactopoda and relatives; white, exclusively present in Tactopoda.



**Extended Data Figure 6.** Relative frequencies of traits in main regions of the panarthropod body plan. Areas in red/blue are unstable/stable components respectively. Areas in grey are not represented. a: Traits observed in Onychophora (o) or Tactopoda (t). b: Traits exclusively observed in Onychophora (o) or Tactopoda (t). The main regions are H: head; M: mouth; E: eye; B: brain; PO: 1st post-ocular; BP: body plan; TT: trunk and tail; A: appendages. Significant differences are identified by *.



**Extended Data Figure 7.** Mapping of a type D triplet along the phylogeny of rhinocerotids. Each trait is represented by a different color. The distribution of trait 27 overlaps with that of trait 99; the distribution of trait 27 overlaps with that of trait 115; however the distributions of trait 99 and 115 are disjoint. 27: Zygomatic/frontal widths|'less than 1.5 '; 99 : Upper molars: crochet|'always present ' ; 115 : M1-2: metastyle|'short '.



# III.3 LINGUISTIC EVOLUTION

The origin of language is linked with the emergence of modern humans (Homo sapiens) some 200 000 years ago (Dediu and Levinson 2013). The emergence of human language drastically changed the character of human society, but we still know little about the details of this process. For a long time biologists and linguists have been noticing surprising similarities between the evolution of life forms and languages, although, the objects studied in biology (e.g genes and genomes) and linguistic (e.g words, languages) are different.

In the 19th century, Charles Darwin (1809-1882) and August Schleicher (1821-1868), a linguist working in Jena (Germany), compared the evolution of languages with the evolution of species (Darwin 1859; Schleicher 1863). August Schleicher propagated what he called the *Stammbaumtheorie* (family-tree theory) (Schleicher 1853a; 1853b), a genealogical classification of language varieties arranged in a genealogical tree (Figure 18). This classification system based on branching trees, was a major development in the study of Indo-European and other language families (Meier-Brügger 2002).

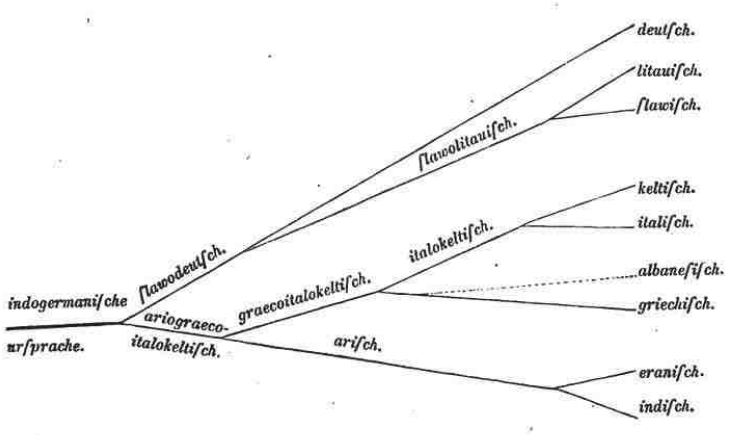



**Figure 18:** The first genealogical tree of the Indo-European languages created by August Schleicher in 1863. (source: Compendium der vergleichenden Grammatik der Indogermanischen Sprachen)

However, soon after the family tree model had first been proposed, many linguistics criticized the tree model for its simplicity. This tree model was reproached to mask the complexity of language evolution (Schmidt 1872). For those tree opponents, language evolution could not be explained simply as a tree-like differentiation, since horizontal transmission often plays an equally important role for the development of languages (Schuchardt 1900). In linguistic, this process is called lexical borrowing which is the transfer of a word from one language to another (Weinreich 1953). This adaptation of foreign

elements is usually a result of language contact, such as contact between speakers of two different languages. Borrowing is not restricted to the concrete integration of foreign words from one language into another, but can also happen purely semantically, if bilingual speakers start integrating structural aspects of one language into the other (Weinreich 1953). Borrowing is similar to horizontal gene transfer in evolutionary biology.

In 1872, Johannes Schmidt introduced the *Wellentheorie* (waves theory) (Schmidt 1872) which states that certain changes spread like waves in concentric circles over neighboring speech communities (Figure 19). He claimed that these waves were independent of each other, and are not necessarily nested. For linguists who investigate the evolution of language varieties, grammatical features, and words, both models (the trees and the waves) are each illustrating one crucial aspect of language evolution.

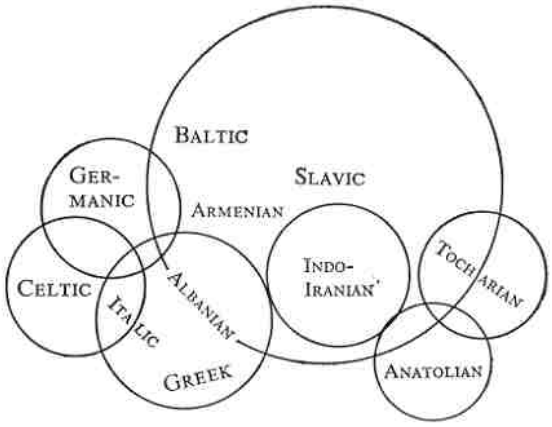



**Figure 19:** Distribution of Indo-European languages seen in term of the waves theory. (After Schmidt and Lehmann).

Many linguists assume that the two models are complementary with the tree model representing the genealogical processes and the wave model representing complex contact relations between languages. Surprisingly, not many attempts were made to combine tree and wave in a common framework (Southworth 1964).

Similarly to what happened in evolutionary biology, an obvious solution, which was also pointed out early by linguists, are network models, which could easily handle both vertical and horizontal relations between languages (Nelson-Sathi et al. 2011). The first explicit network approach was presented by Bonfante in 1931, and tried to depict the historical relations between the major branches of Indo-European (Bonfante 1931). Since Bonfante, several network-based studies have been proposed (Geisler and List 2013).

Linguistic networks are characterized by a high level of abstraction compared to networks in other areas of research. Words can be linked because they share the same context (semantic network), same sounds (phoneme network) or a common ancestor (cognates network) (List et al. 2016a). In addition to representing language history with the help of networks, networks can be used for many additional purposes: They can be employed to search for homologous words in the same way as biologists use them to search for homologous genes, for example, with help of similarity networks. In the previous chapters, it was explained that SSN were useful in the detection of homologous genes forming highly divergent gene families. In linguistics, homologous words, having descended from a common ancestor, are called cognates. Similarity networks can be used to search for highly diverse cognate sets across languages (List et al. 2016b). Moreover, network-based methods can also be used to detect non-tree like aspects of language history, like, for example, compound words, which are similar to composite genes (Figure 20). Many metrics defined for networks exist, which can help investigating the relationships between similar components across or within the same language. They can address the properties of a single node or a pair of nodes, but can be extended to the whole network by averaging.

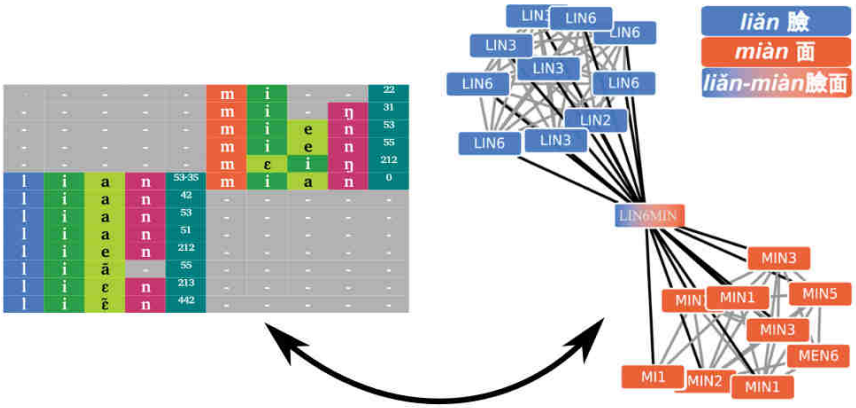



**Figure 20: Similarity networks reconstructed from local alignments for dialect words meaning 'face' in 20 Chinese dialect varieties.**
The data contains three variants, two simple words *liăn* and *mián*, two words of different origin, and one fused form *liăn-mián*. Numbers in the alignment reflect tone patterns, which are characteristic for South-East Asian languages. Edges colored in black differ in their local and global alignments, edges colored in gray show identical alignments for local and global analyses. The fused form serves as a hub connecting the two components.

Words as well as genes evolve through multiple evolutionary processes. Although the evolution of these two different objects share some common processes, there are also specific ones. In the article n° 8, we compared important evolutionary processes in biology and linguistics and identified specific and common processes in these disciplines (Figure 21). We introduced new process-based analogies in biology and linguistics that support the transfer of phylogenetic and network methods, from biology to linguistics, to automatize the detection of common ancestry and multi-level introgressive processes. We showed that interdisciplinary approaches can be fruitful, since methods, models, and research programs can be transferred, creating added values in both disciplines. This article has been accepted and published in the journal "Biology Direct".

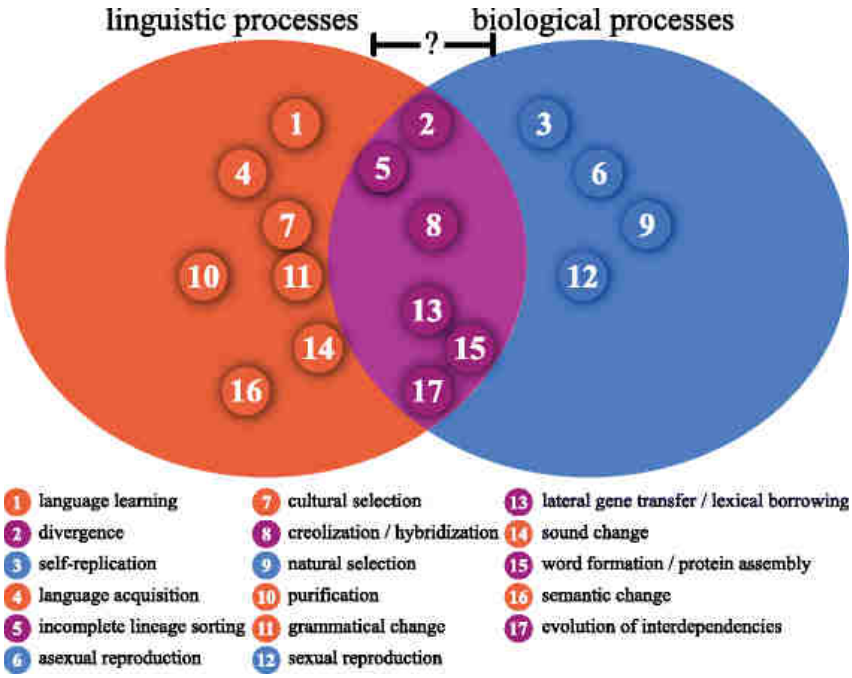



**Figure 21: Contrasting purely linguistic, purely biological, and analogous processes in linguistics and biology.** For Process-Based Analogies, we contrast the biological term with the linguistic term, if both disciplines address the processes in their terminology. See the text for further clarification

List *et al. Biology Direct* (2016) 11:39
DOI 10.1186/s13062-016-0145-2

Biology Direct

REVIEW Open Access

# Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics

Johann-Mattis List[1,2*], Jananan Sylvestre Pathmanathan[2], Philippe Lopez[2] and Eric Bapteste[2]

## Abstract

**Background:** For a long time biologists and linguists have been noticing surprising similarities between the evolution of life forms and languages. Most of the proposed analogies have been rejected. Some, however, have persisted, and some even turned out to be fruitful, inspiring the transfer of methods and models between biology and linguistics up to today. Most proposed analogies were based on a comparison of the research *objects* rather than the *processes* that shaped their evolution. Focusing on *process-based analogies*, however, has the advantage of minimizing the risk of overstating similarities, while at the same time reflecting the common strategy to use processes to explain the evolution of complexity in both fields.

**Results:** We compared important evolutionary processes in biology and linguistics and identified processes specific to only one of the two disciplines as well as processes which seem to be analogous, potentially reflecting core evolutionary processes. These new *process-based analogies* support novel methodological transfer, expanding the application range of biological methods to the field of historical linguistics. We illustrate this by showing (i) how methods dealing with incomplete lineage sorting offer an introgression-free framework to analyze highly mosaic word distributions across languages; (ii) how sequence similarity networks can be used to identify composite and borrowed words across different languages; (iii) how research on partial homology can inspire new methods and models in both fields; and (iv) how constructive neutral evolution provides an original framework for analyzing convergent evolution in languages resulting from common descent (*Sapir's drift*).

**Conclusions:** Apart from new analogies between evolutionary processes, we also identified processes which are specific to either biology or linguistics. This shows that general evolution cannot be studied from within one discipline alone. In order to get a full picture of evolution, biologists and linguists need to complement their studies, trying to identify cross-disciplinary and discipline-specific evolutionary processes. The fact that we found many process-based analogies favoring transfer from biology to linguistics further shows that certain biological methods and models have a broader scope than previously recognized. This opens fruitful paths for collaboration between the two disciplines.

**Reviewers:** This article was reviewed by W. Ford Doolittle and Eugene V. Koonin.

**Keywords:** Process-based analogies, Language evolution, Protein assembly, Word formation, Lateral transfer, Constructive neutral evolution, Similarity networks, Incomplete lineage sorting

*Correspondence: mattis.list@lingpy.org
[1]CRLAO/EHESS, 2 rue de Lille, 75007, Paris, France
[2]Equipe AIRE, UMR 7138, Laboratoire Evolution Paris-Seine, Université Pierre et Marie Curie, 7 quai St Bernard, 75005, Paris, France

## Background

Biological objects on Earth have been evolving for billions of years. The origin of language evolution dates back to only about 200 000 years ago. The specific aspects of the evolution of life forms and the evolution of languages are traditionally investigated by the disciplines of evolutionary biology and historical linguistics. The research objects of the two disciplines differ greatly. Biology deals with substantial objects, that is, objects with a concrete physical manifestation. Languages, on the other hand, are 'products of the human mind' ([1], p. 144). They are intellectual objects ([2], p. 72), that is, objects whose manifestation is based on the interaction between humans. They are realized physically, be it when they are spoken or written down, but their realization is dependent on the existence of individuals who speak and understand them, and in this way, language systems are constantly being reconstructed by new speakers who learn them [3].

Similar models have been developed independently in the history of both disciplines. Both biologists and linguists have a long tradition of using trees to model diversification by a genealogy. Trees were independently popularized by August Schleicher (1821–1868) in 1853 [4] and Charles Darwin (1809–1882) in 1859 [5]. Both fields also share a more recent tradition of using networks to capture reticulation, although early network models of languages [6–9] (see [10, 11]) and life forms [12, 13] (see [14]) even predate the classical family trees [4, 5, 15–17] (see [10, 14, 18], and Fig. 1). Some processual similarities are also reflected in the methods independently developed and applied in both disciplines, such as, for example, cladistic approaches and alignment analyses. In linguistics, approaches for subgrouping based on shared innovations (or shared derived characters) date back to the end of the 19th century ([19], p. 24). In biology they were independently developed in the middle of the 20th century [20]. At about the same time, first approaches to numerical tree reconstruction based on distance data can be found in both disciplines [21, 22]. Although only sporadically applied and never fully automatized, early examples in which linguists aligned corresponding sounds in multiple homologous words can already be found in the early 20th century [23–25]. In biology, automatic methods for sequence alignment were developed from 1970 onwards soon after the rise of molecular biology [26–28]. Both biologists and linguists also struggle with common epistemological limitations, since the processes they investigate lie in the past, which is why uniformitarianism, the assumption that the processes observed today do not differ much from the processes which happened in the past ([29], p. 165), still plays an important role in biology and linguistics [30–32].

Apart from similar models and methods developed independently, there was and is also a considerable amount of explicit transfers between the two disciplines. An early example is the intimate intellectual exchange on Darwin's evolutionary theory and its implications for the study of languages between the biologist Ernst Haeckel (1834–1919) and the linguist August Schleicher (1821–1861) [33]. According to this correspondence, it was Haeckel who brought Schleicher's attention to the work of Darwin. Schleicher was deeply impressed by the similarities of the research objects in such different domains ([34], p. 6). He emphasized, however, also that these parallels would only hold for the essential features, not for the details ([33], p. 29). Haeckel, in turn, took inspiration from Schleicher's language tree diagrams to promote evolutionary tree drawing in biology ([10], p. 300).

In the 20th century, especially the early work on genetics, not long after the correct modelling of the structure of DNA by Watson and Crick [35], was characterized by a strong linguistic influence. This is reflected in the multitude of linguistic terms, like 'alphabet' and 'word' [36] or 'translation' [37], which were used to describe biological phenomena in the biological domain [38]. While, as indicated by Eugene V. Koonin (one of the reviewers of this manuscript), the majority of these terms reflected mere metaphors of which only a minority became later integrated into the standard terminology of biology (see also [39]), we can also find examples for the explicit transfer of linguistic methods and theories to the biological domain. Thus, up to today, the theory of formal grammar [40] plays an important role in addressing certain problems in bioinformatics [41], like RNA folding and protein structure analysis, and it is not uncommon for biological textbooks on sequence comparison to also include a chapter on formal grammars ([42], pp. 233-259). This influence is not restricted to classical models of grammar [43]. Advanced models, like *tree adjoining grammar*, have likewise been used for RNA structure prediction [44], and inherently linguistics methods, like methods for document prediction, have been successfully applied for the task of protein classification [45]. During the last twenty years the direction of interdisciplinary transfer has turned, and many methods originally designed for applications in evolutionary biology have been applied to linguistic data. These include algorithms for phylogenetic reconstruction [46, 47], phylogenetic network approaches [48–52], multiple sequence alignment [53–55], and homolog identification [55, 56].

In the following, we will argue that these transfers can be further enhanced. By shifting from the comparison of research *objects* to the comparison of *processes* affecting the research objects in the disciplines, wrong analogies due to an exaggeration of similarities and a neglection of differences can be avoided. At the same time, the identification of important processes, common to language and biological evolution, can give rise to new, potentially

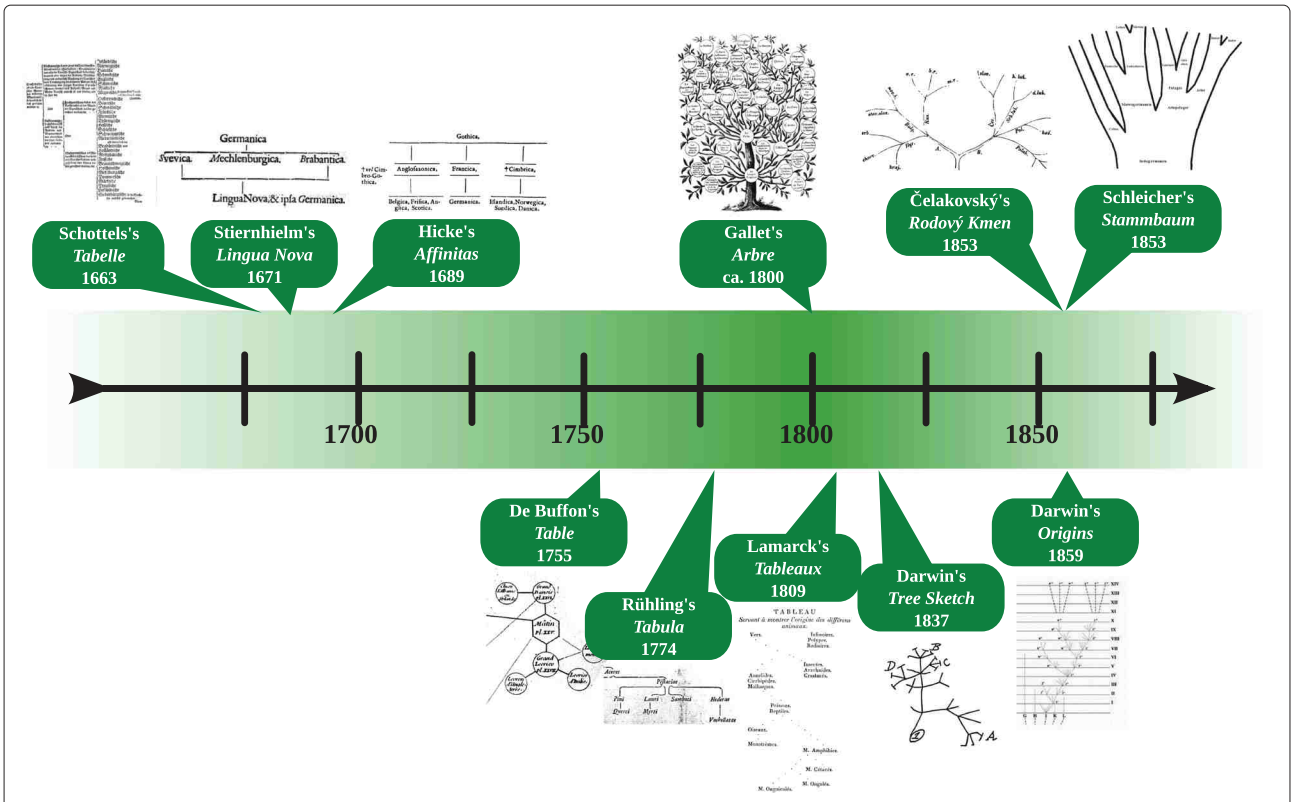



**Fig. 1** Timeline of early tree- and network diagrams in linguistics (*top*) and biology (*bottom*). Schottel's branching table of Germanic languages from 1663 is the earliest we could identify. The three following early diagrams in linguistics by Stiernhielm (1671) [7], Hickes (1689), [9], and Gallet (1800) [8] all contain reticulation, real trees only start with Ćelakovský and Schleicher (1853) [4, 15]. The situation is similar in biology, where the two schemas by Leclerc De Buffon (1755) [12] and Rühling (1774) [13] allow for reticulation, in contrast to Lamarck (1809) [17] and Darwin (1837, 1859) [5, 16]

fruitful analogies. For linguistics, these transfers offer new theoretical and practical ways to explain the mosaic distributions of words across related and unrelated languages, with and without invoking processes of lateral transfer. A new analogy between the process of word formation in linguistics and protein assembly in biology offers a fresh perspective on the idea of a *protein grammar* [57] and can inspire new methods and models in both fields. Invoking a system perspective can further help to demystify the phenomenon of convergent evolution in languages resulting from common descent.

## Process-based analogies

The striking similarities between biological and language evolution opt for a systematic investigation of analogies in the two disciplines. Such an investigation may cumulate in a program whose objectives would be (a) to investigate the isomorphy of processes, methods, and models in the two disciplines, (b) to foster the development of models lacking in either of the disciplines, and (c) to reduce the duplication of effort. Such a program, very close to the one proposed by the Society for General Systems Research in 1954 (as reported by ([58], p. 13)), would further 'promote the unity of evolutionary science through improving communication among specialists' (adapted from ([58], p. 13)). A multitude of analogies between biology and linguistics has been proposed in the past 200 years [59]. Languages have been compared with organisms ([60], p. 16f), species [61], microbes [49, 50], mutualist symbionts [62], and populations [63]. Words have been compared with cells ([33], p. 23f), amino-acids [64], codons [65, 66] and genes [61]. Sounds (phonemes) have been compared with nucleic bases [65, 67] and atoms [64]. Only a small amount of these analogies has received broader attention, many have been rejected quickly after they were first proposed, and only recently, an explicit transfer of methods and models has been initiated [68].

We find two main reasons why the majority of analogies that have been proposed between biology and linguistics have not turned out to be fruitful on the long run. First, most of the proposed analogies are object-based, taking the research objects as their main comparandum. Second, given the different media in which the research objects in the two disciplines manifest, it is well likely that the number of discipline-specific phenomena largely exceeds the number of commonalities. As a result, all

analogies which are proposed between the two disciplines should be rigorously checked, and methods should never be blindly transferred but always carefully adapted to the specific needs of the target discipline [55]. Object-based analogies bear a high risk of overstating similarities in interdisciplinary research and may easily lead to wrong conclusions and inadequate transfer of methods and models. Schleicher, for example, compared languages with organisms and derived from this comparison the hypothesis that languages would also grow old and die [33, 59]. To circumvent this problem we propose to concentrate on analogies between *processes*. *Process-based analogies* (PBA) are explicitly agnostic regarding further analogies between the research objects themselves. In taking processes as our starting point, we build on general approaches to analogy, which usually claim that the core of analogy are similarities of functions [69]. Focusing specifically on processes rather than functions is justified by the evolutionary background of biology and linguistics: processes serve as the major *explanans* in evolutionary research. Identifying analogies between evolutionary processes in these two fields as different as biology and linguistics may thus contribute to a unifying explanatory framework of evolutionary processes. Even when basing analogies on processes, however, we should not forget that we are dealing with very different disciplines, and any methodological transfer should be accompanied by a careful adaptation of methods to the needs of the target discipline. Future research will need to decide whether we the proposed analogies reflect general evolutionary processes or processes specific to the respective disciplines. Our uncertainty regarding the extent to which a unification of evolutionary processes in biology and linguistics is possible is reflected in Fig. 2, where we have marked the degree by which the processes in the disciplines overlap with a question mark.

The focus on processes produces potentially fruitful novel analogies. It can also identify processes that seem to be exclusive to one of these two historical sciences (Fig. 2). Among the exclusively linguistic processes, we identify such processes as *sound change* (Fig. 2:14), *semantic change* (Fig. 2: 16), or *purification* (Fig. 2: 10). Neither of these processes seems to have a biological counterpart: It has been proposed to compare sound change in linguistics with concerted evolution in biology [67], but we think that the analogy between the two processes does not completely hold. In concerted evolution, two traits change in a similar manner. During sound change, the phoneme system of a language changes [70]. An analogous process in biology would be a process in which the canonical amino acids constantly changed during evolution. During semantic change, the associations between words and concepts are restructured ([55], pp. 24–27). One might think of comparing this with changes in the regulation of genes in a genome which may yield drastic changes in function [71]. However, while biological function is still determined and restricted by the nucleic and proteic forms, no necessary limits are imposed on the association between forms and meanings in natural languages: the association is arbitrary in the sense that a substantial link between form and meaning in languages is not necessary [72, 73]. Purification is a process by which language change is actively triggered with the goal to preserve the pure state of one's mother tongue. One paradigmatic example for this kind of change is the Romanian language which was heavily influenced by neighboring Slavic varieties, until, around the end of the 18th century, nationalist movements triggered a purification process by which Slavic loanwords were successively replaced with native Romance words [74].

Exclusively biological processes include, among others, asexual (Fig. 2:6) and sexual reproduction (Fig. 2:12), but most likely also natural selection in a strict sense (Fig. 2:9). Some scholars claim that there is evidence that certain aspects of languages, like their sound systems, correlate with environmental factors [75], while other aspects, like their morphological complexity or the way they change, correlate with demographic factors [76, 77]. But languages are not independent of the ones who use them. They replicate via acquisition (of one's first language, Fig. 2:4) and learning (of a further language, Fig. 2:1). Although we cannot exclude, that selection processes in biology and linguistics are similar and that a common theory of fitness could be derived [78], and that languages, for example, differ regarding the difficulty with which they can be learned, we think it would be premature to draw any process-based analogies here. Linguists tend to avoid the discussion of the fitness of languages due to its political and cultural implications, emphasizing that all natural languages are learnable within the normal time span that children need to acquire a language. There are also no known cases of languages becoming abandoned by their speakers due to their difficulty, since speakers always slightly adjust their languages to fulfil their communicative needs and thus maintain the functionality of their most important communication tool. Even if ease of transmission was a factor potentially influencing language evolution, as suggested by W. Ford Doolittle (the first reviewer of this manuscript), learning difficulty is by no means the sole factor that leads to language spread. The spread of English as a major second and first language, for example, was largely due to political factors, depending on those who carry the language rather than the language itself. It was not the rather simple grammatical structure of English that favored its spread but the fact that large powerful countries in different parts of the world use English as their first and official language. That the speaker size and especially the amount of second language speakers

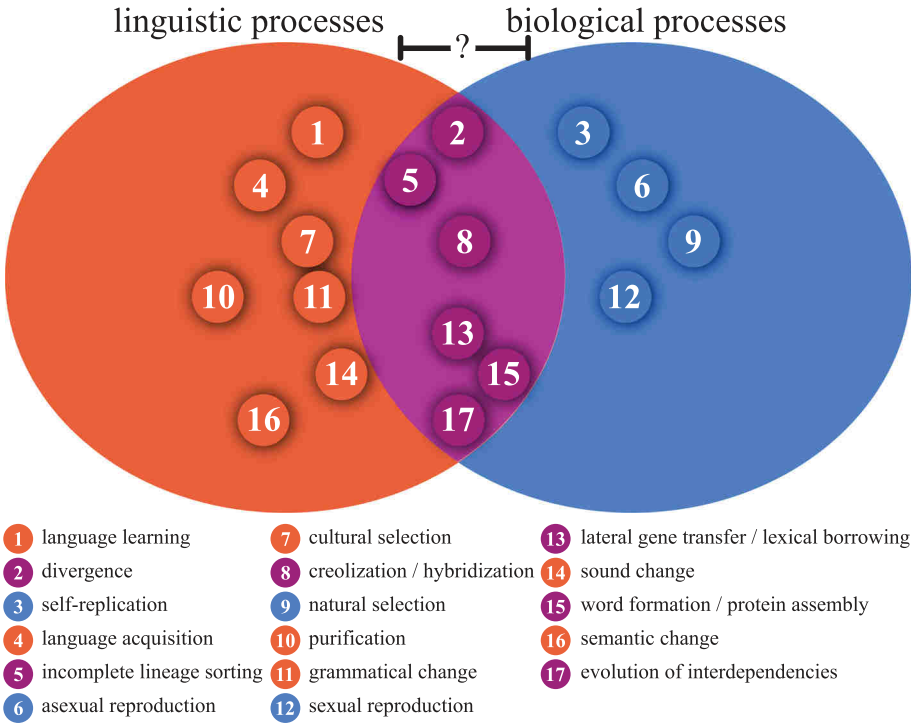



**Fig. 2** Contrasting purely linguistic, purely biological, and analogous processes in linguistics and biology. For Process-Based Analogies, we contrast the biological term with the linguistic term, if both disciplines address the processes in their terminology. See the text for further clarification

may have an impact on the way languages evolve is most likely [76, 77]. In order to be able to assess the various factors more substantially, however, much more research is required in the future, and we are careful in drawing any analogies with biological processes, as we still do not know enough about all the mechanisms involved in language evolution. For this reason, we are careful in identifying a direct counterpart process of natural selection in the linguistic world. There is ample evidence that some kind of selection occurs during language evolution [79, 80]. This selection is often called cultural selection, and we place it among the exclusively linguistic processes (Fig. 2:7).

The large amount of disciplinary-internal processes for which we could not find any counterpart is a challenge for current research in the evolutionary sciences, and a specific challenge for biologists and linguists. One the one hand, future research may show that some of these processes actually have counterparts in the other discipline, on the other hand, we may make progress in explaining *why* those processes are unique to a specific domain. In both cases, we will gain deeper insights into both the unity and the disunity of evolutionary processes across disciplines. But at least as important as the differences are the newly identified commonalities, which we will discuss in detail in the following section.

## New analogies for biology and linguistics

The PBAs which we identified can be roughly divided into three categories, depending on the type of process which is involved. Tree-like processes represent the classical Darwinian framework of descent with independent modification between lineages, like divergence, and drift. Introgressive processes represent a network model of evolution in which lineages can influence each other after divergence, be it lateral transfer and borrowing (Fig. 2:13), hybridization and creolization (Fig. 2:8), or protein assembly and word formation (Fig. 2:15). Systemic processes represent a systemic model of evolution in which the interdependence between the components of evolving objects has a direct impact on the way they change (Fig. 2:17).

### Biological methods can help to automatize the identification of homologous words

While the process of vertical descent is well established in both linguistics and evolutionary biology, it is notoriously difficult to define which words or other linguistic features are historically related across languages. Identifying words of common origin, for example, is of fundamental importance to compare diverging languages. In linguistics, the term *cognate* is used to address those words which share a common origin in which no lateral transfer occurred. So

*cognacy* is, strictly speaking, not the same as *homology* in evolutionary biology [81], although it is often used interchangeably. Just like gene trees can be used to infer species trees in biology, sets of cognate words can be used to infer the relationships between languages [61, 82]. Problematically, the identification of cognates suffers from numerous practical limits. Traditionally, cognates are identified manually in linguistics, without any help of computational methods. But since the classical approaches to cognate identification are notoriously difficult to apply, the number of words used in phylogenetic language comparison is restricted to very small parts of the lexicon which are assumed to be neutral with respect to culture and present in all languages across all times. These basic parts of the lexicon, which are supposed to change slowly, only consist of about 200 words per language [83].

The overall number of words across languages varies drastically, and it is difficult to come up with a reliable statistics. However, given that near-native abilities of second language learners for the major European languages require the knowledge of about 4,000 to 5,000 words [84], it is obvious that cognate sets in computational applications cover an extremely restricted set of words. Despite this extreme restriction, only a fragment of the 7,000 languages spoken today have been thoroughly investigated. Given a large and increasing amount of digitally available data, the discipline can no longer be handled by manual inspection alone.

In evolutionary biology, the problem of identifying processes of vertical transmission in large amounts of data has given rise to a large collection of methods to deal with homolog identification. Some of these methods have already been successfully adapted to linguistic needs [50], thereby showing to biologists that their methods have an even larger application range than assumed by those who originally designed them. In order to enhance these methods further, *sequence similarity networks* could turn out to be very fruitful for historical linguistics (see Fig. 3). In biology, they can be used to identify highly divergent gene families [85]. When adapting the biological similarity scores used in sequence similarity network approaches to linguistic needs, similarity graphs could be used to search for highly diverse cognate sets across languages, and, potentially, even language families, expanding recent automatic approaches to search for deeper relationships among the more than 400 identified language families of the world [86].

### Incomplete lineage sorting as an introgression-free explanans for mosaic cognate patterns

Polymorphisms can create mosaic patterns of homologous genes, but also of cognate words. In linguistics, they may occur on various levels, depending on the data which is used to model language evolution (see Fig. 4). Mosaic patterns can be tentatively explained by introgression (concrete borrowings or language contact in general). In biology, however, another, introgression-free explanans is also commonly considered. This alternative explanans is *incomplete lineage sorting* (ILS, Fig. 2:5). In this process, ancestral polymorphisms are not fully resolved into lineages when rapid divergence occurs ([87], p. 351). ILS was, for example, used to account for the fact that 30 % of the human genes appear more similar to their homologs in Gorilla than to their homologs in Chimpanzee [88]. In the scholarly tradition of historical linguistics, there is no term that might serve as a counterpart. The process, however, is well-known, and was inherently already addressed when linguists like Johannes Schmidt (1843 – 1901) and Hugo Schuchardt (1842 – 1927) refuted Schleicher's family tree theory of language divergence right after it was proposed [89–91]. As shown in Fig. 4, there are various sources for polymorphisms in language evolution. If polymorphisms created from word formation (see below) or lexical replacement are resolved after rapid divergence of the languages, ILS creates patterns quite similar to those observed with genetic alleles in biology. Importantly, phylogenetic methods in biology [92, 93] allow one to reconstruct a lineage tree (i.e. a species tree) taking

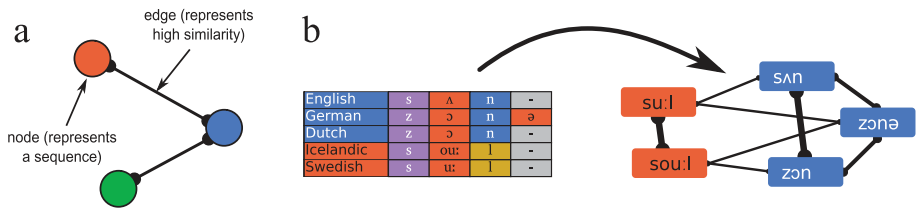



**Fig. 3** Sequence and word similarity networks. **a** In sequence similarity networks, sequences and similarities between sequences are represented in a network. Sequences are represented as nodes, and similarities between sequences are represented as edges if they exceed a certain threshold. Since evolutionary processes leave certain traces in the topology of these networks, they can be identified by applying standard network techniques. **b** Since words can be modeled as sequences of sounds, it is straightforward to create networks which represent the similarity among words. Due to the peculiarities of language evolution, however, similarity measures need to be specifically adapted to linguistic needs. As in biology, linguists start from alignments, as illustrated for words meaning 'sun' in five Germanic languages, but specific scoring functions are used

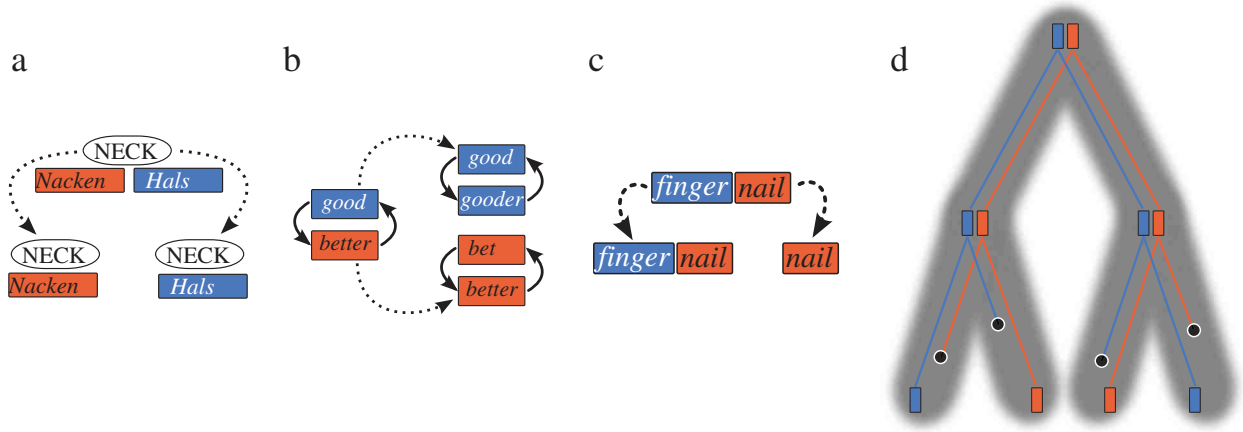



**Fig. 4** Polymorphisms in language evolution. **a** Synonymy: languages have many nearly synonymous words (German *Hals* and *Nacken* both mean 'neck' in English). They can be interchangeably used to express one and the same concept. Near synonymy is often resolved by dropping one of the two words. **b** Analogy: languages with complex morphology (case systems, etc.) often have irregular paradigms which consist of different stems (like *good*, *better*, *best* in English). These paradigms are often resolved retaining only one form and adapting the other forms to this model (e.g., *good*, *gooder*, *goodest*). **b** Derivation: words can be slightly modified by adding affixes (*word derivation*) or merging to words with each other (*compounding*). Often, both the modified or merged forms can still be interchangeably used with the original forms. They can also replace the original forms. **d** Incomplete lineage sorting: if rapid divergence occurs before the polymorphisms are resolved, they may yield patterns that seem to be in contradiction with tree-like divergence

ILS into account. Considering the ILS process and the associated methods could thus directly benefit linguistics. The Indo-European language family is a prominent example. Although the eight main branches of Indo-European are well established, and even the system of the proto-language is rather well understood, scholars have huge problems in determining the exact branching order of the eight groups. In the light of ILS, this may be less surprising. Recent studies on ancient genome-wide data of ancestral Europeans point to a rapid expansion of Indo-European languages in prehistorical times [94]. A careful investigation of the effects of ILS on language data may bring supporting evidence from linguistics.

### Network approaches shed light on introgressive processes in language evolution

In addition to improving the explanation of the complexity produced when intellectual objects of linguistics undergo tree-like evolutionary processes (such as vertical descent or ILS), PBA could also help linguists in their struggles for handling introgressive processes. Introgressive processes are a constitutive part of language evolution. Borrowing of words, the PBA of lateral gene transfer [49–51] (Fig. 2:13), is very frequent and may effect more than 40 % of the stable parts of a language's lexicon [95]. For the task of automatic borrowing identification in linguistic data, sequence similarity networks could again be useful. In biology they are increasingly used to study lateral gene transfer [96–98] and they could be employed in a similar fashion in historical linguistics, as illustrated in Fig. 5a.

Introgressive processes in language evolution are not restricted to processes like borrowing, in which two or more languages interact, but they can also occur in one and the same language. Words are often created from smaller meaningful units from the same language (morphemes) via processes of word formation [11]. Word formation can be roughly divided into two processes: derivation and compounding [99]. While compounding creates new words by merging existing ones, derivation uses affixes which cannot be used in isolation but only when being attached to other words (compare, e.g., the *-ness* in English *sick-ness*). Word derivation and word compounding result in the emergence of word families, that is, groups of words which are cognate within one and the same language. Word families play an important role in lexical organization: by decomposing words into smaller meaningful units (morphemes), speakers can quickly induce the meaning of words, even if they hear them the first time. As a result, speakers can understand between one and three times as many words as they know [100]. The size of word families can vary drastically, be it within one and the same or across several languages. The 60,000 words of the standard lexicon of German, for example, can be assigned to 8,000 word families comprising between 1 and 500 words [102].

The immediate consequence of word families is that cognate words across different languages are not necessarily completely cognate but may often exhibit different degrees of partial cognacy [81]. In Mandarin Chinese, for example, the regular word for 'moon,' *yuè liàng*, consists

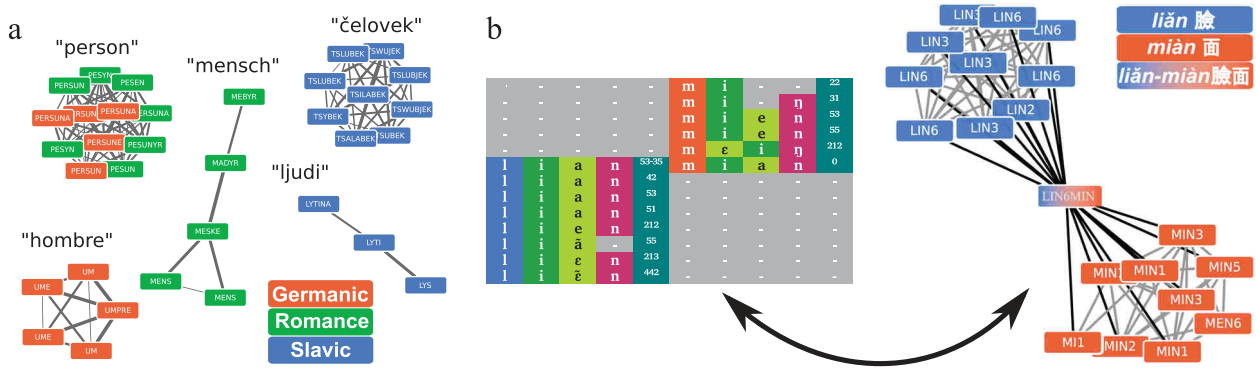



**Fig. 5** Similarity networks applied to linguistic data. **a** Similarity networks are reconstructed from global alignments for words meaning 'person' in Germanic, Romance, and Slavic languages (data taken from [101]). Five large connected components are identified. While three of them are homogeneous regarding the language family and show true cognate sets common in the respective branch of Indo-European, the top-left cluster contains words from all three branches. This cluster mainly shows Romance reflexes of Latin persona 'person'. Slavic and Germanic words occurring in this cluster are all borrowed. **b** Similarity networks are reconstructed from local alignments for dialect words meaning 'face' in 20 Chinese dialect varieties (data taken from [132]). The data contains three variants, two simple words *liăn* and *mián*, two words of different origin, and one fused form *liăn-mián*. Numbers in the alignment reflect tone patterns, which are characteristic for South-East Asian languages. Edges colored in black differ in their local and global alignments, edges colored in gray show identical alignments for local and global analyses. The fused form serves as a hub connecting the two components. Data and code to reproduce the networks is available from the data and material accompanying this article (Additional file 1)

of two morphemes, the first one originally meaning 'moon' in isolation, and the second one meaning 'shine' in isolation. In combination, they now mean simply 'moon'. In Cantonese, the Chinese variety spoken in Hongkong, the regular word for 'moon' is *jyut*$^{6}$ *gwong*$^{1}$, with the first morpheme being cognate with Mandarin *yuè*, but the second element, which means 'light' in isolation, being not cognate with the second element in Mandarin. Although methods for automatic cognate detection have been substantially improved over the last years [55, 103], none of the methods proposed so far is able to handle partial cognates across different languages. Word formation, especially word compounding, however, is very productive in many languages, especially in South-East Asian language families like Sino-Tibetan, Austro-Asiatic, Hmong-Mien, and Tai-Kadai ([104], pp. 62–67) which constitute more than 10 % of the worlds languages [105]. Compounding is not restricted to specific realms of the lexicon but also affects the core vocabulary of languages which is used in phylogenetic approaches. In the Chinese dialects, for example, about 50 % of all nouns and more than 30 % of all words in basic vocabulary are derived from fusion or derivation [106]. In biology, sequence similarity networks have been used to detect composite genes [107]. In a similar manner, word similarity networks could be used to automatically identify compound words, as illustrated in Fig. 5b. In a recent pilot study, it is further shown how a careful adaptation of similarity networks to linguistic needs allows to identify partial homologies (as the one between the Mandarin and Cantonese words for 'moon' shown above) with a high accuracy [106].

**Towards a new linguistics of proteins**

In 2006, Mario Gimona proposed an analogy between the structure of proteins and the syntax of languages, necessitated by the higher complexity of "protein grammar" compared to "DNA grammar" [57]. This idea has been sporadically followed up in the biological literature, where the generation of new functions via the combination of different protein domains in biology is compared with the new meaning that languages produce by combining different words to new sentences [108]. The syntax of a language is usually understood as the set of rules needed to combine words to phrases and sentences which native speakers accept as well-formed examples which are "grammatically correct". However, in linguistics, rule systems by which a set of elements are composed to create elements of a higher order are not restricted to syntax alone, but occur at various levels of organization [109]. There are phonotactic rules that handle the composition of sounds to form well-formed morphemes, there are morphological rules by which morphemes can be combined to form words, and there are even specific rules by which sentences can be combined to form texts [110]. If we take grammar as the cover term for any system of rules which transforms a set of symbols into a sequence of a higher order and function, the question for a grammar of proteins is where to draw the analogy with human languages exactly? Here, we think that a PBA between the process of word formation and the assembly of proteins [111], will be much more fruitful for evolutionary biology than the analogy between syntax and protein structure (see Fig. 6). While the syntax of human languages is

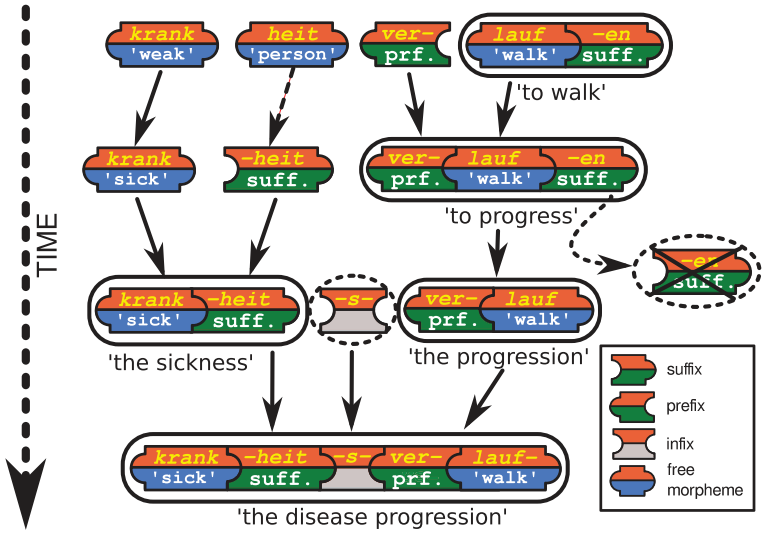



**Fig. 6** Word formation processes in the German language. Word derivation and word compounding are the major processes underlying word formation. Word derivation involves the combination of bound morphemes (suffixes, prefixes, and infixes) with free morphemes (regular words of a language). The graphic shows how the German *Krankheitsverlauf* 'disease progression' has been created in multiple stages by which the adjective krank 'sick' was nominalized with help of the suffix *-heit* and later compounded with the nominal form of the complex verb *verlaufen* 'to progress'. Note that free morphemes may easily turn into suffixes during language change

extremely productive, being capable of creating virtually unlimited numbers of different sentences, the rules underlying word formation are much more restricted. Similar to protein evolution, only a small number of the theoretically possible words is ever realized in a language. Similar to proteins, the words which are realized can also be thought to form a single network of interrelated sequences [112]. A recent study on word formation in English and German further shows that the distribution of morphemes across words resembles the distribution of domains across proteins [113]. Although many aspects still require further research, major processes of word formation are well understood and have been investigated from multiple perspectives, including evolutionary [114] and cognitive aspects [115]. Especially automatic approaches to the unsupervised detection of morphemes date back to the 1950s [116], and many different methods have been proposed over the last decade [117–119]. A closer interdisciplinary exchange between biologists and linguists during which similarities and differences between the processes are identified might inspire new methods and models in *both* biology and linguistics. In biology, first attempts have been made to employ standard methods for natural language processing to study protein domain promiscuity [120, 121]. As these attempts were based on methods originally designed to analyze syntax in natural languages, shifting the methodological transfer to methods designed to analyze word formation might provide biologist with fresh and unexpected insights.

### Invoking a system-perspective to demystify the mysteries of language drift

Almost 100 years ago, Edward Sapir (1884–1939) made the strange observation that language change may produce strikingly similar phases after the divergence of lineages, independent of areal contact or environmental influence [122, 123]. Sapir called this phenomenon of convergence, seemingly conditioned only by common ancestry, drift. Up to today, a more thorough investigation of the phenomenon is lacking, and many linguists even discard it as a mystical observation [124]. If we look at the evolution of systems, that is, the evolution of interdependencies between components of evolving objects as yet another common process in biology and linguistics (Fig. 2:17), we find a possible explanans for this specific kind of language change. Evolutionary biologists distinguish two classes of interdependencies, depending whether they evolved neutrally (as in presuppression) or as a result of some selection. Typically, the evolution of several complex macromolecular machineries (such as the ribosomes or the splicesomes, [125] could be explained by a neutral increase of interdependencies between their elemental components, while convergences in regulatory networks (i.e. the fact that some patterns are more frequent than by chance, such as the feed forward loops in transcription networks) can be explained by considerations on the structure of these networks, e.g. the fact that sets of dependencies between elements stabilize or destabilize the function of the collective system that these elements form [71].

From a linguistic perspective, the use of the systemic perspective as an explanans for linguistic phenomena is by no means new. The structuralist movement, originally initiated by Ferdinand de Saussure (1857–1913) and later popularized by Roman Jakobson (1896–1982) was systemic in its core, assuming that 'each system necessarily manifests as evolution, while, on the other hand, evolution necessarily bears systemic character' ([126], p. 68). In historical linguistics, there is a large amount of literature on system-driven processes of language change. These include work on grammaticalization [127], direction in language change [128], and interaction between the varieties of one given language [129]. Likewise, it might be useful to consider ratchet-like (irreversible) processes which would affect linguistic systems in specific states, just as processes of constructive neutral evolution are assumed to affect biological systems [130]. The common change of languages which once diverged from a common ancestor is thus no longer mystical, but simply a consequence of the interdependencies which they inherited from their ancestor. It is more than likely that the many components of languages present interdependencies affecting their stability and rates of changes. For example, a recent use of sequence similarity networks on phoneme diversity across Chinese dialects revealed that phoneme diversity correlates with the grammatical classes to which these words belong [131]. Hence the internal grammatical structure of languages certainly affects their evolution. Unfortunately, the majority of investigations on interdependencies in linguistics is neither formalized nor quantified. investigations on interdependencies in linguistics is neither formalized nor quantified.

## Conclusion

We reported unities and disunities between evolutionary processes in historical linguistics and evolutionary biology. Common processes encourage the transfer of methods that had not been proposed earlier. The successful methodological transfer between the disciplines in the past encourages us to systematize the efforts of unification while at the same time being careful to not exaggerate the degree of similarity. Given the strong influence of biological approaches to quantitative research in historical linguistics in the past, the still low degree of quantification in historical linguistic research, and the new analogies which we proposed in this paper, it is clear that biologists may have an important role to play, given that their methods have a wider scope than anticipated earlier. On the other hand (following Schleicher's idea proposed in 1863 [33]), given the amount and the subtlety of available historical documentation about the evolutionary processes that triggered linguistic diversity on earth, linguistic data could serve as an additional litmus test for the accuracy of biological methods, and biologists could profit from this advantage in detailed documentation.

In concrete terms, we showed, how biological methods can help to automatize the identification of homologous words in linguistics, how incomplete lineage sorting may serve as an introgression-free explanans for mosaic cognate patterns, and how similarity networks can be used to shed light on introgressive processes in language evolution. Furthermore, by refining the analogy of protein grammar, as a process-based analogy between the processes of protein assembly in biology and word formation in linguistics, both fields could profit from an interdisciplinary exchange and a deeper discussion of similarities and differences between the processes underlying the grammar of proteins and the processes underlying the grammar of words. The increasingly recognized need to account for the systemic dimension of evolution will likely prompt further unification across these fields and further interdisciplinary transfers. In the context of the theory of constructive neutral evolution, it may, furthermore, offer the long missing explanation for the mystical theory of parallel drift in the evolution of diverging languages.

Recalling that – apart from new analogies between evolutionary processes – we also identified processes which are specific to either biology or linguistics, it is important to keep in mind that the use of analogies should always be handled with great care. Not all evolutionary processes accounted for in one discipline necessarily need to have counterparts in other evolutionary disciplines, even if it is possible that future research will add process-based analogies where we failed to identify them. General evolution cannot be studied from within one discipline alone. Although unifying strategies can be fruitful, evolutionary explanations will remain fundamentally *pluralistic* since there is no reason to assume that all processes are common between biology and linguistics. In order to get a full picture of evolution, biologists and linguists need to complement their studies, trying to identify cross-disciplinary and discipline-specific evolutionary processes. If we want to understand how evolution triggered the diversity of substantial and intellectual objects on earth, we need to consider at least these two sister-disciplines.

## Reviewer's comments

*We are very grateful to the reviewers for taking all the time to critically read our manuscript and to comment on it in their reviews.*

### Reviewer's report 1: W. Ford Doolittle, Dalhousie University, Canada

I confess that I put off reviewing this because I feared that I would not understand it, or else would find it unoriginal: how could there be anything new to say about the similarities between historical linguistics and molecular

phylogenetics? But I was wrong: I understand much of the paper and do think it says some important new things.

Basically what the authors propose is that we get even more serious about looking at the cross-applicability of methods and concepts being developed in linguistics and phylogenetics, particularly as these latter focus on evolutionary processes – rather than on the entities that evolve (words and proteins) – and also pay attention to the constraints that give direction to such processes such as syntax and molecular coevolution. Equally useful will be identification of processes that do not appear to be analogous between the domains. The authors suggest sound change, semantic change and purification as purely linguistic processes (the latter involving intent), and asexual/sexual reproduction and natural selection as purely biological.

It would be fun to argue about selection. The authors admit that there might be "cultural selection" (based on "egocentric"? or "content"? bias – see authors' citation 70 [80]) that affect acceptance of certain elements within a language. Might it not also be that certain languages as systems are more likely to persist than others, either because of their ease of transmission (surely some languages are easier to learn than others) or affect on their speakers (surely language structure affects cultural "evolvability" somehow and unwritten languages have obvious limitations)? It may also be that in conceptualizing linguistic natural selection we should accept that evolution by natural selection can result from differential persistence as well as differential reproduction. Frédéric Bouchard (with whom the senior author has worked) has extensively developed this concept for biological evolution.

Authors make a number of observations which seem (to me, in my linguistic ignorance) novel, and well worthy of pursuit. For instance, applying models of incomplete lineage sorting (of alleles) to data in rapidly diverging languages seems a good idea, as does analogizing "the process of word formation in linguistics and protein assembly in biology". It would be good to hear more about this and about using networks to identify composite words, as the senior author has already done for proteins (see their reference 94). It is also amusing that the numbers here are so close. Authors claim that there are about 200 universally conserved "basic parts of the lexicon", and that second language learners need only master 4,000 – 5,000 words. There are maybe 200 universally conserved genes among all genomes, and the average prokaryotic genome has about 5,000 genes!.

Authors show a curious reticence to go all the way in analogizing language and genome evolution. They consider languages to be special since they are 'products of the human mind' and note that "If there was no speaker of the English language, a book containing Shakepeare's Hamlet would just be a collection of paper with ink blots". Actually, probably not. Surely clever Mandarin- (or even Martian-) speaking cryptographers could make some sense of the blots. And anyway, it's analogously true that the sequence of bases in the human genome would only be just a sequence of bases without all the evolved machinery of gene expression and environmentally-affected epigenetic baggage, as opponents of genetic reductionism correctly but so tediously insist.

Authors' response: *We thank the reviewer a lot for the summary. We are glad that despite the initial reservations of the reviewer our manuscript turned out to be comprehensible enough, also for those who are not experts in the field of linguistics. The reviewer mentions that it would 'be fun to argue about selection' in the linguistic domain, pointing to the possibility that persistence of languages is linked to the 'ease of transmission' or 'affect on [...] speakers'. Although in preparing the manuscript, we talked a lot about this issue in our interdisciplinary team, we decided to cut it short in the paper, given not only the difficulty to exhaustively grasp the forces at work in language evolution but also due to the heat with which the topic is discussed in linguistics. We refined the relevant passage by adding some further reasons why we are still careful in drawing the analogy, concluding, that in order to be able to assess the various factors triggering "cultural selection" more substantially, much more research is required in the future. Nevertheless, we agree with the reviewer that it would be very interesting to follow up these questions in more detail and we hope that our paper encourages researchers from different disciplines to increase their interdisciplinary work, looking for solutions to this and other problems related to language evolution. We have slightly modified the relevant passage in the main manuscript, trying to take the reviewer's suggestions more closely into account.*

*Regarding the proposed process-based analogy between word and protein compounding, the reviewer further mentions that it 'would be good to hear more about this and about using networks to identify composite words, as the senior author has already done for proteins'* [107]*. As a matter of fact, we have, while waiting for the reviews of this manuscript, managed to carry out some more detailed pilot studies along these lines, and a manuscript with the title 'Using sequence similarity networks to identify partial cognates in multilingual wordlists' has been accepted for publication in the "Proceedings of the Association of Computational Linguistics 2016 (Short Papers)". In this study, which would have gone beyond the scope of the current paper, we show how a careful adaptation of sequence similarity networks to linguistic needs allows us to identify partial homologies in linguistic datasets with a high accuracy* [106]*. We have now modified the manuscript in such a way that we directly mention this study along with a brief example, thus showing that similarity networks can indeed*

*successfully be used to detect homologies across compound words in different languages.*

*As a final point, the reviewer mentions, with a certain regret, that we 'show a curious reticence to go all the way in analogizing language and genome evolution,' which is definitely correct, but not necessarily since we 'consider languages to be special,' but more since our experience with parallels proposed between the two fields in the past has led us to be rather cautious. In earlier work on the development of the family tree model in the discipline of linguistics, in which the first author was involved* [91]*, it could be shown that – in contrast to the conviction of many scholars – it was an independent development in both disciplines, evoked by the emerging paradigm of uniformitarianism that triggered the development of the tree model rather than interdisciplinary transfer. One could thus argue that – if only the processes are strikingly similar – scholars may sooner or later come up with similar ways to handle them, with or without analogies drawn between disciplines. On the other hand, many of the analogies that were proposed so far, be it the one between languages and organisms by August Schleicher that was mentioned earlier in the manuscript, or the recent one between sounds in languages and nuclein bases in biology, turned out to be disappointing, unfruitful, and at times even completely wrong. While holding back ourselves, we hope, nevertheless, that our idea to start from common processes when searching for potentially fruitful analogies will offer us and our colleagues a tool to channel future methodological transfer across different disciplines. Furthermore, the reviewer has convinced us that our statement that Shakespeare's work would ink blots on paper if there were no speakers of the English language to read it was essentially ill-chosen, not serving the point we wanted to underline, namely, the fact that the medium in which the research objects are realized differs largely in biology and linguistics, and that – in contrast to biology – the aspect of transmission via learning represents a different process of replication and manifestation. We therefore deleted the sentence from the manuscript.*

**Reviewer's report 2: Eugene V. Koonin, NCBI, NLM, NIH, USA**

***Reviewer summary***

The article by List and colleagues draws multiple analogies between evolutionary processes in biology and linguistics. To me, all, rather numerous articles and a few books that I have read on comparisons between biology and linguistics share the same, rather regrettable aspect: they seem very attractive and enticing to begin with but then, disappoint rather sorely. Regrettably, the present article is no exception. Quite frankly, I find that the title of the paper [original title: "Explaining evolution in biology and linguistics using common processes", note by the authors] is a misnomer: nothing is explained here neither in biological evolution nor in the evolution of languages.

I agree that the 'process-based analogy' touted by the authors makes more sense than the (apparently, more traditional) object-based analogy. I can also accept that there is substantial ILS in linguistic evolution and that there is some logic in the analogy between protein folding and word formation. The problem is that, as a student of biological evolution, I cannot formulate the new perspectives or ideas that I get from this article. Sadly, I think that I learned nothing truly new and substantial except for some details on the history of evolutionary linguistics and the interactions between linguists and biologists, in particular Schleicher and Haeckel (these historical details are fascinating). I cannot rule out that linguists do get something fresh out of this but the article has been submitted to a biology journal, so one could expect there to be something biologically relevant and perhaps interesting.

Authors' response: *We thank the reviewer very much for his critical review. First, we agree that the title may have been ill-chosen and changed it accordingly in order to reflect more clearly the scope and content of the manuscript. The new title "Unity and disunity in evolutionary sciences: Process-based analogies open research avenues for biologists and linguists" hopefully gives a much clearer emphasis on what we wanted to discuss in the paper, namely that we face common and distinct processes in the evolutionary sciences, and that a focus on common processes rather than similarities in objects might help better in identifying fruitful analogies between disciplines which may eventually open new possibilities for future research.*

*Second, regarding the reviewer's disappointment that while showing potentially interesting possibilities of methodological transfer from biology to linguistics, we do not offer 'something biologically relevant and perhaps interesting,' we think it is important to emphasize that the scope of this paper regards evolution in general. What we want to show is that neither linguistic nor biological evolution are reducible to one another, even at the level of their processes. Therefore, understanding evolution requires (at least) these two complementary fields, which means that the lessons from biological evolution (and from historical linguistics) will never be self-sufficient to account for what an evolutionist ultimately cares for: evolutionary diversity. As biologists, we are compelled to work closer with linguists if we want to learn about aspects of evolution that are simply – and will otherwise remain – foreign to us. That is one lesson: our biological models are incomplete to account for evolution in general, so it would be not only unfortunate but also wrong-headed to forget about linguistic evolution in our accounts of the history of life. Biology Direct could almost have a section for issues related to evolution in general. As for the linguistic perspective, we have shown that in addition to the biological methods for phylogenetic reconstruction which are now regularly applied by historical*

*linguists, there are many more potentially fruitful analogies which could give rise to methodological transfer (such as lessons from incomplete lineage sorting and sequence similarity networks). So linguists should and usually do care for evolutionary biology. But even if it might not yet seem obvious why linguistics might become methodologically relevant for biologists, we should not forget that quite a few methods have already been transferred from linguistics to biology, especially from the disciplines of computational linguistics and natural language processing* [43]. *Not only classical models of formal grammar (following the hierarchy of the linguist Noam Chomsky* [40]*) are used by biologist, but also advanced models like tree adjoining grammar, which can be used for RNA structure prediction* [44]*, or inherently linguistic methods for document prediction which can be applied in protein classification* [45]*, or stochastic analyses of syntax, being applied to study protein domain promiscuity* [121]. *In order to substantiate this claim, that – despite the many disappointing examples of failed analogies – there are examples for methodological transfer in both directions which could be labelled success stories, we have added further references and elaborated the details in the text.*

*To summarize, we hope that readers will get at least two major ideas from this work: (a) it makes sense to embrace a less biology-centered perspective on evolution in evolutionary studies (that is our ignorabimus); (b) introgressive processes are fundamental to make sense of both linguistic and biological change, so a network perspective constitutes, despite the dissimilarity between both fields, the broadest and most fruitful deep commonality to achieve a form of systemic unification. There is a common core of processes between biology and linguistics, which is why evolutionary biologists and linguists should care about each other's findings. Overall, however, it is true that for all evolutionary sciences such systemic, process-based unifications will remain incomplete. Evolutionary sciences will remain pluralistic in methods and concepts, and another type of unification, i.e. operating in a piecemeal fashion and preserving the singularities of both evolutionary disciplines, will be needed to speak of evolution in general.*

#### *Reviewer recommendations to the authors*

The authors themselves notice that in the early days of genetics, and molecular genetics in particular, linguistic analogies and metaphors have been quite common. Some of these indeed became integral to the molecular biology lingo (transcription, translation), some are used much more sparingly (word, grammar), others have gone practically out of use (suffix, prefix, flexion). Regardless, though, why do these analogies do not really go beyond metaphors? Somehow it appears to me that this is not for the lack of effort on part of those interested in the linguistics-biology comparison. I feel that there is some deep disparity that precludes any substantial cross-fertilization. And here lies my major dissatisfaction with this paper. The problem is not that List et al. fail to find truly productive analogies between linguistics and biological evolutionary processes: many have tried and (at least, in my opinion) they all failed. The regrettable aspect of the paper is its rather careless but baseless optimism. I think the article would have been much improved if the authors embarked on a true critical discussion of these analogies and the reasons they do not appear to come across as genuinely fruitful.

Authors' response: *We agree with the reviewer that many largely disappointing analogies have been drawn between both disciplines, and it is for this reason that we have showed what reviewer 1 called a 'curious reticence to go all the way in analogizing language and genome evolution'. There is a deep dissimilarity between evolutionary biology and historical linguistics, even at the level of processes. There is nonetheless a possiblity of substantial cross-fertilization between both fields, especially around introgressive processes and network-like evolution, and as we can see from the application of formal grammars in biology (mentioned above) and the recent popularity of phylogenetic methods in linguistics, fruitful transfer of methods and models has already taken place in the past and in both directions. Currently, the direction of transfer goes especially from biology to linguistics, and this means that linguists import methods and concepts from biology, adapting them to their needs. Given the rapid growth of computational research in the area of natural language processing, however, it is by no means sure that the situation will always remain as this, and it might well be that even in the nearer future our proposed analogy between word compounding and protein assembly offers biologists who study linguistic approaches and patterns new insights into the phenomena in their discipline. Future will tell whether this claim is careless optimism, or whether exploiting common processes between linguistic and biological evolution will not only turn out to be fruitful but potentially also inspire cross-disciplinary research on a larger scale. But even if our optimism turns out to be unjustified, it will essentially contribute to our understanding of evolutionary processes if we can further narrow down the exact ratio of unity and disunity in the evolutionary sciences.*

*Nevertheless, we understand that we might have been exaggerating our optimism, and we have tried to trim it down to a level which is hopefully acceptable for the reviewer. First, we changed Fig. 2 to reflect more closely that the amount of common processes is presumably much smaller than the general amount of processes (we also try to indicate our own uncertainty by showing a scale with a question mark as value).*

*We also modified the manuscript in several passages to reflect justified scepticism more closely, and we also added references that further substantiate the reviewer's scepticism.*

***Minor issues***

In what sense did Watson and Crick 'detect' DNA? They did not even discover it, they built the correct structural model of DNA that allowed them to explain replication.

Authors' response: *We agree and rephrased the sentence accordingly.*

## Additional file

**Additional file 1:** The supplementary material contains the data and source code needed to reproduce the analyses to retrieve the networks shown in Fig. 5. It can be downloaded at https://zenodo.org/badge/latestdoi/5137/lingpy/process-based-analogies. (PDF 16 kb)

**Abbreviations**

ILS, incomplete lineage sorting; PBA, process-based analogies

**Acknowledgements**

We thank the two reviewers for their challenging critics and helpful advice. We are grateful to David Morrison for sharing his knowledge about early tree- and network models in biology both in personal communication with JML and in multiple blog posts.

**Funding**

EB and JSP are supported by the European Research Council under the European Community's Seventh Framework Programme, FP7/2007–2013 Grant Agreement # 615274. JML is supported by the German Research Foundation, Research Fellowship Programme, Grant # 261553824.

**Availability of data and materials**

The Additional file 1 contains the data and source code needed to reproduce the analyses to retrieve the networks shown in Fig. 5. It can be downloaded at https://zenodo.org/badge/latestdoi/5137/lingpy/process-based-analogies.

**Authors' contributions**

EB initialized the study, JML and EB set up the first draft. JSP and PL commented and revised later versions of the draft. All authors substantially contributed to the redaction of the manuscript and have given final approval of the version to be published. All authors read and approved the final manuscript.

**Authors' information**

JML is post-doctoral research fellow at UPMC (Equipe AIRE) and EHESS (CRLAO) Paris. JSP is a doctoral student in the Equipe AIRE (Adaptation, Integration, Reticulation & Evolution) at UPMC Paris. PL and EB are team leaders of the Equipe AIRE.

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

Received: 22 May 2016 Accepted: 6 August 2016
Published online: 20 August 2016

**References**

1. Popper KR. Three worlds. Tanner Lect Hum Values. 1978;143–167. http://tannerlectures.utah.edu/_documents/a-to-z/p/popper80.pdf.
2. Slingerland E, Collard M. Creating Consilience: Integrating the Sciences and the Humanities. Oxford: Oxford University Press; 2012.
3. Kirby S. The role of I-language in diachronic adaptation. Z Sprachwiss. 2000;18(2):212–25.
4. Schleicher A. Die ersten Spaltungen des indogermanischen Urvolkes [The first splits of the Indo-European prehistoric people]. Allg Monatsschr Wiss Lit. 1853;3:786–7.
5. Darwin C. On the Origin of Species by Means of Natural Selection. London: John Murray; 1859.
6. Schottel JG. Ausführliche Arbeit Von der Teutschen HaubtSprache [Exhaustive Examination of the German Main Language]. Braunschweig: Christoff Friederich Zilligern; 1663.
7. Stiernhielm G. De linguarum origine Præfatio [On the origin of languages] In: Stiernhielm G, editor. D,N Jesu Christi SS. Evangelia Ab Ulfila [The Gospels by Wulfila]. Stockholm: Typis Nicolai Wankif; 1671.
8. Gallet F. Arbre Généalogique des langues mortes et vivantes. Illustration; ca. 1800. http://gallica.bnf.fr/ark:/12148/bpt6k8546015.
9. Hickes G. Institutiones Grammaticae Anglo-Saxonicae et Moeso-Gothicae [Lectures on Anglo-Saxon and Moeso-Gothic grammar]. Oxoniæ: E Theatro Sheldoniano; 1689.
10. Sutrop U. Estonian traces in the tree of life concept and in the language family tree theory. J Estonian Finno-Ugric Linguist. 2012;3:297–326.
11. Zeige LE. Word forms, classification and family trees of languages. Why morphology is crucial for linguistics. Zool Anz – J Comp Zool. 2015;256:42–53.
12. Leclerc de Buffon GL, Vol. 5. Histoire Naturelle Générale et Particulière [General and Specific Natural history]. Paris: Imprimerie Royale; 1755.
13. Rühling JP. Ordines Naturales Plantarum Commentatio Botanica [Botanical Commentary on the Natural Order of Plants]. Goettingae: Abrah. Vandenhoeck; 1774.
14. Ragan M. Trees and networks before and after darwin. Biol Direct. 2009;4(1):43.
15. Čelakovský FL. Čtení O Srovnavací Mluvnici Slovanské [Lectures on Comparative Slavic grammar]. Prague: V komisí u F. Řivnáče; 1853.
16. Darwin C. Notebook on Transmutation of Species; 1837. http://darwinonline.org.uk/content/frameset?viewtype=side&itemID=CULDAR121.-&pageseq=38.
17. Lamarck JB, Vol. 2. Philosophie Zoologique [Philosophy of Zoology]. Paris: Dentu; 1809.
18. Morrison DA. Genealogies: Pedigrees and phylogenies are reticulating networks not just divergent trees. Evol Biol. 2016. doi:10.1007/s11692-016-9376-5.
19. Brugmann K, Vol. 1. Einleitung und Lautlehre: Vergleichende Laut-, Stammbildungs- und Flexionslehre der Indogermanischen Sprachen [Introduction and Phonetics. Comparative Studies of Sound Systems, Stem Formations, and Inflexion Systems of Indo-European Languages], GrundriSS der vergleichenden Grammatik der indogermanischen Sprachen [Foundations of the comparative grammar of the Indo-European languages]. Berlin, Leipzig: Walter de Gruyter; 1886.
20. Hennig W. Grundzüge Einer Theorie der Phylogenetischen Systematik [Foundations of a Theory of Phylogenetic Systematics]. Berlin: Deutscher Zentralverlag; 1950.
21. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull. 1958;28:1409–38.
22. Hymes DH. Lexicostatistics so far. Curr Anthropol. 1960;1(1):3–44.
23. Dixon RB, Kroeber AL. Linguistic Families of California. Berkeley: University of California Press; 1919.
24. Kay M. The Logic of Cognate Recognition in Historical Linguistics. Santa Monica: The RAND Corporation; 1964.
25. Haas MR. The Prehistory of Languages. The Hague and Paris: Mouton; 1969.
26. Needleman SB, Wunsch CD. A gene method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
27. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;1:195–7.
28. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25(4):351–60.
29. Lyell C, Vol. 1. Principles of Geology, Being an Attempt to Explain the Former Changes of the Earth's Surface, by Reference to Causes Now in Operation. London: John Murray; 1830.

30. Christy C. Uniformitarianism in nineteenth century linguistics: Implications for a reassessment of the neogrammarian sound-law doctrine In: Koerner EFK, editor. Progress in Linguistic Historiography. Amsterdam: Benjamins; 1980. p. 249–56.
31. Wells RS. The life and growth of language: Metaphors in biology and linguistics In: Hoenigswald HM, editor. Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective. Philadelphia: University of Pennsylvania Press; 1987. p. 39–80.
32. Croft W. Typology and Universals. Cambridge: Cambridge University Press; 1990.
33. Schleicher A. Die Darwinsche Theorie und die Sprachwissenschaft [The Darwinian Theory and the Science of Languages]. Weimar: Hermann Böhlau; 1863.
34. Hoenigswald HM. On the history of the comparative method. Anthropol Linguist. 1963;5(1):1–11.
35. Watson JD, Crick FHC. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737–8.
36. Gamov G. Possible relation between deoxyribonucleic acid and protein structures. Nature. 1954;173:318.
37. Crick F. The present position of the coding problem. Brookhaven Symp Biol. 1959;12:35–9.
38. Bralley P. An introduction to molecular linguistics. BioScience. 1996;46(2):146–53.
39. Shanon B. The genetic code and human language. Synthese. 1978;39(3): 401–15.
40. Chomsky N. On certain formal properties of grammars. Inform Control. 1959;2:137–67.
41. Searls DB. Linguistic approaches to biological sequences. CABIOS. 1997;13(4):333–44.
42. Durbin R, Eddy SR, Krogh A, Mitchinson G. Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press; 1998.
43. Searls DB. Trees of life and of language. Nature. 2003;426(6965):391–2.
44. Uemura Y, Hasegawa A, Kobayashi S, Yokomori T. Tree adjoining grammars for RNA structure prediction. Theor Comput Sci. 1999;210(2): 277–303.
45. Cheng BYM, Carbonell JG, Klein-Seetharaman J. Protein classification based on text document classification techniques. Proteins: Struct Funct Bioinf. 2005;58(4):955–70.
46. Gray RD, Atkinson QD. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 2003;426(6965):435–9.
47. Ringe D, Warnow T, Taylor A. Indo-European and computational cladistics. T Philol Soc. 2002;100(1):59–129.
48. Nakhleh L, Ringe D, Warnow T. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. Language. 2005;81(2):382–420.
49. Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T. Networks uncover hidden lexical borrowing in Indo-European language evolution. Proc R Soc London, Ser B. 2011;278(1713):1794–803.
50. List JM, Nelson-Sathi S, Geisler H, Martin W. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. Bioessays. 2014;36(2):141–50.
51. List JM, Nelson-Sathi S, Martin W, Geisler H. Using phylogenetic networks to model Chinese dialect history. Lang Dyn Change. 2014;4(2): 222–52.
52. List JM. Network perspectives on Chinese dialect history. Bull Chin Linguist. 2015;8:42–67.
53. Kondrak G. Algorithms for language reconstruction. Toronto: Dissertation, University of Toronto; 2002.
54. Prokić J, Wieling M, Nerbonne J. Multiple sequence alignments in linguistics. In: Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education. Stroudsberg: Association of Computational Linguistics; 2009. p. 18–25.
55. List JM. Sequence Comparison in Historical Linguistics. Düsseldorf: Düsseldorf University Press; 2014.
56. Steiner L, Stadler PF, Cysouw M. A pipeline for computational historical linguistics. Lang Dyn Change. 2011;1(1):89–127.
57. Gimona M. Protein linguistics – a grammar for modular protein assembly? Nat Rev Mol Cell Biol. 2006;7(1):68–73.
58. Von Bertalanffy L. The history and status of general systems theory. Acad Manag J. 1972;15(4):407–26.
59. Percival K. Biological analogy in the study of languages before the advent of comparative grammar In: Hoenigswald HM, editor. Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective. Philadelphia: University of Pennsylvania Press; 1987. p. 3–38.
60. Schleicher A. Zur Vergleichenden Sprachengeschichte [On Comparative Language History]. Bonn: König; 1848.
61. Pagel M. Human language as a culturally transmitted replicator. Nat Rev Genet. 2009;10:405–15.
62. van Driem G. Language as organism: A brief introduction to the Leiden theory of language evolution In: Lin Yc, Hsu Fm, Lee Cc, Sun JTS, Yang Hf, Ho D, editors. Studies on Sino-Tibetan Languages. Taipei: Academia Sinica; 2004. p. 1–9.
63. Mufwene SS. The Ecology of Language Evolution. Cambridge: Cambridge University Press; 2001.
64. Zwick M. Some analogies of hierarchical order in biology and linguistics In: Klir G, editor. Applied General Systems Research: Recent Developments & Trends. New York: Plenum Press; 1978. p. 521–9.
65. Enguix GB, Jiménez-López MD. Natural language and the genetic code: From the semiotic analogy to biolinguistics. In: Proceedings of the 10th World Congress of the International Association for Semiotic Studies (IASS/AIS). La Coruña: Association of Semiotic Studies; 2012. p. 771–80.
66. Jakobson R, Vol. 2. Rapports Internes et Externes du Langage [Internal and External Relations of Language]. Paris: Les Éditions de Minuit; 1973.
67. Hruschka DJ, Branford S, Smith ED, Wilkins J, Meade A, Pagel M, Bhattacharya T. Detecting regular sound changes in linguistics as events of concerted evolution. Curr Biol. 2015;25(1):1–9.
68. Atkinson QD, Gray RD. Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. Syst Biol. 2005;54(4):513–26.
69. Gentner D. Structure-mapping: A theoretical framework for analogy. Cogn Sci. 1983;7:155–70.
70. Bermúdez-Otero R. Diachronic phonology In: de Lacy P, editor. The Cambridge Handbook of Phonology. New York: Cambridge University Press; 2007. p. 497–517.
71. Allen U. Introduction to Systems Biology: Design Principles of Biological Cuircuits. London: Chapman & Hall/CRC; 2007.
72. de Saussure F. Cours de Linguistique Générale [Course on General Linguistics]. Lausanne: Payot; 1916.
73. Merrell F. The Routledge Companion to Semiotics and Linguistics In: Cobley P, editor. London and New York: Routledge; 2001. p. 28–39.
74. Mallinson G. Rumanian In: Harris M, Nigel V, editors. The Romance Languages. London and Sydney: Croom Helm; 1988. p. 391–419.
75. Everett C, Blasi DE, Roberts SG. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. Proc Nat Acad Sci USA. 2015;112(5):1322–7.
76. Lupyan G, Dale R. Language structure is partly determined by social structure. PLoS ONE. 2010;5(1):8559.
77. Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ. Rate of language evolution is affected by population size. Proc Nat Acad Sci USA. 2015;112(7):2097–102.
78. Huneman P. Titles, uses and instruction of use: The status of intention in art and artefacts. Facta Philosophica. 2007;9:3–21.
79. Ghirlanda S, Enquist M, Nakamaru M. Cultural evolution develops its own rules: The rise of conservatism and persuasion. Curr Anthropol. 2006;47(6):1027–34.
80. Tamariz M, Ellison TM, Barr DJ, Fay N. Cultural selection drives the evolution of human communication systems. Proc R Soc London, Ser B. 2014;281(1788):20140488.
81. List JM. Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction. J Lang Evol. 2016;1:. doi:10.1093/jole/lzw006.
82. Atkinson QD, Gray RD. How old is the Indo-European language family? Illumination or more moths to the flame? In: Forster P, Renfrew C, editors. Phylogenetic Methods and the Prehistory of Languages. Cambridge and Oxford and Oakville: McDonald Institute for Archaeological Research; 2006. p. 91–109.
83. Swadesh M. Lexico-statistic dating of prehistoric ethnic contacts. Proc Am Philol Soc. 1952;96(4):452–63.

84. Milton J. The development of vocabulary breadth across the CEFR levels. a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across europe In: Bartning I, Martin M, Vedder I, editors. Communicative Proficiency and Linguistic Development: Intersections Between SLA and Language Testing Research. York: Eurosla; 2010. p. 211–32.
85. 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.
86. Jäger G. Support for linguistic macrofamilies from weighted alignment. Proc Nat Acad Sci USA. 2015;112(41):12752–7.
87. Rogers J, Gibbs RA. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet. 2014;15(5):347–59.
88. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169–75.
89. Schmidt J. Die Verwantschaftsverhältnisse der Indogermanischen Sprachen [The Relations of the Indo-European Languages]. Weimar: Hermann Böhlau; 1872.
90. Schuchardt H. Über die Klassifikation der Romanischen Mundarten. 1319 Probe-Vorlesung, Gehalten zu Leipzig Am 30. April 1870 [On the 1320 Classification of Romance Dialects. Test Lecture, Held at Leipzig on April 1321 30 1870]. Graz. 1900. https://archive.org/details/berdieklassifik01schugoog.
91. Geisler H, List JM. Do languages grow on trees? the tree metaphor in the history of linguistics In: Fangerau H, Geisler H, Halling T, Martin W, editors. Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization. Stuttgart: Franz Steiner Verlag; 2013. p. 111–24.
92. Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 2006;55(1):21–30.
93. Yu Y, Dong J, Liu KJ, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. Proc Nat Acad Sci USA. 2014;111(46):16448–53.
94. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Banffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szecsenyi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature. 2015;522(7555):207–11.
95. Tadmor U. Loanwords in the world's languages In: Haspelmath M, Tadmor U, editors. Loanwords in the World's Languages. Berlin and New York: de Gruyter; 2009. p. 55–75.
96. Halary S, McInerney JO, Lopez P, Bapteste E. EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol. 2013;13:146.
97. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Nat Acad Sci USA. 2013;110(17):1594–603.
98. Bapteste E, Lopez P, Bouchard F, Baquero F, McInerney JO, Burian RM. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Nat Acad Sci USA. 2012;109(45):18266–72.
99. Booij G. The Grammar of Words. An Introduction to Linguistic Morphology. Cambridge: Cambridge University Press; 2005.
100. Nagy WE, Anderson RC. How many words are there in printed school English? Reading Res Q. 1984;19(3):304–30.
101. Wichmann S, Müller A, Wett A, Velupillai V, Bischoffberger J, Brown CH, Holman EW, Sauppe S, Molochieva Z, Brown P, Hammarström H, Belyaev O, List JM, Bakker D, Egorov D, Urban M, Mailhammer R, Carrizo A, Dryer MS, Korovina E, Beck D, Geyer H, Epps P, Grant A, Valenzuela P. The ASJP Database. Version 16. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2013.
102. Augst G. Worfamilienwörterbuch der Deutschen Gegenwartssprache [Dictionary of Word Families in Contemporary German. Tübingen: Niemeyer; 2009.
103. Bouchard-Côté A, Hall D, Griffiths TL, Klein D. Automated reconstruction of ancient languages using probabilistic models of sound change. Proc Nat Acad Sci USA. 2013;110(11):4224–9.
104. Goddard C. Languages of East and Southeast Asia. An Introduction. Oxford: Oxford University Press; 2005.
105. Hammarström H, Forkel R, Haspelmath M, Bank S. Glottolog. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2015. Version 2.7. http://glottolog.org. Accessed 16 July 2016.
106. List JM, Lopez P, Bapteste E. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In: Proceedings of the Association of Computational Linguistics 2016. Short Papers. Stroudsberg: Association of Computational Linguistics; 2016. p. 599–605.
107. Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics. 2013;29(7):837–44.
108. Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure. 2007;15(1):85–99.
109. Stark BR. The bloomfieldian model. Lingua. 1972;30:385–421.
110. de Beaugrande RA, Dressler W. Einführung in die Textlinguistik [Introduction to Text Linguistics]. Tübingen: Niemeyer; 1981.
111. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded proteins. eLife. 2015;4:e09410.
112. Smith JM. Natural selection and the concept of a protein space. Nature. 1970;225(5232):563–4.
113. Keller DB, Schultz J. Word formation is aware of morpheme family size. PLoS ONE. 2014;9(4):93978.
114. Hartmann S. The diachronic change of German nominalization patterns: An increase in prototypicality. In: Selected Papers from the 4th UK Cognitive Linguistics Conference. Lancaster: Cognitive Linguistics Association; 2014. p. 52–171.
115. Heide J, Lorenz A, Meinunger A, Burchert F. The influence of morphological structure on the processing of German prefixed verb In: Onysko A, Michel S, editors. Cognitive Perspectives on Word Formation. Berlin and New York: de Gruyter Mouton; 2010. p. 375–98.
116. Harris ZS. From phoneme to morpheme. Language. 1955;31(2):190–222.
117. Hammarström H. A naive theory of affixation and an algorithm for extraction. In: Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006. Stroudsberg: Association for Computational Linguistics; 2006. p. 79–88.
118. Grönroos SA, Virpioja S, Smit P, Kurimo M. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin and Stroudsberg: Dublin City University and Association for Computational Linguistics; 2014. p. 1177–1185.
119. Griffiths S, Purver M, Wiggins G. From phoneme to morpheme: A computational model In: Baayen H, Jäger G, Köllner M, Wahle J, Baayen-Oudshoorn A, editors. Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics. Stroudsberg: Association of Computational Linguistics; 2015.
120. Basu MK, Carmel L, Rogozin IB, Koonin EV. Evolution of protein domain promiscuity in eukaryotes. Genome Res. 2008;18(3):449–61.
121. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinforma. 2009;10(3):205–16.
122. Sapir E. Language. An Introduction to the Study of Speech. New York: Harcourt, Brace; 1921.
123. Aikhenvald AY. Semantics and pragmatics of grammatical relations in the vaups linguistic area In: Aikhenvald AY, Dixon RMW, editors. Grammars in Contact: A Cross-linguistic Typology. Explorations in linguistic typology. Oxford: Oxford University Press; 2007. p. 237–66.

124. Trask L. Trask's Historical Linguistics, 3rd ed. London and New York: Routledge; 2015.
125. Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW. How a neutral evolutionary ratchet can build cellular complexity. IUBMB Life. 2011;63(7):528–37.
126. Tynjanow J, Jakobson R. Probleme der literatur- und sprachforschung [Problems of literature and linguistic research] In: Viehoff R, editor. Alternative Traditionen [Alternative Traditions]. Braunschweig: Vieweg; 1928. p. 67–9.
127. Heine B, Kuteva T. World Lexicon of Grammaticalizatioin. Cambridge: Cambridge University Press; 2002.
128. Haspelmath M. On directionality in language change with particular reference to grammaticalization In: Fischer O, Norde M, Perridon H, editors. Up and down the Cline – The Nature of Grammaticalization. Typological Studies in Language. Amsterdam and New York: John Benjamins Publishing Company; 2004. p. 17–44.
129. Oesterreicher W. Historizität, Sprachvariation, Sprachverschiedenheit, Sprachwandel [Historicity, language variation, language difference, language change] In: Haspelmath M, editor. Language Typology and Language Universals. Berlin and New York: Walter de Gruyter; 2001. p. 1554–1595.
130. Gray MW, Lukes J, Archibald JM, Keeling PJ, Doolittle WF. Cell biology. Irremediable complexity? Science. 2010;330(6006):920–1.
131. Lopez P, List JM, Bapteste E. A preliminary case for exploratory networks in biology and linguistics: the phonetic network of Chinese words as a case-study In: Fangerau H, Geisler H, Halling T, Martin W, editors. Classification and Evolution in Biology, Linguistics and the History of Science. Concepts – Methods – Visualization. Stuttgart: Franz Steiner Verlag; 2013. p. 181–96.
132. In: Hóu J, editor. Xiàndài Hànyǔ Fāngyán Yīnkù [Phonological Database of Chinese Dialects]. Shanghai: Shànghǎi Jiàoyù; 2004.

Evolutionary biology and linguistics share a more recent tradition of using networks to analyze their respective objects. The article n°9 shows a study case of networks in linguistic context. We used rhyme networks to compare eight different Old Chinese reconstruction systems, in which linguists have tried to detect how the Old Chinese characters were pronounced. Due to the character of the Chinese writing system, which is not phonetic and gives us little evidence to infer how the words have been originally pronounced, the reconstruction of Old Chinese plays an important role, especially in the context of the higher affiliation of Chinese as a Sino-Tibetan language. For our study, we have retrieved rhyme data from the Book of Odes (an old collection of Chinese poems, dating back to the first millennium B.C.) to construct rhyme networks for eight different reconstruction systems, proposed by independent scholars. In our network, nodes represent rhyme words which are linked by an edge whenever they rhyme in the Book of Odes. Rhyming behavior of Old Chinese words plays a crucial role for the reconstruction of Old Chinese pronunciation. Rhyme patterns have been used to test Old Chinese reconstruction systems for consistency and plausibility. From the idea that rhyming in Old Chinese was following the principle of vowel purity, a tendency to disallow rhymes of words with different vowels, we developed a quantitative test using assortativity to check how “pure” each of the given reconstruction systems was with respect to the rhyme patterns. Assortativity measures the similarity between connected nodes regarding their attributes, calculated by the assortativity coefficient. We computed the assortativity coefficients of the rhyme network for all eight reconstruction systems. We have shown that we can easily design quantitative tests that check to which degree different reconstruction systems conform to a given criterion which can be considered as a valuable contribution to the field of Chinese historical linguistics. This article has been accepted and published in the journal "Lingua Sinica".

List *et al. Lingua Sinica* (2017) 3:5
DOI 10.1186/s40655-017-0021-8

Lingua Sinica

SQUIB-B: ISSUES AND NEW PERSPECTIVES Open Access

# Vowel purity and rhyme evidence in Old Chinese reconstruction

Johann-Mattis List[1,2*], Jananan Sylvestre Pathmanathan[2], Nathan W. Hill[3], Eric Bapteste[2] and Philippe Lopez[2]

* Correspondence: mattis.list@lingpy.org
[1]Centre de recherches linguistiques sur l'Asie Orientale, École des Hautes Études en Sciences Sociales, 2 Rue de Lille, 75007 Paris, France
[2]Team Adaptation, Integration, Reticulation, Evolution Université Pierre et Marie Curie, 9 Quai St Bernard, 75005 Paris, France
Full list of author information is available at the end of the article

**Abstract**

Rhyme patterns in Old Chinese poems are important for the reconstruction of Old Chinese pronunciation, as they provide evidence for groups of words which formerly had similar pronunciation. Rhyme patterns can also be used to test Old Chinese reconstruction systems for consistency and plausibility, as reconstruction systems should minimize the conflict with attested rhyme patterns. Here, we build on the idea that rhyming in Old Chinese followed the principle of vowel purity, a tendency to disallow rhymes of words with different vowels, to develop a quantitative test for reconstruction systems of Old Chinese. The test is illustrated by comparing seven different Old Chinese reconstruction systems and by showing that, although the systems differ regarding their degree of vowel purity, the principle seems to hold for Old Chinese rhyme data.

## 1 Introduction

Due to the specific morpheme-syllabic character of the Chinese writing system (Chao 1968: 121), we have considerably fewer clues regarding the original pronunciation of the oldest attested stages of the Chinese language than we do for languages which are written in alphabetic writing systems. As a result, reconstructing the pronunciation of Old Chinese constitutes a challenge in its own right, and quite a few scholars have proposed a variety of reconstructions which differ considerably from one to another (Li 李方桂 1971; Karlgren 1957; Wang 王力 1980; Pan 潘悟云 2000; Starostin 1989; Baxter 1992; Zheng Zhang 郑张尚芳 2003). Apart from the internal structure of Chinese characters, rhyme evidence plays a crucial role in the reconstruction of Old Chinese phonology (Baxter 1992). Based on the fundamental assumption that words which regularly rhyme in older stages of Chinese reflect words with similar pronunciation in their finals, we can systematically investigate Chinese poetry from coherent epochs, assigning words to classes of similar pronunciations. In classical Chinese scholarship, rhyme analysis has a long tradition, going back to scholars like Wu Yu 吳棫 (1100–1154), who was one of the first to systematically assign Chinese characters to specific rhyme classes (He 何九盈 2006: 163).

Up to the end of the 19th century, traditional Chinese rhyme analysis, which was especially devoted to 詩經 *shijing* 'the Book of Odes' (ca. 1050–600 BC), led to the identification of more than 30 distinct rhyme categories (韻部 *yunbu*, see Baxter 1992: 141–150). The classical approach to rhyme analysis, sometimes called 丝贯绳牵法 *siguan shengqian fa* 'link-and-bind method' (Geng 耿振生 2004), or 韵脚系联法

*yunjiao xilian fa* 'rhyme linking method' (Lv 吕胜男 2009) starts from the collection of words which can be shown to rhyme with each other (usually represented by one Chinese character), and then clusters these words into rhyme groups by applying a greedy strategy (Geng 耿振生 2004). This strategy searches exhaustively for connected components in a rhyme network in which rhyme words are modeled as nodes and attested rhyme instances are represented as links between the nodes (List 2017).

The most obvious drawback of the classical rhyme analysis is its *resolution power*; following the idea of connected components blindly will yield very large groups of rhymes and a very small number of distinct categories. The classical analysis favors lumping over splitting, and is furthermore vulnerable to incorrectly identified rhyme patterns and other kinds of errors in the data. The problems of the classical rhyme analysis were explicitly addressed in the Old Chinese reconstruction system of Baxter (1992), which proposed six main vowels for Old Chinese and a total of 52 distinct rhyme groups, thus drastically expanding the number of rhyme categories proposed for Old Chinese by classical scholarship. The choice of a six vowel system was further substantiated by the fact that the reconstruction systems by Sergei A. Starostin and Zheng Zhang Shangfang 郑张尚芳, proposed independently around the same time, also employed six vowels (see Starostin 1989; Zheng Zhang 郑张尚芳 2003). The proposal by Baxter (1992) was further substantiated by a statistical test which tested the likelihood of specific rhyme category groupings to have been occurred by chance. In the recently proposed new reconstruction for Old Chinese by Baxter and Sagart (2014), the rhyme schema by Baxter (1992) was only slightly modified by adding a new coda **-r* for rhyme words which rhyme both with words in coda **-n* and **-j*. This resulted in six additional rhyme categories, one for each of the six main vowels **a*, **e*, **i*, **o*, **u*, and **ə*.

## 2 Vowel purity and rhyme evidence

According to Ho (2016: 176–184), the Old Chinese reconstruction by Baxter and Sagart (2014) contradicts important rhyming principles, especially the principle of vowel purity, according to which rhymes in the Book of Odes were very strict regarding the identity of vowels, while consonant differences could easily be tolerated. According to the author, vowel purity is in conflict in many cases where pronunciations as suggested by the Old Chinese reconstruction by Baxter and Sagart point to different vowels, while the respective words frequently rhyme in the Book of Odes. The argument by Ho (2016) rests on two fundamental assumptions. First, Ho assumes that vowel purity was a key principle in Chinese rhyming. Second, Ho claims that the reconstruction system by Baxter and Sagart is in strong conflict with this principle. Unfortunately, he does not provide any concrete examples, apart from contrasting traditional rhyming categories with the more fine-grained rhyming categories as they were first proposed by Baxter (1992).

Due to the lack of external evidence for Old Chinese pronunciation, the first assumption is very difficult to check. The argument of the author itself rests uniquely on perceived rhyming tendencies in current folk traditions in China. While they may seem suggestive on first sight, they stand in strong contrast to classical rhyme traditions which evolved during the Tang dynasty (618–907) and took the prescriptions in official rhyme books for granted, as well as cross-linguistic tendencies of rhyme production, which may favor similarity in vowels, but not necessarily prescribe identity. This is, for

example, reflected in German rhyme tradition, in which words with vowels [y] and [i] freely rhyme with each other, as in *nieder* [ni:dər] 'down' and *Brüder* [bry:dər] 'brothers', see also Peust 2014: 62)[a]. Another obvious problem of vowel purity is the fact that the Book of Odes from which the rhyme categories are drawn does not reflect a coherent speech variety that was spoken at a single place and time (Baxter 1992: 343–366). On the contrary, the Book of Odes was compiled over a period of at least 400 years (from about 1000 until 600 BC, cf. Kern 2004), and scholars have long suggested that certain passages reflect dialectal rhyme patterns (Baxter and Sagart 2014: 278f). So even when disregarding the problem of overarching rhyme traditions superimposed by society, it would be rather surprising if the system of rhyming showed no stages of transitions and conflicts resulting from language change and dialectal influence.

We can illustrate this further by having a look at concrete poems in the Book of Odes. Table 1 gives Ode 10 as an example, contrasting both what scholars believe reflects the perceived rhyme structure during the time the poem was composed (column rhyme), the traditional opinion regarding the rhyme group to which the rhyme words belong (column group), as well as reconstructions in four different systems (see the table for details). As we can see from this example, stanza 1 shows an impure rhyme in two systems, contrasting the vowels [ə] and [e], namely, those of Pan Wuyun 潘悟云 (Pan 潘悟云 2000) and Wang Li 王力 (Wang 王力 1980). This impure rhyme was also recognized in traditional Chinese phonology, as the traditional rhyme groups 微 *wei* and 脂 *zhi*. The OCBS system (Baxter and Sagart 2014) and the system by Starostin (Starostin 1989) do not show this conflict, as they propose only the vowel [ə] in this group. If we compare across the following stanzas, we can see that all reconstruction systems show specific conflicts regarding the principle of vowel purity, including the traditional classification upon which Ho (2016) bases his criticism. A crucial question for Old Chinese reconstruction is to what degree one should try to avoid impure rhymes, and to what degree one should accept them as reflecting vivid poetry which does not necessarily follow strict rules. How much vowel purity do we need to assume for the Book of Odes?

We cannot directly test the importance of vowel purity for Old Chinese rhyming, as our information regarding Old Chinese vowels relies on reconstructions, and these

**Table 1** Comparing impure and pure rhymes in Ode 10 and how they are reflected in different reconstruction systems

| Text | Stanza | MCH | Pan Wuyun | OCBS | Wang Li | Starostin | Rhyme | Group |
|---|---|---|---|---|---|---|---|---|
| 遵彼汝墳,伐其條枚 | 1.AB | mwoj | mɯɯl | mˁəj | muəi | məj | A | 微 |
| 未見君子,怒如調飢 | 1.CD | tsiX | kril | Cə.kə[j] | kiei | krəj | A | 脂 |
| 遵彼汝墳,伐其條肄 | 2.AB | sijH | ph-ljuds | s-ləp-s | jiet | slhəps | B | 质 |
| 既見君子,不我遐棄 | 2.CD | khjijH | khids | [kʰ]i[t]-s | khiet | khijs | B | 质 |
| 魴魚赬尾 | 3.A | mj+jX | mɯl? | [m]əj? | miuəi | məj? | C | 微 |
| 王室如毀 | 3.B | xjweX | qhʷral? | [ɱ](r)aj? | xiuəi | hʷej? | C | 微 |
| 雖則如毀 | 3.C | xjweX | qhʷral? | [ɱ](r)aj? | xiuəi | hʷej? | C | 微 |
| 父母孔邇 | 3.D | nyeX | mljel? | n[ə][r]? | njiei | n(h)ej? | C | 脂 |

MCH refers to the Middle Chinese reading following Baxter (1992), Pan Wuyun 潘悟云 is the reconstruction following the system of Pan 潘悟云 (2000, available online at http://www.eastling.org/oc/oldage.aspx), OCBS refers to the system by Baxter and Sagart (2014), Wang Li 王力 to the system by Wang 王力 (1980), and Starostin to the system by Starostin (1989), (available online at http://starling.rinet.ru). Rhyme judgments follow Baxter (1992) and Wang 王力 (1980), and group reflects the "traditional rhyme group", the label used in traditional Chinese phonology

reconstructions may well have been proposed with the principle in mind, be it explicitly or intuitively. Whether a given reconstruction system is in strong conflict with the vowel purity principle, on the other hand, *can* be directly tested by inspecting the actual data. Given the restricted corpus of the Book of Odes, an exhaustive investigation of the conflicting cases is possible, and one could compare all Odes in the corpus in different reconstruction systems, just as we have illustrated for Ode 10 in Table 1. Such a *qualitative evaluation* has the obvious disadvantage that it would be very time-consuming, both for the experts who carry it out and for the scholars who read the reports. In order to avoid the problems resulting from manual comparisons, we propose a *quantitative test* that automatically measures the *degree* by which reconstruction systems deviate from the principle of vowel purity. By modeling Chinese rhyme data from the Book of Odes as a weighted network in which rhyme words serve as the nodes and attested rhyme occurrences in the Book of Odes are modeled as links between the rhyme words, we can not only test how well a given reconstruction system conforms to Ho's (2016) vowel purity criterion, but we can even compare alternative reconstruction systems directly with each other.

## 3 Evaluating vowel purity in reconstruction

### 3.1 Materials

#### *3.1.1 Rhyme data*

The rhyme data used for the experiment follow the rhyme assignments for the Book of Odes provided in Baxter (1992) which were digitized and converted into a machine-readable format in List (2017). The data are available online as interactive application, the *Shījīng Rhyme Browser* (http://digling.org/shijing/), where all rhyme decisions can be interactively searched and inspected in the reconstruction systems by Baxter and Sagart (2014) and Pan 潘悟云 (2000). The former is available for download; the latter was taken from the *Thesaurus Linguae Sericae* (Harbsmeier and Jiang 2009). The dataset lists all potential rhyme words in the Book of Odes, which were determined by taking the final character in each line of each stanza across the 305 poems of the Book of Odes. This list of potential rhyme words is contrasted with the actual rhyme words as assigned in Baxter (1992). The interactive application visualizes rhyme annotations by coloring words which are marked as rhyming in the same color, as shown in Table 2 for the poem number 60.

In List (2017), the rhyme data are used to construct a rhyme network of all rhyming words in the Book of Odes. In this network, rhyme words (represented by Chinese characters) are represented as nodes, and links between the nodes are drawn whenever two rhyme words actually rhyme in the Book of Odes. The whole network comprises 1845 nodes and 5266 links between the nodes. The number of recurring links between two nodes is counted and weighted, using specific weighting principles, like (a) counting formulaic (recurring lines in the collection) only once, and (b) by taking the size of the group in which two words rhyme into account when establishing the weights (in order to avoid that large groups of rhyming words are scored more often than smaller ones). As network weighting itself is not of primary importance for the approach presented in this paper, we refer the readers to List (2017) where the rhyme network construction process is described in detail. All data underlying the study are accessible

**Table 2** Example of the structure and display of rhymes of the Book of Odes in the Shījīng rhyme browser

| Section | Rhyme | OCBS | MCH | Final |
|---|---|---|---|---|
| **60.1** | | | | |
| 芄蘭之支 | 60.1.a | ke | tsye | e |
| 童子佩觿 | 60.1.a | $q^{wh}$e | xjwie | e |
| 雖則佩觿 | 60.1.a | $q^{wh}$e | xjwie | e |
| 能不我知 | 60.1.a | tre | trje | e |
| 容兮遂兮 | 60.1.b | sə-lu[t]-s | zwijH | ut |
| 垂帶悸兮 | 60.1.b | [g]$^{w}$i[t]-s | gjwijH | it |
| **60.2** | | | | |
| 芄蘭之葉 | 60.2.a | l[a]p | yep (syep) | ap |
| 童子佩韘 | 60.2.a | l̥[a]p | syep | ap |
| 雖則佩韘 | 60.2.a | l̥[a]p | syep | ap |
| 能不我甲 | 60.2.a | [k]$^{ʕ}$r[a]p | kap | ap |
| 容兮遂兮 | 60.2.b | sə-lu[t]-s | zwijH | ut |
| 垂帶悸兮 | 60.2.b | [g]$^{w}$i[t]-s | gjwijH | it |

Characters shaded in the same color inside the same stanza are judged to rhyme according to Baxter (1992), the labels used by Baxter (1992) are given in the column rhyme. Old Chinese readings (OCBS) for the full words and for the rhymes are given in the reading of Baxter and Sagart (2014). Middle Chinese readings (MCH) follow Baxter (1992)

online at https://zenodo.org/badge/latestdoi/43676744, and we used this data to create the rhyme network for our study. Figure 1 illustrates the structure of the rhyme network by showing a small part of the full graph, corresponding to the codas reconstructed as **-ar*, **-an*, and **-aj* in the reconstruction of Baxter and Sagart (2014).

### *3.1.2 Reconstruction systems*

For all 1845 rhyme words in the network, Old Chinese readings in eight different reconstruction systems were collected from different sources. The system of Baxter and Sagart (2014) is available online for download. Unfortunately, it covers only 1431 characters of the full set of 1845 rhyme words, and 414 readings are missing. The Eastling

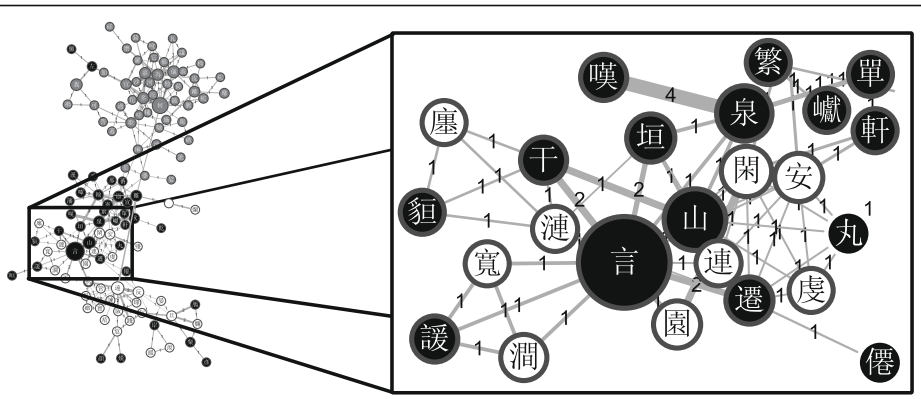



**Fig. 1** Example for a small part of the rhyme network based on the data in List (2017), for rhyme words reconstructed with coda **-ar* (*black nodes*), **-aj* (*gray nodes*), and **-an* (*white nodes*) in the reconstruction system of Baxter and Sagart (2014). *Nodes* correspond to rhyme words, and *edges* indicate whether the nodes they connect rhyme together in the Book of Odes. *Edge weights* represent the frequency of rhyme-word co-occurrences, and *node weights* represent the general frequency by which the words occur in rhyme position in the Book of Odes

project (Shanghai Normal University 上海师范大学 2016, http://www.eastling.org/oc/oldage.aspx) offers Old Chinese reconstructions for various authors, including the systems proposed in Karlgren (1957), Li (1971), Wang 王力 (1980), Zheng Zhang 郑张尚芳 (2003), and the most recent proposals according to the system of Pan 潘悟云 (2000). The Eastling data has a broad coverage, and only 15 out of 1845 readings in the original rhyme data from list (in press) were missing in this collection, thus comprising a total of 1830 readings for each of the five different reconstruction systems. In order to make sure that these different systems are reflected correctly, we compared the Eastling data with original and alternative sources. For Li 李方桂 (1971), we compared the Eastling data with the charts provided in Shen 沈鍾偉 (2005)[b], and for Wang 王力 (1980) and Karlgren (1957), we compared it with the original sources. Given that Pan Wuyun 潘悟云 and Zheng Zhang Shangfang 郑张尚芳 were involved in the creation of Eastling, and that especially the reconstructions of the system outlined in Pan 潘悟云 (2000) are only available online, we assume that the data for these two reconstruction systems are truthfully displayed. Apart from a few incorrect characters in the source by Wang 王力 (1980), which we manually corrected, our comparison did not reveal any errors. In addition to the five reconstruction systems, Eastling also offers readings attributed to William Baxter, but since we could not identify these readings with any known published sources of Baxter corresponding to these readings, we did not use them in our analysis.

The Tower of Babel project (Starostin 2008, http://starling.rinet.ru/) further offers an exhaustive database of character readings following the Old Chinese reconstruction system by Starostin (1989), which was compiled by Sergei Starostin himself from 1991 on and was expanded in the years thereafter. While the original publication by Starostin (1989) lists readings for all rhyme words in the Book of Odes, the online version only offers 1358 character readings for the 1845 characters in our base list, with 487 readings missing. The Old Chinese reconstruction by Schuessler (2007) was collected from a recently published digital version of the book. Unfortunately, only 1224 readings for the 1845 rhyming characters in the Book of Odes could be found, leaving us with 621 missing character readings.

In order to compare the different rhyme systems for vowel purity, the main vowels for all available character readings for the 1845 rhyme words in the rhyme networks were extracted and added as meta-data to each rhyme in the network. The different vowel systems proposed in the different reconstruction systems are shown in Table 3. Although each of our 8 systems has much more than 1200 readings (see column 3 in Table 3), the intersection between all systems is surprisingly low, and if we only retain those readings reflected in all samples, a sample of 875 nodes remains. The data by Schuessler (2007) is missing the largest amount of characters (621 readings), followed by the data of Starostin (1989, 487 readings), and Baxter and Sagart (2014, 414 readings).

It is important to note in this context that missing readings cannot be easily added without the assistance of those who originally created a given reconstruction system. While certain aspects in Old Chinese reconstruction are systematic, allowing us to project attested Middle Chinese readings back to Old Chinese, the projection rules which differ in the reconstruction systems proposed by different scholars do not necessarily allow us to replicate their judgments, as scholars use a range of different types of evidence, including Chinese character structure, evidence from excavated texts, and early

**Table 3** Vowel systems across different Old Chinese reconstructions

| Reconstruction System | No. | Rhymes | Density | a | ɑ | æ | e | ə | o | ɔ | u | ʊ | ɯ | i |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Karlgren (1957) | 1830 | 0.0031 | 0.0026 | x | x | x | x | x | x | x | x | x | x | x |
| Li 李方桂 (1971) | 1830 | 0.0031 | 0.0026 | x | | | x | | | | x | | | x |
| Wang 王力 (1980) | 1830 | 0.0031 | 0.0026 | x | | | x | x | x | | x | | | |
| Zheng Zhang 鄭張尚芳 (2003) | 1830 | 0.0031 | 0.0030 | x | | | x | x | x | | x | | | x |
| Starostin (1989) | 1358 | 0.0035 | 0.0026 | x | | | x | | x | | x | | x | x |
| Pan 潘悟雲 (2000) | 1830 | 0.0031 | 0.0026 | x | | | x | | x | | x | | x | |
| Baxter and Sagart (2014) | 1431 | 0.0038 | 0.0033 | x | | | x | x | x | | x | | | x |
| Schuessler (2007) | 1224 | 0.0041 | 0.0035 | x | | | x | x | x | | x | | | x |

Column Rhymes lists the number of character readings available. Column Density reports the density of the rhyme network, that is, the fraction of the number of attested edges and the number of potential edges

borrowings into neighboring languages (see especially Baxter and Sagart 2014 for a discussion of the different types of evidence used in reconstruction). As a result, we cannot simply add the missing character readings in our comparative dataset without running the danger of incorrectly representing a given reconstruction system. For our comparison, we are left with what we have, and we need to address the problems resulting from gaps in the data. But since we provide all data as an Additional file with this study, we hope that collaborative efforts of the scholarly community may eventually close the gaps in the future.

When comparing across datasets, it is important that we compare samples of the data containing exactly the same nodes, as in smaller or larger samples the basic characteristics, as, for example, the number of edges, may differ, thus giving the reconstruction systems we want to compare different starting chances. The difference is further confirmed by the data on *network density* that is the fraction of the number of edges divided by the number of *potential edges* in a network. The number of potential edges in a network is the number of edges in a network in which all nodes are connected with each other and can be calculated with the help of the formula $(n^2 - n)/2$, where $n$ is the number of nodes in the network[c]. Network density for the different subgraphs is reported in Table 3. As can be seen from the scores, the subgraphs of the different reconstruction systems slightly differ in density depending on the coverage of the data sample, with the smaller datasets showing a higher density.

### 3.2 Methods

We need a measure for the *purity* of clusters in a graph. If the theory of vowel purity holds, we should expect a *high degree of isolation* for those rhyme words which can be grouped by the same vowel. We thus want to compare how well a given external grouping of the nodes in our network (the vowels reconstructed for the rhyme words in a given reconstruction system) conforms to the internal ordering in our network (as reflected by the rhyme relations among the rhyme words). If we accept that we will have a certain degree of vowel impurity in all rhyme networks, be it due to the fact that the poets deliberately decided to tolerate this, or that the underlying data reflects different stages in language history, we would still assume that words rhyme more often with each other if they have the same main vowel.

We can illustrate this notion of purity by creating a fictive dataset of six rhyme words which we label 1, 2,…, 6, and of which 1, 2, and 3 share the same vowel, and 4, 5, and 6 share a vowel, which is different from the vowel of 1, 2, and 3. In Table 4, we display two matrices which contrast different fictive types of rhyme co-occurrence for our six words. If two words rhyme, this is indicated by a cross in the cell of the matrix. Impure rhymes in which two vowels of different quality rhyme with each other are further marked by shading the cell in gray. From the two different matrices, we can easily see that the first one (matrix A) would intuitively reflect a higher degree of vowel purity than the second one (matrix B), simply because the number of impure rhymes is much lower in matrix A.

The same information can be also displayed in a network, in which our words 1, 2,…, 6 are modeled as nodes, and the information, whether they rhyme with each other in the sources (matrices A and B) are displayed by drawing an edge between the nodes. This is illustrated in Fig. 2, and we can see that the network visualization makes it even easier to see the difference between the intuitively rather pure rhyme network in A and the rather impure rhyme network in B. But our intuitive assessment may easily betray us if the data becomes more complex. For this reason, we need a way to measure to which degree a given network structure (the rhyme co-occurrences in the Book of Odes) is in conflict with a given external division of the nodes (the vowels, as annotated in the reconstruction systems of different scholars).

A measure that measures exactly what we want to test is *assortativity* (Newman 2003). Assortativity tests whether nodes sharing connections in a graph are also similar regarding other characteristics. In social network analyses it can, for example, be used to test whether observed patterns in a network, like friendship, come along with properties of the individuals, such as language or gender (ibid.). Assortativity can be measured by calculating the *assortativity coefficient* of a network in which all nodes have a given attribute. The basic idea of this coefficient is to compare the proportion of edges connecting nodes with the same attribute with the proportion of edges connecting nodes with different attributes. Calculating the assortativity coefficient in a network is straightforward. Given a network with nodes and node attributes, one first calculates an *attribute mixing matrix* which indicates the proportion of edges between all attributes. Based on this matrix, the assortativity coefficient can then be calculated with help of the formula:

**Table 4** Rather high and rather low degree of vowel purity in a fictive set of six rhyming words

| A | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | x | x | | | |
| 2 | x | | x | | | |
| 3 | x | x | | x | | |
| 4 | | | x | | x | x |
| 5 | | | | x | | x |
| 6 | | | | x | x | |

| B | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | x | | | | x |
| 2 | x | | x | | | |
| 3 | | x | | x | x | x |
| 4 | | | x | | x | x |
| 5 | | | x | x | | x |
| 6 | x | | x | x | x | |

Tables A and B show six fictive rhyming words, how they rhyme in a set of poems, with a cross in the cell indicating that the words have been shown to rhyme together in at least one poem. Assuming that words 1, 2, and 3 have the same vowels, which is different from the vowels of 4, 5, and 6 (which also share the same vowel), we can find occurrences of impure rhymes whenever one word from the set of 1, 2, and 3 rhymes with one word from the set 4, 5, 6 (indicated by shading the cell in gray). Here, our matrix A reflects a rather "pure" dataset, with only one transition in 3 and 4, while matrix B reflects an impure dataset with as many as four transitions

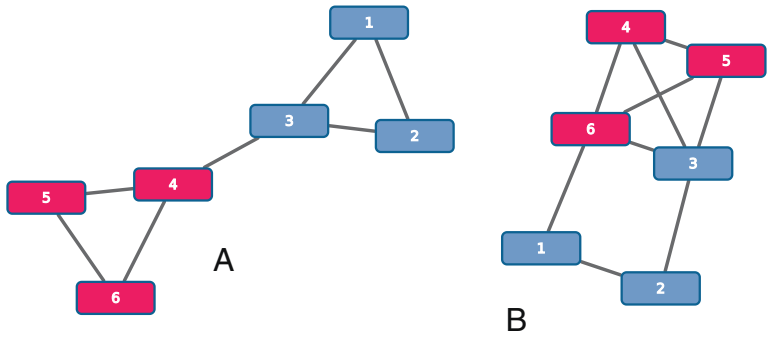



**Fig. 2** Comparing networks with (**a**) high and (**b**) low "purity" regarding the relation of colors and edges

$$r = \frac{\mathrm{Trace}(m) - \|m^2\|}{1 - \|m^2\|}, (1)$$

where $m$ is the attribute mixing matrix, *Trace* is the sum of the diagonal from top left to bottom right, and $||m||$ is the sum of all cells in the matrix (see Newman 2003 for details). An assortativity coefficient equal to 1 indicates full assortativity, with all edges only connecting nodes with the same attributes. 0 indicates no assortativity, and scores between 0 and −1 indicate inverse assortativity in which edges have the tendency to connect nodes with different attributes (ibid.).

As an example on how to calculate the assortativity for a given network, consider again our two networks in Fig. 2. In both networks, colors indicate node attributes, and even from eyeballing, we have already seen above that network A has a high assortativity (as there is only one edge connecting red and blue nodes), while network B has a lower assortativity. In order to calculate the assortativity coefficient for the two networks, we first need to determine the proportion of the edges connecting different types of nodes with each other. Assuming a *directed network*[d], in which we can draw two different edges between two nodes, both indicating the direction (from 1 to 2, or from 2 to 1, as in a one-way street), we have 14 edges (2 × 7) in the first and 18 edges (2 × 9) in the second network (see also Table 4, where the original matrices are given). The proportion of edges linking from red to red, red to blue, blue to red, and blue to blue can then be arranged in a contingency matrix, as illustrated in Table 5, and this matrix is then used as input for formula (1) to calculate the assortativity coefficient $r$. For the networks in Fig. 2, this yields:

**Table 5** Calculating the attribute mixing matrices for the networks from Fig. 1

| A | Red | Blue | Red + blue |
|---|---|---|---|
| Red | 6/14 = 0.43 | 1/14/ = 0.07 | 7/14 = 0.5 |
| Blue | 1/14 = 0.07 | 6/14 = 0.43 | 7/14 = 0.5 |
| Red + blue | 7/14 = 0.5 | 7/14 = 0.5 | 14/14 = 1.0 |
| B | Red | Blue | Red + blue |
| Red | 6/18 = 0.33 | 4/18/ = 0.22 | 10/18 = 0.55 |
| Blue | 4/18 = 0.33 | 4/18 = 0.22 | 8/18 = 0.44 |
| Red + blue | 10/18 = 0.55 | 8/18 = 0.44 | 18/18 = 1.0 |

$$r_A = \frac{0.86–0.5}{1–0.5} = 0.72\,(2)$$

$$r_B = \frac{0.56–0.51}{1–0.51} = 0.1\,(3)$$

We can see from this example that the assortativity coefficient confirms the intuition we might have already had by eyeballing the networks in Fig. 2, namely, that the network structure in network A reflects the coloring of the nodes much better than in network B.

When comparing two or more reconstruction systems with each other, we need to be careful in correctly interpreting the results. If one system has a high assortativity coefficient, this confirms a tendency to produce clusters of high purity. If the assortativity coefficient of another system is lower, however, this could be triggered by the topological structure of the network alone, and not by the reconstruction system. As scholars have chosen their reconstructions independently, assuming different numbers of vowels for their reconstructions, it may well be that the initial number of vowels might favor or disfavor a given analysis. A hypothetical system of one single vowel, for example, would receive the highest assortativity coefficient simply due to the fact that it covers the full network, and in the light of the theory of vowel purity in rhyming, this would also reflect a pure rhyming behavior, as all rhyming instances would show the same vowel.

We need to make sure that the distribution we obtain for a given reconstruction system is not due to chance. More concretely, what is interesting for us, is not only whether the distribution of vowels across a rhyme network is due to chance alone, but also to compare across different reconstruction systems, which system is most unlikely to have arisen by chance. Comparability can be achieved by comparing the results obtained for a given reconstruction system with the results of a *random distribution* obtained for the same dataset. The random distribution can be created by shuffling the node labels (the vowels for each Chinese character in our case). In order to normalize the data, one then compares to which degree the original result differs from the results obtained for the randomized distribution, that is, one compares to how unlikely it is that a given system could have been produced by chance. If we only wanted to test whether a given distribution is likely to be due to chance, we can calculate the $p$ - value, using the formula:

$$p = (S + 1)/(R + 1),\ (4)$$

where $S$ is the number of random distributions with an assortativity coefficient higher than the one we observed, and $R$ is the number of all random distributions we created. The $p$ - value will range between 1 and 0, and the lower the value we obtain, the lesser we would expect that the observed distribution was created by chance. It is customary in the social sciences to set an arbitrary threshold for the $p$ - value, indicating when an experiment is accepted to confirm a hypothesis and when it is rejected. This value is usually 0.05 in psychology and sociology, but much lower in physics.

In addition, since we do not only want to test whether a given reconstruction system is significant with respect to the principle of vowel purity, we also need to find a way to

compare different reconstruction systems with each other. A good score for this difference is to count the number of standard deviations between the mean of the randomized distribution and the non-randomized test (Lopez et al. 2013), which can be done with the help of the formula:

$$\sigma = \frac{r_A - r_E}{s_E}, \quad (5)$$

where $r_A$ is the attested assortativity coefficient, $r_E$ is the mean of the assortativity coefficients in the random sample (the *expected* assortativity), and $s_E$ is the standard deviation. This score, which we will call the *sigma score* in the following, tells us how unexpected a given analysis is with respect to an analysis which was carried out randomly: the higher the score, the lesser we expect an analysis to be due to chance. In the context of vowel purity in Chinese rhyme networks, this means that the higher a score, the more closely it groups the rhymes by vowel quality. By reporting both the sigma scores and the $p$ - values, we further make sure that our results are generally significant.

A further problem mentioned above is the problem of sample size. Since we have a considerable amount of missing readings in our data, we need to make sure that the differences do not influence our results. In order to control this, we apply a straightforward re-sampling procedure by randomly selecting a certain number of nodes from the networks which occur in all reconstruction systems and re-running the complete analysis on these subsets of the data. For this purpose, we created 10 random samples for varying numbers of nodes, ranging from 100 characters up to 800 characters (all random samples as well as the source code to create new random samples are given in the Additional file 1: supplementary material). We ran our basic analysis on all these subsets and averaged the results for a given number of nodes. In this way, we tested the robustness of our approach when dealing with datasets of different sizes and random collections of subsets of the data.

## 4 Results

We computed the assortativity coefficients for the original and the randomized data based on the Book of Odes network for all eight reconstruction systems. The randomized distribution was obtained by shuffling the nodes in each network 1000 times and storing the assortativity coefficient for each run. Thanks to the NetworkX software package (Hagberg 2009), all computations could be carried out in Python, and all source codes to replicate the analyses reported here are given in the Additional file 1: supplementary material. In all cases, our primary question was to which degree the division of the rhyme words in the network according to their reconstructed vowels would reflect the "natural" division of the networks into rhyme classes as represented in the annotated network of rhymes in the Book of Odes. Table 5 shows the results for this experiment for the 875 character readings.

As one can see from the results in Table 6, the reconstruction system by Baxter and Sagart (2014) outperforms all other systems. With an assortativity coefficient of 0.88 and a sigma score of 79, it shows a higher degree of assortativity than the other systems, and a generally high assortativity with respect to vowel purity. The next in order is the system of Starostin (1989), with an assortativity coefficient of 0.84 and a sigma score of 74. The system of Li 李方桂 (1971) performs worse than the other

**Table 6** Results of the analysis for the complete dataset (including all characters reflected in all reconstruction systems), a total of 875 nodes

| Reconstruction System | Assortativity | Randomized Assortativity (Ø) | Standard Deviation | Sigma Score | Rank | *p*-Value |
|---|---|---|---|---|---|---|
| Karlgren (1957) | 0.5824 | -0.0029 | 0.0091 | 64 | 6 | < 0.001 |
| Li 李方桂 (1971) | 0.8230 | -0.0026 | 0.0149 | 56 | 8 | < 0.001 |
| Wang 王力 (1980) | 0.7709 | -0.0026 | 0.0127 | 61 | 7 | < 0.001 |
| Zheng Zhang 鄭張尚芳 (2003) | 0.7435 | -0.0021 | 0.0103 | 72 | 3 | < 0.001 |
| Starostin (1989) | 0.8444 | -0.0025 | 0.0115 | 74 | 2 | < 0.001 |
| Pan 潘悟雲 (2000) | 0.7326 | -0.0020 | 0.0103 | 71 | 4 | < 0.001 |
| Baxter and Sagart (2014) | 0.8765 | -0.0025 | 0.0112 | **79** | **1** | < 0.001 |
| Schuessler (2007) | 0.7244 | -0.0026 | 0.0111 | 66 | 5 | < 0.001 |

systems with a sigma score of 56, followed by the system of Wang 王力 (1980) with a sigma score of 61. As the *p*-values in the last column in Table 6 indicate, all of our experiments are highly significant, and there was no random distribution of vowels in all 1000 which achieved a higher assortativity coefficient than the one we achieved for the observed data. Regardless of the reconstruction system, all reconstructions show a high tendency to reflect vowel purity.

As we mentioned before, due to the large number of missing readings in our data, we need to control for the sample size. As a strategy, we carried out the re-sampling procedure outlined in the end of Section 3.2, in which we split the data into randomly selected samples of varying sizes of 100, 200,…, up to 800 characters, and then applying our basic method to those subsets of the data. The averaged results for the ten different samples we used in each analysis are given in Table 7. For reasons of space, we only report ranks and sigma scores, but all detailed analyses are provided in the Additional file 1: supplementary material. All $p$ values for these analyses were highly significant with $p < 0.01$. As can be seen from the table, all studies on the subsets confirm the tendency we also saw in the full sample from Table 6, and especially the ranks are remarkably stable (the only exception being the analyses by Schuessler and Karlgren in the lower ranks). What one can also see is that the size of the networks has a direct impact on the sigma scores, which is easy to understand keeping in mind that if we select only a small number of nodes the evidence for rhyme co-occurrences will drastically shrink.

**Table 7** Results of the re-sampling test on randomized subsets of the data with varying numbers of characters, and the resulting rankings for all datasets for the respective analysis. The eight re-sampling trials consist of ten randomly selected sets of characters

| Reconstruction System | 100 | # | 200 | # | 300 | # | 400 | # | 500 | # | 600 | # | 700 | # | 800 | # |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Karlgren (1957) | 7 | 5 | 16 | 3 | 23 | 3 | 32 | 3 | 36 | 6 | 45 | 6 | 52 | 5 | 60 | 5 |
| Li 李方桂 (1971) | 7 | 5 | 13 | 8 | 18 | 8 | 26 | 8 | 32 | 8 | 40 | 8 | 45 | 8 | 51 | 8 |
| Wang 王力 (1980) | 7 | 5 | 15 | 7 | 21 | 7 | 30 | 7 | 35 | 7 | 43 | 7 | 49 | 7 | 56 | 7 |
| Zheng Zhang 鄭張尚芳 (2003) | 9 | 2 | 16 | 3 | 23 | 3 | 32 | 3 | 40 | 3 | 49 | 3 | 55 | 3 | 64 | 3 |
| Starostin (1989) | 9 | 2 | 17 | 2 | 25 | 2 | 35 | 2 | 43 | 2 | 51 | 2 | 59 | 2 | 68 | 2 |
| Pan 潘悟雲 (2000) | 9 | 2 | 16 | 3 | 23 | 3 | 32 | 3 | 39 | 4 | 48 | 4 | 54 | 4 | 63 | 4 |
| Baxter and Sagart (2014) | **10** | **1** | **18** | **1** | **27** | **1** | **37** | **1** | **46** | **1** | **55** | **1** | **63** | **1** | **72** | **1** |
| Schuessler (2007) | 7 | 5 | 16 | 3 | 22 | 6 | 31 | 6 | 38 | 5 | 46 | 5 | 52 | 5 | 60 | 5 |

The numbers (100, 200,…, 800) indicate the number of selected nodes, and the cell content of the columns shows the averaged sigma scores. The columns with the hash character (#) reflect the ranking for the respective node selection. Cell content in bold font reflects the highest value(s), cell content shaded in light gray reflects the lowest value(s) in the rank

Apart from the remarkable robustness of the results across different random samples of the data, the difference between the reconstruction systems regarding their individual degrees of vowel purity is also quite striking. This is interesting since scholars have often emphasized the similarities between the more recently proposed reconstruction systems (Behr 1999). Given that we only investigate the main vowels, thus ignoring all other potential disagreements, shows that we are still far away from a *communis opinio* on Old Chinese phonology. The differences between the reconstructions are further illustrated in Fig. 3, where we contrast the reconstructed vowels for 300 characters out of the 1830 character readings in the data. While we can see a rather high agreement in the majority of patterns, especially between the six vowel systems of Old Chinese, it is also easy to identify certain individual differences in the reconstructions. These cases show that it is not one major disagreement triggering the variation, but a notable number of individual reconstructions in which scholars differ.

The assortativity coefficients of all systems and the high significance of our randomized tests indicate that vowel purity plays an important role in Old Chinese

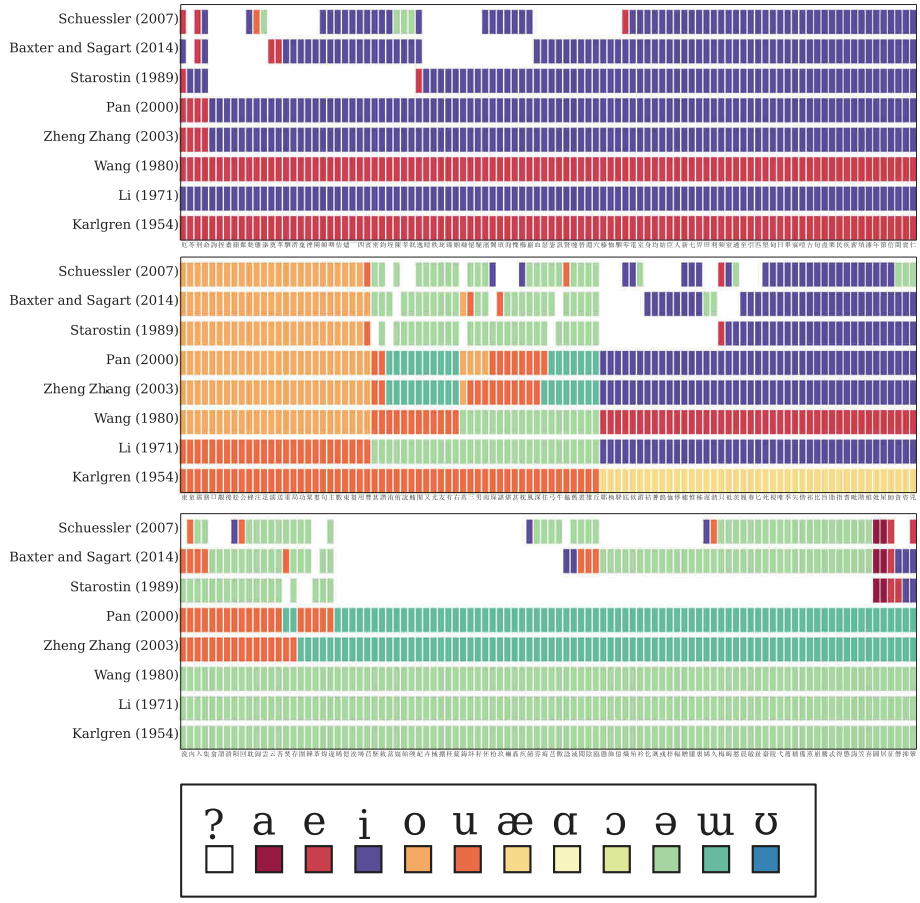



**Fig. 3** Comparing the rhyme patterns across different reconstruction systems. The figure shows three subsets of 100 characters each as they occur in the rhyme data of the Book of Odes; both include missing characters and the respective vowel readings. While a definite structural similarity can be detected, we also find remarkable differences. In the figure, each cell corresponds to one reading for a given character in the row. Since the characters are too small to be readable, we offer a high-resolution version of this figure in the Additional file 1: supplementary material

rhyming. If vowel quality was independent of rhyme decisions, we would expect to find assortativity coefficients to be close to zero, as we found in the random distributions. What this means more concretely is shown in Fig. 4, where we show the full rhyme network in which nodes have been colored according to the system of Baxter and Sagart (2014). From this perspective, we can see that the network is highly structured. Most rhymes which are topographically close from organic groups in the network, as shown by their colors. That one and the same vowel further form multiple distinct clusters is also to be expected, as vowel quality is not the only factor conditioning rhyming. Furthermore, given the overall structure of the network with its one larger component that connects almost all of the characters, we can also see that the rhyme purity assumption is essentially an assumption of degree: we find definite clusters which obviously correspond to words with a very similar if not identical pronunciation in Old Chinese, but we also find obvious transitions between all rhyme groups.

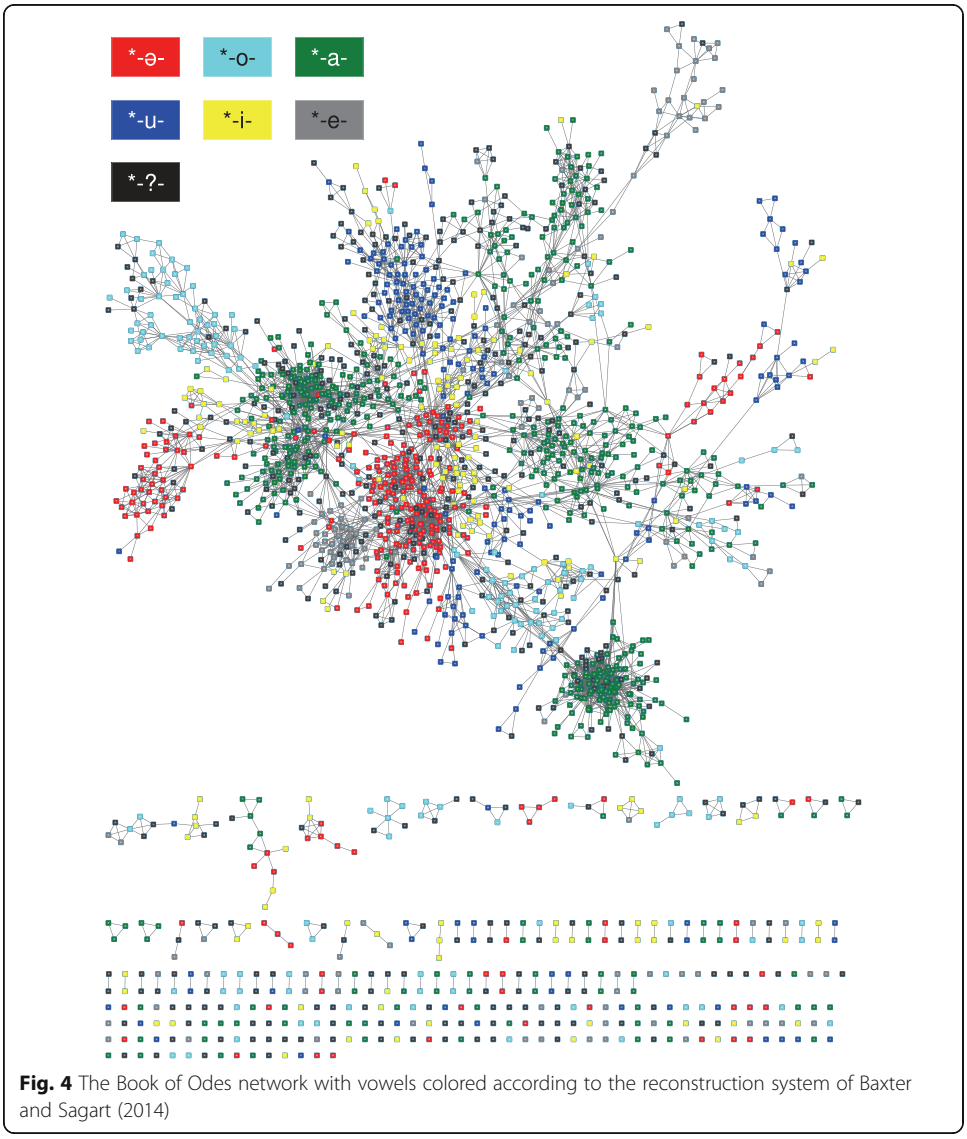



**Fig. 4** The Book of Odes network with vowels colored according to the reconstruction system of Baxter and Sagart (2014)

## 5 Discussion

What can we learn from this experiment? Surprisingly, the reconstruction system of Baxter and Sagart (2014), which was heavily criticized by Ho (2016) for its lack in vowel purity, seems to evince a much higher purity of vowels then all other popular reconstruction systems for Old Chinese, regardless of the number of vowels which these systems actually reconstruct. If vowel identity was indeed a valid criterion for the choice of rhyming words in Old Chinese times, this could be seen as strong evidence for the superiority of the reconstruction system by Baxter and Sagart (2014) closely followed by the system of Starostin (1989). Yet, we should be careful with our conclusions, since vowel purity is surely only one factor that may have contributed to Old Chinese rhyming practice, and we cannot be sure how important this factor was. In order to use the vowel purity criterion to favor or disfavor certain reconstruction systems of Old Chinese, more evidence on the universality or the areal prevalence of this principle in rhyming would be required. Since rhyming practice results from the interaction between language, culture, and cognition, more studies on cross-linguistic and cross-cultural rhyming practices would be needed to clearly use external criteria as evidence for or against a given Old Chinese reconstruction.

Even if we refuse to use the results of this research to rank or evaluate the different reconstruction systems of Old Chinese, we consider it as a valuable contribution to the field of Chinese historical linguistics, as we have shown that we can easily design quantitative tests that check to which degree different reconstruction systems conform to a given criterion. By expanding this principle to the finals of different reconstruction systems, we could, for example, test the general degree of purity with respect to the rhymes in the Book of Odes. As shown in List (2017), we can also use the rhyme networks to resolve uncertainties inside a given reconstruction system. Due to the diversity of poetry collections like the Book of Odes itself, we could further compare rhyming behavior across different partitions of the data, thus testing current hypotheses regarding its development history. Given the crucial role that Chinese plays for the history of the Sino-Tibetan language family, research along these lines may not only have an impact on Chinese historical linguistics, but may also help us to gain new insights into the prehistory of one of the largest language families in the world.

Given that Chinese is not the only language whose older stages are reflected in rhyming, one may even think of applying the method to other languages, such as Tangut (Arakawa 2001) or Egyptian (Peust 2014). When taken with care, network studies on rhyming practice may provide additional evidence for original pronunciation, especially in those situations where the writing system lacks precision in truthfully representing speech in phonetic detail. These methods may also be used to investigate cross-linguistic rhyming tendencies. So far, the vowel purity principle is still a hypothesis rather than a confirmed effect. By adding more data from different languages to the sample, one could investigate whether it reflects a universal tendency rather than a specific tendency in Old Chinese rhyming.

This paper shows that a thorough quantitative comparison can give us new insights into the problems in the reconstruction of Old Chinese, but also into the more general problems of reconstruction in historical linguistics. Instead of dismissing theories or reconstructions by cherry-picking particular examples, a thorough and if possible exhaustive evaluation may often allow us to look at problems from a fresh perspective.

Unfortunately, increasing the amount of data amenable for quantitative investigations is time-consuming. For this reason, the results presented in this paper can only be regarded as preliminary until the existing data are more consistently checked and new data have been added. In order to tackle these problems in the future, collaborative efforts are required, and all scholars should try to contribute by sharing their data as transparently as possible.

## 6 Endnotes

[a]The two words given as example occur as rhyme words in the last stanza of the famous German folk song *Abendlied* ('evening song') by Matthias Claudius' (1714–1840), which originally reads: *So legt euch denn ihr Brüder, In Gottes Namen nieder* ('now lie down you brothers in the name of god').

[b]Note that the original source by Li 李方桂 (1971) does not list all characters of the Book of Odes, and all accounts, be it the one provided by Eastling or the one provided by Shen 沈鍾偉 (2005) apply the principles outlined in Li 李方桂 (1971) independently to Middle Chinese character readings.

[c]For a network of three nodes, we would thus have $(3^2\text{-}3)/2 = 3$ edges (A-B, A-C, B-C, for nodes A, B, and C), and for a network with four nodes, the number of potential edges would amount to $(4^2\text{-}4)/2 = 6$ (A-B, A-C, A-D, B-C, B-D, C-D).

[d]Any undirected network can be transformed to a directed network by replacing all undirected edges between a node pair $n_1$ and $n_2$ with one directed edge from $n_1$ to $n_2$ and one from $n_2$ to $n_1$.

## Additional file

**Additional file 1:** Supplementary data and source code are on-line available for download at: https://zenodo.org/badge/latestdoi/78204684. (PDF 24 kb)

**Acknowledgements**
This research was supported by the DFG research fellowship grant 261553824 "Vertical and lateral aspects of Chinese dialect history" (JML) and by the ERC Synergy Project 609823 ASIA "Beyond Boundaries: Religion, Region, Language and the State" (NWH). EB and JSP are supported by the ERC under the European Community's Seventh Framework Programme, FP7/2007-2013 Grant Agreement 615247. We are specifically indebted to François-Joseph Lapointe for the helpful advice regarding our statistical approach. We are very grateful to Doug Cooper for providing help in preparing the data by Schuessler (2007), and we thank George Starostin, William Baxter, Laurent Sagart, and Pan Wuyun for providing free online access to their data.

**Authors' contributions**
JML initiated the study and wrote the first draft of the paper. JML and NH assembled the data. PL and EB designed the tests on vowel purity. JML and NSP implemented and carried out the tests. All authors approved the final version of the manuscript.

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

## Publisher's Note



**Author details**
[1]Centre de recherches linguistiques sur l'Asie Orientale, École des Hautes Études en Sciences Sociales, 2 Rue de Lille, 75007 Paris, France. [2]Team Adaptation, Integration, Reticulation, Evolution Université Pierre et Marie Curie, 9 Quai St Bernard, 75005 Paris, France. [3]School of African and Oriental Sciences, University of London, Thornhaugh Street, Russell Square, London WC1H 0XG, UK.

Received: 1 November 2016 Accepted: 6 May 2017
Published online: 26 June 2017

**References**

Arakawa, Shintaro 荒川慎太郎. 2001. About the rhymes in Tangut verses 西夏詩の脚韻に見られる韻母について. Linguistic Research of the Kyoto University 京都大学言語学研究 20: 195–224.

Baxter, William H. 1992. A handbook of Old Chinese phonology. Berlin: de Gruyter.
Baxter, William H. and Laurent Sagart. 2014. Old Chinese: A new reconstruction. Oxford: Oxford University Press.
Behr, Wolfgang. 1999. Odds on the Odes. Available at https://web.archive.org/web/20110519085435/http://www.ruhr-uni-bochum.de/gpc/behr/HTML/Excellence.htm. Accessed 24 April 2017.
Chao, Yuen Ren. 1968. A grammar of spoken Chinese. Berkeley and Los Angeles: University of California Press.
Geng, Zhensheng 耿振生. 2004. 20th century's methods in traditional Chinese phonology 20世纪汉语音韵学方法论. Beijing: Peking University Press.
Hagberg, Aric. 2009. NetworkX: High productivity software for complex networks. Available at http://networkx.lanl.gov/index.html. Version 1.11.
Harbsmeier, Christoph, and Shaoyu Jiang. 2009. TLS–Thesaurus Linguae Sericae. A historical and comparative encyclopedia of Chinese conceptual schemes. Available at http://tls.uni-hd.de/home_en.lasso. Accessed 4 April 2012.
Ho, Dah-an. 2016. Such errors could have been avoided. Review of "Old Chinese: A new reconstruction" by William H Baxter and Laurent Sagart. Journal of Chinese Linguistics 44(1): 175–230.
He, Jiuying 何九盈. 2006. History of ancient Chinese linguistics 中国古代语言学史. Beijing: Peking University Press.
Karlgren, Bernhard. 1957. Grammata serica recensa. Bulletin of the Museum of Far Eastern Antiquities 26: 1–332.
Kern, Martin. 2004. Die Anfänge der chinesischen Literatur. In Chinesische Literaturgeschichte, ed. Emmerich Reinhard, 1–87. Stuttgart: Metzler.
Li, Fang-kuei 李方桂. 1971. Studies on archaic Chinese phonology. Tsinghua Journal of Chinese Studies 清華學報 9(1–2): 1–61.
List, Johann-Mattis. 2017. Using network models to analyze Old Chinese rhyme data. Bulletin of Chinese Linguistics 9(2): 218–241. doi: 10.1163/2405478X-00902004.
Lopez, Philippe, Johann-Mattis List, and Eric Bapteste. 2013. A preliminary case for exploratory networks in biology and linguistics: The phonetic network of Chinese words as a case-study. In Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization, ed. Heiner Fangerau, Hans Geisler, Thorsten Halling, and William Martin, 181–196. Stuttgart: Franz Steiner Verlag.
Lv, Sheng-nan 吕胜男. 2009. A brief study of the methodology of the study of ancient rhyme-andnd Concurrently on the study of the rhyme of "Jinwen Shangshu" 古韵研究方法论发微-兼论今文《尚书》用韵研究. Journal of Nanyang Normal University (Social Sciences) 南阳师范学院学报(社会科学版) 8(2): 57–61
Newman, Mark EJ. 2003. Mixing patterns in networks. Physical Review E 67(2): 1–13.
Pan, Wuyun 潘悟云. 2000. Chinese historical phonology 汉语历史音韵学. Shanghai: Shanghai Educational Publishing House.
Peust, Carsten. 2014. Towards a typology of poetic rhyme with observations on rhyme in Egyptian. In Egyptian-Coptic linguistics in typological perspective, ed. Eitan Grossman, Martin Haspelmath, and Tonio S Richter, 341–386. Berlin and Munich: de Gruyter Mouton.
Schuessler, Axel. 2007. ABC etymological dictionary of Old Chinese. Honolulu: University of Hawaii Press.
Shanghai Normal University 上海师范大学. 2016. Eastling. Old Chinese phonology 东方语言学.上古音查询. Available at http://www.eastling.org. Accessed 12 April 2016.
Shen, Zhongwei 沈鍾偉. 2005. Chart of Li Fanggui's Old Chinese reconstructions 李方桂上古音韻表. In Essays in Chinese Historical linguistics: Festschrift in memory of professor Fang-Kuei Li on his centennial birthday 漢語史研究: 紀念李方桂先生百年冥誕論文集, ed. Pang-Hsing Ting and Anne O Yue, 571–588. Taipei: Academia Sinica.
Starostin, George S. 2008. Tower of Babel: An etymological database project. Available at http://starling.rinet.ru. Accessed 12 April 2016.
Starostin, Sergei A. 1989. Rekonstrukcija drevnekitajskoj fonologičeskoj sistemy Reconstruction of the phonological system of Old Chinese. Moscow: Nauka.
Wang, Li 王力. 1980. Rhyme readings in the Book of Odes 詩經韻讀. Shanghai: Shanghai Guji Press.
Zheng Zhang, Shangfang 郑张尚芳. 2003. Old Chinese phonology 上古音系. Shanghai: Shanghai Educational Publishing House.

# IV. CONCLUSION

During these three years, I used and studied network methods which I have found to be efficient for comparing very large datasets (up to millions of objects), in particular for structuring genetic diversity into clusters. This provided me with a rather broad view on the genetic diversity that is the result of evolution. Because these approaches were very generic, I could address a diversity of evolutionary issues. Throughout my thesis I also developed several bioinformatics tools to achieve my aims.

The first kind of issues were directly related to the evolution of primary sequences. CompositeSearch allowed me to quantify the gene families associated with particular evolutionary events such as transition to animal multicellularity. Moreover, beyond this relatively classic clustering of molecular sequences, CompositeSearch provided insights on a set of more complex objects: families of composite genes. This double angle returned a more comprehensive description of the changes affecting individual genomes and metagenomes because evolution could be investigated at the level of full sized genes as well as at a subgenic level. These two levels may be coupled: for example, during the transition to animal multicellularity, there was both an increase in the number of novel gene families in the ancestor of animals and an increase in the combinations of genetic fragments associated in these taxa. The processes leading to composite genes were less known than the evolution of entire novel gene families and logically the effect of these combinatorial processes deserved a better description.

My work on plasmids also shows that composite genes can be plasmid encoded and affect host cell biology. My work on soil microbiomes showed that the abundance of composite genes can be correlated with pollution levels. These two examples highlight that important biological novelties and adaptations can result from processes acting at subgenic levels. I believe that this work is just the tip of the iceberg for the contribution of composite genes to the evolution of gene families, genomes and organisms. Further tests will of course be necessary to confirm these in silico predictions.

An additional computational development will also be necessary to provide a more comprehensive description of the composite genes. Typically, CompositeSearch could only detect composite genes that combine fragments form distinct gene families. As such, it didn't

detect tandem repeats for instance. Moreover, the composite genes that CompositeSearch detected were only genes along which at least two components families match. As such, it could not detect extended genes, i.e. genes that presented original terminal extensions or insertions with respect to their homologs. Detecting these additional gene remodeling emerges as a natural perspective from my PhD research. I have started implementing a tool called ExtensionSearch to this endeavor. ExtensionSearch is a parallelized program implemented in C++ that uses SSN to detect extended genes and extended gene families. It is able to analyze SSN that are composed of several million nodes and hundreds of millions of edges. Based on the CompositeSearch algorithm, it is capable of capturing not only the conserved extended gene family forming cliques but also the less conserved ones forming quasi-cliques. There are two different case of extension that can be detected are shown in Figure 20.

The first case is the extension that happened within a gene family (Figure 20), where an extra segment might originate from either a terminal extension or loss of a terminal segment. The second case is an extension that happened in a remodeled genes. Figure 20 shows an example of a composite gene where the two component are not fused in a consecutive position and are separated by an insertion.

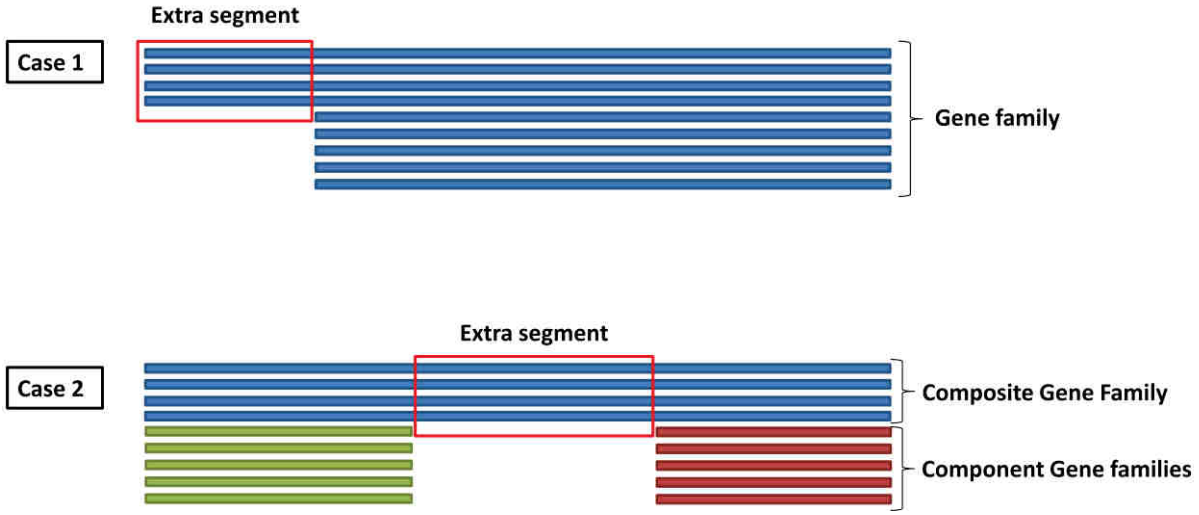



**Figure 22:** Different type of gene extension detected by ExtensionSearch.

My lab mates are expected to take over the completion of the project in order to analyze in particular gene extensions associated with animal evolution. Indeed, I found this question of high interest as I started collaborating with James McInerney, Mary O'Connell and Ray Moran, who were among the first external users of CompositeSearch. Together we

initiated an ambitious and comprehensive analysis of the events of molecular evolution that have affected animal genomes. This work currently features a study of a potential molecular clock for gene family creation, gene family duplication and the evolution of composite genes (Annex 2). We utilized novel algorithms in phylogenomics and sequence similarity networks to understand how and to what extent linear processes, such as gene duplication and mutation in gene families, and non-linear processes, that merge gene families, facilitate the emergence of new genes and novel phenotypes/functions. While we understand that mutational molecular clocks tend to tick with complex but increasingly well-understood rates (Lynch 2010), we have not yet been able to understand how, when, and especially at what rate, gene remodeling has impacted animal proteomes. Until recently, no comprehensive dataset of proteins was available to investigate such questions. Our preliminary results demonstrate that the processes involved in the evolution of protein coding genes strikingly differ depending on which part of the animal tree of life you examine. We show that remodeled genes are widely prevalent in animals and are in fact a major conduit for genetic innovation. Finally, we show that in extant species, like human, remodeled genes tend to reuse and recycle already existing remodeled material. On a broader scale, this work provides a first glimpse into how the protein coding elements of animals have evolved through time and how this correlates with major evolutionary transitions in phenotypes from sponges to humans. Before the corresponding manuscript is submitted, we will add to this work a study of the gene extensions that are found in these taxa.

While most of my PhD thesis was focused on simple SSNs, I also contributed to develop additional types of network. On the one hand, I implemented scripts that ease the construction and analysis of bipartite graphs, which are becoming increasingly popular since the teams of Fernando Baquero (Spain), Eugene Koonin (USA) and Tal Dagan (Germany) are also using comparable approaches. With my scripts, lateral gene transfers in prokaryotes can be investigated, even by biologists who are not primarily programmers. For example, my collaborator Alex Jaffe, currently a PhD student in metagenomics, has introduced this method in the Banfield lab. On the other hand, I also contributed to develop a new type of co-occurrence network, called trait network. Here the challenge was not so much the avalanche of data but rather the need to investigate carefully selected data under a different perspective, to get even more out of the current datasets. The construction of trait networks purposely relies on simple rules, reflecting the distribution of traits in taxa. They are of immediate interest to paleontologists, but because these networks are very generic, they will also be of use in the near future to investigate the distribution of diverse components in diverse

biological systems (proteins in organellar proteomes, OTUs in environmental samples and cultural items in human societies).

Across all these studies arose a common theme. It is possible to model the evolution of entities be they genes, genomes, organisms and languages in a way that accounts for more of the actual complexity of these objects. Namely, aspects of the modular nature of evolved forms can be captured using networks (composite genes, introgressed genomes, versatile body plans and compound words). Another pleasant outcome of network studies is that they allow to connect issues from different fields (paleontology, molecular evolution, linguistics) in a fruitful way. I am very thankful to all my great collaborators.

That being said, bioinformatics remains at the center of network analyses, which means that evolutionary biologists that are trained in bioinformatics could make significant contributions if they succeed in overcoming the following remaining challenges. First, in the post-genomic era, we have now access to large molecular data with considerable genetic diversity from genomic and metagenomic projects. During my thesis, I was led to construct and study very large similarity networks, for example with the study of gene remodeling in the polluted environments, where the SSN was composed of 3,166,706 nodes and 282,789,792 edges. The construction of large similarity networks remains a computational challenge in terms of memory and processing time. Computing pairwise similarity is a fundamental task in the construction of SSN, especially for composite gene detection where it needs information about the position of the alignments on the sequences. While a quadratic complexity might seem good enough, in practice the exponential growth of biological sequences calls for speedup of sequence alignment tools such as BLAST. One alternative that I propose is to use new alignment software such as DIAMOND. These new variants of BLAST approach have been using flexible-length seeds on a reduced amino acid alphabet. They considerably reduce the computational time on a desktop machine, and should be used in the future, especially for large metagenomics datasets. Moreover, since pairwise comparison can be easily parallelized, promising low-cost solutions could come from the Hadoop framework, in order to parallelize existing software such as HAMOND or HBLAST, respectively for DIAMOND and BLAST.

Another perspective to accelerate comparisons would be to adopt an approach which reduces the complexity of the data using the alignment-free methods based on *k*-mers. These methods are based on the comparison of subsequences of length *k* shared between sequences..

Using alignment-free tools we could quickly screening the huge data to filter the sequences in order to reduce computational time and memory usage during the all versus all sequence comparison.

Recently, graph databases have started to appear, easing the mining of very large networks. Such databases uses graph structures for semantic queries with nodes, edges and properties to represent and store data (Angles and Gutierrez 2008). With the emergence of big data in biology, several researchers have started to also use graph databases for biological network analyses (Henkel et al. 2015; Lysenko et al. 2016; Mullen et al. 2016). However, to our knowledge, there is no example of graph databases in evolutionary biology. A neat future development would then be to incorporate graph database management systems, such as Neo4j (Webber 2012), to our network-based studies of molecular evolution.

# V. REFERENCES


Adai AT, Date SV, Wieland S, Marcotte EM. 2004. LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks. Journal of Molecular Biology 340(1):179-190.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO. 2013. Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc Natl Acad Sci U S A 110(17):E1594-603.

Angles R, Gutierrez C. 2008. Survey of graph database models. ACM Comput. Surv. 40(1):1-39.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310(2):311-25.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. 2009. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4(2):e4345.

Avelar GM, Schumacher RI, Zaini PA, Leonard G, Richards TA, Gomes SL. 2014. A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus. Curr Biol 24(11):1234-40.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37(Web Server issue):W202-8.

Bapteste E, Lopez P, Bouchard F, Baquero F, McInerney JO, Burian RM. 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci U S A 109(45):18266-72.

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L et al. . 2013. Networks: expanding evolutionary thinking. Trends Genet 29(8):439-41.

Bashton M, Chothia C. 2002. The geometry of domain combination in proteins. J Mol Biol 315(4):927-39.

Bolten E, Schliep A, Schneckener S, Schomburg D, Schrader R. 2001. Clustering protein sequences--structure prediction by transitive homology. Bioinformatics 17(10):935-41.

Bonfante G. 1931. I dialetti indoeuropei. Annali del R. Istituto Orientale di Napoli 4:69-185.

Brandes U, Robins G, McCranie ANN, Wasserman S. 2013. What is network science? Network Science 1(1):1-15.

Brennan G, Kozyrev Y, Hu SL. 2008. TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. Proc Natl Acad Sci U S A 105(9):3569-74.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523(7559):208-11.

Brown T. 2002. Genomes 2nd edition. Oxford: Wiley-Liss; Chapter 16, Molecular Phylogenetics.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Meth 12(1):59-60.

Cheng S, Karkar S, Bapteste E, Yee N, Falkowski P, Bhattacharya D. 2014. Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Frontiers in Ecology and Evolution 2(72).
Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. Science 300(5626):1701-3.
Clune J, Mouret JB, Lipson H. 2013. The evolutionary origins of modularity. Proc Biol Sci 280(1755):20122863.
Corel E, Lopez P, Meheust R, Bapteste E. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol 24(3):224-37.
Courseaux A, Nahon JL. 2001. Birth of two chimeric genes in the Hominidae lineage. Science 291(5507):1293-7.
Dagan T. 2011. Phylogenomic networks. Trends Microbiol 19(10):483-91.
Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A 105(29):10039-44.
Daniel R. 2005. The metagenomics of soil. Nat Rev Microbiol 3(6):470-8.
Darwin C. 1859. On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life. London :: John Murray.
Dediu D, Levinson SC. 2013. On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. Frontiers in Psychology 4(397):1-17.
Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. 2011. Accessing the soil metagenome for studies of microbial diversity. Appl Environ Microbiol 77(4):1315-24.
Diestel R. 2006. Springer Science & Business Media.
Dolinski K, Troyanskaya OG. 2015. Implications of Big Data for cell biology. Mol Biol Cell 26(14):2575-8.
Duboule D, Wilkins AS. 1998. The evolution of 'bricolage'. Trends Genet 14(2):54-9.
Duranti S, Ferrario C, van Sinderen D, Ventura M, Turroni F. 2017. Obesity and microbiota: an example of an intricate relationship. Genes Nutr 12:18.
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460-1.
Edgar RC, Batzoglou S. 2006. Multiple sequence alignment. Curr Opin Struct Biol 16(3):368-73.
Eisenstein M. 2015. Big data: The power of petabytes. Nature 527(7576):S2-4.
Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402(6757):86-90.
Enright AJ, Ouzounis CA. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol 2(9):RESEARCH0034.
Fondi M, Fani R. 2010. The horizontal flow of the plasmid resistome: clues from inter-generic similarity networks. Environ Microbiol 12(12):3228-42.
Geisler H, List JM. 2013. Do languages grow on trees? The tree metaphor in the history of linguistics. In: Fangerau H, Geisler H, Halling T, Martin W, editors. Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization. Stuttgart: Franz Steiner Verlag. p. 111-124.
Gilbert W. 1978. Why genes in pieces? Nature 271(5645):501.
Gimmler A, Korn R, de Vargas C, Audic S, Stoeck T. 2016. The Tara Oceans voyage reveals global diversity and distribution patterns of marine planktonic ciliates. Sci Rep 6:33555.
Grau-Bove X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. Elife 6.

Guinane CM, Cotter PD. 2013. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. Therap Adv Gastroenterol 6(4):295-308.
Hagel JM, Facchini PJ. 2017. Tying the knot: occurrence and possible significance of gene fusions in plant metabolism and beyond. J Exp Bot.
Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci U S A 107(1):127-32.
Halary S, McInerney JO, Lopez P, Bapteste E. 2013. EGN: a wizard for construction of gene and genome similarity networks. BMC Evol Biol 13:146.
Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. 2007. The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8(4):319-30.
Hauswedell H, Singer J, Reinert K. 2014. Lambda: the local aligner for massive biological data. Bioinformatics 30(17):i349-i355.
Henkel R, Wolkenhauer O, Waltemath D. 2015. Combining computational models, semantic annotations and simulation experiments in a graph database. Database: The Journal of Biological Databases and Curation 2015:bau130.
Huson DH, Xie C. 2014. A poor man's BLASTX—high-throughput metagenomic protein database search using PAUDA. Bioinformatics 30(1):38-39.
Jablonski D, Shubin NH. 2015. The future of the fossil record: Paleontology in the 21st century. Proc Natl Acad Sci U S A 112(16):4852-8.
Jachiet PA, Colson P, Lopez P, Bapteste E. 2014. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. Genome Biol Evol 6(9):2195-205.
Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. Bioinformatics 29(7):837-44.
Jacob F. 1977. Evolution and tinkering. Science 196(4295):1161-6.
Jaffe AL, Corel E, Pathmanathan JS, Lopez P, Bapteste E. 2016. Bipartite graph analyses reveal interdomain LGT involving ultrasmall prokaryotes and their divergent, membrane-related proteins. Environ Microbiol 18(12):5072-5081.
Jones CD, Begun DJ. 2005. Parallel evolution of chimeric fusion genes. Proc Natl Acad Sci U S A 102(32):11373-8.
Jones CD, Custer AW, Begun DJ. 2005. Origin and evolution of a chimeric fusion gene in Drosophila subobscura, D. madeirensis and D. guanche. Genetics 170(1):207-19.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res 20(10):1313-26.
Kawai H, Kanegae T, Christensen S, Kiyosue T, Sato Y, Imaizumi T, Kadota A, Wada M. 2003. Responses of ferns to red light are mediated by an unconventional photoreceptor. Nature 421(6920):287-90.
Kaznadzey A, Alexandrova N, Novichkov V, Kaznadzey D. 2013. PSimScan: Algorithm and Utility for Fast Protein Similarity Search. PLoS ONE 8(3):e58505.
Kennedy J, Flemer B, Jackson SA, Lejon DP, Morrissey JP, O'Gara F, Dobson AD. 2010. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. Mar Drugs 8(3):608-28.
Kim M, Lee KH, Yoon SW, Kim BS, Chun J, Yi H. 2013. Analytical tools and databases for metagenomics in the next-generation sequencing era. Genomics Inform 11(3):102-13.
King N. 2004. The unicellular ancestry of animal development. Dev Cell 7(3):313-25.
King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I et al. . 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451(7180):783-8.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. Molecular Biology and Evolution 28(2):1057-1074.
Kodzius R, Gojobori T. 2015. Marine metagenomics as a source for bioprospecting. Marine Genomics 24, Part 1:21-30.
Kumar S. 1996. A stepwise algorithm for finding minimum evolution trees. Mol Biol Evol 13(4):584-93.
Lanza VF, Tedim AP, Martinez JL, Baquero F, Coque TM. 2015. The Plasmidome of Firmicutes: Impact on the Emergence and the Spread of Resistance to Antimicrobials. Microbiol Spectr 3(2):PLAS-0039-2014.
Lewontin RC. 1970. The Units of Selection. Annual Review of Ecology and Systematics 1(1):1-18.
Li FW, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, Ruhsam M, Sigel EM, Der JP, Pittermann J et al. . 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc Natl Acad Sci U S A 111(18):6672-7.
List J-M, Lopez P, Bapteste E. 2016a. Using Sequence Similarity Networks to Identify Partial Cognates in Multilingual Wordlists.
List J-M, Pathmanathan JS, Lopez P, Bapteste E. 2016b. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. Biology Direct 11(39):1-17.
Liu M, Grigoriev A. 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? Trends Genet 20(9):399-403.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet 4(11):865-75.
Luef B, Frischkorn KR, Wrighton KC, Holman HY, Birarda G, Thomas BC, Singh A, Williams KH, Siegerist CE, Tringe SG et al. . 2015. Diverse uncultivated ultra-small bacterial cells in groundwater. Nat Commun 6:6372.
Lynch M. 2010. Evolution of the mutation rate. Trends Genet 26(8):345-52.
Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. 2016. Representing and querying disease networks using graph databases. BioData Mining 9(1):23.
Ma Y, Paulsen IT, Palenik B. 2012. Analysis of two marine metagenomes reveals the diversity of plasmids in oceanic environments. Environ Microbiol 14(2):453-66.
Mandal RS, Saha S, Das S. 2015. Metagenomic surveys of gut microbiota. Genomics Proteomics Bioinformatics 13(3):148-58.
Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428):751-3.
Marsh JA, Hernandez H, Hall Z, Ahnert SE, Perica T, Robinson CV, Teichmann SA. 2013. Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153(2):461-70.
Marsh JA, Teichmann SA. 2010. How do proteins gain new domains? Genome Biol 11(7):126.
McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philos Trans R Soc Lond B Biol Sci 370(1678):20140332.
Meheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc Natl Acad Sci U S A 113(13):3579-84.
Meier-Brügger M. 2002. Indogermanische Sprachwissenschaft. Berlin and New York: de Gruyter.

Melo D, Porto A, Cheverud JM, Marroig G. 2016. Modularity: Genes, Development, and Evolution. Annual Review of Ecology, Evolution, and Systematics 47(1):463-486.
Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P et al. . 2016. EBI metagenomics in 2016--an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 44(D1):D595-603.
Mitra K, Carvunis AR, Ramesh SK, Ideker T. 2013. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14(10):719-32.
Mullen J, Cockell SJ, Woollard P, Wipat A. 2016. An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. PLoS ONE 11(5):e0155811.
Nasir A, Kim KM, Caetano-Anolles G. 2014. Global patterns of protein domain gain and loss in superkingdoms. PLoS Comput Biol 10(1):e1003452.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443-53.
Nei M, Kumar S, Takahashi K. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci U S A 95(21):12390-7.
Nelson-Sathi S, List J-M, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T. 2011. Networks uncover hidden lexical borrowing in Indo-European language evolution. Proceedings of the Royal Society of London B: Biological Sciences 278(1713):1794-1803.
Nesme J, Achouak W, Agathos SN, Bailey M, Baldrian P, Brunel D, Frostegard A, Heulin T, Jansson JK, Jurkevitch E et al. . 2016. Back to the Future of Soil Metagenomics. Front Microbiol 7:73.
Newman M. 2010. Networks: An Introduction. Oxford University Press, Inc.
Nie Y, Liang J, Fang H, Tang YQ, Wu XL. 2011. Two novel alkane hydroxylase-rubredoxin fusion genes isolated from a Dietzia bacterium and the functions of fused rubredoxin domains in long-chain n-alkane degradation. Appl Environ Microbiol 77(20):7279-88.
Nozue K, Kanegae T, Imaizumi T, Fukuda S, Okamoto H, Yeh KC, Lagarias JC, Wada M. 1998. A phytochrome from the fern Adiantum with features of the putative photoreceptor NPH1. Proc Natl Acad Sci U S A 95(26):15826-30.
Nutman AP, Bennett VC, Friend CRL, Van Kranendonk MJ, Chivas AR. 2016. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. Nature 537(7621):535-538.
O'Driscoll A, Belogrudov V, Carroll J, Kropp K, Walsh P, Ghazal P, Sleator RD. 2015. HBLAST: Parallelised sequence similarity--A Hadoop MapReducable basic local alignment search tool. J Biomed Inform 54:58-64.
Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature 284(5757):604-7.
Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences11Edited by J. Thornton. Journal of Molecular Biology 273(1):349-354.
Parry LA, Edgecombe GD, Eibye-Jacobsen D, Vinther J. 2016. The impact of fossil data on annelid phylogeny inferred from discrete morphological characters. Proc Biol Sci 283(1837).
Patthy L. 1999. Genome evolution and the evolution of exon-shuffling--a review. Gene 238(1):103-14.
Pearson WR. 2013. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics Chapter 3:Unit3 1.
Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8):2444-8.

Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. 2011. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res 21(4):599-609.
Promponas VJ, Ouzounis CA, Iliopoulos I. 2014. Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey. Brief Bioinform 15(3):443-54.
Proulx SR, Promislow DE, Phillips PC. 2005. Network thinking in ecology and evolution. Trends Ecol Evol 20(6):345-53.
Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al. . 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59-65.
Reisz RR, Sues H-D. 2015. The challenges and opportunities for research in paleontology for the next decade. Frontiers in Earth Science 3(9).
Rodrigues MV, Borges N, Henriques M, Lamosa P, Ventura R, Fernandes C, Empadinhas N, Maycock C, da Costa MS, Santos H. 2007. Bifunctional CTP:Inositol-1-Phosphate Cytidylyltransferase/CDP-Inositol:Inositol-1-Phosphate Transferase, the Key Enzyme for Di-myo-Inositol-Phosphate Synthesis in Several (Hyper)thermophiles. Journal of Bacteriology 189(15):5405-5412.
Rogers RL, Hartl DL. 2012. Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster. Molecular Biology and Evolution 29(2):517-529.
Rosenberg MS. 2005. Evolutionary distance estimation and fidelity of pair wise sequence alignment. BMC Bioinformatics 6:102.
Rosenberg MS. 2009. Sequence Alignment

Methods, Models, Concepts, and Strategies. University of California Press.
Ruiz-Trillo I, Burger G, Holland PW, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet 23(3):113-8.
Salim HMW, Koire AM, Stover NA, Cavalcanti ARO. 2011. Detection of Fused Genes in Eukaryotic Genomes using Gene deFuser: Analysis of the Tetrahymena thermophila genome. BMC Bioinformatics 12:279-279.
Salim HMW, Negritto MC, Cavalcanti ARO. 2009. 1+1=3: A Fusion of 2 Enzymes in the Methionine Salvage Pathway of Tetrahymena thermophila Creates a Trifunctional Enzyme That Catalyzes 3 Steps in the Pathway. PLoS Genetics 5(10):e1000701.
Sangwan N, Lata P, Dwivedi V, Singh A, Niharika N, Kaur J, Anand S, Malhotra J, Jindal S, Nigam A et al. . 2012. Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels. PLoS One 7(9):e46219.
Sankoff D. 1972. Matching Sequences under Deletion/Insertion Constraints. Proceedings of the National Academy of Sciences of the United States of America 69(1):4-6.
Schleicher A. 1853a. Die ersten Spaltungen des indogermanischen Urvolkes. Allgemeine Monatsschrift für Wissenschaft und Literatur 3:786-787.
Schleicher A. 1853b. O jazyku litevském, zvlástě ohledem na slovanský. Čteno v posezení sekcí filologické král. České Společnosti Nauk dne 6. června 1853. Časopis Českého Museum 27:320-334.
Schleicher A. 1863. Die Darwinsche Theorie und die Sprachwissenschaft. Weimar: Hermann Böhlau.
Schmidt J. 1872. Die Verwantschaftsverhältnisse der indogermanischen Sprachen. Hermann Böhlau.
Schuchardt H. 1900. Über die Klassifikation der romanischen Mundarten. Probe-Vorlesung, gehalten zu Leipzig am 30. April 1870. Graz.

Sebe-Pedros A, Degnan BM, Ruiz-Trillo I. 2017. The origin of Metazoa: a unicellular perspective. Nat Rev Genet 18(8):498-512.
Sellers PH. 1974. An algorithm for the distance between two finite sequences. Journal of Combinatorial Theory, Series A 16(2):253-258.
Sharpton TJ. 2014. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5:209.
Smith FW, Boothby TC, Giovannini I, Rebecchi L, Jockusch EL, Goldstein B. 2016. The Compact Body Plan of Tardigrades Evolved by the Loss of a Large Body Region. Curr Biol 26(2):224-9.
Smith MR, Caron JB. 2015. Hallucigenia's head and the pharyngeal armature of early ecdysozoans. Nature 523(7558):75-8.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147(1):195-7.
Song N, Joseph JM, Davis GB, Durand D. 2008. Sequence similarity network reveals common ancestry of multidomain proteins. PLoS Comput Biol 4(4):e1000063.
Sorensen SJ, Bailey M, Hansen LH, Kroer N, Wuertz S. 2005. Studying plasmid horizontal transfer in situ: a critical review. Nat Rev Microbiol 3(9):700-10.
Southworth FC. 1964. Family-tree diagrams. Language 40(4):557-565.
Suetsugu N, Mittmann F, Wagner G, Hughes J, Wada M. 2005. A chimeric photoreceptor gene, NEOCHROME, has arisen twice during plant evolution. Proc Natl Acad Sci U S A 102(38):13705-9.
Suga H, Chen Z, de Mendoza A, Sebe-Pedros A, Brown MW, Kramer E, Carr M, Kerner P, Vervoort M, Sanchez-Pons N et al. . 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. Nat Commun 4:2325.
Suga H, Dacre M, de Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. 2012. Genomic Survey of Premetazoans Shows Deep Conservation of Cytoplasmic Tyrosine Kinases and Multiple Radiations of Receptor Tyrosine Kinases. Science Signaling 5(222):ra35.
Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A et al. . 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348(6237):1261359.
Takahashi K, Nei M. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17(8):1251-8.
Tamminen M, Virta M, Fani R, Fondi M. 2012. Large-Scale Analysis of Plasmid Relationships through Gene-Sharing Networks. Molecular Biology and Evolution 29(4):1225-1240.
Tan J, Kuchibhatla D, Sirota FL, Sherman WA, Gattermayer T, Kwoh CY, Eisenhaber F, Schneider G, Maurer-Stroh S. 2012. Tachyon search speeds up retrieval of similar sequences by several orders of magnitude. Bioinformatics 28(12):1645-1646.
Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28(1):33-6.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278(5338):631-7.
Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol 4(4):466-85.

Thomson TM, Lozano JJ, Loukili N, Carrio R, Serras F, Cormand B, Valeri M, Diaz VM, Abril J, Burset M et al. . 2000. Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene. Genome Res 10(11):1743-56.
Tordai H, Nagy A, Farkas K, Banyai L, Patthy L. 2005. Modules, multidomain proteins and organismic complexity. FEBS J 272(19):5064-78.
Torruella G, de Mendoza A, Grau-Bove X, Anto M, Chaplin MA, del Campo J, Eme L, Perez-Cordon G, Whipps CM, Nichols KM et al. . 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. Curr Biol 25(18):2404-10.
Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol Biol Evol 29(2):531-44.
Tsoka S, Ouzounis CA. 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26(2):141-2.
Van Melderen L. 2010. Toxin-antitoxin systems: why so many, what for? Curr Opin Microbiol 13(6):781-5.
Van Melderen L, Saavedra De Bast M. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet 5(3):e1000437.
Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD, Bailey MJ, Nalin R, Philippot L. 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Micro 7(4):252-252.
von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P. 2003. Genome evolution reveals biochemical networks and functional modules. Proc Natl Acad Sci U S A 100(26):15428-33.
Wake MH. 2008. Organisms and Organization. Biological Theory 3(3):213-223.
Wang W, Zhang J, Alvarez C, Llopart A, Long M. 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster. Mol Biol Evol 17(9):1294-301.
Wang WL, Xu SY, Ren ZG, Tao L, Jiang JW, Zheng SS. 2015. Application of metagenomics in the human gut microbiome. World J Gastroenterol 21(3):803-14.
Webber J. 2012. A programmatic introduction to Neo4j. Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity. Tucson, Arizona, USA: ACM. p. 217-218.
Weinreich U. 1953. Languages in contact. With a preface by André Martinet. The Hague and Paris: Mouton.
Wilson SJ, Webb BL, Ylinen LM, Verschoor E, Heeney JL, Towers GJ. 2008. Independent evolution of an antiviral TRIMCyp in rhesus macaques. Proc Natl Acad Sci U S A 105(9):3557-62.
Wu YC, Rasmussen MD, Kellis M. 2012. Evolution at the subgene level: domain rearrangements in the Drosophila phylogeny. Mol Biol Evol 29(2):689-705.
Yoon B-J. 2009. Hidden Markov Models and their Applications in Biological Sequence Analysis. Current Genomics 10(6):402-415.
Yu J, Blom J, Sczyrba A, Goesmann A. 2017. Rapid protein alignment in the cloud: HAMOND combines fast DIAMOND alignments with Hadoop parallelism. J Biotechnol 257:58-60.
Zhang H, Ning K. 2015. The Tara Oceans Project: New Opportunities and Greater Challenges Ahead. Genomics Proteomics Bioinformatics 13(5):275-7.
Zhang YJ, Li S, Gan RY, Zhou T, Xu DP, Li HB. 2015. Impacts of gut bacteria on human health and diseases. Int J Mol Sci 16(4):7493-519.

Zhao Y, Tang H, Ye Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28(1):125-126.
Zmasek CM, Godzik A. 2011. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol 12(1):R4.

# ANNEX 1

*Article n°10: Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery*

*(submitted to the journal "Genome Biology and Evolution")*

**Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery**

Guillaume Bernard[1], Jananan Pathmanathan[1], Romain Lannes[1], Philippe Lopez[1] and Eric Bapteste[1,*]

[1]Sorbonne Universités, UPMC Université Paris 06, Institut de Biologie Paris-Seine (IBPS), Paris F-75005, France.

*Author for Correspondence: Eric Bapteste, Institut de Biologie Paris-Seine (IBPS), UPMC Université Paris 06, Paris, France, Phone :+33144272164, E-mail: eric.bapteste@upmc.fr

**Abstract**

Microbes are the oldest and most widespread, phylogenetically and metabolically diverse life forms on Earth. However, only relatively recently have they been discovered and has their diversity started to become seriously investigated. For these reasons, microbial studies that unveil novel microbial lineages and processes affecting or involving microbes deeply (and repeatedly) transform knowledge in biology. Considering the quantitative prevalence of taxonomically and functionally unassigned environmental sequences in metagenomics datasets, and that of uncultured microbes on the planet, we propose that unraveling the microbial dark matter should be identified as a central priority for biologists. Based on former empirical findings of microbial studies, we sketch a logic of discovery with the potential to further highlight the microbial unknowns.

**Keywords:** metagenomics, eukaryogenesis, microbial evolution, tree of life, web of life, CPR bacteria

## Introduction

Microbial studies are fascinating. Not only their findings can deeply transform knowledge in a broad range of scientific fields (from evolutionary biology to zoology and medical and environmental sciences), but also, whereas philosophers of sciences debate whether there is such thing as a logic of scientific discovery (Schickore 2014), microbial studies provide biologists with a set of empirical rules to enhance one's chances to discover novel and unexpected life forms. This unique potential of microbial studies to reshape knowledge has been recognized relatively recently. If the laymen

nowadays appreciate that microbes impact our everyday life (i.e. via their fermentative roles in food production), and know that microbes also impacted our recent human histories (i.e. via their contribution to major pandemics (Diamond 1997)), from a scientific perspective, microbes are nonetheless rather novel objects of studies. There are both technical and conceptual reasons for this late yet broad recognition of microbes, as we will highlight below, while providing an empirical recipe for further insights into the microbial dark matter.

In 1619, the famous astronomer Galileo, whose observations of the moons of Jupiter had threatened the geocentric theory, modified a telescope to magnify nearby terrestrial objects. Although he clearly was a revolutionary thinker, he found these observations of the minute world of limited interest, and, only 6 years later, did his friends name *microscopio* the strange inverted telescope Galileo had invented (Falkowski 2015). By contrast, Robert Hooke, an English polymath scientist, and, later, Anton van Leeuwenhoek, who did not belong to the academic world, were much more excited by describing their microscopic observations. In 1671, van Leeuwenhoek, who had substantially changed the design of the microscope to enhance its magnifying power, initiated a series of striking findings: microscopic lifeforms are abundant and everywhere to be seen. Microbes, who had populated Earth for over 3.5 billion years, were for the first time exposed to the human eye (Falkowski 2015). Both a technical progress and an uncommon ability to delve into an unseen world were critical components of that progress. However, since biological theory at the time considered the living world was distributed into two major groups: plants and animals, van Leeuwenhoek naturally assumed he was observing populations of minute animals (with tiny organs), when microbes were mobile, rather a new kind of living beings. In that sense, the unveiled microbiological world was first rationalized in ways that fit within pre-existing theoretical categories derived from the known living world. Importantly, neither Hooke nor van Leeuwenhoek had immediate scientific successors. Arguably, it took another 200 years (Falkowski 2015), and several novel conceptual and technological developments to

formulate an issue, currently at the forefront of microbial studies: « is it possible that unknown microorganisms, with different properties than those currently associated with the known living world, are thriving in nature? ».

The potential theoretical importance of such 'known unknowns' and even 'unknown unknowns' of the microbial world (e.g. unknown genes, genomes, functions, organisms, processes and communities associated with uncultured microbes and virus), that were often popularized under the catch-phrase 'microbial dark matter', should not be underestimated. Much of the extant knowledge in biology, i.e. about biological entities and biological processes, heavily relies on analyses conducted on macro-organisms and on cultured microbes. Yet, 99% of the microbial diversity is impossible to culture (Staley and Konopka 1985). Unraveling the microbial dark matter could thus have led to two (nonexclusive) types of observations. Either the discovery of hidden microbes will show that microbes unveiled from the microbial dark matter are comparable in terms of genetic diversity, ecological roles, abundance, evolutionary history and affected by processes similar to those affecting cultured microbes, in which case our current knowledge of microbes is representative of what is really happening in nature (e.g. we will simply find more of what we already knew by mining the microbial world) ; or the microbial dark matter will prove to host entities and processes that differ from those already described, with the major consequence that scientific knowledge will not only need to be completed but also corrected as microbiologists gain access to this still hidden microbial world in order to consider new phenomena, poorly explained in extant theories. Such significant theoretical transformations have arguably occurred when i) microbiologists looked for life in extreme environments, ii) detected life under unexpected forms, and iii) unveiled new processes involving microbes, which allows us to stress some key features for the success of a scientific research oriented toward the discovery of microbiological novelty.

**Searching life in extreme environment: a few lessons**

The developments of molecular markers and sequencing techniques were instrumental for the discovery of extremophiles. By unveiling the archaea, a novel early branching Domain of life, possibly sister-group to eukaryotes, Carl Woese's phylogenetic studies of the 16SRNA revolutionized the views on the entire biological world (Woese and Fox 1977; Woese, et al. 1990). Woese argued that, rather than being partitioned into two major groups, the eukaryotes and the prokaryotes, the living world encompassed a much broader microbial diversity, justifying its classification into 3 Domains of life. Subsequently, Woese and his colleagues (referred to as 'the Woese army' by Lynn Margulis (Doolittle 2013)) actively promoted this position, bringing the newly termed 'archaea' into full light, while intending to ban the use of the 'older' term 'prokaryotes' (Pace 2006). Importantly, this comparative approach of molecular phylogenetics was later coupled to a phase of exploratory science (Waters 2007) (i.e. a strategy of data mining, which goes from the data to the hypotheses (Burian 2013), in strong contrast with the then classic hypothetico-deductive strategy, which operated from the hypotheses to the data, heralded by Karl Popper). Since exploratory science is not first aimed at rejecting (or confirming) pre-established hypotheses (thus deepening current knowledge), it can potentially produce novel, unexpected knowledge, or simply fail, making the financial and scientific investment in exploratory studies especially risky. Fortunately, the pioneering approach, largely based on the development of metagenomics, which bypassed the need for culture studies, thereby lifting a blind spot imposed by culture-based investigation to comparative analyses, produced remarkable results when microbiologists turned their eyes to extreme regions (in terms of temperature, pH, pressure, mineralization, radiations) that many considered *a priori* devoid of life (Pikuta, et al. 2007). The seemingly counter-intuitive idea to sample lifeforms in environments hostile to life unveiled a broad diversity of extremophiles in the 3 Domains. Microbiologists realized that life was possible at temperature > 113-200 Celsius degree, at negative pH (!) and at pH> 11, at pressures exceeding 1200

atmospheres ; that microbes could be resurrected after 20-40 millions of years of dormancy, survive 2.5 years of travel in space, and thrive within rocks as well as in the terrestrial stratosphere (at > 44km of altitude) (de los Rios, et al. 2003; Pikuta, et al. 2007)(see for example: https://www.slideshare.net/AnjaliMalik3/extremophiles-imp-1). They concluded nonetheless that some of these statistics were so unexpected that Pikuta *et al.* (Pikuta, et al. 2007), summarizing the ongoing knowledge on extremophiles, drew too short axes for temperature, pH and salinity on plots showing the physico-chemical conditions compatible with life. Some environmental microbes were definitely outliers with respect to the majority of known creatures. This counter-intuitive search for extremophiles likely reaches his summit in astrobiological studies, which search for life beyond Earth, seeking to define biomarkers in exoplanetary analogs and to train to detect these biomarkers in regions of the universe that currently fit the minimal requirements for life in C, H, N, O, P, S, liquid water and energy (Olsson-Francis and Cockell 2010). No one knows whether extraterrestrial microbes will ultimately be discovered this way, but, at least, ironically terrestrial microbes have increased chances to spread in space (Checinska, et al. 2015).

**Searching life under unusual forms: a few lessons**

In as much as metagenomics enhance microbial dark matter studies, e.g. by unraveling extremophiles, it also raises issues, since metagenomics has its own blind spots. The selection of samples, markers and the many filtering decisions and heuristics in the subsequent bioinformatic treatments imposed by the wealth of metagenomic data, as well as the increased standardization of the methods and questions of metagenomic studies (a logical scientific development for a comparative science (Vigliotti, et al. 2017)) raise the risk that the most unexpected of life forms, even if already sequenced, remain drowned under this deluge of data. This risk has notorious roots: our observations are strongly constrained by

what our theory makes us prone to expect, and therefore by former perspectives informing various criteria in the sampling process. This limit is obvious in the process of size-fractioning associated with metagenomics analyses, such as the one conducted in the Tara expedition, which *a priori* optimized the net sizes of its filter to capture different taxa of marine microbes (Karsenti, et al. 2011). This procedure entails the inherent risk that important players of the microbial world may be overlooked if their sizes do not satisfy these filtering conditions. For example, 10 years ago, few (or even no) microbiologists nor virologists would have assumed that bacteria smaller than 0.2 microns and viruses larger than 0.2 microns existed (Council 1999). This view radically changed with the discovery of ultra-small bacteria, aka nanoorganisms, such as the CPR in 2015 (Brown, et al. 2015) or some DPANN in 2010 (Baker, et al. 2010), and with the discovery of giant viruses, such as Mimiviridae, in 2003 (La Scola, et al. 2003). These two taxa are now found in diverse environments, albeit at low abundance. Moreover, CPR discovery further required acetate amendment, i.e. a technologically-induced modification of the environment during the sampling process (Brown, et al. 2015). CPR are remarkably phylogenetically diverse, representing up to 15% of the bacterial domain, and present an unusual biology (i.e. 16SRNA with insertion, lack of essential metabolic genes), which suggests that all CPR depend on other life forms. Mimivirus biology is not less striking. In particular, they are hosts to yet another new kind of viruses : virophages, i.e. viruses of giant viruses (Boyer, et al. 2011). The phylogenetic position of these relatively newcomers, especially regarding how deep CPR and giant viruses branch (if they do) with respect to the other Domains of life, is heavily debated (Colson, et al. 2012; Hug, et al. 2016; Moreira and Lopez-Garcia 2015). Such debates illustrate that attempts to establish novel groups inevitably (and logically) arise resistances, but no one questions that an accurate picture of the microbial world and its evolution can any longer satisfactorily be achieved without including nanoorganisms and giant viruses.

Metagenomics has not merely unraveled new microbial lineages, it has also reported new gene families (Lok 2015) and unusual gene forms. In principle, newly sequenced environmental genes could fall into one of 4 groups (Figure 1).

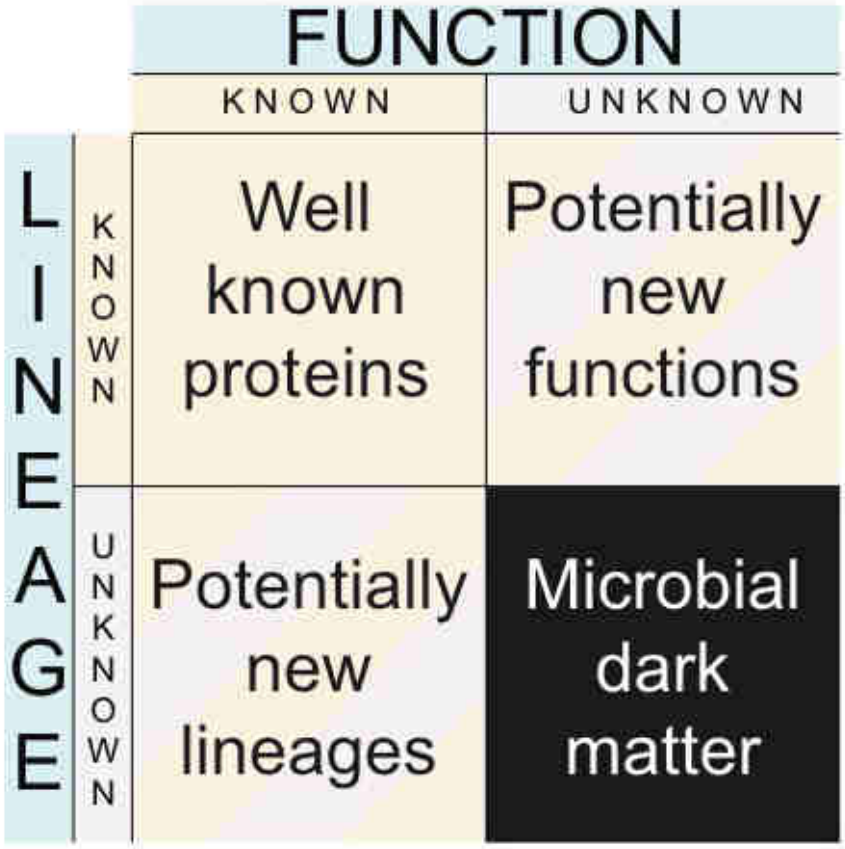



**Figure 1: Four types of environmental sequences.** Environmental sequences can be classified based on their taxonomical annotation (horizontal line) and their functional annotation (vertical column), which defines 4 categories. The cells in purple and black correspond to categories that are not readily explained based on current biological knowledge.

The *in silico* functional and taxonomical annotations of environmental genes using existing ontologies (here, applied to 339 metagenomes (Fondi, et al. 2016), sampling a diversity of environments, i.e. soil, seawater, inland-water, wastewater, host, air, bioremediation, biotransformation, and sludge waste) indicates that most environmental genes have unknown functions, and belong to uncharacterized microbial lineages (Figure 2). In fact, when the minimum %ID threshold is set at 95%, >50% of these genes are neither functionally nor taxonomically annotated,

and at 50%, >30% of these genes are neither functionally nor taxonomically annotated, which stresses the genuine abundance of microbial dark matter in metagenomic data.

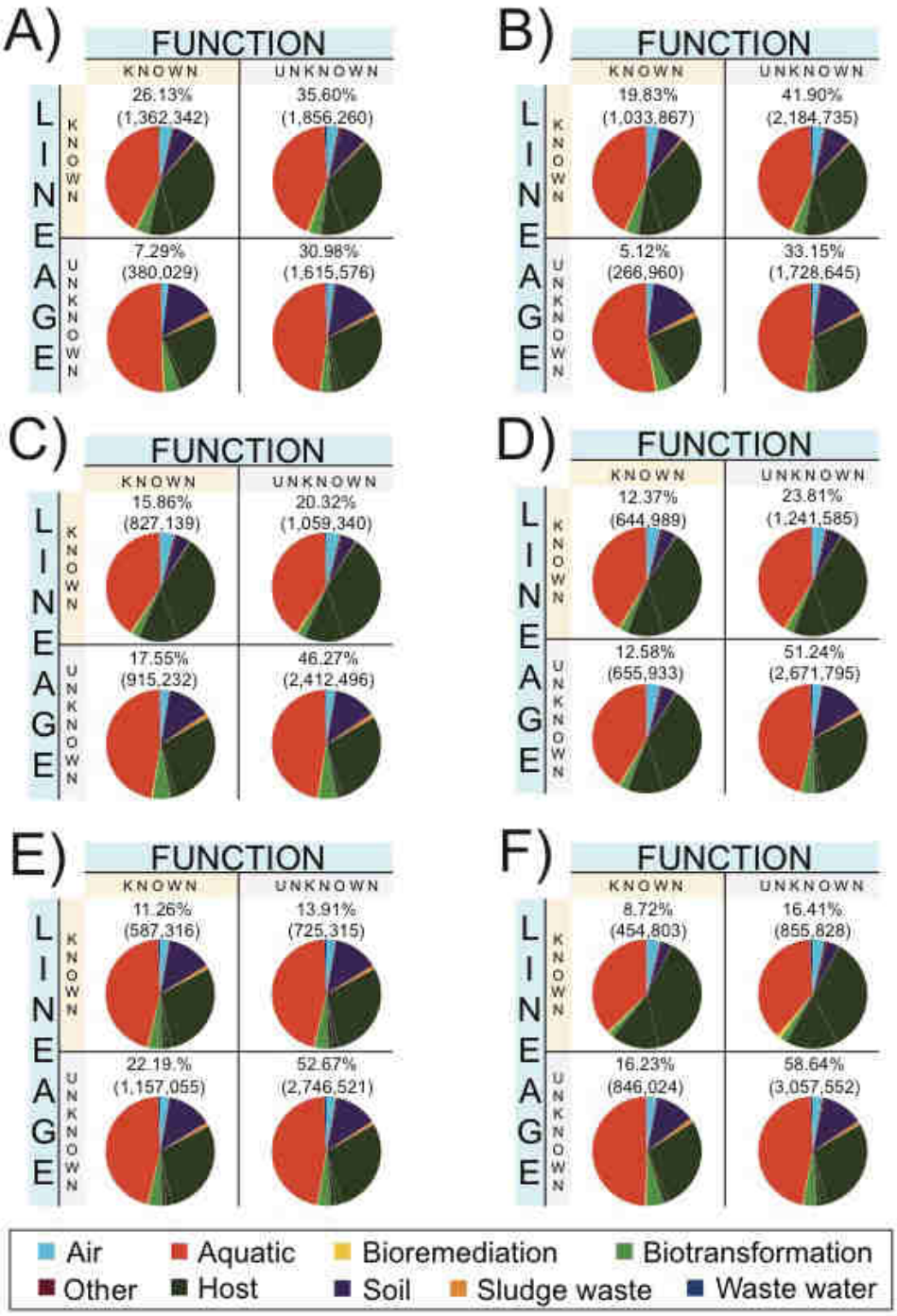



**Figure 2: Microbial dark matter across a diversity of environmental samples.** Proteins inferred (with FragGeneScan (Rho, et al. 2010)) based on Metagenomic sequences from (Fondi, et al. 2016), clustered based on their taxonomy (using MEGAN 6 (Huson, et al. 2016)) and functional (using

EggNOG-mapper (Huerta-Cepas, et al. 2017)) annotation. The pie charts represent the proportion of proteins from each type of environment. The taxonomy annotation was performed using three minimum percentage of identity: 50% (panels A and B), 85% (panels C and D) and 95% (panels E and F). In panels A, C and E the proteins were clustered based on their functional annotation including the category S ('Function unknown'). Panels B, D and F were clustered with the exclusion of the category S.

Bioinformatics developments are currently designed to associate these unknown genes to reference gene families. For example, the search for highly divergent homologs using sequence similarity networks (Lopez, et al. 2015) highlighted that a large majority of the ancient gene families that are well-conserved in cultured microbes have extremely divergent homologs in nature. Lopez *et al.* proposed that at least some of these very divergent homologs might sign the existence of deep branching yet unseen major divisions of life (Lopez, et al. 2015). Discovering environmental deeper lineages, branching below the currently recognized prokaryotic domains, could re-open the debate on the number of Domains of life, questioning our fundamental knowledge in terms of biological classifications and regarding early life evolution. Bioinformatics studies however need to be complemented by another type of experimental evidence, i.e. individual sequences of genomes from putative very early branching microbes, or even isolations of these organisms. Thus, so far, despite the actual high number of environmental 'known unknowns' no major scientific journal has yet been convinced that enough evidence for new candidate Domains is available (Parks, et al. 2017).

**Microbial processes as a yet unexhausted source of knowledge**

At the same time that microbes left the realm of microbial dark matter, our knowledge on processes involving or affecting microbes evolved substantially. The focus on interactions and the use of networks rather than trees to frame microbial studies is emerging as a major trend. It is becoming obvious that simple tree-based models, aiming at reconstructing the divergence of lineages from a last common ancestor, are not fully doing justice to the diversity and complexity of the processes explaining microbial evolution. Introgressive processes such as lateral gene transfer stress the

collective dimension of microbial evolution (Bapteste, et al. 2012). Likewise, the discovery of environmental microbes with genuinely incomplete genomes (i.e. lacking essential genes) and of syntrophic consortia insists on the importance of metabolic, ecological, and evolutionary scaffolding in the microbial world (Brown, et al. 2015; Caporael, et al. 2013; DeLong 2007; Ereshefsky and Pedroso 2015; Morris, et al. 2012; Sachs and Hollowell 2012). The claim that in nature microbes depend on other microbes to survive, contrasts strongly with the notion that natural selection ultimately favors individual optimized lineages via the success of the fittest cells amongst large and phylogenetically homogeneous microbial populations. It matches however well with the empirical observation that pure culture fails for most microbes (Staley and Konopka 1985), and in fact provides an explanation for this great plate anomaly. Microbes belong to collectives rather than they live alone. Other striking interactions are also unveiled as scientists dig further into the microbial world. For example, unheard forms of communication impact microbial and viral population dynamics (Erez, et al. 2017). Microbiomes and their hosts co-construct a broad range of animal and plant phenotypes (Gilbert, et al. 2015), to the point that some propose to introduce holobionts (the emergent associations of hosts and microbes) as a novel kind of central evolutionary player (Bordenstein and Theis 2015; Moran and Sloan 2015; Theis, et al. 2016). At an even broader scale, in the environment, microbes, most of which are unknown, are now assumed to affect the geochemical processes that shape our planet (Guidi, et al. 2016) and, by a process called niche construction (Laland, et al. 2016), these microbes are considered likely to impact ecosystems and the future of life. All these processes (lateral gene transfer, scaffolding, communication, microbial co-construction, and niche construction), while widespread in the microbial world, are still rather peripheral in biological explanations. Introducing the processes of microbial dark matter within biological theory thus requires revising the relative priority currently attributed to concepts in scientific explanations, which is likely to be a slow and tedious epistemic process. For example, prokaryotic biology, especially when considering microbiomes, appears in fact so different from the biology of model eukaryotic organisms that several

evolutionary biologists and theoreticians have independently suggested that key aspects of the classic Darwinian theory and of the Modern Synthesis would have been very different had microbial studies been more central during the early development of the evolutionary theory. Others however disagree that the structure and content of the evolutionary theory requires to be reshaped, even in the light of this new knowledge in microbiology (Wray 2014). Yet, debates around the gene content, nature and phylogenetic position of Asgard archaea (Saw, et al. 2015; Zaremba-Niedzwiedzka, et al. 2017) (Da Cunha, et al. 2017) powerfully illustrates that an enhanced knowledge of the microbial dark matter has unquestionably the potential to transform central elements in the evolutionary theory. If Asgard archaea, currently only known via assemblies of environmental reads, prove to be sister-groups of eukaryotes, this should (at least) impact the very notion of a tree of life, the number of Domains of life, and, depending on the intimate structural biology and metabolisms of these Asgard, it will also help testing amongst competing hypotheses for the origin of eukaryotes (Koonin 2015; Sousa, et al. 2016).

**Conclusion**

The discovery of an increasing number of types of microbes has consistently shown that our planet hosts microbes with properties that were not simply identical to the ones formerly described. Studies of the microbial dark matter have brought forward the existence of novel entities (e.g. nanoorganisms, giant viruses, virophages, etc.) and novel relationships within the microbial world (e.g. viral languages, high divergence, scaffolding, etc.). This formerly dark microbial matter has not been unraveled randomly. To sum up its logic of discovery, it has required : to think outside the box (e.g. Woese's invention of a novel Domain ), to take scientifically and financially risky decisions (e.g. sampling sites where life was unlikely), to develop novel methods pushing back the limits of detection (e.g. better microscopes, inclusive networks), to prepare one's mind to detect unknowns and unexpected forms (e.g. biomarkers), to identify and to seek to explain anomaly (e.g. the great plate count anomaly), to

change perspectives (e.g. embracing the notion of nanoorganisms, or of multiple prokaryotic domains), to use analogies to uncover new microbial systems (e.g. for the study of extremophiles in space), to purposely depart from normal scientific practices and background knowledge (e.g. network studies of divergent gene forms, exploration of increasingly extreme environments), to be willing to create novel groups (e.g. Archea, CPR, Mimiviridae,…), and finally to convince (e.g. by banning competing notions, or by establishing new attractive fields, such as metagenomics). Indeed, many of these discoveries presented in this work generated resistances. These resistances are perfectly explainable. Unraveling the unknown is especially difficult, because although we could empirically sketch a logic of scientific discovery, at the time each novel finding was made, their inventors could not yet rely on a standard method but essentially they had to convince the rest of the community that both their unusual approaches and finding were relevant. Convincing its own peers is finally essential, and possibly one of the largest and commonest challenge for microbial dark matter studies, and this seems especially difficult even for creative outsiders. Van Leeuwenhoek's pioneering example offers indeed a great reminder that extraordinary results can easily be forgotten.

**References**

Baker BJ, et al. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A 107: 8806-8811. doi: 10.1073/pnas.0914470107

Bapteste E, et al. 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc Natl Acad Sci U S A 109: 18266-18272. doi: 10.1073/pnas.1206541109

Bordenstein SR, Theis KR 2015. Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. PLoS Biol 13: e1002226. doi: 10.1371/journal.pbio.1002226

Boyer M, et al. 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci U S A 108: 10296-10301. doi: 10.1073/pnas.1101118108

Brown CT, et al. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523: 208-211. doi: 10.1038/nature14486

Burian RM. 2013. Exploratory Experimentation. In. Encyclopedia of Systems Biology Springer New York. p. 720-723.

Caporael L, Griesemer J, Wimsatt W. 2013. Scaffolding in Evolution, Culture, and Cognition. MIT Press.

Checinska A, et al. 2015. Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. Microbiome 3: 50. doi: 10.1186/s40168-015-0116-3

Colson P, de Lamballerie X, Fournous G, Raoult D 2012. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology 55: 321-332. doi: 10.1159/000336562

Council NR editor.; 1999 Washington, DC.

Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P 2017. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet 13: e1006810. doi: 10.1371/journal.pgen.1006810

de los Rios A, Wierzchos J, Sancho LG, Ascaso C 2003. Acid microenvironments in microbial biofilms of antarctic endolithic microecosystems. Environ Microbiol 5: 231-237.

DeLong EF 2007. Microbiology. Life on the thermodynamic edge. Science 317: 327-328. doi: 10.1126/science.1145970

Diamond J. 1997. Guns, Germs, and Steel: The Fates of Human Societies: W. W. Norton.

Doolittle WF 2013. Carl R. Woese (1928-2012). Curr Biol 23: R183-185.

Ereshefsky M, Pedroso M 2015. Rethinking evolutionary individuality. Proc Natl Acad Sci U S A 112: 10126-10132. doi: 10.1073/pnas.1421377112

Erez Z, et al. 2017. Communication between viruses guides lysis-lysogeny decisions. Nature 541: 488-493. doi: 10.1038/nature21049

Falkowski P 2015. Leeuwenhoek's Lucky Break. Discover: 1-5.

Fondi M, et al. 2016. "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of Gene Sharing in Environmental Samples through Network Analysis. Genome Biol Evol 8: 1388-1400. doi: 10.1093/gbe/evw077

Gilbert SF, Bosch TC, Ledon-Rettig C 2015. Eco-Evo-Devo: developmental symbiosis and developmental plasticity as evolutionary agents. Nat Rev Genet 16: 611-622. doi: 10.1038/nrg3982

Guidi L, et al. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature 532: 465-470. doi: 10.1038/nature16942

Huerta-Cepas J, et al. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. doi: 10.1093/molbev/msx148

Hug LA, et al. 2016. A new view of the tree of life. Nat Microbiol 1: 16048. doi: 10.1038/nmicrobiol.2016.48

Huson DH, et al. 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol 12: e1004957. doi: 10.1371/journal.pcbi.1004957

Karsenti E, et al. 2011. A holistic approach to marine eco-systems biology. PLoS Biol 9: e1001177. doi: 10.1371/journal.pbio.1001177

Koonin EV 2015. Archaeal ancestors of eukaryotes: not so elusive any more. BMC Biol 13: 84. doi: 10.1186/s12915-015-0194-5

La Scola B, et al. 2003. A giant virus in amoebae. Science 299: 2033. doi: 10.1126/science.1081867

Laland K, Matthews B, Feldman MW 2016. An introduction to niche construction theory. Evol Ecol 30: 191-202. doi: 10.1007/s10682-016-9821-z

Lok C 2015. Mining the microbial dark matter. Nature 522: 270-273. doi: 10.1038/522270a

Lopez P, Halary S, Bapteste E 2015. Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol Direct 10: 64. doi: 10.1186/s13062-015-0092-3

Moran NA, Sloan DB 2015. The Hologenome Concept: Helpful or Hollow? PLoS Biol 13: e1002311. doi: 10.1371/journal.pbio.1002311

Moreira D, Lopez-Garcia P 2015. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? Philos Trans R Soc Lond B Biol Sci 370: 20140327. doi: 10.1098/rstb.2014.0327

Morris JJ, Lenski RE, Zinser ER 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3. doi: 10.1128/mBio.00036-12

Olsson-Francis K, Cockell CS 2010. Experimental methods for studying microbial survival in extraterrestrial environments. J Microbiol Methods 80: 1-13. doi: 10.1016/j.mimet.2009.10.004

Pace NR 2006. Time for a change. Nature 441: 289. doi: 10.1038/441289a

Parks DH, et al. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol.

Pikuta EV, Hoover RB, Tang J 2007. Microbial extremophiles at the limits of life. Crit Rev Microbiol 33: 183-209. doi: 10.1080/10408410701451948

Rho M, Tang H, Ye Y 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 38: e191. doi: 10.1093/nar/gkq747

Sachs JL, Hollowell AC 2012. The origins of cooperative bacterial communities. MBio 3. doi: 10.1128/mBio.00099-12

Saw JH, et al. 2015. Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. Philos Trans R Soc Lond B Biol Sci 370: 20140328. doi: 10.1098/rstb.2014.0328

Schickore J. 2014. "Scientific Discovery". In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy.

Sousa FL, Neukirchen S, Allen JF, Lane N, Martin WF 2016. Lokiarchaeon is hydrogen dependent. Nat Microbiol 1: 1-3.

Staley JT, Konopka A 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39: 321-346. doi: 10.1146/annurev.mi.39.100185.001541

Theis KR, et al. 2016. Getting the Hologenome Concept Right: an Eco-Evolutionary Framework for Hosts and Their Microbiomes. mSystems 1. doi: 10.1128/mSystems.00028-16

Vigliotti C, Lopez P, Bapteste E. 2017. Microbial diversity studies: the (paradoxical) challenge to have a broad view with metagenomics. In: Maurel PGMC, editor. Evolution and Biodiversity: ISTE Editions. p. in press.

Waters CK 2007. The nature and context of exploratory experimentation: an introduction to three case studies of exploratory research. Hist Philos Life Sci. 29: 275-284.

Woese CR, Fox GE 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74: 5088-5090.

Woese CR, Kandler O, Wheelis ML 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87: 4576-4579.

Wray GAea 2014. Does evolutionary theory need a rethink? No, all is well. Nature 514: 161-164.

Zaremba-Niedzwiedzka K, et al. 2017. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541: 353-358. doi: 10.1038/nature21031

**Acknowledgements**

R.L, J.P, G.B and E.B. are funded by the European Research Council (FP7/2017-2013 Grant Agreement #615274). We thank Dr. Karen Olsson-Francis and Dr. Lucie Bittner for stimulating discussion.

# ANNEX 2

*Article n° 11: Major protein-coding innovation by gene remodeling in the animal kingdom*

*This article is in preparation and will be submitted to the journal "Molecular Biology and Evolution".*

# Major protein-coding innovation by gene remodeling in the animal kingdom.

**Moran RJ, Pathmanathan JS, Bapteste E, Lopez P, Creevey CJ, Sui Ting K, McInerney JO and O'Connell MJ**

***In Preparation***

## 1 Introduction

Understanding the origin of genetic and functional novelty across the *Metazoa* (i.e. all multicullar animals) is a fundamental problem in modern biology. Although there has been a lot of research carried out investigating the role of tree-like mechanisms (such as gene duplication and loss) in creating novel genes, little is understood about the role of non-tree-like mechanisms (such as gene remodelling) in creating novel genes. By performing analyses of sequence similarity networks (SSNs) and phylogenomics across 63 Metazoan genomes, we assess the contribution of two interrelated mechanisms of protein coding evolution to the diversity of animal protein coding gene families. Firstly, we use a novel phylogenetic approach to plot gene duplication, loss and point mutation in the evolution of gene families on the animal tree of life. Secondly, we use a network approach to study novel gene family evolution by gene remodeling. In this article, we focus on gene fusion/fission as the mechanism of gene remodeling. We show that gene remodeling is present right across animal life, and is a major source of novel protein-coding gene family genesis. In general, we see that the rate and specific mechanism involved in the generation of novel protein-coding gene varies significantly depending on the lineage. For example, the *Deuterostomia* as compared to the *Protostomia* have a much higher incidence of gene remodeling (specifically fusion). Bilaterian animals are either deuterostomes (in development first opening becomes the anus) or protostomes (first opening becomes the mouth). On a broader scale, this work provides insight into how novel protein-coding gene families have evolved through time and contributed to the diversity we see across the *Metazoa*.

Genome sequencing is revealing the existence of an enormous repertoire of protein coding genes in animal genomes (Consortium 1998, Adams et al. 2000, Holt et al. 2002, Consortium 2004). Recombinogenic processes and transcription-mediated read-thorough create remodeled genes that likely contribute novel protein coding genes to genomes (Zhou and Wang 2008, Kaessmann et al. 2009, Wu et al. 2013, Agaram et al. 2015). Indeed, given the diversity of protein domain combinations, it is reasonable to assume that protein remodeling has made a contribution to the whole-organism diversity observed in *Metazoa*. Well-understood and well-characterized examples of gene remodeling include *Jingwei*, a remodeled *Drosophila* gene derived 2 MYA from a fusion of a retrotransposed copy of an *Adh* locus and the 5' end of the *yande* gene. The novel phenotype conferred by the resultant remodeled protein is a new specificity towards long-chain primary alcohols (Wang et al. 2000, Long et al. 2003). The *Kua-UEV* fusion gene in human is remodeled from two adjacent genes (*Kua* and *UBE2V1*)(Thomson et al. 2000) . The functional impact of this remodeling event is that the ubiquitin conjugating enzyme *UBE2V1,* which normally has activity localized solely to the nucleus, now has novel activity localized to the cytoplasm (Thomson et al. 2000). While we understand that mutational molecular clocks tend to tick with complex but increasingly well-understood rates (Lynch 2010), we have not yet been able to understand how, when, at what rate, and to what extent remodeling has impacted on animal proteomes. Until recently, we have not had available a comprehensive dataset of proteins to determine the details and genome wide impact of gene remodeling processes.

Animals exhibit significant diversity in development, morphology and indeed body plan. We define major transitions as events that have allowed a lineage to radically change their environment, a biological function, and/or phenotype. From studies such as McLean et al. (2011) (on the lack of penile spines in Humans) and D'Apice et al. (2004) (on the cause of Progeria) we know that even small changes at the genetic level can cause major phenotypic effect (D'Apice et al. 2004, McLean et al. 2011). While phenotypic transitions in the Metazoa such as the emergence of the mesoderm, mineralized skeleton and chordate have been well documented (Bell 2015), the underlying genetic changes contributing to these major phenotypic transitions are generally quite poorly understood. Major questions in theoretical evolutionary biology that are addressed in this article include: are these major phenotypic

transitions fuelled by the emergence of novel protein coding gene families, and, has gene remodeling contributed to these novel families at a steady rate over time or in a punctuated manner across the fossil record.

# 2 Methods

### 2.1 Data acquisition

We retrieved our data from the OMA database (Altenhoff et al. 2014). We only used coding DNA sequences (CDS). All data was passed through our initial quality filter (Section 2.2). From the data that passed this filter, we took representatives across the *Metazoa* for each major phylum of the tree (Figure 1). Finally, form all the genomes that passed the quality filter, we selected 63 of these Metazoan species as representatives of all major groupings within the *Metazoa* (Figure 2). Some pre-computed Smith-Waterman alignments were available for download for ~50% of the species comparisons (Altenhoff et al. 2014) and we used these pre-computed alignments in our analysis.

### 2.2 Quality check filter

Data quality is of paramount importance in any analysis. We carefully researched all aspects of data quality. For example, commonly used statistics such as contig/scaffold N50 and fold coverage are not good measures of data quality as they are not always easily accessible and do not directly correlate with data quality (Bradnam et al. 2013). Therefore, we used the raw sequence data and from it extrapolated our own statistics/metrics and subsequently assigned data quality measures. This was challenging due to the sheer amount of storage needed for the data (data for a single species was in the region of many Gigabytes) but we had some working solutions in place. However, even more challenging was the acquisition of raw genome data. Although there are hundreds of animal genomes sequenced, the raw data is not always available. In addition to these challenges we discovered that our assumption that high sequencing quality would correlate to high protein coding data quality was simply not supported by the data – i.e. high-quality sequences can be poorly annotated. Therefore, we explored an alternative procedure for assessing data quality.

We developed the following procedure to provide us with the necessary metrics on data quality. The procedure involves two data filters - both of which are based on sets of protein coding genes that are present across: (1) all of life, and (2) all of Metazoa. The number of conserved protein coding genes present in a genome acts a proxy for the quality of that genome and its annotation. A genome with all (or at least 70% of -) the orthologs is deemed high quality and genomes with large portions of missing orthologs are lower quality (*i.e.* <70% of orthologs). The first filter uses a gene set of 412 orthologs that are found across all Metazoa (Powell et al. 2012). Using a reciprocal BLASTp (Altschul et al. 1990b), we identify the distribution of the 412 ortholog families in each genome.

**Figure 1: Relationships and divergence times of the 63 Metazoan species sampled.**

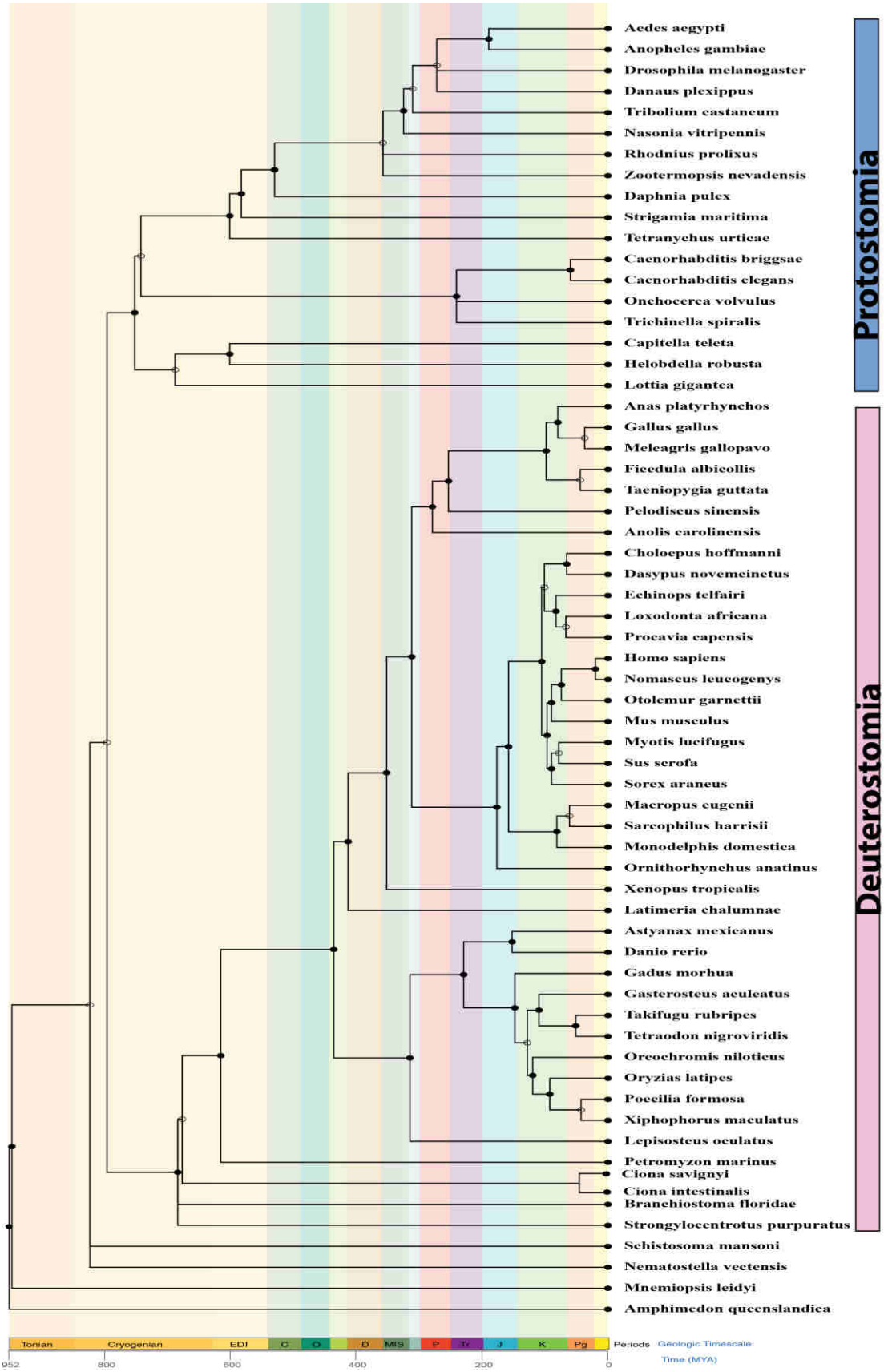


Figure 1: The phylogenetic relationship and dates of all 63 Metazoa represented in our dataset are shown as generated using TimeTree *(Kumar et al. 2017)*. Branches in the phylogeny are scaled according to divergence times. The geological periods from the Tonian period (~952 MYA) to present day are color-coded and are scaled in Millions of Years.

**Figure 2 The distribution of our 63 genomes into their major taxonomic groupings across the Metazoa**

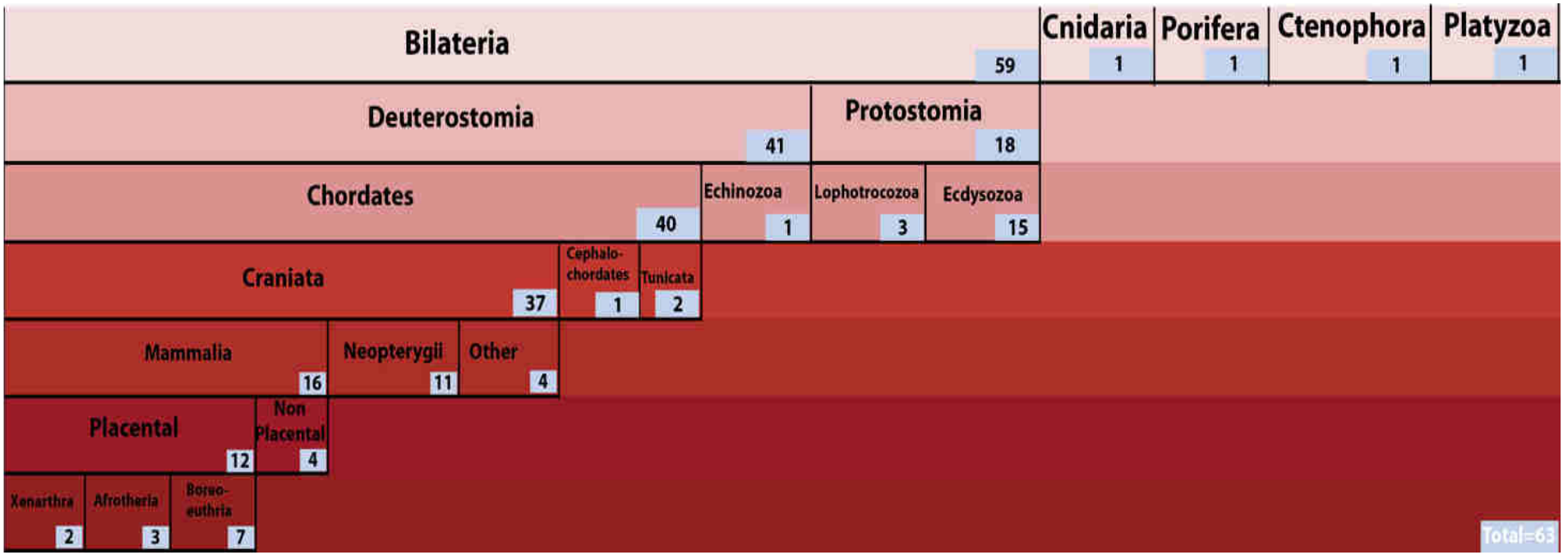



Figure 2: The distribution of the 63 Metazoa used in this study into their major taxonomic groupings. The numbers provided are post-filtering and are ordered by major taxonomic category. Each row represents a level of taxonomic grouping on the tree. Each subsequent row (from top to bottom) describes the more minor taxonmic groupings within the major taxonmoic groupings. For example, Bliateria is a major taxonomic grouping that can be further divded into either Deuterostomia or Protostomia.

The second filter is the stricter of the two filters. The second filter uses the 40 highly conserved orthologous families that are found in all of life as the query in the homology search (Ciccarelli et al. 2006). As there are 412 genes in the first filtering step, it is expected that there will be a reasonable level of variation in distribution (in terms of the number matches in each genome), this is useful in that we can then rank the genomes by quantity of genes present. However, this is not a strict filter and by random chance some genes be missed even in higher quality genomes. The 40 highly conserved genes are present in all of life and are highly conserved in their sequences as well as in their distribution. Therefore, we can query one of the 40 genes against a particular species to determine if it is present, and we can assess the level of conservation of the orthologous sequence. This combination of filters, carried out in this order, allowing the highest quality data possible to be retained while also accounting for variation in sequence and annotation quality.

**2.3 Metazoan topology and dating**

Sampling across Metazoa is guided by phylogeny - this is particularly challenging as there is no agreed species phylogeny for the Metazoa. The large number of alternative topologies, and therefore the large number of contentious yet critical nodes, increases the number of permutations necessary to represent the evolution of the animal kingdom. In addition, some lineages of metazoan life are understudied and poorly sampled (e.g. *Porifera*) whilst others (e.g. the mammals) are well studied and densely sampled. To this end, a dataset capable of addressing the major transitions in animal evolution was constructed with multiple representatives from before and after each transition.

We used the topology and node dates from TimeTree (Kumar et al. 2017) (latest access on 23/09/2016 ). In total 51/63 of the species in our sampling were represented in TimeTree. For the remaining 12 species that were not present in the TimeTree database, we searched for their closest neighbours (sister taxa) in TimeTree. In most cases the time estimate was available for a member of the same genus.

**2.4 Gene gain and loss analysis – OMA**

The major frameworks for inferring orthology are either graph-based or tree-based. Graph-based methods use graph theory to create a network where genes/sequences are represented as nodes on the network connected by edges representing evolutionary

relationships between nodes. Usually there are two parts to graph-based orthology inference. Initially an orthology network is created where nodes (genes) are connected by edges based on a statement of orthology. The OMA algorithm (Roth et al. 2008) we employ here to define gene birth and death, bases the edges/connections on Smith-Waterman alignments. Following on from this the orthologs are clustered into groups or gene families. In OMA the orthologous groups are Hierarchical Orthologous groups or HOGs (Altenhoff et al. 2014). HOGs are defined as groups of genes that have descended from a common ancestor in a given taxonomic range. Traditionally, gene/tree reconciliation is used to identify HOGs. However, OMA employs a graph-based method to identify HOGs based directly on the orthology graph it generates. The removal of traditional gene/species tree reconciliation from the inference process significantly reduces the computational cost. As well as this, OMA also reports several other advantages over standard bidirectional best hit approaches: it uses evolutionary distances instead of scores, considers distance inference uncertainty, includes many-to-many orthologous relations, and accounts for differential gene losses. (Roth et al. 2008). Using OMA v 1.1.2 (Altenhoff et al. 2014) and the default parameters (as discussed with the authors of OMA) we carried out our gene family evolution analyses. All available pre-computed data for our database was downloaded from OMA (http://omabrowser.org (Altenhoff et al. 2014)). We used the *familyanalyzer* python module (Altenhoff et al. 2014), made available from the authors, to analyse gene events across our topology according to the HOGs produced by OMA.

**2.5 Remodeled gene detection using CompositeSearch**

By definition a remodeled gene is formed through a recombinogenic process such as gene fusion, where segments of the remodeled gene are derived from different gene families. Sequence similarity networks (SSN) where each node represents a unique sequence and each edge represents the similarity between sequences, appear to be well suited to identify and study this genetic mosaicism (Alvarez-Ponce et al. 2013, Bapteste et al. 2013).

We constructed a SSN using the results of an all-against-all BLASTp (Altschul et al. 1990a) of 1.2 million protein coding sequences (i.e. all genes within our database of 63 genomes). In this undirected network, two proteins are connected based on their

similarity scores (E-value <= 1e-5, %Identity >= 30%). The SSN was made symmetrical by keeping only the best match of each pairwise comparison. We detected the remodeled genes and their families in this SSN using the software package CompositeSearch (*under review* Pathmanathan JS et al, 2017). The structure of the SSN captures much of the history of the evolution of the gene, such as divergence by point mutations and also recombinogenic events like fusions/fission events (Adai et al. 2004, Jachiet et al. 2014). Typically, gene families form sub-graphs with high connectivity, in which connected sequences display significant BLAST E-values ≤ 1E-5, mutual covers ≥ 80%, %Identity ≥ 30%.

The results of CompositeSearch were parsed to retain only remodeled gene families with more than one remodeled gene and having no overlapping contributing sequences or components. We removed any singleton remodeled gene families, i.e. those with only a single member, as these were more potentially false positives. This removed 44,453 of the 53,456 remodeled gene families (~83% of total remodeled gene families were removed in this step). We also removed any remodeled gene families where only 1 member was remodeled. This removed 1,065 (~10%) of the leaving a total of 7938 remodeled gene families.

**2.6 Remodeled Gene Family Classification using CompositeClassifier**

Remodeled gene families were classified based on their origin and that of their components. This classification allowed us to identify if a remodeled gene family was formed by the fusion of pre-existing or entirely new protein coding gene families. All gene families were then placed on the reference tree and their last common ancestor was inferred using parsimony (Farris 1977). The type of remodeled gene family can be inferred by comparing its position on the phylogeny to that of its components. For example, a remodeled gene family is classified as old if its component gene families evolved before its emergence on the same path. The categories are as follows: Old refers to instances when components of a remodeled family are found at ancestral nodes only. Mixed refers to instances when components of a composite family are found ancestrally in the same lineage and also at the current node of comparison (a mix of old and new components). Complex refers to instances when components for a composite family are found in another path on the tree (not in a common ancestor). Contemporary refers to instances when a components of a composite gene family are

found at the present node (all components arose at the same node on the tree as the composite). Subsequently remodeled refers to instances when components of a composite gene family are found at younger time points in the tree that the composite family (this is gene fission). Undefined are instances of families that cannot be categorized by these rules. This approach aims to show the evolutionary combinatorial processes under which genes evolve (Figure 3.5(B))

**2.7 Functional Enrichment Analysis**

Using the stats.hypergeom function from the python SciPy package (Jones et al. 2014a), the genes at each node on the tree were assessed for enrichment of Gene ontology (GO) functions using a Bonferroni multiple testing correction (Weisstein 2004). Domains from Pfam (Finn et al. 2016) and their associated GO terms were retrieved from the gene ontology website (Consortium 2015). We represented each family defined in the CompositeSearch (Pathmanathan JS et al, 2017) analysis by its common Pfam domains and GO terms. Our criteria were that the GO term must be in > 50% of all genes in the family and a Pfam domain had to be ubiquitous within the family. For example, Family_A has 100 member genes. Qualitatively, all members have Domain_W and Domain_X, 4 members also have Domain_Y and 62 have Domain_Z. The first filter states that each ontology must be present in the majority (>50%) of the members to be included as a representative. This filter would exclude Domain_Y as it is not a majority. It has 4 associated Pfam domains, 12 have term_A, 100 have terms B, and C, and 70 have term D. As the criteria requires 100% of the genes to have a term only terms B and C pass the filter.

**Figure 3: Phylogenetic tree of our metazoan sampling with internal nodes labelled**

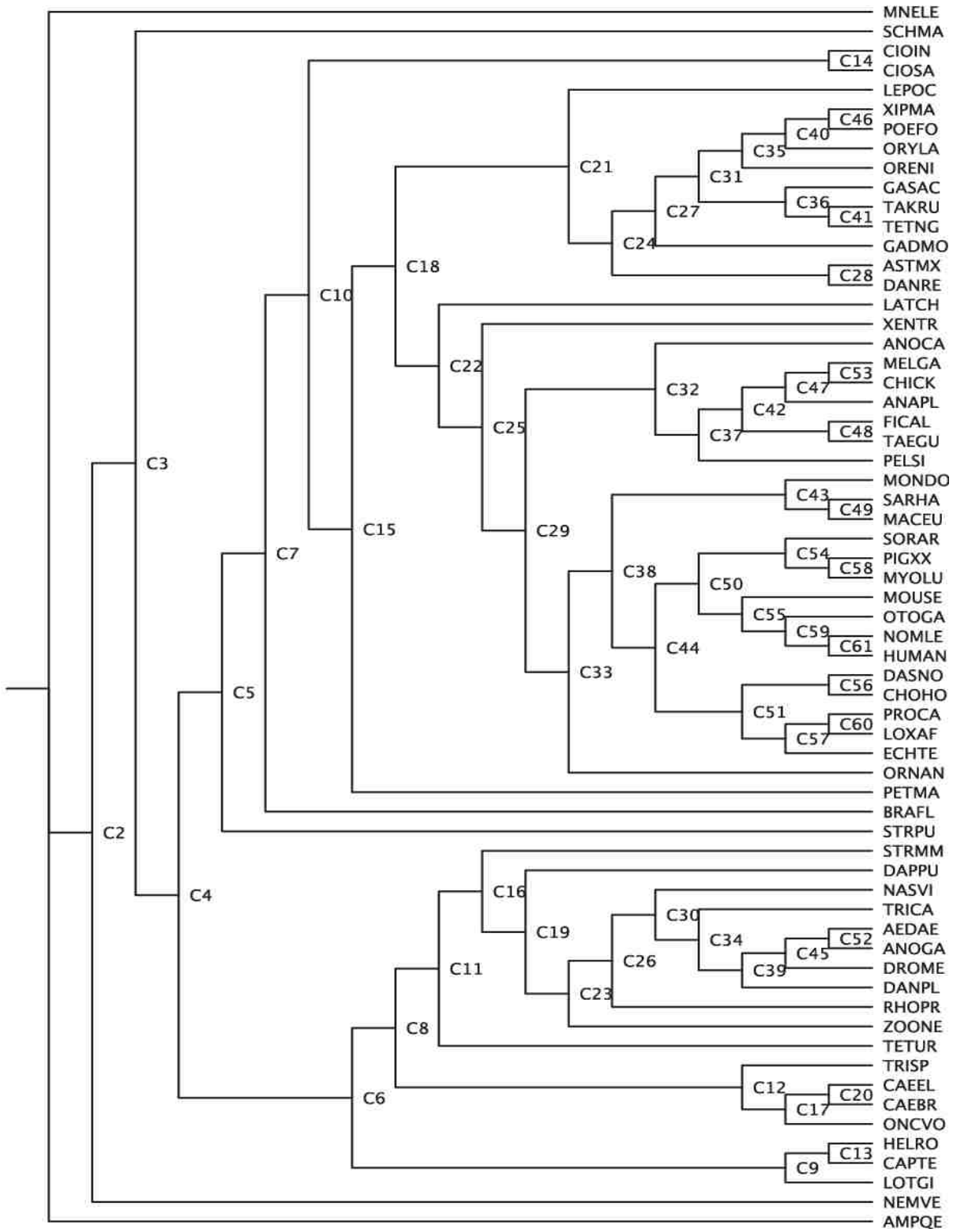



Figure 3: A phylogenetic tree (cladogram) of our sampling of the *Metazoa* with the internal nodes labeled for future reference when describing results.

The next filtering step works on the average of each Pfam domain per gene in the family. So, if all genes have 1 copy of Domain_W; 48 members have 2 copies of Domain_X, 40 members have 3 copies of Domain_X, and 12 have 1 copy of Domain_X; of the 62 members containing the passing Domain_Z 40 members contain 2 copies of Domain_Z and 22 only have 1 copy. The most common filtered domain sums per gene would be:

Domain_W: 1 copy

Domain_X: Average = 2.28 copies or 2 copies

Domain_Z: Average=1.65 copies or 2 copies

In essence, this removes gene families that show a high probability of being homoplastic and give the representative domains present in a gene family.

# 3 Results

### 3.1 Novel protein coding gene families emerge throughout the Metazoa and primarily by gene remodeling.

The orthology network created by OMA was based on Smith-Waterman alignments and subsequently identified orthologous families using a hierarchical clustering method (Altenhoff et al. 2014) (Section 3.2.4). The Hierarchical Orthologous groups (HOGs) produced by this method were defined as groups of genes that descended from a common ancestor in a given taxonomic range. These groups allowed us to identify where gene gain, loss or duplication arose in time. Novel gene families are those that had not been found prior to this point on the tree and what type of new gene family they are (*e.g.* remodeled or non-remodeled). The analysis of gene gain and loss identified 45,612 instances of novel genes at internal nodes on the animal tree. Of this cohort of novel genes 36,948 (81%) are remodeled (Figure 3.4). The majority of internal nodes (57/61) have more novel remodeled gene families than novel non-remodeled gene families. The average number of novel genes per node in the phylogeny is 760 and the median is 390 (Standard deviation = 1003). Most nodes that have above average number of novel gene families are major transitional nodes, including the following (*Clade* (total number of novel genes, % of novel genes that are remodeled at each node)): *Eumetazoa* (2267, 68%); *Bilateria* (3005, 65%); *Protostomia* (2179, 92%); *Euteleostomi* (3674, 73%); *Sarcopterygii* (1584, 85%), and

*Neopterygii* (3026, 89%). In the *Protostomia* there are a total of 2,179 novel genes and in the *Deuterostomia* there are 957. However, on average both *Protostomia* and *Deuterostomia* have the same number of novel non-remodeled genes per node (118 in both cases). The *Protostomia* have more novel remodeled genes (797) per internal node than *Deuterostomia* (493) (Table 3.1).

### 3.2 Gene remodeling is prevalent across the *Metazoa*, particularly at nodes of major phenotypic transition

Using a sequence similarity network (SSN) approach employed in CompositeSearch (Pathmanathan JS et al, 2017) we identified a total of 71,460 gene families in animal evolution. The analysis spans 63 Metazoan species representing all major groups of animals and 20,801 million cumulative years of animal evolution (Figure 3.1). On the SSN, remodeled gene families are represented as nodes that hold otherwise unconnected gene families together on the graph and we identify a total of 48,985 nodes with this feature (Figure 3.5). Using the canonical species phylogeny (Section 3.2.3) each of the 71,460 gene families were mapped to their node of origin. Each internal node (61 in total) in the phylogeny contained remodeled gene families and 49/61 of the internal nodes had more remodeled than non-remodeled gene families indicating that for the majority of internal nodes more novel gene families emerge by gene remodeling than other mechanisms.

Next, we wished to determine if the genesis of novel gene families by remodeling is distributed equally across the phylogeny or are there particular nodes that have a higher instance of novel gene family genesis by gene remodeling when compared to the other nodes in the tree. In particular we identified the internal nodes that contained the largest number of gene families (Figure 4): *Eumetazoa* (3913 families– 84% remodeled); *Bilateria* (8075 families– 87% remodeled); *Deuterostomia* (1019 families– 86% remodeled); *Vertebrata* (1500 families – 85% remodeled); *Euteleostomi (*7723 families- 84% remodeled); *Sarcopterygii* (1057 – 78% remodeled); and *Amniota* (2267– 75% remodeled). Each of these nodes represents a major transition in metazoan life history. In contrast, a large number of new gene families also emerge on two more recent nodes on the tree: (1) the ancestral node of *Caenorhabditis briggsae* and *C. elegans* has 4621 new gene families, 30% of which

**Table 1: Gene counts for each node in the Metazoan tree from the OMA and CompositeSearch analyses.**

| Name | #genes | #duplicated | #lost | #novel/singleton(leaf) | #Comp Fams | #Genes in Comp Fams |
|---|---|---|---|---|---|---|
| SCHMA | 11404 | 715 | 7428 | 7972 | 44 | 136 |
| STRPU | 26882 | 2896 | 9697 | 17563 | 520 | 2213 |
| BRAFL | 28464 | 3582 | 10079 | 18015 | 465 | 1688 |
| CIOSA | 13936 | 387 | 2724 | 6557 | 52 | 115 |
| CIOIN | 16500 | 459 | 1706 | 8077 | 55 | 130 |
| C14 | 9867 | 1578 | 11156 | 1789 | 351 | 897 |
| PETMA | 10766 | 1188 | 14587 | 4340 | 53 | 130 |
| XENTR | 19291 | 2365 | 12098 | 4461 | 36 | 122 |
| ORNAN | 19730 | 1452 | 12789 | 5786 | 58 | 120 |
| MACEU | 15262 | 374 | 6289 | 1174 | 7 | 14 |
| SARHA | 19337 | 650 | 3741 | 2518 | 13 | 51 |
| C46 | 20143 | 1548 | 2093 | 10 | 8 | 16 |
| MONDO | 16844 | 1284 | 7898 | 2458 | 26 | 231 |
| C40 | 21394 | 3345 | 7240 | 144 | 30 | 78 |
| CHOHO | 12329 | 477 | 10551 | 1378 | 3 | 6 |
| DASNO | 23533 | 1905 | 2960 | 3991 | 43 | 130 |
| C53 | 21218 | 2788 | 4656 | 12 | 14 | 28 |
| ECHTE | 16499 | 650 | 7780 | 2089 | 12 | 25 |
| LOXAF | 21050 | 1427 | 2677 | 2149 | 5 | 12 |
| PROCA | 16002 | 275 | 5806 | 1034 | 6 | 12 |
| C58 | 20603 | 1282 | 1875 | 3 | 8 | 16 |
| C54 | 21761 | 1403 | 3408 | 28 | 87 | 180 |
| C47 | 24333 | 2936 | 4761 | 80 | 239 | 531 |
| OTOGA | 19514 | 926 | 2852 | 1030 | 2 | 4 |
| HUMAN | 30808 | 1314 | 1200 | 11464 | 72 | 192 |
| NOMLE | 18717 | 341 | 2631 | 1453 | 6 | 21 |
| C61 | 19699 | 629 | 1613 | 160 | 86 | 232 |
| C59 | 20773 | 677 | 1389 | 27 | 31 | 82 |
| MOUSE | 25679 | 1876 | 3332 | 5945 | 16 | 40 |
| C55 | 21743 | 1193 | 4869 | 46 | 89 | 202 |
| SORAR | 13096 | 641 | 12154 | 1502 | 5 | 11 |
| PIGXX | 21452 | 1849 | 4653 | 3222 | 14 | 28 |
| MYOLU | 19862 | 1704 | 4771 | 1679 | 18 | 46 |
| C60 | 21813 | 1212 | 2137 | 15 | 13 | 27 |
| C56 | 23270 | 1251 | 3304 | 10 | 11 | 23 |
| C48 | 25853 | 3399 | 3507 | 140 | 175 | 515 |
| C41 | 27309 | 2633 | 2008 | 1181 | 1226 | 5246 |
| C35 | 26558 | 1721 | 1601 | 1516 | 1205 | 5809 |
| C31 | 25639 | 2128 | 3135 | 1019 | 495 | 2978 |
| PELSI | 18318 | 1001 | 5757 | 2710 | 27 | 62 |
| CHICK | 15504 | 246 | 1927 | 1394 | 11 | 59 |

| | | | | | | |
|---|---|---|---|---|---|---|
| MELGA | 14627 | 105 | 2735 | 1408 | 13 | 26 |
| C57 | 15887 | 441 | 1604 | 46 | 47 | 110 |
| ANAPL | 15753 | 177 | 3469 | 1899 | 13 | 39 |
| C49 | 17212 | 650 | 2255 | 84 | 47 | 109 |
| TAEGU | 17104 | 849 | 2503 | 2440 | 33 | 72 |
| FICAL | 15383 | 179 | 2463 | 1124 | 5 | 10 |
| C50 | 16608 | 739 | 3027 | 186 | 54 | 116 |
| C42 | 19036 | 679 | 2328 | 347 | 142 | 414 |
| C36 | 20650 | 937 | 2004 | 317 | 82 | 340 |
| ANOCA | 18029 | 1172 | 7119 | 2524 | 29 | 174 |
| C32 | 21789 | 2717 | 7261 | 974 | 113 | 301 |
| C27 | 26527 | 1166 | 684 | 1306 | 1696 | 10214 |
| C24 | 25211 | 1091 | 1123 | 480 | 435 | 3482 |
| LATCH | 20358 | 2376 | 10990 | 4608 | 58 | 176 |
| C21 | 25165 | 3008 | 3881 | 1584 | 825 | 7158 |
| DANRE | 27499 | 2379 | 3795 | 6085 | 69 | 353 |
| ASTMX | 23079 | 1162 | 4677 | 3488 | 23 | 48 |
| C28 | 23595 | 5085 | 6072 | 259 | 143 | 316 |
| ORENI | 22257 | 1644 | 5579 | 2129 | 9 | 19 |
| ORYLA | 20499 | 683 | 6986 | 3245 | 30 | 110 |
| XIPMA | 20370 | 180 | 3469 | 928 | 2 | 4 |
| POEFO | 25163 | 1541 | 1817 | 3284 | 29 | 65 |
| C51 | 22814 | 1068 | 1854 | 215 | 125 | 262 |
| C43 | 23807 | 827 | 1308 | 79 | 43 | 108 |
| C37 | 24559 | 1613 | 3091 | 145 | 101 | 337 |
| GASAC | 21773 | 1167 | 4645 | 2710 | 6 | 12 |
| TETNG | 20020 | 691 | 3155 | 2447 | 19 | 48 |
| TAKRU | 22942 | 585 | 3219 | 5448 | 29 | 64 |
| C44 | 20345 | 801 | 3040 | 55 | 79 | 186 |
| C38 | 22888 | 1329 | 4484 | 14 | 19 | 44 |
| C33 | 26592 | 1711 | 861 | 392 | 299 | 1053 |
| GADMO | 20479 | 1424 | 9187 | 2741 | 48 | 113 |
| C29 | 26050 | 2695 | 2655 | 819 | 404 | 1219 |
| C25 | 26402 | 2130 | 823 | 1293 | 668 | 3398 |
| LEPOC | 18893 | 1326 | 9101 | 2383 | 14 | 29 |
| C22 | 24764 | 5720 | 7132 | 3026 | 496 | 3206 |
| C18 | 25622 | 3470 | 558 | 3674 | 6486 | 103008 |
| C15 | 20330 | 3381 | 1151 | 768 | 1280 | 33573 |
| C10 | 18274 | 1678 | 1457 | 534 | 802 | 13561 |
| C7 | 18008 | 2138 | 1306 | 728 | 728 | 14607 |
| C5 | 17107 | 5332 | 2340 | 959 | 874 | 11631 |
| TRISP | 15661 | 575 | 4319 | 11333 | 287 | 919 |
| CAEBR | 21610 | 1268 | 942 | 8020 | 132 | 376 |
| CAEEL | 20800 | 1838 | 674 | 6508 | 109 | 455 |
| C19 | 13750 | 1311 | 991 | 5381 | 1350 | 5597 |
| ONCVO | 12948 | 447 | 2662 | 6806 | 73 | 237 |

| | | | | | | |
|---|---|---|---|---|---|---|
| C16 | 8535 | 1073 | 1382 | 953 | 404 | 1461 |
| C11 | 8307 | 1645 | 8255 | 368 | 77 | 304 |
| STRMM | 14888 | 1166 | 8600 | 7353 | 182 | 549 |
| DAPPU | 30088 | 2143 | 7108 | 20308 | 843 | 2727 |
| ZOONE | 14336 | 866 | 6895 | 5987 | 46 | 216 |
| RHOPR | 15045 | 835 | 7535 | 7546 | 109 | 490 |
| NASVI | 16986 | 1080 | 6647 | 8899 | 370 | 1021 |
| TRICA | 14798 | 1037 | 5887 | 6555 | 93 | 242 |
| DANPL | 16232 | 727 | 4847 | 7971 | 103 | 333 |
| DROME | 14506 | 1349 | 3120 | 5430 | 59 | 157 |
| ANOGA | 12499 | 692 | 2202 | 3016 | 27 | 63 |
| AEDAE | 15129 | 1843 | 1600 | 4420 | 103 | 280 |
| C52 | 11248 | 1056 | 1528 | 772 | 404 | 949 |
| C45 | 11332 | 854 | 2196 | 388 | 315 | 976 |
| C39 | 12660 | 1043 | 1459 | 131 | 84 | 284 |
| C34 | 13408 | 872 | 1304 | 231 | 175 | 493 |
| C30 | 14000 | 919 | 1289 | 260 | 194 | 628 |
| C26 | 14513 | 694 | 839 | 276 | 184 | 570 |
| C23 | 14689 | 989 | 2087 | 738 | 476 | 2087 |
| C20 | 15457 | 1180 | 1132 | 451 | 162 | 759 |
| C17 | 15432 | 1344 | 913 | 631 | 218 | 1258 |
| TETUR | 18019 | 1847 | 9326 | 11282 | 230 | 1255 |
| C12 | 14846 | 1427 | 1655 | 454 | 124 | 1450 |
| C8 | 15156 | 1816 | 4064 | 520 | 116 | 790 |
| HELRO | 23263 | 1638 | 6737 | 13900 | 389 | 1618 |
| CAPTE | 31325 | 2472 | 3070 | 17704 | 635 | 2473 |
| C13 | 15099 | 2133 | 2633 | 414 | 78 | 233 |
| LOTGI | 23514 | 2332 | 5381 | 11331 | 290 | 1785 |
| C9 | 16028 | 3844 | 5370 | 1421 | 134 | 402 |
| C6 | 17525 | 5619 | 2769 | 2179 | 407 | 2455 |
| C4 | 14426 | 1612 | 67 | 3005 | 7001 | 123510 |
| C3 | 10402 | 638 | 205 | 308 | 612 | 15143 |
| NEMVE | 26036 | 3299 | 3195 | 17192 | 197 | 625 |
| C2 | 9832 | 1920 | 34 | 2267 | 3301 | 92486 |
| AMPQE | 28464 | 3053 | 987 | 21286 | 979 | 4148 |
| MNELE | 16020 | 992 | 2650 | 11929 | 263 | 1020 |

Table 1: For each node in the tree (Col1) we have shown the counts for each node describing the following: 1) the number of genes present in the genome, 2) the number of gene duplication events, 3) the number of gene loss events, 4) the number of novel/singleton(leaf nodes), 5) the number of composite gene families emerging and 6) the number of composite genes emerging.

are the result of gene remodeling, and (2) the common ancestor of *Ciona savignyi* and *C. intestinalis* has 916 new gene families with 37% the result of gene remodeling.

The protein-coding elements that contribute to a remodeling event are known as components and can be of different ages or can themselves be the result of gene remodeling (Figure 5). To extract more detail on each case of gene remodeling detected we used CompositeClassifier from CompositeSearch (Pathmanathan JS et al, 2017) (Section 2.6). We categorised the components of every remodeled gene family based on their phylogenetic placement as: old, mixed, complex, undefined, contemporary and subsequently remodeled (Figure 5). In general, we see that most remodeling events on the tree are categorized as old. This means that most gene remodeling occur using only genetic material that is ancestral.

In general, the emergence of remodeled gene families is more prevalent within Deuterostomes than Protostomes (501 as compared to 288 remodeled gene families per internal node on average) (Figure 5). However, in Section 3.1 above we show that Deuterostomes have less novel remodeled genes than Protostomes indicating that Protostomes rely on gene remodeling as a mechanism to create novel genes more than Deutrostomes. The most prevalent category of remodeling in Metazoa is to reuse ancestral genetic protein coding elements (old category) with 50% and 51% of remodeling events in Protostomes and Deuterostomes respectively the result of old remodeling events (Figure 5). Therefore protein-coding gene families that are already established, or segments thereof, are used most often to create new gene families.

The large number of remodeled gene families predicted may be due to rapid turnover throughout the tree. We calculated the consistency index (CI) for remodeled and non-remodeled gene families (Kluge and Farris 1969) (where the maximum CI of 1 indicates that a family is gained/lost only once). Remodeled gene families have an average CI of 0.4 as compared to 0.7 for non-remodeled gene families suggesting that remodeled gene families are gained/lost more readily than non-remodeled gene families.

**Figure 4: Proportion of remodeled and non-remodeled events in novel gene family genesis**

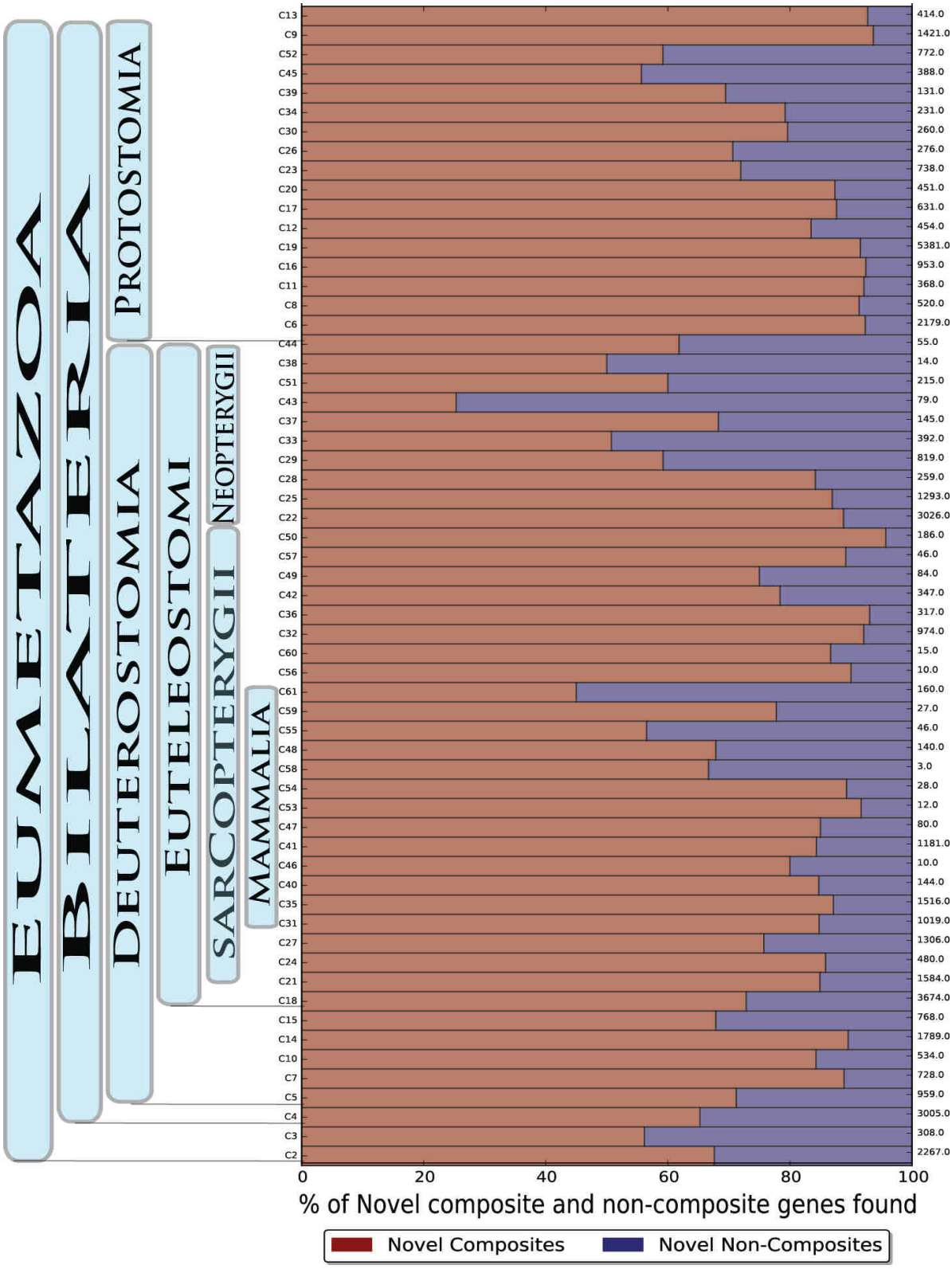
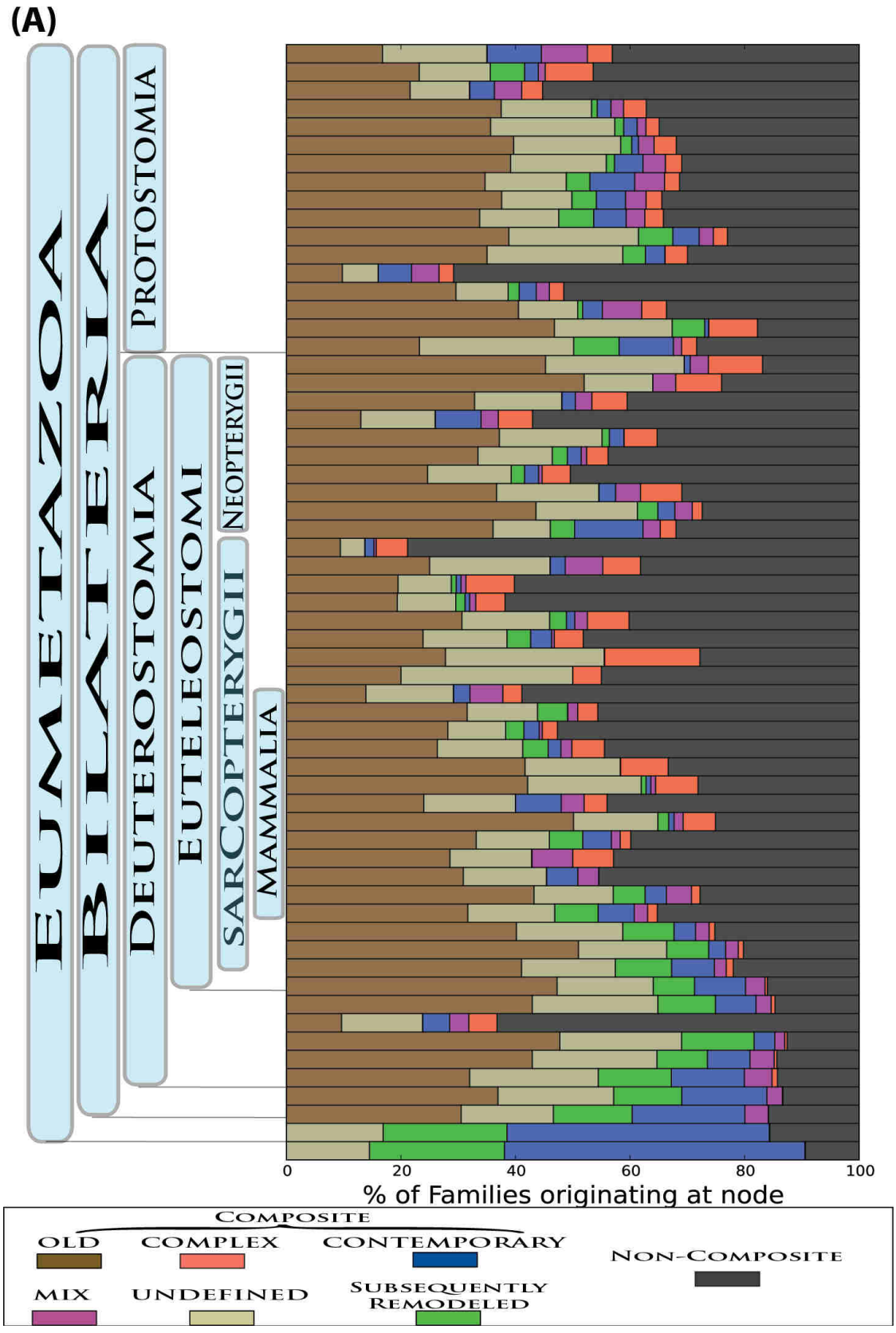



Figure 4: Each bar represents the proportion of novel genes that arose at each internal node on our tree (found in our OMA analysis) in each category: composite or non-

composite (determined from our CompositeSearch analysis) (each bar represents 100%). The number in black on the right Y-axis represents the number of novel genes that originate at this node in the tree. The red bar represents the proportion of novel genes that are composite and the blue bar represents the proportion of novel genes that are non-composite. The left Y-axis represent the label we have given to internal nodes of the tree (Figure 3).We have outlined the major taxonomic groupings.

**Figure 5: Gene remodeling across the *Metazoa***



**(B)**

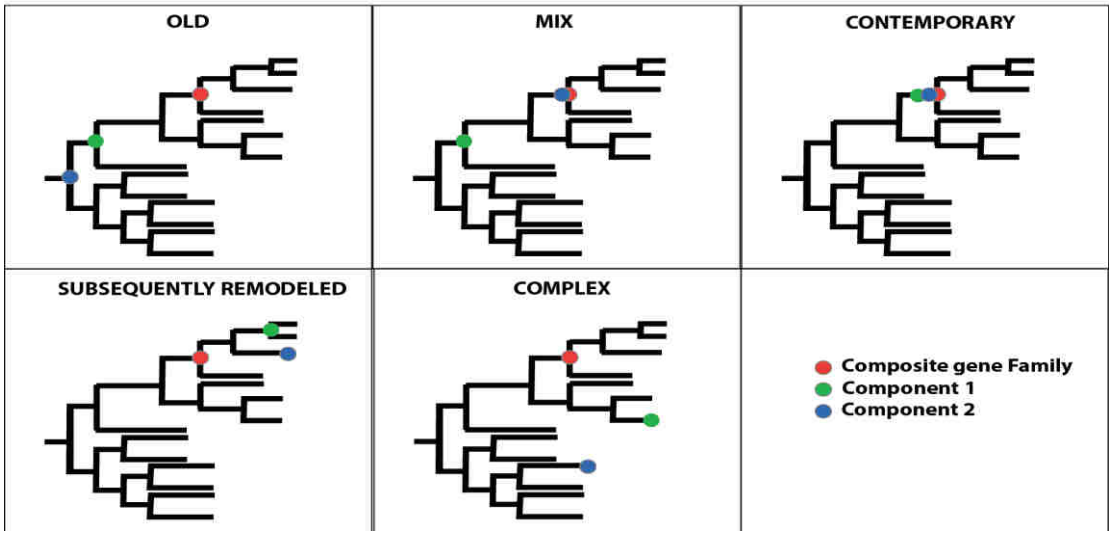



Figure 5: **(A)** Each bar represents the proportion of each category of family from the CompositeSearch analysis (each bar represents 100%) for each internal node of our tree. All colored bars are subcategory of composite gene families, black represents the proportion of gene families that are not composite. We have outlined major taxonomic groups. The node labelling system is illustrated in Figure 3. (**B**) We categories the components of every remodeled gene family based on their phylogenetic placement as: old, mixed, complex, undefined, contemporary and subsequently remodeled.

### 3.3 The rate of novel gene genesis across the *Metazoa* is not strictly clocklike

To determine the rate at which novel genes are emerging across the *Metazoa* we compared the rate of novel gene genesis for remodeled and non-remodeled genes. In general, we find that the rate at which novel gene families arise from gene remodeling is higher than the emergence of novel genes from other mechanisms (Figure 6). The average number of novel remodeled genes per node per million years (MY) is 13.0, and for novel non-remodeled genes it is 3.0. While there are some minor fluctuations (e.g. *Bilateria)* in the rate of generation of novel non-remodeled genes, the rates remain relatively similar across nodes (standard deviation = 5.7 from the mean). This is not the case for novel remodeled genes that have a comparatively high average standard deviation of 17.9 from the mean. Some major nodes in the animal phylogeny show a relatively high rate of emergence of novel gene genesis by gene remodeling, *Bilateria* (71.5 novel remodeled genes per MY); *Sarcopterygii* (60.2/MY); *Theria* (72.0/MY); *Protostomia* (46.2/MY), and *Ecdysozoa* (47.5/MY) are all examples of this.

Overall, novel remodeled genes have emerged at a faster rate than novel non-remodeled genes. But certain time points in metazoan evolution show higher than expected rates of emergence of novel gene families by both remodeling and non-remodeling mechanisms. One such node is the *Bilateria* node, at ~797 MYA (Kumar et al. 2017), arguably one of the most significant transitions in the *Metazoa* representing the origin of the third germ layer (the mesoderm) and increased morphological complexity (Martindale et al. 2002). The *Bilateria* node has on average 109 novel gene families emerge per MY. Another example of a high rate of novel gene family genesis is the origin of placental mammals (Crompton and Jenkins Jr 1979) (82 novel genes per MY).

### 3.4 Gene remodeling impacts the functional landscape at major phenotypic transitions in the *Metazoa*

The potential functional roles of the remodeled genes (at the level of domains) was assessed using Pfam domain data(Finn et al. 2016). For each internal node on the tree we established a list of significant functions gained at that time point (Section 2.7). Functional analysis of remodeled gene families at the *Euteleostomi*

**Figure 6: The rate of novel gene genesis is not strictly gradual**

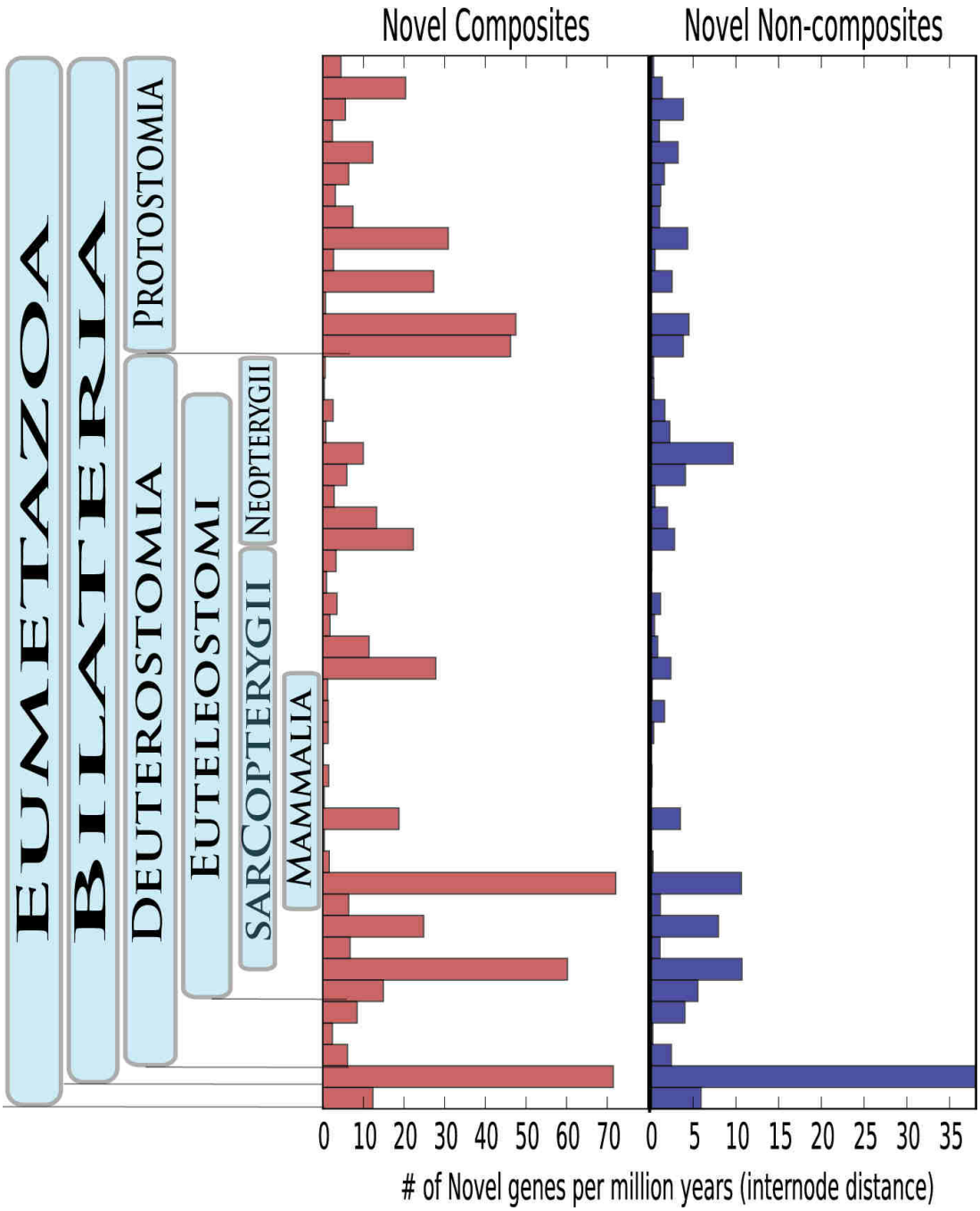



Figure 6: The bar charts represent the number of novel genes that originate at internal nodes divided by the internode distance(time) between the node and its closest ancestor. This gives the number of genes per unit of time for each node. Nodes that have a short internode distance (<10 million years) were not included on this as the short period of time skews the data. The Red bars represents the rate of novel composite genes and the blue bars represent the rate of novel non-composite genes. We have outlined the major taxonomic groupings

**Table 2: Sample of Functional enrichment for novel remodelled genes found at some *Metazoa* transition nodes.**

| Enriched Gene | Corrected P-value | Tree Node |
|---|---|---|
| MHCII(Todd et al. 1988) | 3.8e-05 | *Euteleostomi* |
| RAG-2 involved in the initiation of V(D)J recombination during B and T cell development (Shinkai et al. 1992) | 5e-06 | *Euteleostomi* |
| Fibrinogen | (3.9e-07) | *Euteleostomi* |
| Ribosomal_protein_L44 | 4.2e-07 | *Eumetazoa* |
| Ribosomal_protein_L21e | 2.9e-09 | *Eumetazoa* |
| Ribosomal_L27e_protein_family | 2.0e-08 | *Eumetazoa* |
| Ribosomal_protein_S17 | 3.5e-06 | *Eumetazoa* |
| DHODH)(Fang et al. 2013), | 3.2e-05 | *Eumetazoa* |
| DHFR(Schnell et al. 2004), | 7.3e-08 | *Eumetazoa* |
| GPK(Wu et al. 2004) | 3.2e-05 | *Eumetazoa* |
| NDPK(Almgren et al. 2004) | 5.1e-21 | *Eumetazoa* |
| WNT | 5.8e-05 | *Deuterostomia* |
| Lipoxygenase | 2.0e-05 | *Deuterostomia* |
| Hydroxymethylglutaryl-coenzyme A reductase | 3.8e-07 | *Deuterostomia* |
| GDP dissociation inhibitor | 8.3e-07 | *Deuterostomia* |
| GrpE | 1.5e-08 | *Deuterostomia* |
| Peptidase M41 | 4.2e-05 | *Deuterostomia* |
| MOSC | 1.0e-05 | *Deuterostomia* |
| GPI transamidase subunit PIG-U | 3.8e-06 | *Deuterostomia* |
| Cytochrome b | 5.7e-06 | *Chordata* |
| Cytochrome C and Quinol oxidase polypeptide I | 5.7e-10 | *Chordata* |
| V-ATPase subunit | 5.7e-06 | *Chordata* |
| ATP synthase protein 8 | 2.9e-07 | *Chordata* |
| Glycosyltransferase_family_6 | 6.1e-05 | *Chordata* |
| Tight Junction protein | 4.4e-05 | *Chordata* |
| Nuclear receptor coactivator | 3.3e-06 | *Chordata* |

Table 2: The table shows examples of novel remodeled genes (enriched gene) that were found to be significantly enriched (Corrected p-value) for a particular function at particular nodes in the tree (Tree node column). All nodes shown in this example represent nodes on the animal tree where major phenotypic changes have occurred.

ancestral node, identifies that many immune system related functions are introduced at this point (Table 2) and this node of course represents a major transition in the emergence of the adaptive immunity (Flajnik 2014). At the origin of the *Eumetazoa* novel gene families gained by gene remodeling have significant enrichment for ribosomal protein related functions and for enzyme functions related to cell proliferation (Table 2). The origin of the *Deuterostomia* has significant enrichment in functions related to cell signaling, development and metabolism (Jones et al. 2014b). The origin of *Chordata* shows significant gains in a number of key processes (Jones et al. 2014b) such as the remodeling of proteins involved metabolism and generating cellular energy and protein packaging and transport (Table 2). In summary, there are a plethora of significantly enriched functions at most internal nodes, with some nodes containing functions that correlate with a major phenotypic transition at that node.

# 4 Discussion

This chapter gives an insight into the role of composite gene remodeling (gene fusions/gene fissions) in the evolution of novel protein coding genes across the *Metazoa*.

It has been established that modular proteins have an important role in the evolution of the *Metazoa.* For example, Patthy (2003) shows that a large proportion of proteins involved in the extracellular matrix of multicellular animals are a result of chimeric or gene fusions (Patthy 2003). However, it is generally believed that events to create a gene fusion/fission are rare (Jachiet et al. 2013). Fusion genes have been well documented in animals (Buljan et al. 2010, Marsh and Teichmann 2010). In humans, fusion genes are often linked with cancer (Soller et al. 2006, Soda et al. 2007, Lawson et al. 2011). However, it has not been fully established as to how this composite gene (fusion/fission gene) mechanism drives the evolution of novel proteins and phenotypes right across the *Metazoa*. We have shown that composite genes are indeed present in all major groups across the *Metazoa* (Figure 5). We have shown that they quantitatively form a major part of metazoan protein coding families. Furthermore, we have found that the majority of composite gene events occur using ancestral protein coding elements within the *Metazoa*. Until now, there has been no research into this aspect of composite gene formation.

In addition to this, we wanted to understand not only the prevalence of composite genes, but also how they impact the creation of novel proteins across the *Metazoa*. It has been established that fusion genes can indeed create novel proteins (Long 2000, Thomson et al. 2000). However, the extent to which this process creates novel proteins has not been documented. Our findings suggest that composite gene formation is a major mechanism for creating novel genes in the *Metazoa* (Figure 4). We find that in the vast majority of our sample animal species, more than >50% of novel genes are created through gene remodeling events. This result gives an insight into the important role composite gene formation has in genetic innovation. However, there are examples of fusion genes making their parent genes redundant. If this occurred often, the number of non-composite novel genes that we find would be diminished as they would not be found in our search if they became functionally redundant.

After establishing that composite genes are prevalent across all major groups in the *Metazoa* and do have a major role in creating novel proteins, we wanted to gain an insight to the rate of composite gene formation through time in the evolutionary history of animals. There has been much debate on the rate of evolution. Two strongly supported hypothesis of evolutionary rate are phyletic gradualism and punctuated equilibrium (Gould 1972). Phyletic gradualism refers to slow, gradual changes that accumulate over time to create new species (within intermediate species present). Punctuated equilibrium argues that evolution occurs in bursts of evolution (bursts of high rate) that are tied to speciation events to create new species (Gould 1972). Our results indicate that the rate of composite gene evolution is not strictly clocklike (Figure 6). We show that novel composite genes have emerged at a faster rate than novel non-remodeled genes. Interestingly, we see that certain time points in metazoan evolution show higher than expected rates of emergence of novel gene families by both remodeling and non-remodeling mechanisms. For example, we found a high rate of novel gene genesis at the *Bilateria* node which represents a major transition in the *Metazoa* where the third germ layer (the mesoderm) was introduced, allowing for increased morphological complexity (Martindale et al. 2002).

In order to gain an insight into the functional importance of the novel composite genes we found across the *Metazoa*, we carried out a functional enrichment analysis.

The wide distribution and abundance of composite genes in the *Metazoa* suggests that these genes are not restricted to a single functional pathway. Literature shows examples of very different pathways and functions being carried out by composite genes (Long 2000, Soller et al. 2006, Demichelis et al. 2007, Soda et al. 2007, Lawson et al. 2011, Agaram et al. 2015). Our functional analysis supports this. We show many composite genes that are enriched for functions and pathways at each node of the *Metazoa* such as immune system genes at the *Euteleostomi* node – a point in animal history where adaptive immunity originates.

One possible reason for the higher level of apparent homoplasy that we found in the remodeled gene families (as compared to non-remodeled gene families) is the presence of epaktologs causing interpretation errors. Epaktalogs are multidomain gene families that share sequence similarity through the independent acquisition of the same domains rather than being homologous due to a common ancestry. The classical types of homologs that algorithms detect are orthologs (homologous genes derived from the same gene in a common ancestor), paralogs (homologous genes derived from a duplicate copy of the same gene) and pseudoparalogs (homologous genes in a genome where at one of the genes was transferred from another species). It is difficult to distinguish between epaktologs and paralogs. This can lead to interpretation errors, where epaktologs are treated as paralogs. In other words, trying to cluster a group of epaktologs as a family with a single point of origin on the tree is incorrect because the epaktologous genes are not directly related through descent. They are only related due to homology shared by containing the same domain (Nagy et al. 2011).

Lastly, our approach relies on high quality data as annotation and sequencing errors can cause incorrect inferences. To diminish the impact of this we have used strict filtering parameters and high-quality genomes. This work can be built on as more high-quality genomes become available, particularly for non-vertebrates.

# 5 Conclusion

In summary, we have utilized novel data driven methods to assess the contribution of tree-like and non-tree-like mechanisms in the creation of novel protein coding

elements across the *Metazoa* using 63 high quality genomes. We have illustrated that gene remodeling is prevalent across the entire *Metazoa* and has a significant contribution to novel gene genesis from protein coding elements. We have shown that the rate of novel gene genesis for remodeled genes is not clocklike and is higher than novel gene genesis of non-remodeled genes. Finally, we have given an insight into how gene remodeling may have had a significant impact in driving adaptive evolution at nodes of major phenotypic transition.

# REFERENCES


Adai, A. T., S. V. Date, S. Wieland and E. M. Marcotte (2004). "LGL: creating a map of protein function with an algorithm for visualizing very large biological networks." J Mol Biol **340**(1): 179-190.

Adams, Mark D, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins and Richard F Galle (2000). "The genome sequence of Drosophila melanogaster." science **287**(5461): 2185-2195.

Agaram, Narasimhan P., Hsiao-Wei Chen, Lei Zhang, Yun-Shao Sung, David Panicek, John H. Healey, G. Petur Nielsen, Christopher D. M. Fletcher and Cristina R. Antonescu (2015). "EWSR1-PBX3: A novel gene fusion in myoepithelial tumors." Genes, Chromosomes and Cancer **54**(2): 63-71.

Almgren, Malin AE, K Cecilia E Henriksson, Jennifer Fujimoto and Christina L Chang (2004). "Nucleoside Diphosphate Kinase A/nm23-H1 Promotes Metastasis of NB69-Derived Human Neuroblastoma11NIH RO1 CA78241 and RO1 CA78241S grants (CL Chang)." Molecular cancer research **2**(7): 387-394.

Altenhoff, Adrian M, Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, Bartlomiej Tomiczek, Steven Müller and Henning Redestig (2014). "The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements." Nucleic acids research: gku1158.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990a). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers and David J. Lipman (1990b). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Alvarez-Ponce, D., P. Lopez, E. Bapteste and J. O. McInerney (2013). "Gene similarity networks provide tools for understanding eukaryote origins and evolution." Proc Natl Acad Sci U S A **110**(17): E1594-1603.

Bapteste, E., L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie and J. Whitfield (2013). "Networks: expanding evolutionary thinking." Trends Genet **29**(8): 439-441.

Bell, Graham (2015). The Evolution of Life, Oxford University Press.

Bradnam, Keith R., Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A. Chapman, Guillaume Chapuis and Rayan Chikhi (2013). "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience **2**(1): 1-31.

Buljan, Marija, Adam Frankish and Alex Bateman (2010). "Quantifying the mechanisms of domain gain in animal proteins." Genome biology **11**(7): R74.

Ciccarelli, Francesca D., Tobias Doerks, Christian Von Mering, Christopher J. Creevey, Berend Snel and Peer Bork (2006). "Toward automatic reconstruction of a highly resolved tree of life." science **311**(5765): 1283-1287.

Consortium, Gene Ontology (2015). "Gene ontology consortium: going forward." Nucleic acids research **43**(D1): D1049-D1056.

Consortium, International Human Genome Sequencing (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-945.

Consortium, Sequencing (1998). "Genome sequence of the nematode C. elegans: A platform for investigating biology." science **282**: 2012-2018.

Crompton, AW and Farish A Jenkins Jr (1979). "Origin of mammals." JA Lillegraven, Z. Kielan− Jaworowska, and WA Clemens, Jr.(eds.), Me− sozoic Mammals: The First Two− thirds of Mammalian History: 59-73.

D'Apice, MR, R Tenconi, I Mammi, J van den Ende and G Novelli (2004). "Paternal origin of LMNA mutations in Hutchinson–Gilford progeria." Clinical genetics **65**(1): 52-54.

Demichelis, Francesca, K. Fall, S. Perner, Ove Andrén, F. Schmidt, S. R. Setlur, Y. Hoshida, J. M. Mosquera, Y. Pawitan and C. Lee (2007). "TMPRSS2: ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort." Oncogene **26**(31): 4596-4599.

Fang, JingXian, Takeshi Uchiumi, Mikako Yagi, Shinya Matsumoto, Rie Amamoto, Shinya Takazaki, Haruyoshi Yamaza, Kazuaki Nonaka and Dongchon Kang (2013). "Dihydro-orotate dehydrogenase is physically associated with the respiratory complex and its loss leads to mitochondrial dysfunction." Bioscience reports **33**(2): e00021.

Farris, James S (1977). "Phylogenetic analysis under Dollo's Law." Systematic Biology **26**(1): 77-88.

Finn, Robert D, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi and Amaia Sangrador-Vegas (2016). "The Pfam protein families database: towards a more sustainable future." Nucleic acids research **44**(D1): D279-D285.

Flajnik, Martin F (2014). "Re-evaluation of the immunological Big Bang." Current Biology **24**(21): R1060-R1065.

Gould, Niles Eldredge-Stephen Jay (1972). "Punctuated equilibria: an alternative to phyletic gradualism."

Holt, Robert A, G Mani Subramanian, Aaron Halpern, Granger G Sutton, Rosane Charlab, Deborah R Nusskern, Patrick Wincker, Andrew G Clark, JoséM C Ribeiro and Ron Wides (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." science **298**(5591): 129-149.

Jachiet, P. A., P. Colson, P. Lopez and E. Bapteste (2014). "Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network." Genome Biol Evol **6**(9): 2195-2205.

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): 837-844.

Jones, Eric, Travis Oliphant and Pearu Peterson (2014a). "{SciPy}: open source scientific tools for {Python}."

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell and Gift Nuka (2014b). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Kaessmann, Henrik, Nicolas Vinckenbosch and Manyuan Long (2009). "RNA-based gene duplication: mechanistic and evolutionary insights." Nature Reviews Genetics **10**(1): 19-31.

Kluge, Arnold G and James S Farris (1969). "Quantitative phyletics and the evolution of anurans." Systematic Biology **18**(1): 1-32.

Kumar, Sudhir, Glen Stecher, Michael Suleski and Hedges S Blair (2017). "TimeTree: A resource for timelines, timetrees, and divergence times." Molecular biology and evolution.

Lawson, Andrew RJ, Guy FL Hindley, Tim Forshew, Ruth G Tatevossian, Gabriel A Jamie, Gavin P Kelly, Geoffrey A Neale, Jing Ma, Tania A Jones and David W Ellison (2011). "RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology." Genome research **21**(4): 505-514.

Long, Manyuan (2000). "A new function evolved from gene fusion." Genome research **10**(11): 1655-1657.

Long, Manyuan, Esther Betrán, Kevin Thornton and Wen Wang (2003). "The origin of new genes: glimpses from the young and old." Nature Reviews Genetics **4**(11): 865-875.

Lynch, Michael (2010). "Evolution of the mutation rate." Trends in Genetics **26**(8): 345-352.

Marsh, Joseph A and Sarah A Teichmann (2010). "How do proteins gain new domains?" Genome biology **11**(7): 126.

Martindale, Mark Q, John R Finnerty and Jonathan Q Henry (2002). "The Radiata and the evolutionary origins of the bilaterian body plan." Molecular phylogenetics and evolution **24**(3): 358-365.

McLean, Cory Y, Philip L Reno, Alex A Pollen, Abraham I Bassan, Terence D Capellini, Catherine Guenther, Vahan B Indjeian, Xinhong Lim, Douglas B Menke and Bruce T Schaar (2011). "Human-specific loss of regulatory DNA and the evolution of human-specific traits." Nature **471**(7337): 216-219.

Nagy, Alinda, László Bányai and László Patthy (2011). "Reassessing domain architecture evolution of metazoan proteins: major impact of errors caused by confusing paralogs and epaktologs." Genes **2**(3): 516-561.

Patthy, László (2003). "Modular assembly of genes and the evolution of new functions." Genetica **118**(2): 217-231.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, Thomas Rattei, Ivica Letunic and Tobias Doerks (2012). "eggNOG v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges." Nucleic acids research **40**(D1): D284-D289.

Roth, Alexander CJ, Gaston H Gonnet and Christophe Dessimoz (2008). "Algorithm of OMA for large-scale orthology inference." BMC bioinformatics **9**(1): 518.

Schnell, Jason R, H Jane Dyson and Peter E Wright (2004). "Structure, dynamics, and catalytic function of dihydrofolate reductase." Annu. Rev. Biophys. Biomol. Struct. **33**: 119-140.

Shinkai, Yoichi, Kong-Peng Lam, Eugene M Oltz, Valerie Stewart, Monica Mendelsohn, Jean Charron, Milton Datta, Faith Young, Alan M Stall and Frederick W Alt (1992). "RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V (D) J rearrangement." Cell **68**(5): 855-867.

Soda, Manabu, Young Lim Choi, Munehiro Enomoto, Shuji Takada, Yoshihiro Yamashita, Shunpei Ishikawa, Shin-ichiro Fujiwara, Hideki Watanabe, Kentaro Kurashina and Hisashi Hatanaka (2007). "Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer." Nature **448**(7153): 561-566.

Soller, Maria Johansson, Margareth Isaksson, Peter Elfving, Wolfgang Soller, Rolf Lundgren and Ioannis Panagopoulos (2006). "Confirmation of the high frequency of the TMPRSS2/ERG fusion gene in prostate cancer." Genes, Chromosomes and Cancer **45**(7): 717-719.

Thomson, Timothy M, Juan José Lozano, Noureddine Loukili, Roberto Carrió, Florenci Serras, Bru Cormand, Marta Valeri, Víctor M Díaz, Josep Abril and Moisés Burset (2000). "Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene." Genome research **10**(11): 1743-1756.

Todd, John A, Hans Acha-Orbea, John I Bell, Nelson Chao, Zdenka Fronek, Chaim O Jacob, Michael McDermott, Animesh A Sinha, Luika Timmerman and Lawrence Steinman (1988). "A molecular basis for MHC class II-associated autoimmunity." science **240**(4855): 1003-1010.

Wang, Wen, Jianming Zhang, Carlos Alvarez, Ana Llopart and Manyuan Long (2000). "The origin of the jingwei gene and the complex modular structure of its parental gene, yellow emperor, in drosophila melanogaster." Molecular biology and evolution **17**(9): 1294-1301.

Weisstein, Eric W (2004). "Bonferroni correction."

Wu, Chia-Chin, Kalpana Kannan, Steven Lin, Laising Yen and Aleksandar Milosavljevic (2013). "Identification of cancer fusion drivers using network fusion centrality." Bioinformatics: btt131.

Wu, Guoyao, Yun-Zhong Fang, Sheng Yang, Joanne R Lupton and Nancy D Turner (2004). "Glutathione metabolism and its implications for health." The Journal of nutrition **134**(3): 489-492.

Zhou, Qi and Wen Wang (2008). "On the origin and evolution of new genes—a genomic and experimental perspective." Journal of Genetics and Genomics **35**(11): 639-648.

**Abstract**

Over the recent years, it has become clear that molecular evolution proceeds not only by divergence from a common ancestor, but also by combining parts from evolving objects of different origins, through processes that are called introgressive. Lateral gene transfers are probably the most well-known of these processes, but introgression has been shown to also happen at various levels of biological organization. As a result, most biological evolving objects (genes, genomes, communities) can be composed of parts from different phylogenetic origins and can be described as composites. Such modular evolution is inadequately modeled by trees, since composite objects are not merely the result of divergence from a common ancestor only. Networks on the other hand are much more suited for handling modularity, and graph theory can be used to search networks for patterns that are characteristic of such reticulate evolution. During this PhD, I developed a piece of software, *CompositeSearch*, that can efficiently detect composite genes in massive sequence dataset, comprising up to millions of sequences. This algorithm was used to identify and quantify the abundance of composite genes in polluted soil environments, and in prokaryotic plasmids. These studies show that important biological novelties and adaptations can result from processes acting at subgenic levels. However, as shown in this manuscript, networks provide a framework that goes well beyond the boundaries of molecular evolution and I have applied them to other evolving entities, such as animals (trait networks) morphology and languages (word networks). In both cases, modularity appears to be a major evolutionary outcome, following rules that remain to be investigated.

**Résumé**

Au cours des dernières années, il est devenu manifeste que l'évolution moléculaire procède non seulement par divergence depuis un ancêtre commun, mais aussi en combinant des fragments d'objets évoluant d'origines différentes, par des processus appelés introgressifs. Les transferts horizontaux de gènes sont probablement les plus connus de ces processus, mais l'introgression affecte aussi d'autres niveaux d'organisation biologique. En conséquence, la plupart des objets biologiques évoluant peuvent être composés de partie d'origines phylogénétiques différentes et peuvent être décrits comme composites. Une telle évolution modulaire se modélise mal par des arbres, puisque les objets composites ne sont pas seulement le résultat d'une divergence depuis un ancêtre. Les réseaux en revanche sont bien plus aptes à modéliser la modularité, et la théorie des graphes peut être utilisée pour chercher dans ces réseaux des patrons caractéristiques d'une évolution réticulée. Pendant cette thèse, j'ai développé le logiciel *CompositeSearch* qui détecte efficacement les gènes composites dans des jeux de données de séquences massifs, jusqu'à plusieurs millions de séquences. Cet algorithme a été utilisé pour identifier et quantifier l'abondance des gènes composites dans des environnements de sols pollués ainsi que dans les plasmides des procaryotes. Les résultats montrent que d'importantes adaptations et nouveautés biologiques découlent de processus œuvrant au niveau subgénique. Par ailleurs, comme je le montre ici, les réseaux fournissent un cadre conceptuel dont l'utilité va bien au-delà de l'évolution moléculaire et je les ai appliqués à d'autres objets évoluant, comme les animaux (réseaux de traits morphologiques) et les langues (réseaux de mots) . Dans les deux cas, la modularité se révèle être une conséquence évolutive majeure, et obéit à des règles qui restent à préciser.