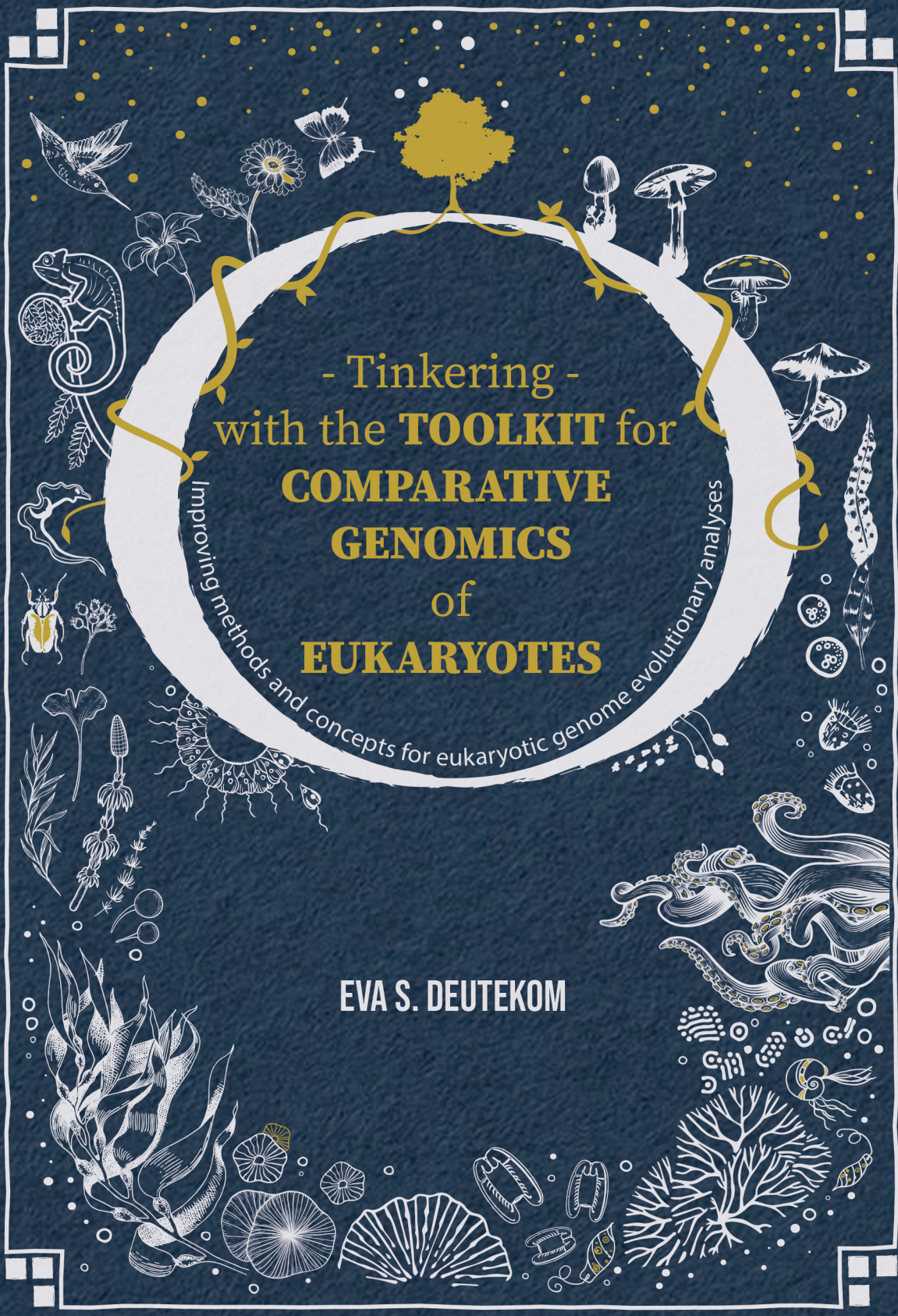

- Tinkering -
with the TOOLKIT for
COMPARATIVE
GENOMICS
of
EUKARYOTES
Improving methods and concepts for eukaryotic genome evolutionary analyses
EVA S. DEUTEKOM

TINKERING WITH THE TOOLKIT FOR COMPARATIVE GENOMICS OF EUKARYOTES

Improving methods and concepts for eukaryotic genome evolutionary analyses

Eva Stephanie Deutekom

The studies in this thesis were financially supported by the Dutch Research Council (NWO), grant number: 016.160.638

ISBN: 978-94-6416-831-0
Cover: *Sea me sparkle*, designed by Eva S. Deutekom
Provided by thesis specialist Ridderprint, ridderprint.nl
Printing: Ridderprint
Layout and design: Birgit Vredenburg, persoonlijkproefschrift.nl

TINKERING WITH THE TOOLKIT FOR COMPARATIVE GENOMICS OF EUKARYOTES

Improving methods and concepts for eukaryotic genome evolutionary analyses

Sleutelen aan de gereedschapskist voor vergelijkende genoom analyse van eukaryoten

Verbeteren van methoden en concepten in eukaryote genoom evolutie analyses

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Utrecht Universiteit op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op

maandag 21 maart 2022 des middags te 4.15 uur

door

Eva Stephanie Deutekom

geboren op 23 maart 1988

te Curaçao, Nederlandse Antillen

Promotor: Prof. dr. B. Snel

Copromotor: Dr. T.J.P. van Dam

CONTENTS

CHAPTER 1.

General Introduction

FROM THE VERY BEGINNING

When we think about life, we often think of large organisms that are easy to spot, such as humans, trees, fish and insects. That said, just as our universe does not solely consist of the stars (including our sun) that we can see, life is not just the organisms we can easily spot. Often living next, in or on, the larger organisms are tiny organisms we can maybe see with a magnifying glass and even tinier organisms that we can only see with a microscope. Quite breathtakingly, all these organisms are related.

Organisms are governed by their genetic material, which is like a roadmap of how to exist as a living thing. The genetic material is a collection of giant molecules inside the organism's cell(s), called deoxyribonucleic acid (DNA). DNA is a large chain of four smaller molecules, called nucleotides, that we can represent with four[1] letters, A, T, G, and C. The nucleotides follow a particular order to make meaningful chunks of sequences on the DNA, called genes. Each gene encodes a specific protein. Proteins are also chains of smaller molecules called amino acids. There are around 20 amino acids that we can represent with 20 letters[2]. Before a gene can be decoded and translated into a protein, another unique protein first transcribes a gene's nucleotide sequence to a messenger ribonucleic acid molecule (mRNA). mRNA is like a postcard with an instruction written onto it. The instruction is again given by four nucleotides[3]. The postcard with the instruction is sent off to where another unique protein will read and translate the mRNA to a protein.

What I shortly described above is the flow of genetic information within a biological system. Simply summarised, this generally[4] means DNA makes mRNA, mRNA makes protein, and the protein does something. Proteins are essential players in the functioning of cells and the functioning of organisms. To obtain all the information stored in the DNA sequences, we need to annotate the genome or gene sequences. In other words, we

1 There have been more types of nucleotides found, but these are the universal ones.
2 This counts for most organisms, but some have more amino acids.
3 In this case it is the nucleotides AUGC, where U replaces the T. The U is a form of T, except energetically less expensive to make and less stable than T. Since mRNA is short-lived (until it makes a protein) it doesn't have to be as stable and T in DNA.
4 Of course, it is not so simple. For instance, DNA is also be copied to another DNA molecule (replicating of cells). DNA can also be synthesised from RNA (which happens when we get infected by viruses).

need to extract the information from the multitudes of different patterns that arise from the letters in DNA. For instance, identifying which parts of the DNA encode functional products or proteins (protein prediction). From there we can also find what particular function the proteins have in cells by looking at, for instance, protein domains that often form the functional units of a protein.

The differences in the sequences of genes and how they translate to proteins can lead to vastly different organisms. Although these organisms are very different, they are all related to each other by evolution. The study of these evolutionary relationships is the subject of phylogeny. We can infer phylogenetic relationships between organisms by comparing the nucleotide sequences of genes, the amino acid sequences of proteins, or both. More specifically, we can compare the letters that represent the sequences.

The phylogenetic comparison of different species led to identifying distinct lineages of organisms, that were catalogued into "the domains of life" [1]. These domains are Bacteria, Archaea and Eukarya (represented by the tree in Figure 1). The domains of life are thought to have developed from a common ancestor, something called the Last Universal Common Ancestor (LUCA) of all organisms.[5]LUCA existed about 3.5 to 4.5 billion years ago [2]. Eventually, LUCA evolved into the distinct ancestors that all led to the descendants in the domains of life, Bacteria, Archaea and Eukarya.[6]

There is no fossil evidence of LUCA's existence. It lived a few hundred million years before the earliest fossil evidence of any life. Nevertheless, hypotheses about LUCA are possible by comparing the DNA and proteins of the many living species in all the domains of life. After all, these species are all LUCA's descendants. Comparing species like this is much like comparing parts of a family tree. There are brothers and sisters, aunts and uncles, but also distant relatives in distant countries.

5 LUCA might not have been the only life that existed, just the one that has living descendants.

6 Viruses are not included here, because they are not cells but little packages made from proteins containing small pieces of DNA or RNA. It is still debated what the origin is of viruses and some interesting hypothesis exist.

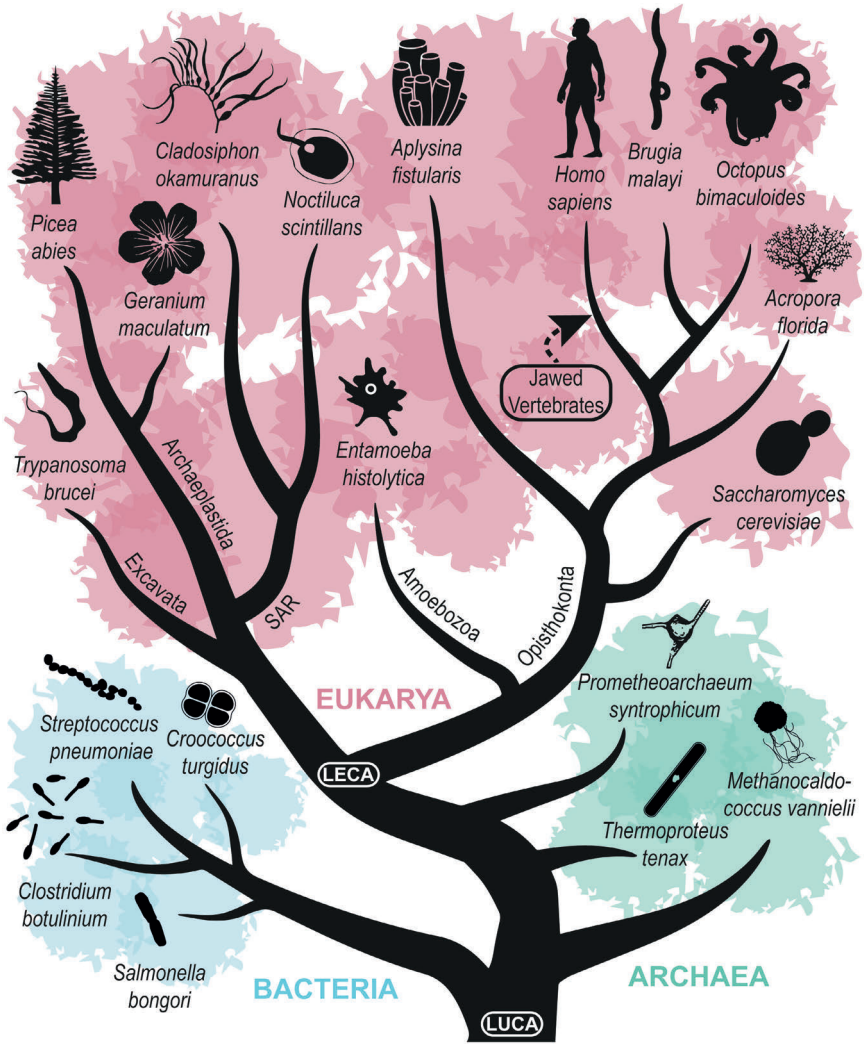


Figure 1. The domains of life. For Bacteria, some examples of disease-causing bacteria (such as, salmonella, botulism and pneumonia) and the cyanobacteria *C. turgidus*. For Eukarya, African sleeping sickness-causing *T. brucei*; Conifer tree *P. abis*; Wild geranium *G. maculatum*; Edible kelp species (like wakame) *C. okamuranus*; Sea sparkle *N. scintillans*; Amoebiasis-causing amoebozoan *E. histolytica*; Yellow tube sponge *A.fistularis*; Human (*H. Sapiens*); Lymphatic filariasis-causing *B. malayi*; Californian two-spot octopus *O. bimaculoides*; Bakers (or budding) yeast *S. cerevisiae*. Eukaryotic supergroups according to the species tree used in this thesis are written next to the branches. Jawed vertebrates emerged less than 400 million years ago (dotted arrow) and include birds, amphibians, reptiles, bony fish and mammals. For Archaea, a methane producing archaeon *M. vannielii* and a heat loving (thermophilic) sulphur-dependent archaeon *T. tenax*. A third archaeon has a pivotal position in the tree as a close relative to eukaryotes, *P. syntrophicum* [3].

There are considerable differences between the domains of life. The differences are mainly in the cellular structures of the organisms residing in each domain. Eukaryotic cells differ from those of Bacterial and Archaean cells because they are typically larger and more complex. Eukaryotic cells also contain multiple compartments that have different functions, called organelles. These compartments include the mitochondria[7], chloroplasts[8], and the cell nucleus that encapsulates the genetic code in densely packed DNA. Eukaryotes can thank the nucleus for their name, as it comes from the Greek word translating to "good kernel" (or core). Collectively Bacteria and Archaea are called prokaryotes, it comes from the Greek word that translates to "before the kernel".

Almost all organisms we can see with the naked eye are multicellular eukaryotes. Think of animals, trees, or a healthy wakame salad next to your sushi. That salad is a kelp species and should definitely not be mistaken for a plant. But there are also unicellular (or single-celled) eukaryotes. Think of the sea sparkles that light up the ocean on hot summer nights, or think of the yeast we use when making bread, beer or wine. Many disease-causing eukaryotes exist as well. After all, bacteria and viruses do not monopolise disease. Examples of eukaryotes that cause diseases you might have heard of are the malaria-causing *Plasmodium falciparum*, lymphatic filariasis-causing *Bruga malayi*, African sleeping sickness-causing *Trypanosoma brucei*, and amoebiasis-causing amoebozoan *Entamoeba histolytica*.

Many characteristics can differ between more closely related eukaryotes as well. For example, animals can have legs, flippers or wings. They can have eyes, some other form of an optical organ, or no eyes at all. Sometimes they have a brain. Sometimes they have nine brains, three hearts and blue blood like the octopus. Meanwhile, trees can photosynthesise and take up nutrients from the ground using complex root systems. Surprisingly, or not surprisingly at all, trees can also communicate [4,5]. Eukaryotes can be highly different, but they are still related by their distinct cellular structure and phylogenetic history.

---

7 "The powerhouse of the cell" is something everyone probably remembers learning in high school. The mitochondria have a very important function in cells, namely to provide organisms with energy, by fuelling important biochemical reactions.

8 Chloroplasts are found mainly in plants and algal cells and through photosynthesis capture the energy of sunlight.

We can represent the phylogenetic history of eukaryotes by their own tree branch shown in an expanded form in the phylogenetic tree in Figure 1. Like the ancestor LUCA, eukaryotes also share a common ancestor called the Last Eukaryotic Common Ancestor (LECA). LECA is at the base of the eukaryotic tree and sister to Archaea and is estimated to have lived around 1.6 to 1.8 billion years ago [6]. LECA was likely a cell, or community of cells [7], and more complex than LUCA. Currently we think LECA was already compartmentalized in a way that make eukaryotic cells so unique. A question that arises here is how LECA became more complex than its prokaryotic brothers and sisters. Ultimately, this increase in cell complexity gave life as we know it the kick it needed to evolve into forms beyond those of simple cells to more complex multicellular organisms.

What exactly happened is still highly debated. What we do know with some certainty is that LECA evolved from its simpler prokaryotic relatives. We know that it started with at least an organism related to the Archaea that at one point gobbled up a bacterium belonging to the Alpha proteobacteria [8]. The bacterium was so useful to the cell that the cell kept the bacteria alive and eventually became dependent on it. This event is how mitochondria were born. Multiple events like this happened over the evolution of eukaryotes. A similar and important event led to the photosynthetic organisms. The cells that now carried around mitochondria again munched on another organism, namely a photosynthetic bacterium belonging to the Cyanobacteria (Figure 1). This beneficial symbiosis led to the chloroplasts in the plants and algae and is how they photosynthesise. This well-established theory on how eukaryotic cells came to be is called the endosymbiotic theory and is the leading theory on the origin of eukaryotic cells. We find evidence for this theory in the organelles of the eukaryotic cells themselves, namely that the organelles still contain DNA molecules that are similar to the DNA of bacteria that the eukaryotic ancestor likely consumed.

As is with LUCA, there is no fossil evidence of the existence of LECA. By looking at many diverse eukaryotic species and their genetic code, we can trace back LECA's contents and estimate its genetic code, genes, and even how it might have looked. Comparing the complete DNA (or genomes) of different species and making phylogenetic trees like the one in Figure 1 allows us to infer any common ancestor between species. For instance,

we can find that we once in the not-too-distant past[9] shared a common ancestor with that nine brained, three hearted, blue-blooded octopus.

Besides looking at whole genomes, we can also look at the evolutionary history of the individual genes or proteins in different eukaryotic species. Comparing the genes of proteins, or the proteins themselves, in different species can clarify how these proteins came to be. An example protein you might know of is haemoglobin. Haemoglobin is the protein in your red blood cells that reversibly binds and transports oxygen from the lungs to the rest of your body using iron ions. The iron gives blood the red colour. Haemoglobin consists of four protein subunits, each containing a haem molecule with iron that binds oxygen. The four protein subunits subdivide into two types, two alpha and two beta subunits. It turns out that an organism that existed before the ancestor to the jawed vertebrates (Figure 1) 400 million years ago had one subunit[10] for oxygen transport [9]. That subunit then copied (duplicated) into two, the alpha and beta subunit. The subunits later evolved to bind to each other into a ball of four that we see in current-day mammals, like us. These genetic mutations altered the way mammals can bind oxygen more efficiently. The increased efficiency made it possible for mammals in all shapes and sizes to acquire enough oxygen while going about their daily business.

But why does the octopus have blue blood? If we skip over all the other parts why the octopus is so different from mammals, it is because of the protein called haemocyanin. Unlike haemoglobin, haemocyanin uses copper molecules to transport oxygen and is dissolved in the blood directly, rather than bound by a cell. It is copper that gives the distinct blue colour. Haemocyanin evolved independently from haemoglobin. In other words, it evolved separately from haemoglobin to perform a similar function.[11] Even though the two proteins' functions are similar, the proteins do not share any traceable common origins. Both haemocyanin and haemoglobin evolved long after the ancestor of octopuses and humans diverged to occupy their own branch in the eukaryotic tree.

9 In the grand scheme of things, but it was still hundreds of million years ago.
10 The subunit was AncMH monomeric protein, ancestor to monomeric myoglobin and haemoglobin subunits [9]. Myoglobin is also still present in skeletal muscle tissue.
11 We call this convergent evolution.

The sequence analysis and comparison of the haemocyanin and haemoglobin proteins and their phylogeny played an essential part to elucidate both the origin and evolution of these proteins [9,10]. Most of what I described in this short introduction is the field of phylogenomics, or finding the phylogeny and evolutionary relationships of genomes and genes by comparing their sequences. Over the past few decades, many genomes of different species have become available due to the scientific advancements in obtaining these genomes by multiple high-tech sequencing methods. High-tech sequencing has increased the data available to do large scale phylogenomic studies. Nevertheless, increased data does not necessarily mean you can gain more knowledge from it. You first have to know how to get the knowledge out of the data. If the data becomes too large to handle by hand, you need programming and computers, which is where bioinformatics comes in.

This thesis describes the work done using a large genome set of many diverse eukaryotic species to improve the computational methods available to analyse and understand eukaryotic evolution. There are still many unsolved questions, eukaryotic genome evolution is complex and many forces drive it. Research of the past decades has shown that eukaryotic evolution is not shaped by genomes becoming more complex with more genes that ultimately gives rise to more complex and varied eukaryotes. Surprisingly, how many different genes there are in a eukaryotic genome is not as relevant as first thought. More relevant in genome evolution is the duplication of genes and subsequent loss of genes that give vastly different species. One such example is again seen with the protein haemoglobin. While the octopus has blue blood and human red, that of the Antarctic crocodile icefish is white! The icefish is the only known white-blooded vertebrate. Its blood is white because it has lost its haemoglobin [11]. A big question is how important gene loss is in shaping eukaryotes and how much is this number impacted by our ability to predict or find the genes and their protein products?

Another big question relevant in eukaryotic genome evolution is what the function is of the genes in the genome. The function of many genes is still unknown. Can we find effective ways to predict these functions using large scale analyses, rather than painstakingly set up single experiments for each gene? This and the above questions are all intertwined with the question of how eukaryotes evolved from LECA. LECA already had many different

genes, which duplicated and got lost many times. Genes often got lost in modular fashion, in other words got lost together, indicating the loss of whole functional pathways or protein complexes. How effective are we at finding these evolutionary patterns that drove the evolution from LECA to current day eukaryotes? In this thesis I will touch upon all these questions.

THE DATA DELUGE WITH A SILVER LINING

It was merely 25 years ago that the first eukaryotic genome sequence was completed through a worldwide collaborative effort, the genome of baker's yeast *Saccharomyces cerevisiae* [12]. The sequencing of genomes of many other organisms followed, including the genome of humans. Today we can sequence genomes in almost a blink of an eye. At the moment of writing this introduction, for eukaryotes there are 109 unique genomes completely sequenced (155 including unique strains) and 7895 unique genomes partially sequenced (draft genomes) of a total of 16067 genome entries (NCBI 01-02-2021[12]).

An unwanted side effect of this increase in (genomic) data is the data deluge. In other words, the sheer amount of data produced is surpassing the limit of institutions to manage it. Many genomic datasets are under-utilised due to the lack of proper methodologies or resources, or the unavailability of people that know how to handle and prepare large datasets. Consequently, there is a growing gap between the world of well-known and well-annotated genes and the world of unknown and un-annotated ones [13].

In addition, there are often many errors in genome and gene annotation. Errors already in the data often accumulate in databases and are hard to correct [14,15]. Counterintuitively, this problem seems to increase rather than decrease with improved sequencing technologies [16,17]. Among other things, this has to do with the difficulty of automatically annotating draft genomes. Automatic annotation is needed to handle the large amounts of genomes currently being generated and does not compare to the gold standard of manual annotation. Automatic annotation is difficult due to the complexity of eukaryotic genes, genome sequencing errors, and protein prediction errors (Chapter 1).

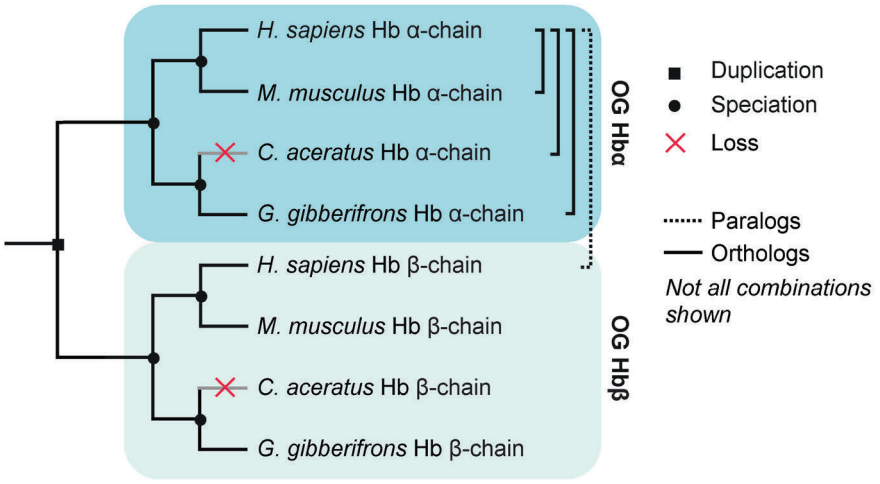
12 https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/

Multiple automatic annotation approaches exist. One is ab initio, which uses mathematical models to identify genes in assembled DNA. Another is homology, or sequence similarity, gene or protein prediction, and it is a strategy often used to (functionally) annotate genes and proteins. Homology (Box 1) is also a strategy often used to infer sequence evolution, while simultaneously is based on sequence evolution principles [18]. Observations made in sequence evolution are that functionally important parts of genomes and proteins tend to evolve slower than functionally unimportant ones [19]. This means that functionally important (parts of) sequences will tend to stay similar through time. This is even more so for closely related species. These observations are now principles and are what lie at the base of homology-based gene predictions: identifying matching parts of an unknown gene or protein to parts of a gene or protein that does have a functional annotation. Matching of genes like this becomes increasingly harder the more distantly related species are.

Box 1 Homology, orthology and paralogy

Homology is an evolutionary relationship between two biological features. This relationship can be between morphological features (such as wings) or genetic features (such as proteins or protein domains). Two features are homologous if they share a common ancestral origin. Homology is a qualitative trait, which means that something is either homologous or not. Homology is also a transitive trait. For instance, if two proteins X and Y are homologous, and Y and Z are homologous, X and Z are also homologous. There are exceptions to this rule that can complicate homology inference. Some proteins can have parts that are homologous, such as proteins that contain certain domains that are homologous to domains of another protein, while the rest of the domains are not shared or homologous. For instance, this happens with fusion (combined) proteins or fission (split) proteins. Homologous proteins can also be missed due to one protein undergoing domain shuffling.

Proteins, genes, or both, can be related by homology in two ways. If two homologous proteins originate from a speciation event resulting in two species with the same protein, they are orthologous (Box Figure 1). If homologous proteins originate from a duplication event resulting in two proteins in a single species, they are paralogous. Proteins in a group of species are orthologous if they shared a common ancestral protein in an ancestral species. These orthologous proteins will form an orthologous group (OG). An example is given in the box Figure 1 for haemoglobin α and β chains in human, mouse and crocodile icefish.



Box Figure 1. An ancestral protein was duplicated to form the haemoglobin α and β chains. After the duplication, speciation events occurred that eventually gave rise to *Homo sapiens* (human), *Mus musculus* (mouse), *Chaenocephalus aceratus* (blackfin crocodile icefish), and *Gobionotothen gibberifrons* (humped rockcod). *H. sapiens* haemoglobin chain α is orthologues to that of chain α in *M. musculus*, *C. aceratus*, and *G. gibberifrons*. *H. sapiens* haemoglobin chain α is paralogous to *H. sapiens* haemoglobin chain β. *C. aceratus* lost both haemoglobin chains. Closely related to *C. aceratus* is *G. gibberifrons*, which still has both haemoglobin chains. Haemoglobin α (or β) stems from a single gene in the common ancestor of *H. sapiens*, *M. musculus*, *C. aceratus* and *G. gibberifrons*.

The *orthology conjecture* states that orthologous proteins have a biologically equivalent function in different species, while paralogous proteins typically diverge in function after duplication [20]. Computational methods can use this assumption to predict protein function. By finding a protein's orthologous counterpart with an inferred function, we can assume that the unknown protein has the same function [20,21].

However, defining protein function is not a trivial task, and function cannot be easily quantified [14]. Function transfer from a known gene or protein to an unknown one, assumes that the function is retained in genes or proteins that have similar sequences and that the gene or protein arose from a common ancestral one (otherwise they are different units entirely). If you are lucky, the gene of interest hits orthologs (Box 1) in other species in a database search. The complexity rises again if you also have paralogs. Interestingly, the distinction between paralogs and orthologs was already introduced in 1970 [22]. However, many erroneous functional annotations because of improper usage of homology still propagate in sequence databases [14].

This also highlights another challenge, the case of the dark matter in genomes. In other words, genes with an unknown function or function that cannot be (readily) inferred because they do not have similar features with any annotated proteins in a database. These so-called hypothetical genes are treasure troves of knowledge that are often left aside. The increase in genomic data also means an increase in the number of (functionally) orphaned genes. The often exotic nature of newly sequenced genomes also complicates this further, since many (of their) genes are not yet known or annotated in sequence databases, and have no apparent homologs.

With many new sequencing projects already on their way, the data wave is quite quickly growing into a tsunami. For some projects, the main focus is even sequencing as many eukaryotic species as possible, like the ambitious Earth BioGenome Project [23] and the 1000 Plants project [24]. The increase of available genomes does not necessarily mean increasing knowledge about genes' function or evolution, especially if we cannot link them to other genes with known functions.

Comparative genomics offers a solution to analyse large groups of genomes and is a versatile and useful approach to investigate the function and the evolution of large sets of genes. Not so long ago, data was the limiting factor in comparative and evolutionary genomics. That data bottleneck is no longer the issue with the exponential increase of available genomes of many diverse eukaryotic species and lesser-known or under-represent phyla [25]. The data increase has multiple positive effects for comparative and evolutionary genomics as well [14]. First, the ultimate net effect will be that more sequences increase the chance of finding homologous genes. This is will be only directly beneficial if these homologous genes are (functionally) annotated as well. Second, more complete sequences of distantly related species increase the representation of conserved gene families. This is one thing that has contributed to, for instance, the knowledge of an ever-expanding gene repertoire of LECA. If more diverse eukaryotic species share a gene, the gene was likely present in LECA. Third, more sensitive methods are developed to search and analyse the large amounts of data. The positive effects will go hand in hand with biologists needing to rely increasingly on bioinformatics-based tools to generate and analyse the data.

COMPARATIVE GENOMICS AND GENOME EVOLUTION OF EUKARYOTES

From the first comparative genome analyses done on eukaryotes [26,27], our understanding of the evolution of eukaryotic genomes has increased tremendously and has shown us how complex and flexible eukaryotic genomes are. Eukaryotic species phylogeny has improved and is still in a constant state of adjustment and change, mainly due to the new additions of phylogenetically relevant genomes and transcriptomes [25].

The current and most recent consensus is that the eukaryotic tree of life can be taxonomically divided into nine supergroups [25]. A supergroup that might be known to many readers from the members it contains is the Amorphea. This supergroup now includes two clades that were previously considered their own supergroups, Opisthokonta (animals, fungi and respective unicellular relatives) and Amoebozoa (Amoeba and most slime moulds). Another supergroup with well-known members is the Archaeplastida (land plants and algae). Eukaryotes differ considerably in their morphology, lifestyles, cycles and habitats [28]. They can be unicellular or multicellular, reproduce asexually or sexually, or live a free-living, parasitic or symbiotic life. Nevertheless, eukaryotes all share complex, unique and essential cellular structures that distinguish them from prokaryotes.

It is a widespread belief lingering in many minds that eukaryotes and their genomes gradually increased in complexity over time. Supposedly, this gradual increase is the main driving force of species-rich clades, such as animals and plants [29,30]. This thought is not surprising, given the knowledge that the very first organisms were reasonably simple compared to the multitude of complex multicellular forms we witness around us today. Primary efforts to explain evolutionary mechanisms behind genomes becoming more complex went towards gene inventions, gene duplications, whole-genome duplications, and gene loss associated with the redundancy of genes after gene and genome duplications [31].

The increased availability of genomes and their comparisons showed that diverse eukaryotic species share various genes. The picture emerged that the eukaryotic ancestor was more complex than expected and already contained many genes found in present-day species [31]. The theory of inflation and streamlining was put forth to explain the evolutionary

dynamics of genomes that contradict the notion that eukaryotes increased their genome complexity gradually [30]. The theory posits that genomes complexify rapidly, e.g., after a whole-genome duplication (inflation) [29,32,33]. The inflation is followed by lineages gradually losing genes, often reciprocally, that were present in their ancestor (streamlining). The reciprocal loss of ancestral genes is the leading cause for the species-rich clades of today.

The evidence for the prevalent and reciprocal loss of genes is found in the heterogeneous presences and absences of genes, in other words, the patchy patterns of genes in eukaryotic lineages. Typical examples of proteins showing patchy patterns are the wingless (Wnt) protein family in animals [31], and the kinetochore protein complex [34,35], ciliary proteins[36], and Rab GTPases [37] in eukaryotes. Large scale studies also showed that the loss of genes is even more prevalent than the gain of genes [31,37–39].

Patchy gene patterns form because functionally linked genes tend to co-evolve [31]. For instance, some clear examples of genes that are functionally linked and co-evolve are DNA (homologous recombination) repair genes [40] and important eukaryotic reproduction genes [41]. The co-occurrence of genes suggests that the genes co-evolved. In turn, these patterns of co-presence and co-absence of proteins can help infer protein function with the principle of "guilt-by-association" [42,43]. If proteins are similarly present or absent in a set of species, i.e., co-occur or are co-absent in a set of species, they are likely functionally linked.

The presence and absence of genes in genomes, or gene content, is a strong phylogenetic signal for genome evolution [44,45]. Inferring the presence and absences of proteins and their evolution in species often starts with a search for the same protein in multiple species, i.e., orthologs in other species [46] (Box 1). If we want to predict and understand the function and evolution of genes, we should start by looking for orthologs. However, finding orthologs has its challenges and systematic errors in orthology inference can affect the inferred phylogenetic signal [47].

ORTHOLOGY PREDICTION

Making a distinction in orthologous (and paralogous genes) (Box 1) is relevant in a wide range of applications, most notably inferring protein function and

understanding gene evolution[13]. Genes and genomes are subject to strong evolutionary forces that drive their evolution. For instance, horizontal gene transfer from one species to another (other than inheritance) and gene loss is a major evolutionary force in prokaryotic genome evolution. Although less predominant, horizontal gene transfer has also shown to occur in eukaryotes [48]. It is still highly debated [49,50] how much horizontal gene transfer contributed to the overall eukaryotic genome evolution. In eukaryotes, the major evolutionary forces that influence genomes and gene content are manifold. For instance, gene domain dynamics (losses, gains and duplications), endosymbiosis events, gene duplications, whole genome duplications, gene loss, and gene fission and fusion. All these events obscure clear orthologous relationships between genes.

Biology is not the only factor making it challenging to discern orthologous groups from any group of homologous genes. Many technical issues I mentioned in previous sections are also at play. For instance, incomplete gene annotation or assembly of genomes that cause incomplete or incorrect sets of orthologous genes [51]. Also, high evolutionary rates of genes make it harder to infer homologous relationships between sequences, and consequently orthologous relationships between them, warranting more sophisticated methods to find these orthologs [52].

Manual detection of orthologs [46] might be accurate and effective enough for a small number of genes. However, automatic detection of orthologs is needed when many genes and genomes come into play. This is especially the case for large scale comparative genome studies that, for instance, want to observe the dynamics for individual genes and how they compare to entire gene sets and species sets. Multiple automatic orthology inference methods or orthology databases exist (reviewed recently by [53]). Multiple methods exist, because there is still no consensus or common solution for dealing with complex evolutionary behaviour of genes and genomes.

The different orthology methods make use of different or multiple strategies to deal with the complexity of genome evolution. Not surprisingly, tests and benchmarks to evaluate and increase the accuracy of these methods are available as well, e.g., Quest for Orthologs [54,55],

13 Most notably for this thesis. One other relevant application is for instance phylogenetic tree inference.

Orthobench [56]. Benchmarking orthology inference methods is often based on manually curated sets of known orthologous protein families. However, benchmarking various methods remains a difficult problem in its own right, since there are multiple conceptual and practical difficulties to overcome. One such practical difficulty is often the small size of manually curated orthologous groups to benchmark with. Small sets are hard to generalise with and prone to overfitting.

Orthology methods might also be tailored to certain biological questions or phylogenetic species ranges. For instance, the orthologous group databases dedicated to a particular phylogenetic range, like animal orthologues groups in Treefam [57]. If we are interested in ancestral genes in animals, then Treefam might be a good first choice. For other questions, the choice is much less clear. The importance of choosing the right methods and databases for your data and biological question we evaluate in Chapter 2, where we look at the outcome and behaviour of different orthology methods when used to look at phylogenetic patterns at the level of eukaryotes, particularly LECA. If we are interested in ancestral genes in eukaryotes, then we might be better off selecting datasets or methods that (can) include a diverse range of eukaryotes with preferably genomes that take pivotal positions in the evolution of eukaryotes, such as eggNOG [58], Ancestral Panther [59] or inferring our own orthologues groups.

PHYLOGENETIC PROFILING

Orthologous genes can function as the genetic or operational unit to infer a gene's presence or absence (gene content) across multiple species, since orthologues genes in multiple species should stem from a single gene in their last common ancestor (Box 1). One way to capture the gene content and gene patterns in genomes is to use phylogenetic profiles. Phylogenetic profiling is a seemingly straightforward approach. The profiles in their simplest form are no more than binary vectors that contain 1's for presences and 0's for absences of proteins in multiple species. Consequently, this makes the method highly scalable to large data and particularly useful for large sets of species and genes.

The presences and absences of genes in phylogenetic profiles are a result of evolutionary processes that together with a species tree can represent evolutionary gene trajectories (Figure 2) [28]. By comparing phylogenetic

profiles, we can deduce whether two or more proteins have similar trajectories and are co-evolving, i.e., if two or more proteins co-occur or are co-absent. If the profiles are very similar, it might suggest that the proteins are functionally related or possibly interdependent for their function. One straightforward example is shown for haemoglobin α and β that co-occur or are co-absent in multiple jawed vertebrates (Figure 2) The profile similarity between both of the haemoglobin chains indicates their functional link, which we know is to form the functional haemoglobin protein complex. If profiles are opposite each other, it might suggest that different proteins have (developed) similar functions in different species. This is the case for the profiles of haemoglobin with haemocyanin.

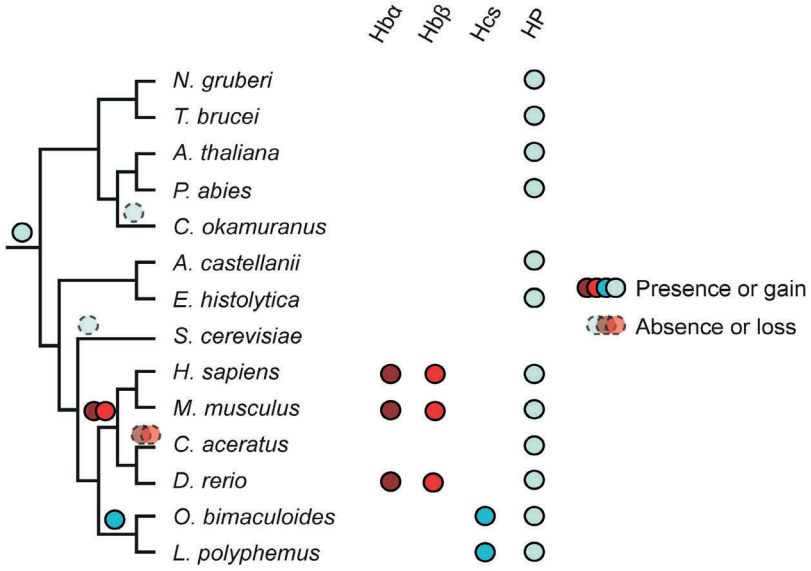


Figure 2. Phylogenetic profiling indicating co-evolution and possible functional interactions. Phylogenetic profiles represent the presences and absences of genes across a selection of (eukaryotic) species. For example, the genes for haemoglobin α and β (Hb α and β) are always present together in the jawed vertebrates (*H. sapiens, M. musculus and D. rerio*), unless the genes are lost like in the *C. aceratus*. This correlation of profiles indicates co-evolution and that the proteins depend on each other for their function, which we know to be the case. Haemocyanin (Hcs) is only present in Mollusca (*O. bimaculoides*) and Arthropoda (*L. polyphemus*). Anti-correlating profiles like that of Hb and Hcs might indicate that different proteins perform similar functions in different organisms, which we also know to be the case for these two proteins. HP is a hypothetical protein that is found in most of the eukaryotic species represented here and is an example of a gene that was likely already in LECA and lost twice independently (in *S. cerevisiae* and *C. okamuranus*).

Phylogenetic profiling has proven successful to infer protein function. A recent example predicted the function of CENATAC, a protein that plays a role in human Aneuploidy and causes several human defects [60]. With phylogenetic profiles, patterns can emerge that indicate which proteins have co-evolved, been lost together, or arose due to duplications. In turn, co-evolution can shed light on protein function. If a functionally uncharacterised protein co-evolves with a functionally characterised protein, these proteins' functions are likely similar. For instance, they cooperate in the same process or interact.

With the need to infer the function of the ever-growing dark matter of genomes, phylogenetic profiling provides a relatively straightforward way to do so. The method is independent of expensive and intensive experimental methods. There are of course some shortcomings to this method. First, when used in large scale analysis there is no manual curation like we could do with the species tree in Figure 2. The method does not take phylogenetic relationships between the species into account (even though the name says otherwise). Secondly, since the method is a "post orthology-inference" technique, it is highly dependent on the (correct) inference of orthologs. Orthology inference is at this point a bottleneck for both computational cost and performance. Phylogenetic profiling can benefit from certain species and orthologous group selection procedures (Chapter 3 and 4), particularly when handling eukaryotes with complex evolutionary behaviour.

IN THIS THESIS

Biologists are increasingly relying on bioinformatics-based tools and methods to generate or analyse biological information. This era's bottleneck is not the amount of data, but its annotation, quality and control, management, and extracting useful biological insights. Bioinformatic tools and methods are in constant development and improvement to handle large amounts of data accurately and efficiently, and while doing so making it possible to observe trends in genome evolution shaping the genomes of eukaryotes.

For evolutionary genomics, development and improvement of even the very basics underlying tools and methods is still in continuous motion. For instance, the ongoing discussions regarding a core theory

in evolutionary genomics the orthology conjecture (Box 1) [20,21,61–63]. But this also includes the improvement of biological knowledge, such as finding or correcting the root of the (eukaryotic) tree of life [25]. Also, the improvement of methodological knowledge, such as helping increase the accuracy of orthology prediction methods [55]. There are still many complicated issues to deal with if we want to properly study eukaryotic genome evolution, such as the difficulty of automatically predicting proteins, orthologs, and protein function. To be able to still move forward in genome evolutionary analyses, we use the bioinformatic toolkit we have and also continue to improve on the tools in it and the inferences we make with them. In this thesis I will focus on improving multiple methodological aspects and inferences made in evolutionary genomics. In other words, I will tinker with the toolkit for comparative genomics of eukaryotes.

Gene loss has been an unexpected big trend coming out of computational genomics research done in the last decade. However, we also know that genome data and annotation is not perfect. The number of genomes that do not exceed draft quality is increasing, which results in errors in annotating and gene numbers [64]. These errors suggest that gene loss to a certain extent could be the result of genes that are falsely inferred to be absent. In **Chapter 2** we explore the effect of gene prediction errors on gene loss estimates in eukaryotes. There are multiple non-biological reasons why a gene might be absent in a genome. In Chapter 2 we explore one of these reasons, the imperfect automatic prediction of proteins from their genes in DNA. We also evaluate if species-specific absences should be regarded as suspicious from the start.

To be able to do large scale analyses of genome evolution and the evolution of genes, it is needed to infer orthologs on a large scale and use tools that automatically infer orthologs. In **Chapter 3** we will dive deeper into automated orthology prediction. We compare multiple automatic orthologous group inference methods by their ability to recapitulate several observations and trends in eukaryotic genome evolution. Certain automated orthology methods might be more appropriate for certain biological questions. We evaluate the gene content of the Last Eukaryotic Common Ancestor's (LECA's), the pervasiveness of gene loss, protein interaction prediction with phylogenetic profile similarity (co-occurrence), and the overlap with manually determined orthologous groups that were present in LECA.

Phylogenetic profiling in eukaryotes still shows good promise to predict and study functional relationships between proteins. A fair amount of research has been done on mainly reference species selection for phylogenetic profiling in prokaryotes. However, it proves difficult to obtain a high performance in eukaryotes due to the complexity of eukaryotic genome evolution. We were able to get a relatively high performance in Chapter 3. In **Chapter 4** we build on the results and knowledge gained in Chapter 3. We will look at species, orthologous group and reference interactome selection for the use in large scale phylogenetic profiling in eukaryotes for protein interaction prediction. In Chapter 4 we evaluate our choices made in Chapter 3 to understand why certain meta parameters influence phylogenetic profiling in eukaryotes.

In **Chapter 5** we not only want to understand why certain meta parameters influence phylogenetic profiling in eukaryotes, but also increase the performance of protein interaction prediction further using phylogenetic profiling with machine learning. Machine learning can help identify patterns that are not clearly visible to us through the analyses done in Chapter 4, such as complex interactions between (sets of) species. However, machine learning is a relatively hard, its success often depends on the right algorithm for the data at hand. In Chapter 5, I will explore how to effectively use machine learning on our data to increase the performance of protein interaction prediction using phylogenetic profiling.


1

CHAPTER 2.

Measuring the impact of gene prediction on gene loss estimates in eukaryotes by quantifying falsely inferred absences

Eva S. Deutekom, Julian Vosseberg, Teunis J.P. Van Dam, Berend Snel

PLoS Computational Biology 15(8): 1–15 (2019).

Availability and implementation

All data can be reproduced with the code provided at:
https://github.com/ESDeutekom/ImpactGenePrediction

ABSTRACT

In recent years it became clear that in eukaryotic genome evolution gene loss is prevalent over gene gain. However, the absence of genes in an annotated genome is not always equivalent to the loss of genes. Due to sequencing issues, or incorrect gene prediction, genes can be falsely inferred as absent. This implies that loss estimates are overestimated and, more generally, that falsely inferred absences impact genomic comparative studies. However, reliable estimates of how prevalent this issue is are lacking.

Here we quantified the impact of gene prediction on gene loss estimates in eukaryotes by analysing 209 phylogenetically diverse eukaryotic organisms and comparing their predicted proteomes to that of their respective six-frame translated genomes. We observe that 4.61% of domains per species were falsely inferred to be absent for Pfam domains predicted to have been present in the last eukaryotic common ancestor. Between phylogenetically different categories this estimate varies substantially: for clade-specific loss (ancestral loss) we found 1.3% and for species-specific loss 16.88% to be falsely inferred as absent. For BUSCO 1-to-1 orthologous families, 18.3% were falsely inferred to be absent. Finally, we showed that falsely inferred absences indeed impact loss estimates, with the number of losses decreasing by 11.8%.

Our work strengthens the increasing number of studies showing that gene loss is an important factor in eukaryotic genome evolution. However, while we demonstrate that on average inferring gene absences from predicted proteomes is reliable, caution is warranted when inferring species-specific absences.

AUTHOR SUMMARY

To understand the evolution of eukaryotic species, we can look at the differences and similarities in their genomes. Since the first genomes were sequenced, scientists have, among other things, been studying these differences and similarities by evaluating the presences and absences of genes, and they have been trying to understand how these patterns explain the evolution of different eukaryotic species from their last common ancestor. It is now known that the evolution from the last eukaryotic common ancestor was dominated by gene loss and duplications.

Here we want to take the presence and absences patterns of genes in 209 diverse eukaryotic species as a guideline to estimate the loss of genes in these different species after they evolved from their common ancestor. Following this, we want to quantify how this loss estimate and the inferred absences of genes are influenced by faulty gene predictions by comparing absences of predicted proteins and absences of genes in genomes. A difference in these absences will indicate that some of them are not absent at all. It is important to quantify how many genes are falsely inferred as absent due to prediction problems and if this can be estimated from certain suspicious patterns in absences.

Our results show that overall gene absences are inferred reliably. However, suspicious absences in a species, i.e., absences that are species specific and not supported by absences in other closely related species, have a higher chance of being falsely inferred.

INTRODUCTION

During the evolution of eukaryotic genomes, the number of gene loss events is estimated to be higher than gene gains [31,37–39] and this high loss gives rise to patchy phylogenetic patterns of gene occurrence. A high level of gene loss suggests a gene rich ancestor of eukaryotes. Alternatively, patchy phylogenetic patterns of genes could also be indicative of horizontal gene transfer (HGT) from prokaryotes to eukaryotes or from eukaryotes to eukaryotes [39]. Nevertheless, studies showed that generally these patchy patterns are better explained by differential gene loss and gene presence in the Last Eukaryotic Common Ancestor (LECA) [38,39] and not by HGT. It has been proposed that in evolution new genes and functional repertoires originate in rapid genome expansions, followed by adaptive genome streamlining, or gene loss, giving rise to divergent species [30,33].

A small number of highly debated reports on gene losses [65,66] turned out to have incorrectly inferred genes as lost [67–69]. In fact, there are many reasons to presume that not all inferences of gene loss are equally trustworthy. The number of genomes published that do not exceed draft quality is increasing, resulting in annotation errors and errors in the number of genes found in the genome [64]. This suggests that the reported high number of loss events to some extent could result from genes whose absences have been falsely inferred. Genes can be inferred as absent for multiple reasons: due to technical difficulties in genome sequencing [68], due to misassembly of draft genomes, due to faulty protein prediction [64], due to insensitivity/bias in sequence similarity detection [70], or they are a bona fide loss. Recently, partial Pfam domain hits were in part attributed to incomplete gene models, yet another type of gene prediction and annotation problem [71].

Measuring the absences of genes that are expected to be universally conserved in organisms is a popular measure of genome annotation quality and completeness. The CEGMA pipeline [72] and later the BUSCO tool [73] successfully implemented this principle using near-universal single-copy orthologs. Absences of these single-copy orthologs are considered to be suspect and are widely used to quantitatively assess annotation quality and genome completeness.

While analysing the kinetochore protein complex and the absences of its subunits in eukaryotes, we recently showed that 10.9% of these absences could be found in six-frame translated DNA [74]. These falsely inferred absences in the kinetochore included important subcomplexes that would have otherwise been assumed to be absent in multiple species. One example was KNL1, a two sub-unit complex consisting of Knl1 and Zwint1, which plays a crucial role in microtubule attachment to the centromeres during mitosis. The KNL1 complex was wrongly inferred as completely or partially absent in 19 out of 109 species due to prediction problems. For 3 out of 19 species the complete complex was incorrectly inferred as absent, for 5 species the subunit Knl1 was falsely inferred as absent, and for 11 species the subunit Zwint1 was falsely inferred as absent. The study also showed that absences have a higher chance of being falsely inferred when they were species-specific absences or made little biological sense due to e.g. functional restrictions in protein complexes [74].

There is ample anecdotal evidence that poor gene annotation will influence gene loss analyses [65,74]. However, we here aim to systematically quantify the impact of gene prediction on the estimated gene loss by reanalysing absences inferred from predicted proteomes by analysing six-frame translated DNA. In particular, we hypothesize that absences that are not supported by absence in sister taxa are more likely to be false. Therefore, additional to the overall analysis of absences, we test the hypothesis that species-specific absences will be more likely falsely inferred as absent. We find that gene prediction in general is trustworthy and that loss remains an evolutionary important factor in eukaryotic genome evolution, with the caveat that suspicious, or species-specific, absences have a substantially higher chance of being falsely inferred.

RESULTS

Loss of inferred ancestral Pfams

To measure the impact of gene prediction on apparent gene loss in eukaryotes, we first inferred a list of proteins that indicate loss of these proteins in present-day species. For this we first estimated their presence in the Last Eukaryotic Common ancestor (LECA). Instead of utilizing orthologous relations, we used the Pfam domain family database [70] to detect homologous protein domains in the predicted proteomes of present-day species. Pfam domains have the advantage that they are clearly defined

units for detecting protein homology, whereas other databases would make it necessary to differentiate between paralogs and orthologs of partial hits in the DNA or make it necessary to call fusion and fission relationships of genes, which is easily subject to error and remains one of the largest problems within bioinformatics [75]. Another advantage of using Pfam is that it allows us to compare our loss and LECA estimates to previous work that analysed eukaryotic genome evolution on the scale of protein domains [38].

We analysed the presence of Pfam domains in 209 proteomes from a diverse set of eukaryotic species that can be divided into six supergroups: Amoebozoa, Archeaplastida, Cryptophyta/Haptophycea, Excavata, Opisthokonta and SAR (consensus species tree shown in Supplementary Figure 1 and species summarized in Supplementary Table 1). We then inferred potential LECA domain presences using the Dollo parsimony method and consequently inferred losses. In this method, domains can only be gained once and domain losses are minimised. The resulting LECA domain content consists of 5479 Pfams (Table 1, Proteome data), which is comparable to the LECA content of Pfam domains as previously estimated by [38] using a similar method. Our estimate of LECA content is higher than the previously estimated LECA content (5479 versus 4431) as we use more species, as well as more evolutionary distant species.

The LECA gene content inferred from naïve Dollo parsimony is very sensitive to horizontal gene transfers (HGT). If there were independent HGT events from bacteria to multiple lineages at both sides of the root in the eukaryotic tree, the Dollo parsimony method would incorrectly infer presence of that domain in LECA and thus infer many incorrect loss events. Therefore, we subsequently removed Pfam domains that were likely HGTs from bacteria to increase the reliability of our LECA estimate. These possible HGT Pfams were inferred based on a phylogenetic position of eukaryotic sequences among prokaryotic sequences or on being present in a small subset of eukaryotic species. Upon removal of these Pfams, the LECA content decreased to 4182 Pfams (Table 1, Proteome data. Pfams shown in Supplementary Table 2). The 4182 LECA domains were inferred to be lost 111320 times, with a median of 26 losses per domain in our set of 209 species (Table 1). Our results are in line with previous reports, which also find a large number of gene loss [37–39].

Table 1. Summary of data and results from the proteome and six-frame translated genomes.

	Proteome data (N = 209)		
	BUSCOs	**Pfams**	
		Non-strict LECA[a]	**Strict LECA[b]**
Total (LECA) domains	303 (47874)	5479 (1145111)	4182 (874038)
Species-specific absences	5791	97655	71559
Clade-specific absences	n/a	419323	218203
Absences	6055	516978	289762
Loss	-	162671 Median: 30	111320 Median: 26
	Six-frame translated genome data		
	BUSCOs (N = 158)[c]	**Pfams (N = 199)[c]**	
Found species-specific absences	1093 (18.9%) Median: 18.3%	-	13111 (18.3%)[d] Median: 16.88%
Found clade-specific absences	n/a	-	4301 (1.97%)[d] Median: 1.3%
Found total absences	1093	-	17412 (6%)[d] Median: 4.61%
Loss	n/a	-	98209 Median: 23

(a) LECA inferred with non-strict Dollo parsimony criteria, similar as [38]. (b) LECA inferred with stricter Dollo parsimony criteria that includes removed horizontal gene transfers. (c) Only genomes with more than 5 BUSCO absences were added for further calculations, leaving 158 genomes. Due to unforeseen tool crashes during six-frame translation, 199 genomes were left for analysis with the Pfam set. (d) Between brackets are the percentages of found absences compared to total absences of that category in the proteome data.

Quantifying falsely inferred absences and possible differences between clade- and species-specific absences

An absence might be falsely inferred as a loss due to sequencing issues, genome assembly issues or incorrect gene prediction. While we are unable to correct for the sequencing and assembly issues, we are able to identify possible falsely inferred absences. We performed a hmmsearch of LECA Pfam domains against six-frame translated genomes of the proteomes that were initially analysed. Two examples of falsely inferred absences are schematically shown in Supplementary Figure 2. Not all Pfam hits were

expected to be true presences, since the residual homology of pseudogenes can also lead to the detection of a Pfam. Therefore, we excluded hits containing stop codons in the alignments as they are likely pseudogenes and instead inferred an absence. Following this, our pipeline retrieved hits for 6% of all previously inferred Pfam absences (16878), which thus represent potentially falsely inferred absences (Table 1), with a median of 4.61% over all our 199 six-frame translated genomes ("Pfam total" in Figure 1.B.). This estimate provides an upper estimate for this problem and as shown below is largely driven by a specific subset of false absences.

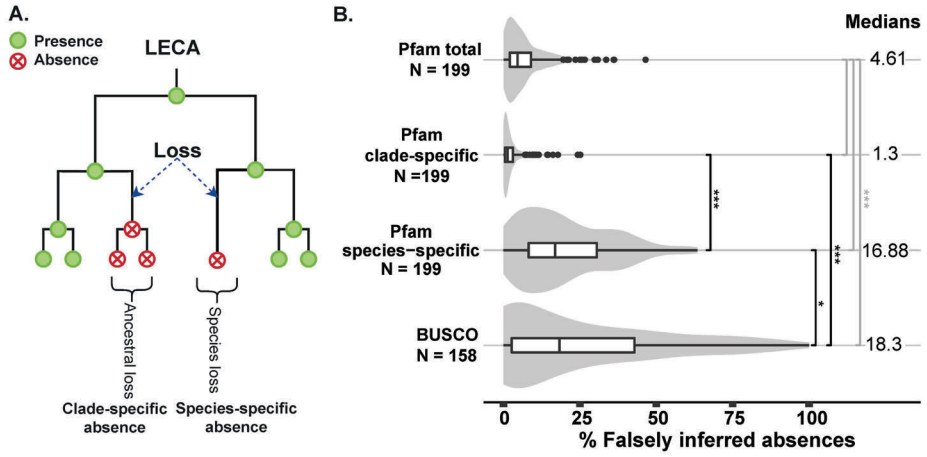


Figure 1. False inference of different absences. (A) Graphical representation of two different types of absences and loss. Clade-specific absences are phylogenetically supported by an ancestral loss. Neighbouring species, i.e., the clade, have the same absence. Species-specific absences are not phylogenetically supported by an ancestral loss, or in other words it is a single loss. A loss is independent of previous losses, in other words the first time a gene is lost. (B) Percentages of falsely inferred absences in different absence groups across genomes. From top to bottom the violin plots show: the percentages of falsely inferred absences in the total Pfam set absences, clade-specific absences and species-specific absences, and the BUSCO set absences. Since the BUSCO set contains a small number of domains (303), only the genomes with more than five absences (N = 158) were added to this figure. Note that the Pfam results are based on 199 species (N = 199) due unforeseen tool crashes during the analysis (see Materials and Methods and Supplementary Table 1). Significance levels of pairwise comparisons between groups are given with black asterisks and comparisons between total absences and the rest of the groups in grey. Significance levels are *** for $p \leq 0.001$ and * for $p \leq 0.05$ (Wilcoxon signed rank test). Data is summarized in Table 1. Violin plots are scaled to have the same maximum width.

Previous analyses suggested that not all absences were equally likely to be correct [74]. Absences that made little biological sense, i.e., were suspicious, tend to have a higher chance of being falsely inferred as absent. Often a suspicious loss was an observed single absence in a single species amidst a larger clade. To explore if there is a difference in detecting falsely inferred absences between suspicious and non-suspicious absences, we defined two categories of absences: clade-specific and species-specific (Figure 1.A.). Clade-specific absences are supported by an ancestral loss, meaning they are supported by absences in one or more directly related species with independently sequenced and annotated genomes. Species-specific absences are not supported by losses in directly related species. Absences in the BUSCO domain set (see Introduction) can be classified as suspicious absences as well, since all eukaryotes are assumed to have these single-copy orthologs. BUSCO therefore functions as an additional independently derived measurement of species-specific absences. We quantified to what extent these two types of absences are falsely inferred.

We found significant differences between species-specific and clade-specific absences in terms of their likelihood to be found in six-frame translated DNA. We found hits for 18.3% of the species-specific absences (Table 1), with a median of 16.9% per genome (Figure 1.B.). We found hits for 18.9% of the BUSCO absences, with a median of 18.3% per genome (Table 1 and Figure 1.B.). The median of falsely inferred species-specific absences in the Pfam set is surprisingly similar to that of the BUSCO absences, despite a weak positive correlation between these two sets (Supplementary Figure 3). In contrast, we found substantially (and significantly) lower hit percentages for clade-specific absences, with only 1.97% found for the clade-specific absences, with a median of 1.3% per genome. This 10-fold difference between found clade- and species-specific absences demonstrates that a species-specific absence has a higher chance to be a false absence than an absence that is supported by sister lineages. Moreover, it is this high rate of falsely inferred species-specific absences which significantly raises the overall rate of found absences to 6%.

Additionally, we focussed more on the phylum taxonomic level to see if there is a change in falsely inferred absences when looking at different phyla (Supplementary Figure 4). We observe the same trend as that on the level of LECA, with species-specific and BUSCO absences being more falsely inferred as absent than clade-specific absences. There seems to be

no specific trends in the individual phyla. However, it clearly shows that certain phyla are overrepresented. These results provide a straightforward, but effective way of guiding the detection of possible falsely inferred absences in both large- and small-scale evolutionary analyses.

From Figure 1.B. it is also clear that several species have a higher percentage of falsely inferred absences, shown by the outliers (black points). For these species, this could signify that they have either lesser quality genomes or predicted proteomes. In Figure 2 this is highlighted by high instances of red in a particular genome, which indicates a high number of found species-specific absences, or dark green, which indicates a high number of found clade-specific absences. This is also highlighted by the number of BUSCO absences found for the same genome (bar chart Figure 2). The genome specific values depicted in Figure 2 can be found in Supplementary Table 1.

We also took a subset of genomes, that can be considered model organisms for evolutionary studies, to analyse if any methodological differences between model and non-model organisms have an effect on falsely inferred absences (Supplementary Figure 5). For this subset of model organisms (N = 35), we can observe the same trend as that of the whole dataset, with species-specific absences being more falsely inferred as absent. Surprisingly, for the BUSCO set (N = 21) the median of falsely inferred absences per genome lies higher in the model organism subset, 34.29 % compared to 18.3%. Additionally, looking at N50 values of all the genomes, a proxy for genome assembly quality, we can see no significant link between falsely inferred absences and N50 values (Supplementary Figure 6). Therefore, rather unexpectedly, it appears that completeness of sequencing or assembly problems are not an indication for higher expected false absences.

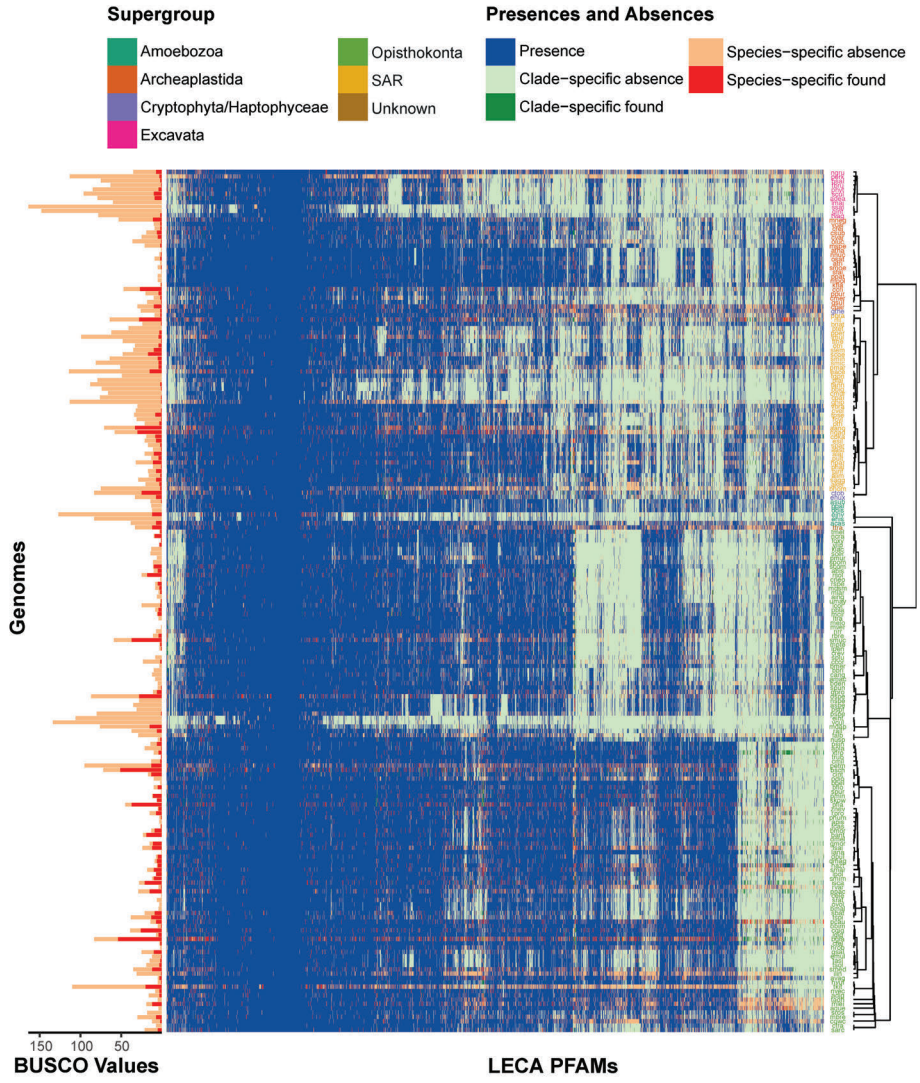


Figure 2. Presences and absences of all LECA Pfams in all 199 species. The barchart (top) shows the BUSCO absences and found BUSCO absences. The large matrix shows presences and all types of absences as shown in the coloured legend. Species are clustered according to the species tree (Supplementary Figure 1) shown by the dendrogram. Pfams are clustered with hierarchical (complete-linkage) clustering. Pfam labels are left out for clarity.

Another effect did become apparent during the analysis: short Pfam domains have a higher chance to be falsely inferred as absent. Supplementary Figure 7 shows Pfam lengths of the top 100 highest numbers of falsely inferred Pfam absences, showing a significant difference (almost twice as much) between the median of the Pfam lengths of the 100 most found Pfam absences versus the rest. This trend is potentially explained by short single domain proteins that fall just below the commonly used cut-off length of 100 amino acids in genome annotation pipelines for proteins with only *in silico* evidence [76,77].

Impact of incorrect gene prediction on gene loss estimates in eukaryotes

To answer the question whether incorrect gene prediction could influence genome evolution inferences, we re-analysed the loss events and corrected our initial estimated domains loss by including the hits we found in six-frame translated DNA. Figure 3 shows the loss corrected with the Pfam domains found in six-frame translated genomes (coloured bars) and the uncorrected loss according to proteomes (white bars). The number of times a LECA Pfam is lost in general shifts to lower values (Figure 3 inset). The number of Pfams with many loss events decreased. Notably, the Pfam domains that were conserved in all species, i.e., lost zero times, showed the largest increase, from 138 to 186 domains.

The found hits decrease the amount of loss by 11.8%, from 111320 to 98209, reducing the median loss per Pfam from 26 to 23 (p-value < $2.2\text{-}10^{-16}$ Wilcoxon signed rank test) (Table 1 and Figure 3 vertical lines in histogram). The reason for this relatively higher impact on loss estimates, despite the smaller percentage of 6% falsely inferred absences, is that every species-specific loss is counted equally as a clade-specific loss (Figure 1.A.). Since species-specific absences are much more likely to be falsely inferred as absent than clade-specific absences, they have relatively more impact on the amount of loss. Thus, species-specific loss and their higher likelihood for being falsely inferred as absent is a significant issue in comparative genomics studies on gene loss.

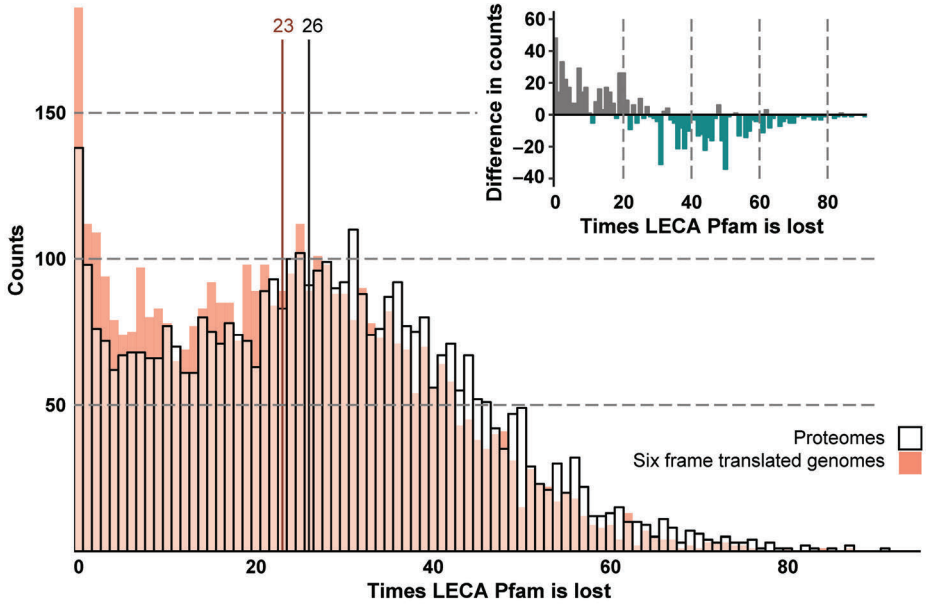


Figure 3. Times LECA Pfam is lost. Distribution of the estimated loss of LECA Pfam domains in proteomes shown by white bars, with the median loss given by the black vertical line. The Dollo parsimony approach places 4182 Pfams in LECA. These LECA Pfams have been lost independently 111320 times. A large number of Pfams are conserved in all current day species (never lost). Distributions of the corrected loss of LECA Pfam domains from six-frame translated genomes are shown by orange-coloured bars, with the corrected median loss given by the red vertical line. The inset shows the difference in distributions of the six-frame translated genomes minus the proteomes.

DISCUSSION

Eukaryotic genome evolution is dominated by gene duplication and gene loss [30,31,37–39]. However, absences of genes in predicted proteomes do not always indicate that these genes are truly lost. During the past few years high profile reports of specific cases of gene loss (peptide hormone ghrelin in soft-shell and sea turtle [65] and multiple genes in birds [66]) were disproven [67–69]. Falsely inferred absences could greatly influence conclusions drawn when analysing genome evolution, the evolutionary trajectory of proteins or protein complexes and adaptation of organisms. In our study, we showed that per genome 4.61% of absences are falsely inferred to be absent. Additionally, we showed that for the two different types of absences these percentages differ significantly: clade-specific

absences are only falsely inferred as absent 1.3% of the time, but species-specific are falsely inferred as absent 16.88% (Pfam) and 18.3% (BUSCO) of the time.

Our estimates rely on the specific design of our analysis, such as the use of Pfam HMMs and the Dollo parsimony approach. The Dollo parsimony approach is a simplified way of describing eukaryotic genome evolution: only one domain gain is allowed and the number of losses is minimized, effectively ignoring HGT events. The importance of HGT in eukaryotes remains controversial and is still an active area of study [49,50]. Nevertheless, the usage of Dollo parsimony allows direct comparisons with a similar approach previously described in Zmasek & Godzik [38], as well as give a straightforward way of defining LECA for identifying patterns in absences and identifying clade- and species-specific absences. Even though we are not trying to infer the gene content of LECA, we want to estimate the LECA Pfam content as accurate as possible because otherwise we cannot reliably interpret absences in terms of loss. Therefore, we additionally added a stricter criterion for accepting Pfams as LECA Pfams and removed possible HGT using a phylogenomics approach.

Our combined Dollo parsimony and phylogenetic HGT filtering approach, yields a LECA size in terms of Pfam domains comparable to that of Zmasek & Godzik [38] and in terms of genes to that of Wolf & Koonin [30]. It would be expected that the increased sampling in our work of more diverse genomes, such as free-living heterotrophs and poorly sampled taxa, would increase the number of inferred LECA Pfams compared to these earlier studies. At the same time, the phylogenetic approach for removing suspected HGT families, decreases the number of inferred LECA Pfams.

The number of losses might be influenced by uncertainties in the tree of life and its topology: a clade-specific absence might become a species-specific absence and vice versa due to minor rearrangements in the used tree topology. However, we expect that this will not significantly influence the results, since in general the leaves of the tree are confidently assigned and the uncertainties often lie in the specific hierarchy in higher-level taxonomy, such as the location of the root of the eukaryotic tree of life [78–80].

It is important to note that over the years improvements in species sampling and sensitivity in homology detection have led to drastically expand the gene content of LECA and in turn increase in loss events to the high numbers now reported [30,31,37–39]. However, the gene prediction problem is not the only technical issue influencing gene loss estimates. Other (technical) issues could also artificially increase gene loss estimates. For example, domain profiles (HMMs) can be insufficiently sensitive due to biased/limited sequence sampling or due to strict bit score cut-offs chosen [70] due to an (understandable) focus on avoiding false positives. Especially in lineages with rapidly evolving genes, unrecognized homologs can be the cause of falsely inferred absences and consequently higher loss estimates. Improving the sensitivity of HMMs of protein domains has anecdotally been shown to improve domain detection [74]. Another issue is incomplete genome assemblies, which preclude genes from being found. For instance, many gene absences in bird genomes were shown to stem from genome assemblies with stretches of strongly decreased coverage due to GC-rich regions [68]. Genes that are falsely inferred as absent due to incomplete sequencing of certain genomes can also not be found by simply searching the DNA sequence for homologs, as is done here. The combined effect of all these issues in addition to gene prediction is not known yet, but could further lower gene loss estimates.

With this study, we want to provide some guiding estimates of the extent of one particular technical problem, i.e., unpredicted genes present in genome sequences. This problem is in practice known, but to our knowledge has never been systematically quantified. Our results show that in general gene prediction is of good quality and inferred absences are likely not false. However, there is more than a 10-fold difference between the number of falsely inferred clade-specific absences (1.3%) versus species-specific absences (16.88% for Pfam and 18.3% for BUSCO). This is directly in line with the observation that ghrelin was already reported in the red-eared slider turtles and later indeed correctly inferred to be present (and not absent) in the genome of soft-shell and sea turtle [67]. The importance of gene loss for eukaryotic molecular evolution is fundamentally not impacted by falsely inferred absences and remains a dominant factor in shaping eukaryotic gene repertoires. Still, loss decreases by 11.5% due to falsely inferred absences that can be found in six-frame translated DNA and our study clearly demonstrates that biologically suspicious absences should invite additional technical scrutiny.

Conclusion

The results of our study show that when absences are surprising and/or suspicious they have a higher chance of being falsely inferred as absent. This result is especially important for the evolutionary analyses of proteins and their domains and estimating their loss. It provides a cautionary tale that if an absence appears suspicious there is a good reason to investigate this further and conclusions should not only rely on automated gene prediction alone.

Our findings agree with existing notions of gene prediction problems, but no study as of yet has quantified to what extent gene prediction influence gene loss estimates. Our simple but effective approach described in this study provides a straightforward way to analyse gene absences and quickly assess their reliability in large- and small-scale evolutionary analyses

MATERIALS AND METHODS

To measure the impact of gene prediction on gene loss estimates we first needed to establish gene content in the last eukaryotic common ancestor (LECA) to infer loss patterns from LECA to current day species. We did this by analysing the presences and absences of protein domains in current day species, and then inferring LECA content using these presence/absences profiles. With this LECA content we inferred the loss of LECA domains in current day species. Following this, we looked at protein domains that are not found in the proteomes to see if they were encoded in the genomes of the respective species. Supplementary Figure 8 schematically shows the procedures and the following sections describe these procedures in more detail.

Compiling the database

To study the presences and absences of genes across the eukaryotic tree of life we used predicted proteomes and genomes of 209 phylogenetically diverse eukaryotic organisms from multiple supergroups: 122 Opisthokonta, 6 Amoebozoa, 23 Archaeplastida, 3 Crypto-/Haptophyceae, 13 Excavata, 41 SAR, and 1 unidentified (species summarized in Supplementary Table 1). We chose these species to represent a broad eukaryotic diversity. The predicted proteomes and genomes were obtained from a variety of sources (Supplementary Table 1).

To examine if absences in the proteomes could still be found in the genomes of their respective species, we used the tool Transeq (Translate nucleic acid sequences) from EMBOSS [81] to translate the genomes in six open reading frames to protein sequences with the default codon table. For ciliate species, we used the ciliate codon table (table 6). We successfully analysed 199 of the 209 genomes (Supplementary Figure 1). One genome (human) could not be translated due to its large size and Transeq crashing as a consequence, two species did not have available genomes and seven translated genomes could not be analysed due to an unknown error in the hmmsearch tool (Supplementary Table 1).

Protein domain content in proteomes and translated genomes

The protein domain repertoire was determined with the hmmsearch alignment tool from the HMMER package 3.1b2 (dated February 2015) [82] using sequence profiles, HMMs (Hidden Markov Models), from the Pfam 31.0 database [83] and the BUSCO eukaryota database (*odb9*) [84]. We took HMM specific quality scores for Pfam (gathering cut-offs) and BUSCO domains to validate the hits in the alignments.

Some Pfam domains could be absent from predicted proteomes because they are (part of) a non-functional gene, i.e., a pseudogene. We therefore removed pseudogenes from our hits in six-frame translated genomes with a custom-built script that removed hits with stop-codons in their sequences. Best scoring non-overlapping hits were considered for further analysis in presence/absence profiles.

Approximating domain content of the last eukaryotic common ancestor with Dollo parsimony

We used the Dollo parsimony approach for the ancestral state reconstruction, i.e., the domain content of LECA, using presence/absence profiles of Pfam domains in the predicted proteomes and projecting them on a bifurcating species tree. The species tree is a consensus tree combined from literature, which is summarized in Supplementary File 1 and the species tree shown in Supplementary Figure 1). The Dollo parsimony code was updated and translated to python from [42]. This approach allows for a gene/domain to be gained only once through a phylogenetic tree, which may require an arbitrary number of subsequent losses, and traces presences/absences back to the root (LECA) of the tree. We added additional criteria to increase the accuracy of our LECA estimate by only considering

Pfam domains that are present in at least 3 supergroups and are left and right of the root (Supplementary Figure 1).

To remove Pfams that are in LECA due to possible horizontal gene transfer (HGT), we used a phylogenomics based approach. We inferred and analysed phylogenetic trees based on Pfam sequences containing sufficient phylogenetic signal from a diverse set of prokaryotes and eukaryote to identify possible HGT Pfams as follows. The eukaryotic database described above was supplemented with the prokaryotic proteomes in eggNOG4.5 [58] and the Asgard archaeal predicted proteomes from [85]. Pfam domains were detected with hmmsearch as described above. Reduction of the number of sequences was necessary to make it computationally feasible to apply sequence alignment and phylogenetic reconstruction. To reduce the number of sequences to be used in phylogenetic inference, kClust 1.0 [86] (clustering threshold 2.93) was performed on the eggNOG prokaryotic sequences and a ScrollSaw-like method [2] was applied to the eukaryotic sequences. The sequences in bidirectional best BLAST 2.6.0+ [87] hits (BBHs) between sequences from different sides of the eukaryotic root were selected. For each Pfam the selected prokaryotic and eukaryotic sequences were aligned (mafft v7.310 [88] auto option); these alignments were trimmed (trimAl v1.4.rev15 [89] gap threshold 10%). Phylogenetic trees were inferred with IQ-TREE 1.6.4 [90] (LG4X model, 1000 ultrafast bootstraps [91]). The resulting trees were analysed using the ETE3 toolkit [92].

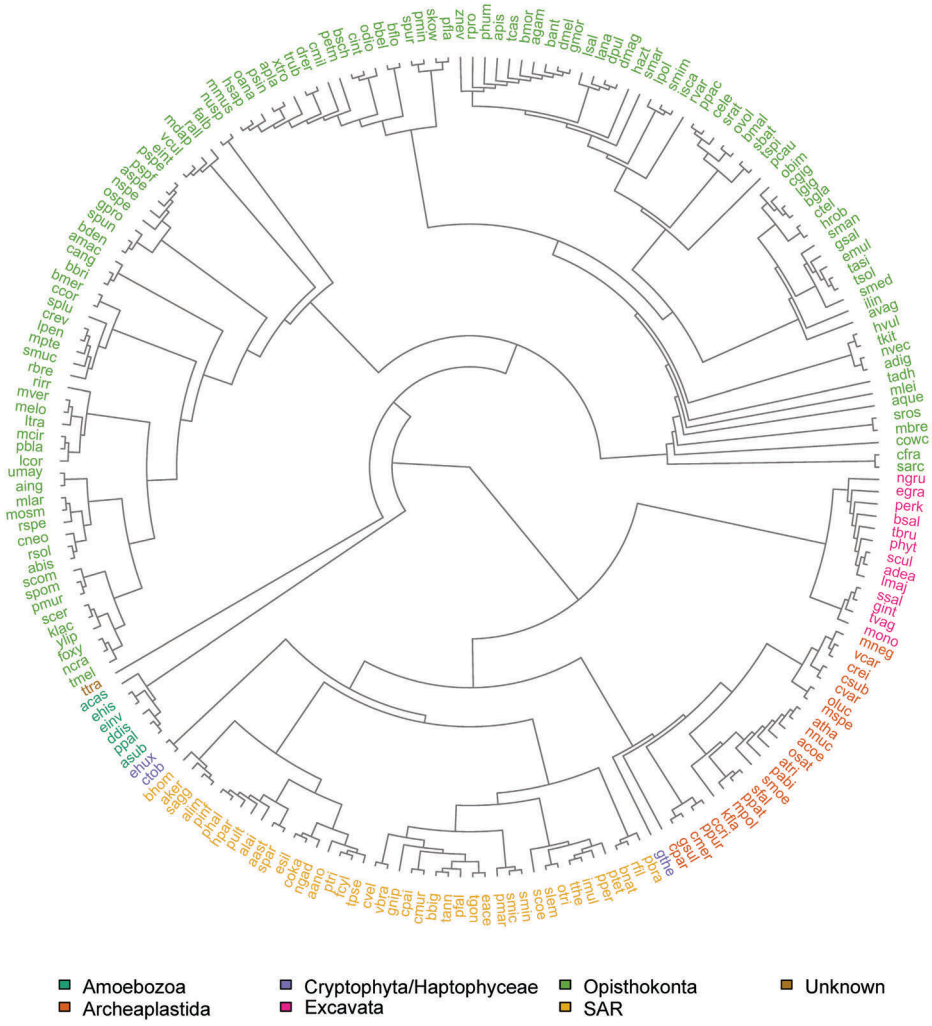
For each monophyletic eukaryotic clade in a tree, it was first checked if there were species from both sides of the eukaryotic root present in that clade. If at least one such potential LECA clade was present in the tree, the information from the eukaryotic sequences not in the BBHs, and therefore not in the tree, was incorporated. By assigning these sequences to their best representing hit in the tree, the percentage of species in which a homolog from that clade was present was calculated for five supergroups: Excavata, SAR + Haptista, Archaeplastida + Cryptista, Amoebozoa and Opisthokonta + Apusozoa. If the mean of these percentages was at least 15%, the clade was annotated as a LECA clade. If there was at least one LECA clade in a tree, the Pfam was annotated as present in LECA. Having a set of trusted LECA Pfams allowed us to remove the non LECA Pfams resulting from horizontal gene transfer, contamination or the chloroplast endosymbiosis from our LECA set.

We also defined two different groups of absences, clade- and species-specific. Clade-specific absences are supported by an ancestral loss of a domain, while species-specific absences are not (see Figure 1). We analysed events in the leaves of Pfam domain trees generated by Dollo parsimony. Leaves with ancestral losses (Pfam loss in parent node) are defined as clade-specific absences. Leaves with single (independent) losses (Pfam present in parent node) are defined as species-specific losses.

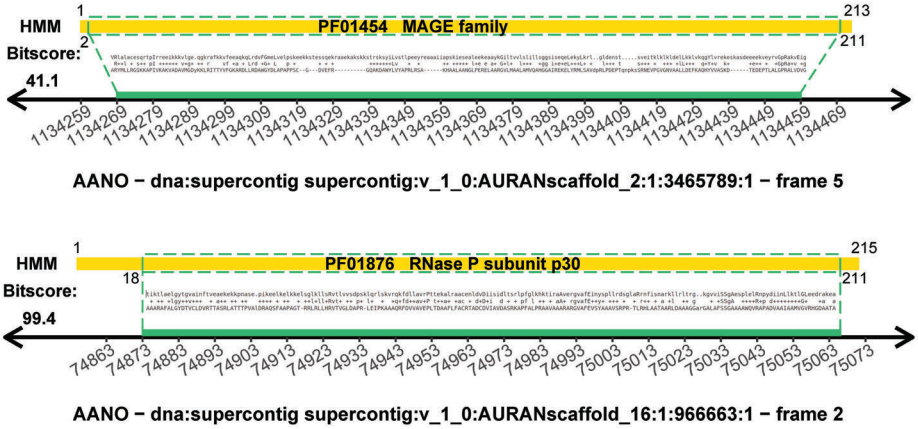
ACKNOWLEDGEMENTS

We thank Jolien van Hooff for collecting and compiling the species tree and the Theoretical Biology and Bioinformatics group for commenting on and discussing the manuscript. We want to thank Eelco Tromer for the extensive initial analysis of the kinetochore proteins. We would also like to thank Amir Masoud Abdol for revising and discussing the manuscript and improving the design of the figures.

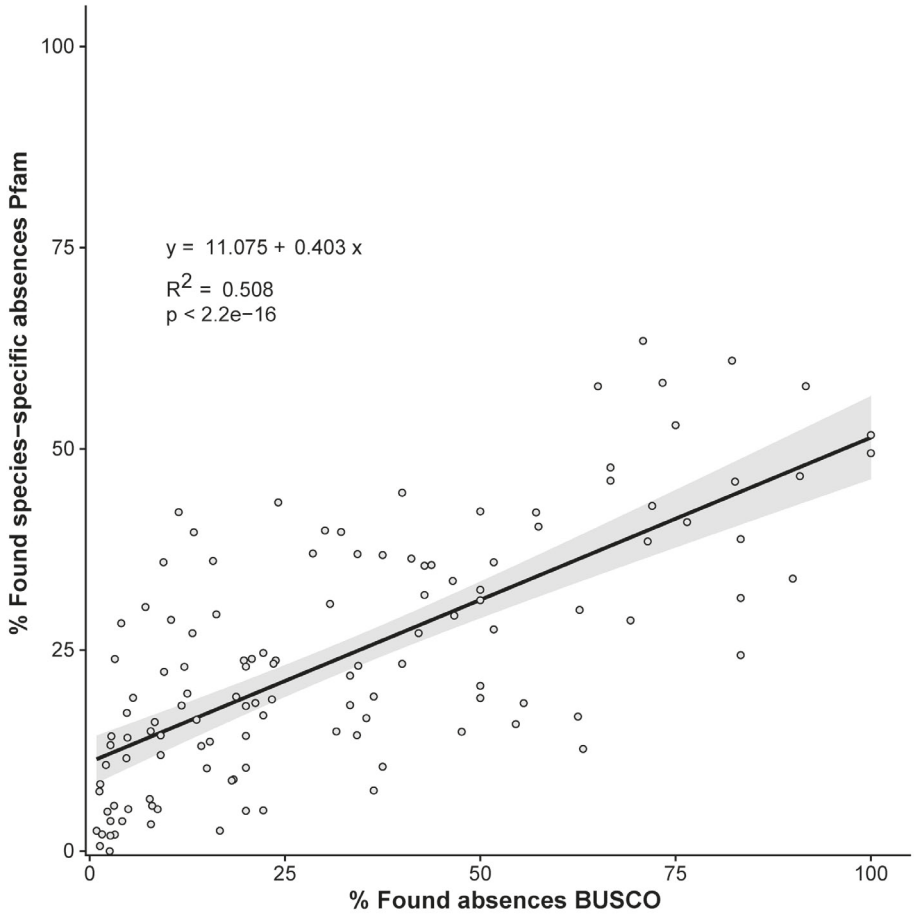
SUPPLEMENTARY FIGURES



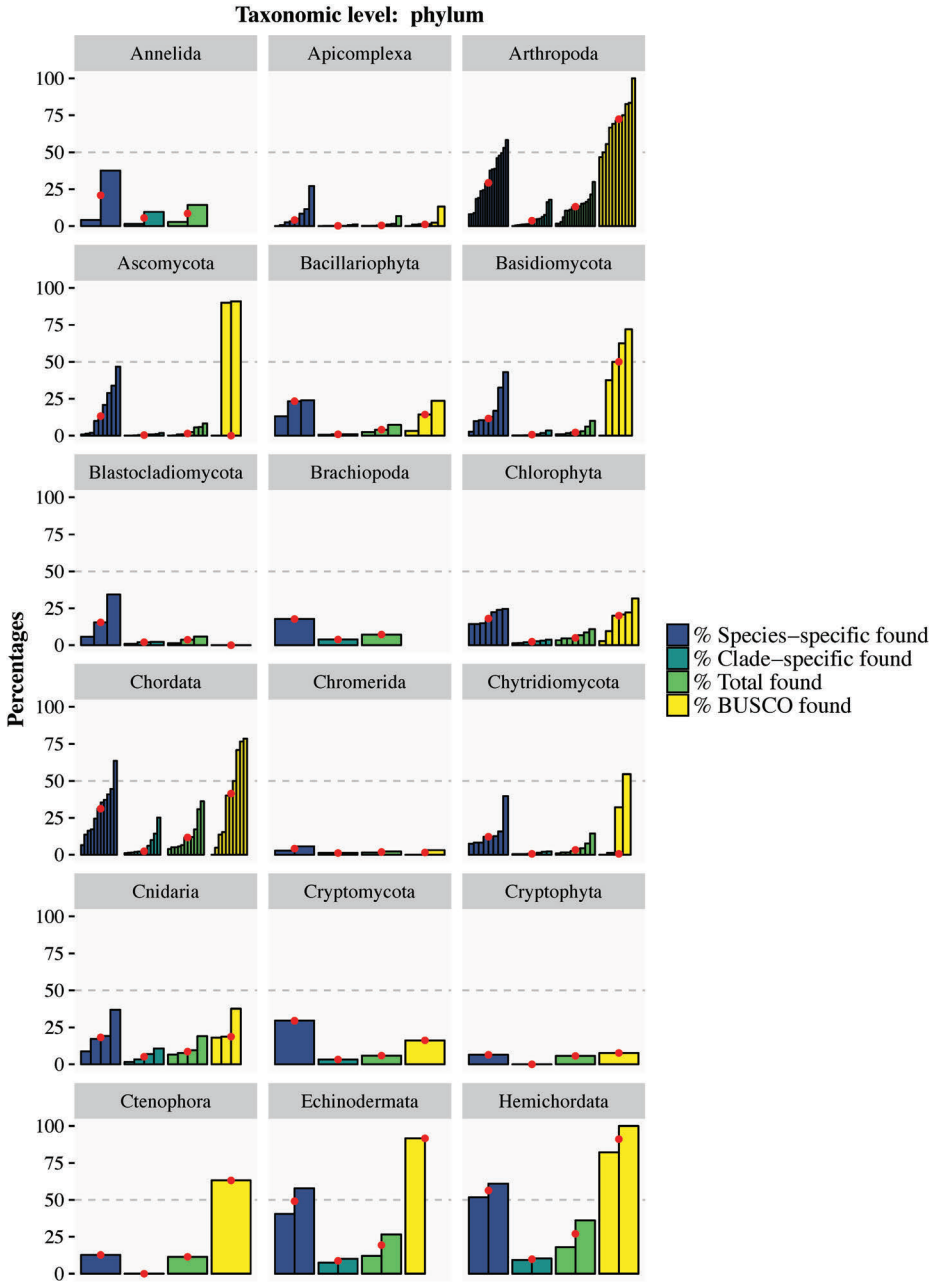
Supplementary Figure 1. Species Tree. A phylogenetic tree of the species used in this analysis. Supergroups are given indicated the legend and the full names that belong to the abbreviations can be found in Supplementary Table 1. Species with asterisks were used to estimate the LECA Pfam content with Dollo parsimony, but for multiple reasons (e.g., no genome available) they could not be used to quantify falsely inferred absences (see Results, Materials & Methods and Supplementary Table 1).



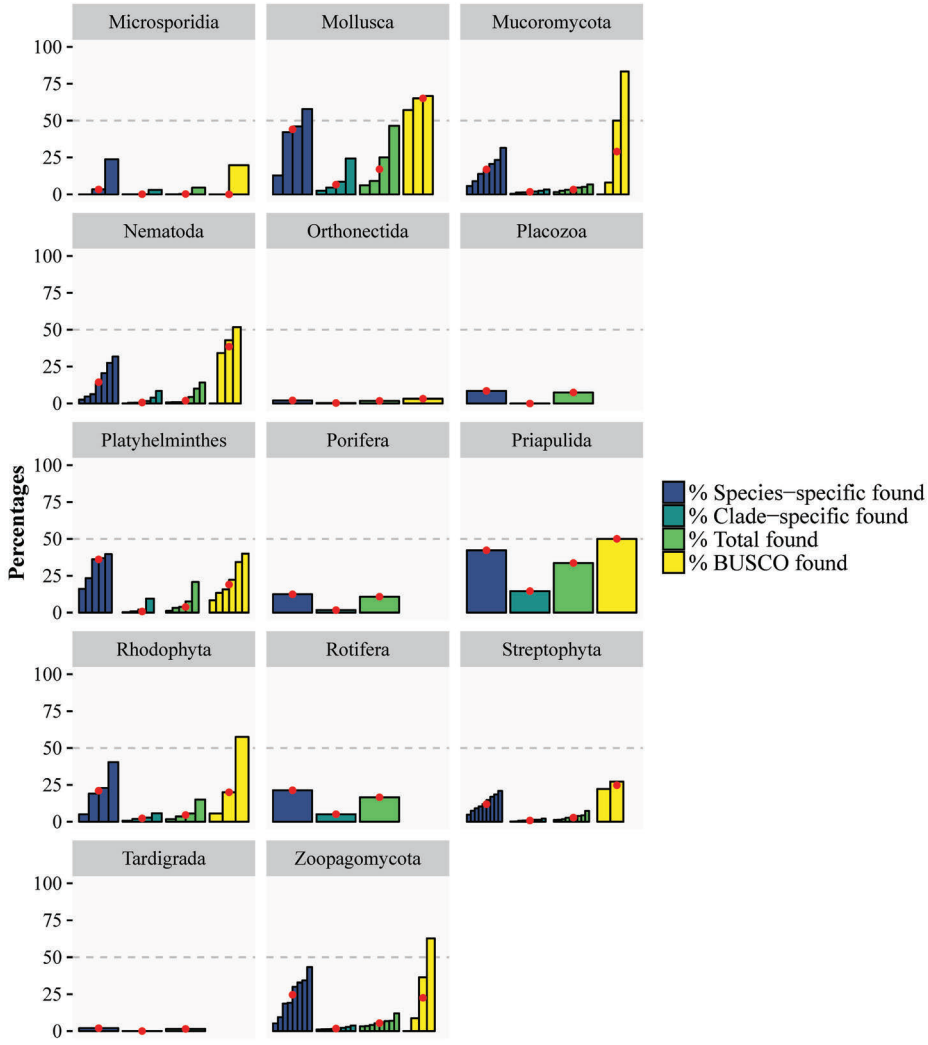
Supplementary Figure 2. Example of two hits falsely inferred absences in Aureococcus anophagefference. Two Pfam domains previously inferred as absent are found in six frame translated DNA. Shown are the HMM overlapping with the scaffolds (x-axis) together with the bitscore. The hmmsearch tool "sequence output" is shown between the HMM and scaffold.



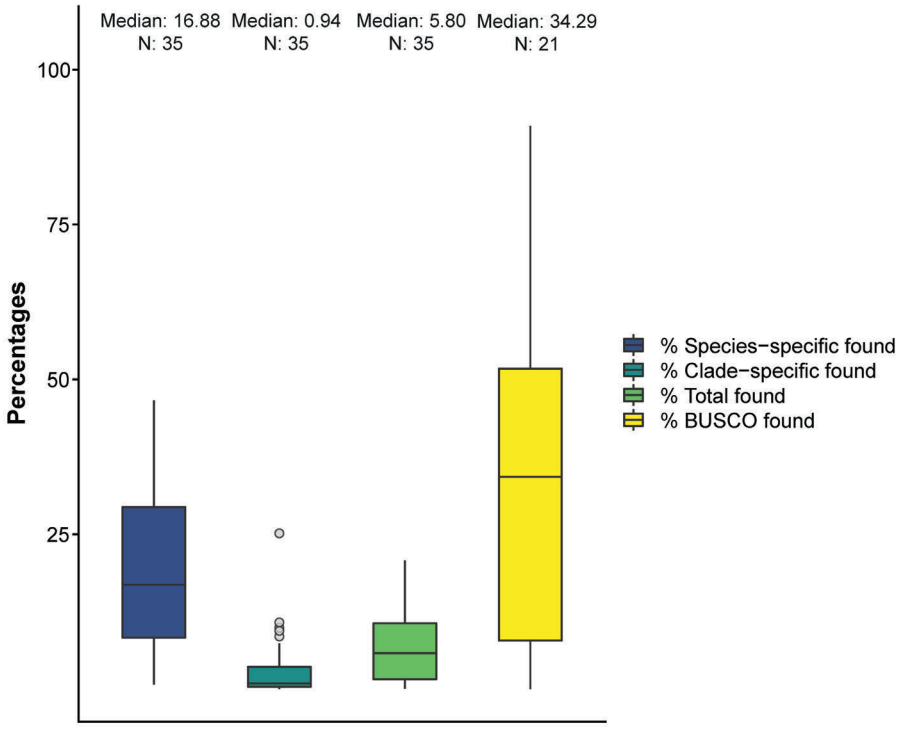
Supplementary Figure 3. Percentages of found species-specific Pfam absences vs. BUSCO absences per genome. We fitted a linear model (black line), shown in the graph with a 95% confidence interval (shaded area).



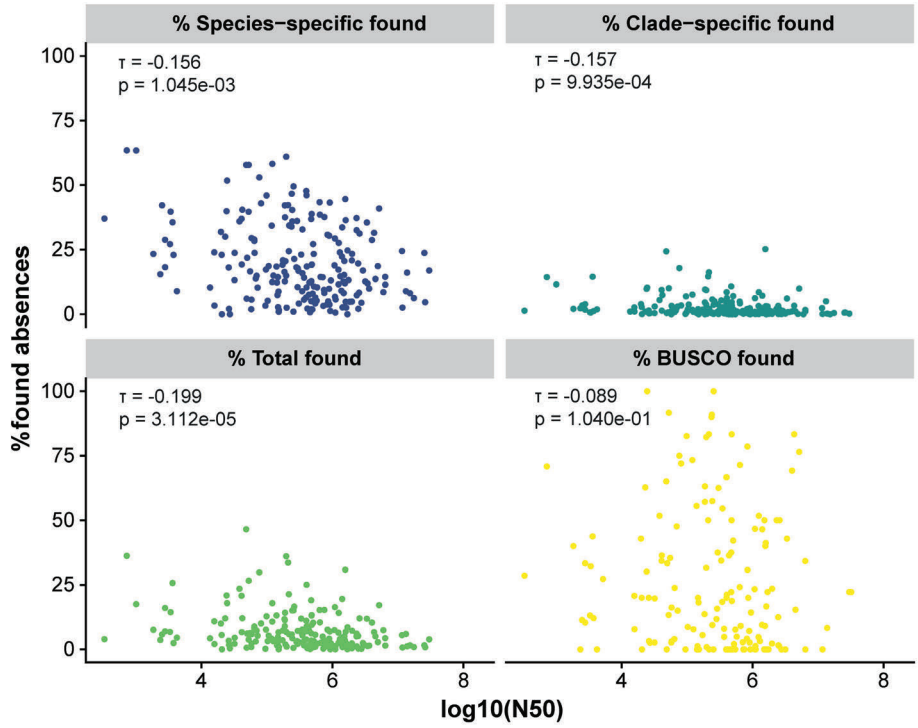
Supplementary Figure 4. Continues on next page.



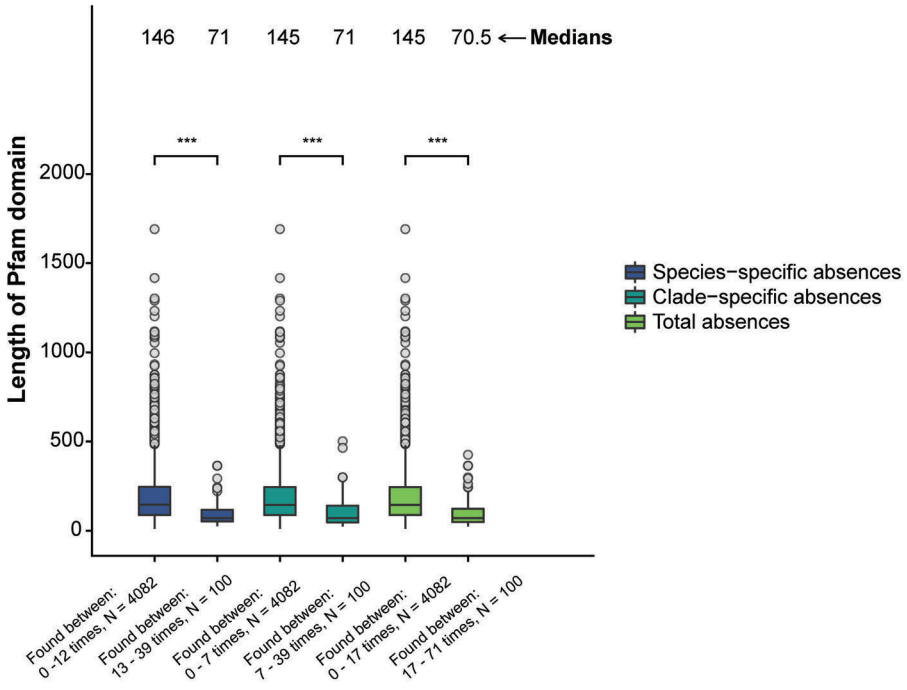
Supplementary Figure 4. Percentages falsely inferred absences found per genome, grouped per phylum. For all the genomes containing a phylum taxonomic annotation (N = 152), the genomes were grouped per phylum in a bar chart, showing percentages falsely inferred absences coloured by four absence groups. Median values are given by the red points (unless there is only one genome the red point is equal to the result) and for clarity grey dotted lines show 50% falsely inferred values. Individual phyla can highly differ in the number of genomes sampled, with Arthropoda having the highest number.



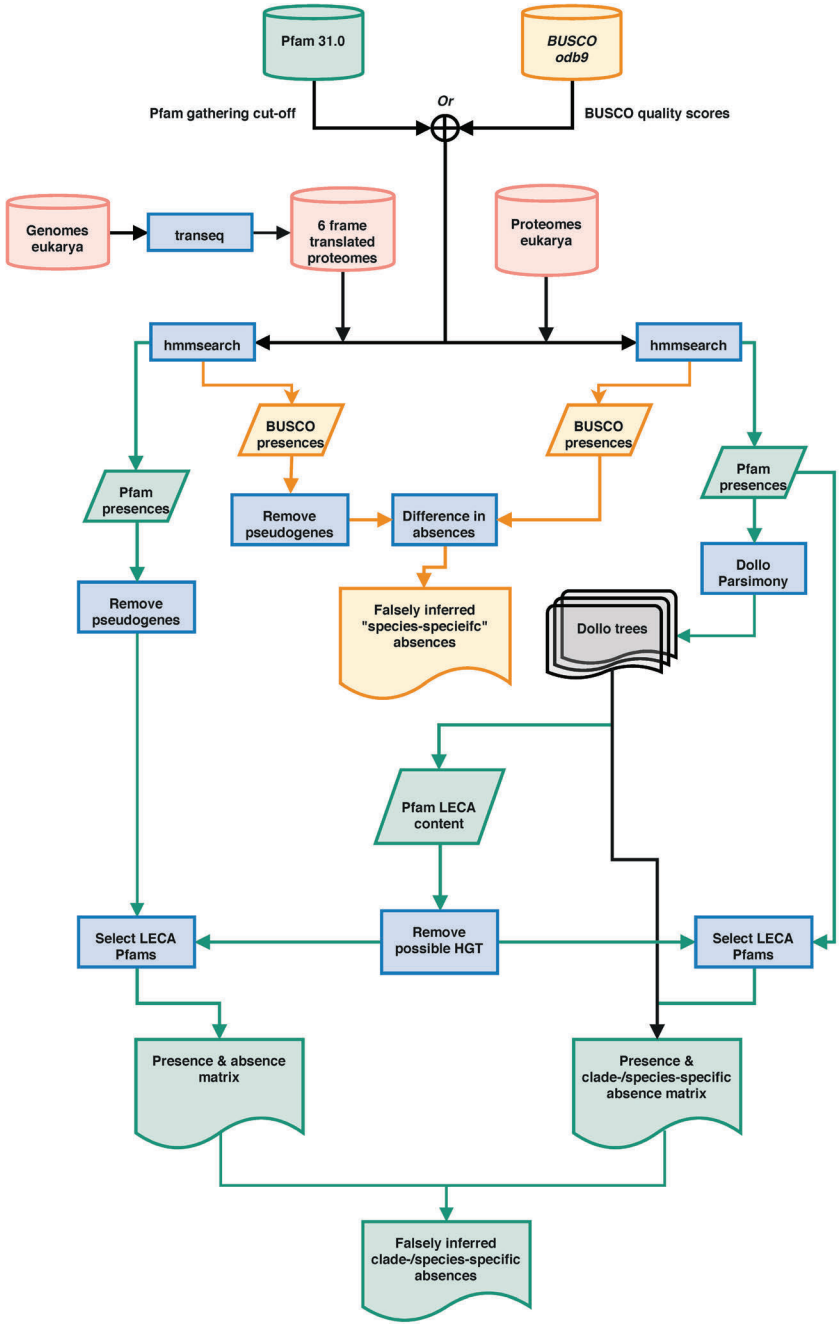
Supplementary Figure 5. Percentages falsely inferred absences found in model organisms. Percentages of falsely inferred absences in different absence groups in a subset of genomes representing model organisms (N= 35). The BUSCO set contains a small number of domains (303), only the genomes with more than five absences (N = 21) were added to this figure. Significance levels of pairwise comparisons between groups are given with black asterisks Significance levels are *** for $p \leq 0.001$ and * for $p \leq 0.05$ (Wilcoxon signed rank test).



Supplementary Figure 6. Comparing found falsely inferred absences with genome assembly quality (N50). The different panels show the different absence groups versus log(N50) values. In the upper left corner of every panel the correlation coefficient t is shown and corresponding p-value (Kendall rank correlation). There is little association found between the two values in either of the categories of falsely inferred absences.



Supplementary Figure 7. Pfam hmm lengths of found absences. Hmm lengths are compared in three different absence groups: all, clade- and species-specific, for the 100 highest numbers of absences vs. the rest. Medians are shown at the top of the graph and significance (Wilcoxon rank sum test) is shown above the comparisons.



Supplementary Figure 8. The workflow for quantifying falsely inferred absences. The BUSCO data is given in yellow, the Pfam data is given in green and processes are given in blue.

SUPPLEMENTARY TABLES

Supplementary Table 1. Per species data. Information on the species used in this study, including taxonomic information of each species, counts of absences and found absences, and download locations of genomes/proteomes.

https://github.com/ESDeutekom/ImpactGenePrediction/blob/main/SupplementaryTableS1.xlsx

Supplementary Table 2. Pfam data. Information on the inferred LECA Pfams, including lengths and counts of absences and found absences.

https://github.com/ESDeutekom/ImpactGenePrediction/blob/main/SupplementaryTableS2.xlsx

Supplementary Table 3. Species tree resources. The file contains a list of resources used for the reconstruction of the species tree (Supplementary Figure 1). The tree is used in this analysis to project presences (and absences) in the Dollo parsimony approach.

https://github.com/ESDeutekom/ImpactGenePrediction/blob/main/SupplementaryFileS1.pdf

CHAPTER 3.

Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes

Eva S. Deutekom, Berend Snel*, Teunis J.P. van Dam*
*These authors contributed equally to this manuscript

Briefings in Bioinformatics 22(3): 1–9 (2020)

Availability and implementation

The data and code underlying this article are available in github and/or upon reasonable request to the corresponding author: https://github.com/ESDeutekom/ComparingOrthologies.

ABSTRACT

Insights into the evolution of ancestral complexes and pathways are generally achieved through careful and time-intensive manual analysis often using phylogenetic profiles of the constituent proteins. This manual analysis limits the possibility of including more protein-complex components, repeating the analyses for updated genome sets, or expanding the analyses to larger scales. Automated orthology inference should allow such large-scale analyses, but substantial differences between orthologous groups generated by different approaches are observed.

We evaluate orthology methods for their ability to recapitulate a number of observations that have been made with regards to genome evolution in eukaryotes. Specifically, we investigate phylogenetic profile similarity (co-occurrence of complexes), the Last Eukaryotic Common Ancestor's gene content, pervasiveness of gene loss, and the overlap with manually determined orthologous groups. Moreover, we compare the inferred orthologies to each other.

We find that most orthology methods reconstruct a large Last Eukaryotic Common Ancestor, with substantial gene loss, and can predict interacting proteins reasonably well when applying phylogenetic co-occurrence. At the same time derived orthologous groups show imperfect overlap with manually curated orthologous groups. There is no strong indication of which orthology method performs better than another on individual or all of these aspects. Counterintuitively, despite the orthology methods behaving similarly regarding large scale evaluation, the obtained orthologous groups differ vastly from one another.

INTRODUCTION

Gaining insight into the evolution of eukaryotic pathways and protein complexes is often obtained by careful and intensive manual efforts of inferring orthologous groups (OGs), often using phylogenetic profiles [34,93–95]. The reconstructions of the evolution of these pathways is changing our view of eukaryotic genome evolution. With the increase of genomic data, specifically of divergent eukaryotes, the pervasiveness of gene loss became a clear pattern for genetic variation [31,96,97] that has been observed in many eukaryotic functional pathways, protein complexes and metabolic pathways [34,93–95,98,99]. From these manual reconstructions it is becoming more apparent that the loss of parts of these complexes or pathways between the Last Eukaryotic Common Ancestor (LECA) and extant species is mostly non-random, as whole (sub)complexes are often co-lost or, at least, co-absent. A final theme in these studies is that more and more processes and pathways are inferred to have likely been present in LECA and thus we are faced with an ever expanding LECA [100].

In addition to the insights offered for the study of specific pathways, large scale orthologies should provide benefits that overcome the limits in case studies. Manually curated OGs and phylogenetic profiles are laborious and don't scale well when new or updated genomes become available. The development of accurate computational methods to automatically infer orthologies to the same degree as manual analysis is challenging since the orthology inference algorithms need to take into account complex histories of genes, such as duplications, losses and gains of genes and/or their domains and horizontal gene transfer [46,75,101,102]. Readily available curated orthology databases are often very useful for more specific evolutionary questions, e.g. TreeFam database containing mainly animal gene families [57], or OrthoDB containing mainly vertebrate, arthropod, fungi, plant and bacterial gene families [103]. For other evolutionary questions these databases are not always practical, as the species covered in these databases might not be suitable for a particular evolutionary question, e.g., due to (limited) taxonomic range or over/under-sampling of species.

In an effort to progress this field, a large group of researchers have come together to work on a collaborative quest for orthologs (QfO) and have derived a suite of benchmarking tools to determine the performance of orthology methods, both old and new, in a systematic manner [54]. The QfO Benchmarking suite is a very powerful way to objectively evaluate orthology methods in a generic way and has thereby paved the way for advancement in the field of orthology. QFO focusses on a single metric, namely how well an orthology method can recapitulate OGs in a gold standard fixed set of reference proteomes. The benchmark results of orthology methods thus depend on this reference set.

Complementary to the QFO, we here investigate different orthology inference methods and how well they can recapitulate a number of observations made regarding genome evolution in a large diverse set of eukaryotic genomes. We specifically investigate loss patterns and numbers, co-occurrence and LECA gene content. We expand our analysis by investigating qualitative differences between the obtained orthologies and their inferred OGs, and how they capture high quality manually curated OGs.

We find that most orthology methods reconstruct a large LECA, with substantial gene loss, and can predict interacting proteins reasonably well when applying phylogenetic co-occurrence. However, derived OGs show imperfect overlap with manually curated OGs. There is no strong indication of which orthology method performs better than another on either or all of these aspects. Counterintuitively, despite the orthology methods behaving similarly regarding large scale evaluation, the obtained OGs differ vastly from one another.

METHODS AND MATERIALS

Inferring (LECA) orthologs in a large-scale dataset

To investigate different automated orthology methods we inferred orthologous groups (OGs) with 167 proteomes (2865661 sequences) from a diverse set of eukaryotes ([104], Supplementary Figure 1 and Supplementary Table 1). Since we have a large set of proteomes, the orthology inference methods used in this study were chosen based on reasonable computational time, ease of parallelizing the process on multiple CPU's, ease of projection on our set of proteomes, and development activity. Also, the methods

must be able to infer OGs of genes: multiple orthology inference methods we found have pairwise species comparisons, and not the multispecies comparisons that give OGs. Table 1 lists the orthology inference methods and databases chosen for this study and a brief description of each. There are other orthology inference methods that were considered [105–108], but due to various reasons were not used in this study. They are listed in Supplementary Table 2 with a short description.

The eukaryotic eggNOG [58] hmm profile database (version 4.5.1) was used to annotate our eukarya dataset with hmmsearch (version 3.2.1 | June 2018 in HMMER 3.1b2 package February 2015) and a hit cut-off e-value of 10^{-3}. Additionally, the eggNOG protein database (version 4.5.1) was used to annotate the eukarya dataset with DIAMOND and a hit cut-off e-value of 10^{-3}. For both strategies we used the emapper annotation tool (version emapper-1.0.3) provided by eggNOG. We wanted to see if there exist any (large) differences between these strategies.

We applied Orthofinder (version 2.1.2) [109] with default values using both all-vs-all BLAST (hit cut-off e-value 10^{-3}) and DIAMOND (hit cut-off e-value 10^{-3}) as sequence aligners. We chose DIAMOND as a high-throughput aligner that is multiple orders of magnitude faster than BLAST [110]. Also, here we wanted to see if there exist any (large) differences between the alignment strategies.

We additionally used Broccoli (version 1.0) [111] with default parameter settings. SonicParanoid (version 1.3.0) [112] was used with default parameter settings in the sensitive mode (MMseqs2 sensitivity parameter is set to s = 6.0 by SonicParanoid), which is suggested for more distantly related species such as our diverse set of eukaryotes. Since we were not able to get SonicParanoid running on our device, we obtained the results through personal communication with the authors. SwiftOrtho [113] was run using all-vs-all BLAST (hit cut-off e-value 10^{-3}) and the default parameter settings.

Table 1. Orthology inference methods and databases used in this study.

Tool/Dataset	Prediction type	Description and notes
Ancestral Panther http://ancestralgenomes.org	Database	Ancestral genomes dataset contains reconstructed ancestral genomes based on gene family trees from the PANTHER database, from which HMM profiles were built.
Broccoli https://github.com/rderelle/Broccoli	De novo prediction	K-mer pre-clustering to simplify proteomes, followed by a similarity search (DIAMOND) and phylogenetic analysis (FastTree2). Orthologous groups are inferred using a machine learning algorithm, LPA. Extremely fast when run on a large dataset.
EggNOG (DIAMOND and hmmer http://eggnog5.embl.de/#/app/home	Database	Manually curated sequence sets ran with (1) seed ortholog assignments (DIAMOND) and (2) HMM profile searches (hmmer).
Orthofinder (DIAMOND and BLAST) https://github.com/davidemms/OrthoFinder	De novo prediction	Uses both (1) DIAMOND or (2) BLAST as an aligner. Has a sequence length and phylogenetic distance normalized bit-score cut-off between pairs of genes, which function as edge weights in the orthogroup graph. Clustering of genes is done with the MCL method.
SonicParanoid http://iwasakilab.bs.s.u-tokyo.ac.jp/sonicparanoid/	De novo prediction	Uses MMseqs2 as an aligner. The algorithm of InParanoid is used as backbone, with changes to the core algorithm that reduce the execution time and increase the usability of the tool. Relies on cumulative alignment score of groups and avoids using thresholds based on confidence score between pairs of genes. Clustering of genes is done with the MCL method.
SwiftOrtho https://github.com/Rinoahu/SwiftOrtho	De novo prediction	Taking the same approach as OrthoMCL for normilzed bit-score cut-off between pairs of genes, which function as the edge weights in the graph. Clustering of genes is done with the MCL method. SwiftOrtho is optimized for speed and memory usage when applied to large-scale data.

In order to investigate loss patterns for each orthology and to manage the amount of orthogroups in further analysis steps, following the orthogroup inference by each method we used the Dollo parsimony approach [38,114,115] with an additional strict inclusion criteria [104] as a heuristic to infer orthogroups that were likely present in LECA. Briefly, the Dollo parsimony method assumes genes can be gained only once and losses are minimized. To be called a LECA OG, the genes belonging to the OG must be in at least three eukaryotic supergroups distributed over the Amorphae and Diaphoretickes (previously opimoda and diphoda) species [116].

Additionally, we used Panther ancestral genes from the Ancestral Genomes Resource database [59] that has ancestral genes up to the Last Universal Common Ancestor. Gene trees and corresponding multiple sequence alignments were obtained through personal communication with the authors. We acquired eukaryotic ancestral genes by traversing the gene trees with an inhouse script using the ete3 package and obtaining genes in the eukaryotic ancestral nodes, the leaves of these ancestral nodes we define as LECA genes. Additionally, we followed the same criteria for the other orthologies above to select LECA genes and require inclusion of at least three eukaryotic supergroups distributed over both Amorphae and Diaphoretickes species to avoid including possible issues with in/outparalogy resulting from erroneous tree inference. We trimmed the provided multiple sequence alignments with an inhouse script using the biopython package to remove empty columns that were left after obtaining LECA genes. Next, we made hmm profiles from the multiple sequence alignments (hmmbuild 3.2.1 | June 2018 in HMMER 3.1b2 package February 2015) (http://hmmer.org/) that were subsequently aligned to our eukarya set using hmmsearch (version 3.2.1 | June 2018 in HMMER 3.1b2 package February 2015) with a cut-off e-value of 10^{-3}.

Measuring co-occurrence with phylogenetic profiles and (non) interacting proteins

For all orthologies we constructed phylogenetic profiles by defining the presence (1) and absence (0) of all orthologs in the 167 species for a given orthology. For evaluating (non-)interacting proteins we obtained multiple protein interaction datasets. For interacting proteins we used the human BioGRID interaction dataset that contains physical interactions between proteins [117] (version 3.5.172 May 2019). We filtered this set to obtain pairs found in at least five independent publications (PubMed ID's) as a measure

of how thoroughly these proteins were investigated and how amenable they are to high-throughput measurements.

We defined a pseudo negative interaction set from BioGRID by taking pairs of the proteins that were found to be interacting at least five times, but not with each other. We applied these criteria so that the negative set only contains proteins that were found in other interactions and thus exclude the possibility of the interaction not being observed due to a myriad of technical reasons. Additionally, we used a previously compiled negative interaction dataset [118] and cross referenced this set with interactions reported in BioGRID to remove recently found interacting proteins. Finally, we defined a random interaction set from pairs of random OGs. Only OGs that contained a human protein were included. Since the positive and negative sets can contain an OG participating in multiple (non-) interaction pairs, to enable similar properties in this random set, OGs were drawn with replacement and cannot be paired with themselves, or form the same pair multiple times.

Table 2. Different distance (D) or correlation (C) measures used to calculate the distances between phylogenetic profiles of OGs.

Distance or Correlation measure	Values interval	Conversion
Braycurtis	[0,1]	-
Cityblock (Manhattan)	[0,∞]	D / max(D) = D
Cosine	[0,1]	-
Dice	[0,1]	-
Euclidean	[0,∞]	D / max(D) = D
Jaccard	[0,1]	-
Kendalltau	[-1,1]	(1-C) / 2 = D
Kulsinski	[0,1]	-
Rogertanimoto	[0, 1]	-
Russellrao	[0,1]	-
Sokalmichener	[0,1]	-
Spearman	[-1,1]	(1-C) / 2 = D
Yule	[0,∞]	D / max(D) = D

To evaluate the (dis)similarities between phylogenetic profiles of (non-) interacting proteins we calculated the distances between profiles of (non-) interacting OGs using 13 distance and correlation measures (Table 2). The correlation measures were converted to distances (1 - correlation). Distance measures that had values other than between 0 to 1 were converted by dividing by the maximum value found for that distance. We chose the cosine distance and the pseudo negative set that had the best predictive power for further analysis (see Results and Supplementary Figure 2).

Additional to the distance between the phylogenetic profiles of interacting proteins within a method, we calculated the cosine distances between phylogenetic profiles of orthologous groups of different orthology methods. The groups compared are the ones that mapped back to the same human gene, which ideally should be a similar/the same OG.

Comparing against manually curated orthology sets and all-vs-all inference methods

To benchmark the automatically inferred OGs against a set of high quality OGs, we took manually curated sets of protein complexes constructed previously by members of our lab and mapped them to our latest dataset version. The manually curated orthology set contains a total of 125 OGs (Table 3).

Table 3. Manually curated OGs.

Complex/Pathway	Number of OGs	Study/Reference
Intraflagellar transport complex	26	(van Dam et al., 2013)
Kinetochore	91	(van Hooff et al., 2017; Tromer et al., 2019)
TBP-associated factors	8	(Antonova et al., 2019)

We used two cluster overlap measures to compare the overlap between the manually curated and automatically inferred OGs. The F-Grand K-Clique Score (FGKCS) [119] matches cliques of (in our case) OG members within the set of all possible cliques between OGs from automated methods and our manually defined set of OGs and determines the performance using the F-Grand metric.

The Adjusted Rand Score (ARS) [120], roughly, by counting how many pairs in each cluster occur together in the same cluster between methods and is adjusted for chance (and not on overlap/intersection like Jaccard similarity index). Because the ARS focusses on pairwise agreement between clusters, methods that define and/or generate singleton clusters (orthologous groups of size 1) are penalised more than methods that do not. To avoid over–penalising, we only compared LECA orthologous groups, thus removing singleton clusters from the analysis.

To better understand how ARS and FGKCS compare to actual cluster similarity between the inferred and manual OGs, we first stepwise shuffled a percentage of labels for one automated orthology definition and compared it with the same but unshuffled definition. We measured the relationship of ARS and FGKCS compared to the OG similarity as the percentage of unshuffled members.

To visualize how the manually curated OGs were represented in the automatically inferred OGs, we looked at the number of proteins overlapping between each manual OG and automatically inferred OGs. We divided this overlap number by the total amount of proteins in the manual OGs. This in turn gave us a matrix of overlap fractions that could be visualized for each manually vs. automatically inferred OG. We additionally calculated the percentage under– and oversplit sequences. We did this by counting the number of sequences that were not in the OG containing the most sequences (under the assumption that this OG is the correct one) per row (oversplitting) or column (undersplitting). This number we divided by the total amount of sequences in this row or column. We calculated the assignment of the manual OGs by taking the total amount of sequences assigned from the manual OGs to the automatically inferred LECA OGs and divided it with the total amount of sequences in the manual OG set (i.e., 5852).

RESULTS AND DISCUSSION

Comparison between orthology methods show similar behaviour after inferring LECA orthologous groups

To investigate different orthology methods and how well they recapitulate a number of observations regarding genome evolution, we first inferred orthologies on a diverse set of 167 eukaryotic proteomes [104] using eight different inference methods. Supplementary Table 3 shows an overview of the statistics of every orthology inference method and their complete set of inferred orthologous groups (OGs).

In the original orthology definition, Fitch [22] described that if the history of a single gene is as the history of the species, they should be called orthologous, i.e. they are a single gene in the ancestor of the two species and resulted from a speciation there rather than a duplication earlier. Extrapolated for eukaryotic species this should in principle imply that an inferred orthology relation for a set of diverse eukaryotes should equate to the presence of a gene in LECA. However, despite labelling themselves as orthology methods, most methods do not explicitly perform an analysis to ascertain that each inferred OG indeed represents an ancestral presence. Hence, in addition to inferring the OGs with the methods as is, we estimated LECA OGs using a slightly more strict extension (see Methods and Materials) of the Dollo parsimony method [38,114,115] and calculated their (independent) loss to extant species.

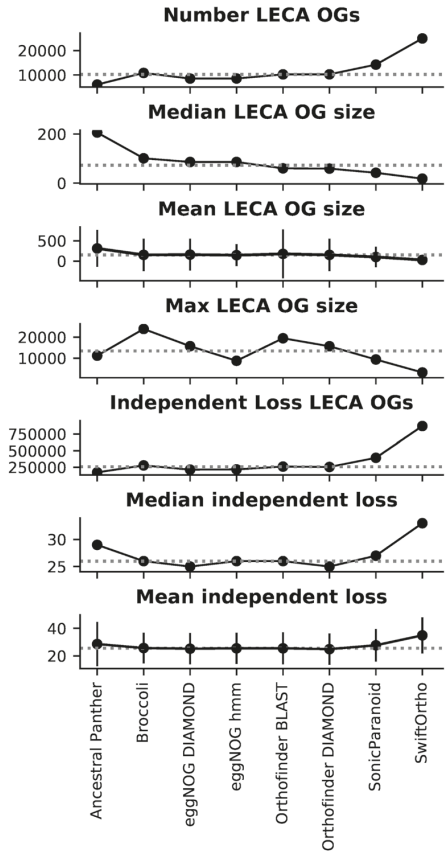


Figure 1. Comparison of different orthology methods after inferring LECA OGs. Shown are the statistics of the number of LECA OGs, size of the OGs and independent loss measures. Grey dotted line shows the median value of a given metric. The vertical lines in the mean size OGs and mean independent (indep.) loss OGs are standard deviations.

Except for the outlier values of Ancestral Panther and SwiftOrtho, the inferred orthologies are similar between the methods while comparing the statistics for inferred LECA OGs (Figure 1). The high amount of automatically derived LECA OGs is consistent with previous large scale studies [38,59] and the observation from multiple manual analyses where individual OGs are found to be in fact in LECA OGs [121]. The inferred retained LECA OGs for each species also follow a similar pattern in the different methods (Supplementary Table 4 and Supplementary Figure 3).

To evaluate the selection criteria of LECA OGs on the results presented above, we additionally inferred LECA OGs using less stringent two

supergroup and more stringent four supergroup criteria (Supplementary Figure 4). The behaviour of the different methods is highly similar to that of the three supergroup criterium, save for the logically higher number of OGs with the two supergroup criteria, and lower number of OGs with the four supergroup criteria.

Independent loss distributions behave relatively similar between orthologies (Figure 2), with little differences in median, means and standard deviations (Figure 1), except for Ancestral Panther, SonicParanoid and SwiftOrtho. Ancestral Panther reports the lowest number of LECA OGs, but the median OG size is the highest of all the methods (Figure 1), indicating that the Panther OGs are very inclusive. This in turn results in a broader distribution of independent loss, and thus higher median loss, for these OGs. SwiftOrtho reports the highest number of LECA OGs, with a smallest median OG size (Figure 1), indicating SwiftOrtho is very strict. This inflates the independent loss distribution, giving SwiftOrtho the highest independent loss off all the methods.

Nevertheless, the number of independent loss is high for all inferred orthologies, as is expected [38,115]. The number of independent loss has been shown to be influenced by gene prediction problems causing falsely inferred gene absences [104], where a suspicious absences can often be found back in the DNA. However, this is likely an equally big problem for all methods, since many orthology methods require predicted proteomes and cannot run on DNA/six frame translated DNA.

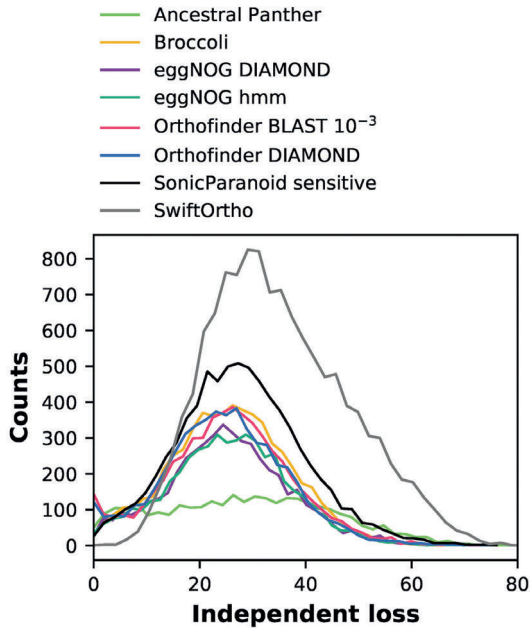


Figure 2. Loss distributions of LECA OGs estimated from different orthology inference methods. Except for Ancestral Panther, SonicParanoid and SwiftOrtho, the loss distributions show similar patterns. For eggNOG hmm vs. Orthofinder BLAST/Broccoli and eggNOG DIAMOND vs. Orthofinder DIAMOND, there is no significant difference between distributions (Kruskal–Wallis H test p-value > 0.001).

Co-occurrence of interacting proteins is predicted similarly and fairly well by most of the orthology methods.

Independent gene loss is not random [31]. In fact, there are countless observations that co-occurring (or co-lost) proteins tend to interact [34,43,93–95]. This creates an additional opportunity for evaluating orthology methods, namely how well different methods can predict co-occurrence of interacting proteins. This metric is also of great relevance for phylogenetic profile methods [43]. For this reason, we calculated distances between phylogenetic profiles of LECA OGs for every method and assessed how well protein-protein interactions were predicted by that method. These evaluations will identify orthology methods that are able to capture a high(er) degree of co-occurrence between interacting proteins and that are more capable of identifying 'functional' orthologous relationships over an arbitrary set of species.

We used the human BioGRID protein interaction dataset [117] as a positive interaction set, and constructed a negative interaction set from BioGRID by selecting proteins that are well studied (i.e. detected to be interacting at least with five other proteins), but have not been detected to interact with each other (zero interactions in BioGRID). We tested 13 distance measures (Table 2). For all distances we observed a clear signal for both the positive and negative interaction set (Supplementary Figure 2 and Supplementary Figure 5), with a weaker signal for SwiftOrtho. The cosine distance had the best area under the curve (AUC) values compared to the other distances measured for most methods (Supplementary Figure 2).

There is no large difference in the predictive power for co-occurrence of interacting proteins from the different orthology inference methods (Figure 3). Orthofinder BLAST has a marginally higher AUC value than the other methods. SwiftOrtho has noticeable lower predictive power for co-occurrence.

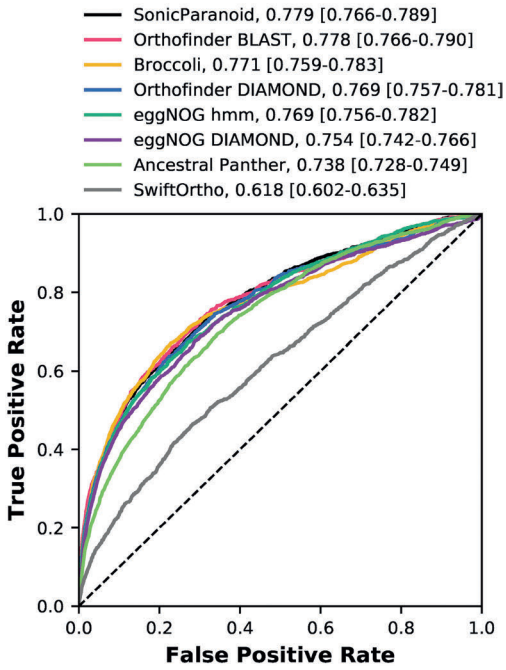


Figure 3. ROC plot comparing the predictive power of for co-occurrence of interacting proteins from different orthology inference methods, using the cosine distance. The legend additionally shows the AUC values for each curve and the confidence intervals (bootstrap n = 1000).

Our ability to predict protein-protein interactions by phylogenetic profiling (as measured by AUC) is corroborating a more elaborate phylogenetic profile method dedicated paper [122], combining phylogenetic profiles in conjunction with the MinHash technique. Most importantly, most orthology methods successfully recapitulate the observation that loss is not random but co-occurs between interacting proteins.

A logical (and desirable) explanation for the similar performance between methods would be that human proteins are assigned similar phylogenetic profiles across different methods. However, when comparing OGs mapped to the same human protein, the phylogenetic profiles generated from the OGs by the different methods differed substantially from each other (Supplementary Figure 6.).

OGs inferred by different methods show imperfect overlap in between methods and manually curated OGs.

From previous work [34,93–95] we collected a set of 125 manually curated OGs at LECA level (Table 3). The comparison to manually curated OGs aligns our evaluation with other evaluation strategies, such as Quest For Orthologs, while the comparison of the inferred OGs to one another gives us a view of how similar the OGs are to each other.

The comparison of the inferred OGs with the manually curated OGs shows overall an imperfect, but decent, overlap for all methods (Figure 4 for Adjusted Rand Score (ARS), and Supplementary Figure 7 for F-Grand K-clique Score (FGKCS)). The overlap between the manual set against all inferred OGs is very high (mean ARS of 0.85) compared to the all-vs-all comparisons (mean ARS of 0.52) between the inferred OGs. This is due to the fact that the manual OG set is a smaller subset of sequences and thus more easily included wholly into the inferred larger OGs, but also because errors in the inferred OGs are measured one-sided to the manual OGs, compared to the all-vs-all comparison where errors are measured two-sided. To understand how the ARS relates to actual OG similarity between the manual and automatically inferred OGs see Supplementary Figure 8. (Supplementary Figure 9 for FGKCS).

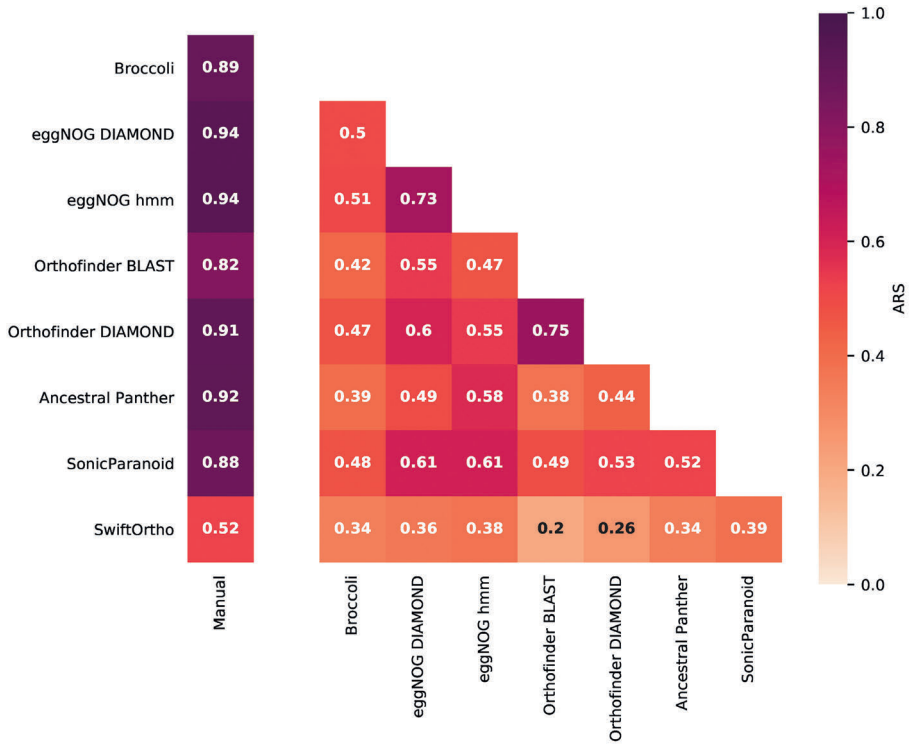


Figure 4. All-against-all comparisons of the different orthology methods and the manual set using the (Adjusted Rand Score) ARS. An ARS of zero would indicate that two OGs vary at a level that would be expected by chance, while an ARS of 1 is a perfect OG overlap.

An important and rather problematic observation is that the comparison between the inferred OGs of different methods shows there is low overlap, or little consistency, between the methods. As expected, similar methods have similar OGs (eggNOG hmm vs. eggNOG DIAMOND and Orthofinder BLAST with Orthofinder DIAMOND). Nevertheless, even between these methods there is still a considerable difference. To evaluate if some of the OGs were consistent between the methods, we collected OGs that overlapped with 100% of their sequences. Also here, there is little overlap between the methods (Supplementary Figure 10 and Supplementary Table 5), indicating that most OGs are difficult to annotate consistently (e.g. due to many duplications). It is important to uncover which types of errors are made in each automated method.

Representation of the manually curated OGs in automatically inferred OGs - where are the differences?

In order to investigate what types of errors are made by the different orthology methods, we wanted to see how the various manually curated OGs are represented in the inferred OGs. Although manually curated OGs are not always perfect, nor their inference without any caveats [46], this issue at this time is unsolvable and also counts for automated methods. Manually curated OGs are the most reliable at the moment.

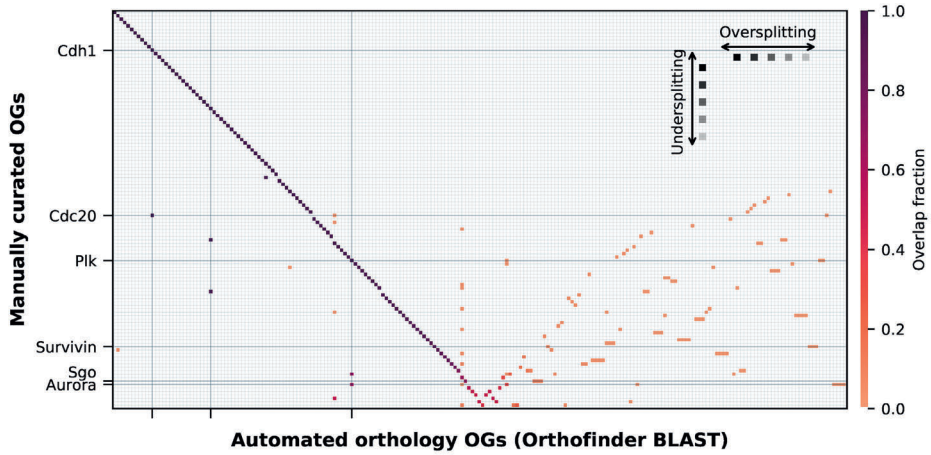


Figure 5. Heatmap showing the fraction of overlap of the clusters between the manually curated OGs (only examples are labelled) and the inferred OGs from, in this case, Orthofinder BLAST as example. The manually curated OGs (y-axis) are sorted from highest to lowest overlap with an Orthofinder OG (x-axis), creating a diagonal (dark purple). The non-diagonal part (lower right corner) is clustered with the weighted clustering method. On the left side of the diagonal there are clear examples of undersplitting of the manually curated OGs. To the right of the diagonal (lower right corner) there are clear examples of oversplitting, or misclassifications. The colorbar shows the fraction of overlap between the manual OG sequences to the automated OG sequences.

We visualized the fraction of proteins per manual OG overlapping with matching proteins assigned to the inferred OGs of the different methods, to the total amount of manual proteins in a given manual OG (Figure 5). Multiple grid points in the same column mean that multiple manually curated OGs are contained within a single automatically inferred OG, indicating an undersplitting in the orthology inference, due to for instance unrecognized (pre-LECA) (out-)paralogs [46]. A clear example of undersplitting is Cdc20 and its outparalog Cdh1 arising from a, by

most of the methods unrecognized, pre-LECA duplication [95] that as a consequence are consistently lumped together in the automated orthology inference methods in various degrees (Figure 5 and Supplementary Figure 11). The same applies for the mitotic kinases Aurora and Plk.

Multiple grid points in the same row of Figure 5 means that a single manually curated OG is contained within multiple inferred OGs, indicating an oversplitting in the automated orthology, due to for instance misclassification, unrecognized homology, or lineage specific duplications (in-paralogs) [46]. A few examples of oversplitting are of the OGs Aurora, Cdc20, Plk, Sgo, Survivin (Figure 5) and MadBub (Supplementary Figure 11), which are consistently fragmented to a higher degree in most automated orthologies (Supplementary Table 6). These proteins have elevated copy numbers, due to possibly recurrent duplications (and subfunctionalization) [34]. Although not apparent in our set, loss, gain or misprediction of protein domains could also cause oversplitting.

The bulk of the automated OGs assigned correctly to the manually curated OGs (diagonal with dark purple grid points (Figure 5). Although, under- and oversplitting can be seen in different degrees between the methods. To compare the degree of under- and oversplitting of the manual OGs between automatically inferred orthologies we calculated the percentage of sequences that are under- and oversplit in the different methods (Figure 6). This shows that, overall, SwiftOrtho has the least undersplitting, but has the lowest assignment and most oversplitting of all the methods. This is in line with their claim that SwiftOrtho is a high precision and low recall method.

EggNOG hmm and eggNOG DIAMOND have the least undersplitting and Orthofinder DIAMOND the least oversplitting. These values indicate that eggNOG is good at recognizing more distant homology, while Orthofinder DIAMOND is better in correctly classifying orthologs, and detecting recent duplication, but the numbers between the methods are fairly similar. Looking at the total amount of manual OG sequences that are assigned to a LECA OG in the automatically inferred orthologies, we see that Ancestral Panther has the highest percentage of manual OG sequence assignment.

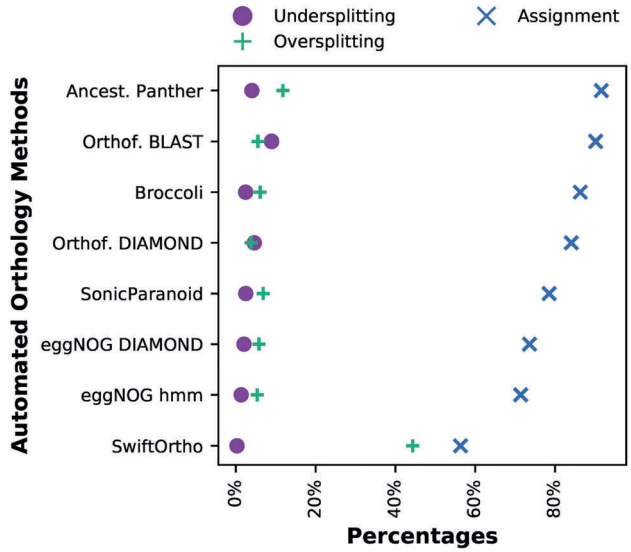


Figure 6. Degree of under- and oversplitting of the manual OGs in the different orthology methods. The methods are sorted from highest (top) to lowest (bottom) assignment of the manual OG sequences in the automated orthologies.

CONCLUSIONS AND OUTLOOK

Orthology prediction is difficult, but pivotal, in comparative genomics. Many methods and tools have been developed to capture orthologous relationships between genes as accurately as possible. We evaluated several orthologies created by automated orthology methods with different underlying algorithms for their ability to capture a number of key observations on eukaryotic genome evolution.

We show that co-occurrence of interacting proteins is predicted similarly well by all orthology methods that were tested and most are similar in behaviour when determining gene loss patterns in eukaryotic evolution. However, they show imperfect overlap with manually curated OGs. Important to note is that, although all orthology inference methods used in this study are similar when describing general patterns in eukaryotic genome evolution, they show large differences amongst the inferred orthologies themselves.

Different methods can provide more optimal solutions compared to others when studying different evolutionary scenarios, such as coevolution, grouping more distant orthologs or more recent duplications. However, many challenges remain in orthology inference, both biological (orthology vs. paralogy) and practical (data increase and computational resources), all of which we have experienced throughout this study.

Nevertheless, our results show that (automated) orthology methods show similar behaviour with respect to large scale evolutionary observations, such as loss patterns, but caution is warranted when looking at the smaller scale, such as single OGs of interest.

Automatic and manual orthology methods are complementary. Leveraging this complementarity could improve comparative genomics in the near future. This means that (orthology) databases should aim to provide OG assignments in a convenient way to accommodate manual analyses and vice versa to be able to use manual assigned OGs as seeds in automated orthology analyses, or to guide development of orthology inference tools.

KEY POINTS

We compared multiple orthology inference methods by looking at how well they perform in recapitulating multiple observations made in eukaryotic genome evolution.

Co-occurrence of proteins is predicted fairly well by most methods and all show similar behaviour when looking at loss numbers and dynamics.

All the methods show imperfect overlap when compared to manually curated orthologous groups and when compared to orthologous groups of the other methods.

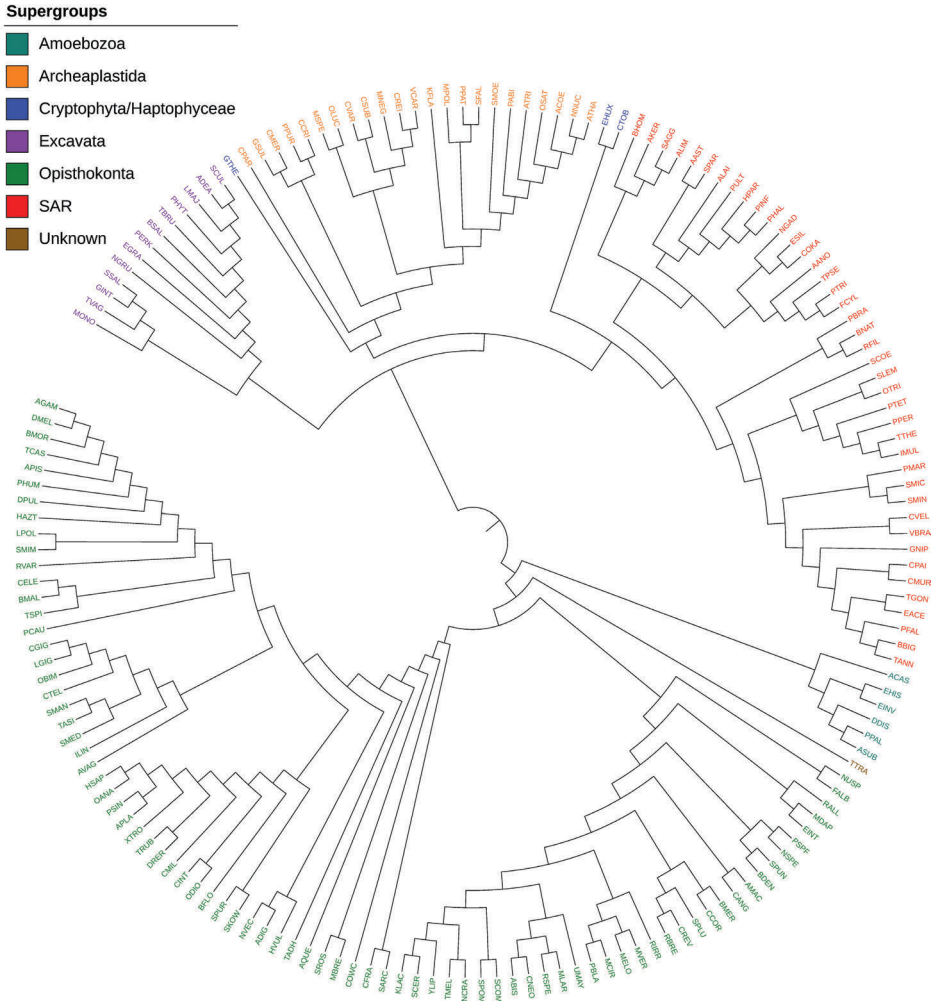
Differences are compared between methods by looking at how the inferred orthologies represent a high-quality set of manually curated orthologous groups.

We conclude that all methods behave similar when describing general patterns in eukaryotic genome evolution. However, there are large differences within the orthologies themselves, arising from how a method can differentiate between distant homology, recent duplications, or classifying orthologous groups.

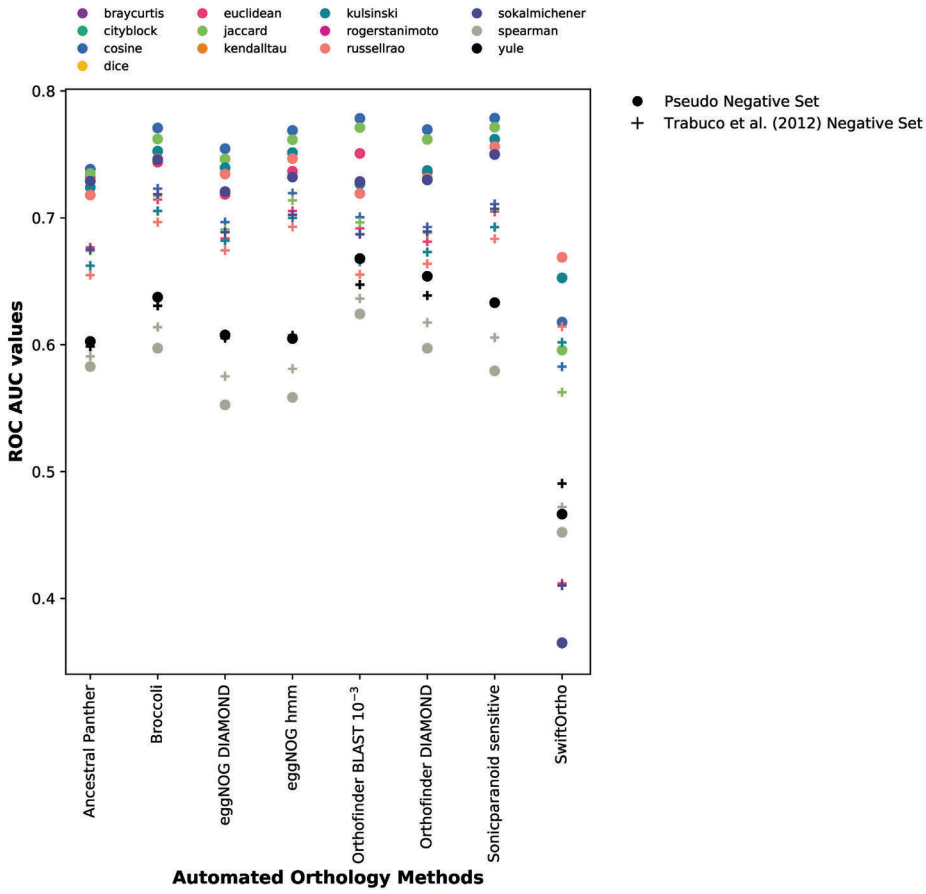
ACKNOWLEDGEMENTS

We would like to thank the developers of Ancestral Genomes, Xiaosong Huang and Paul Denis Thomas, for providing us with the additional data used in this study. We would also like to show our gratitude to Salvatore Cosentino, who provided us with the SonicParanoid results on our dataset.

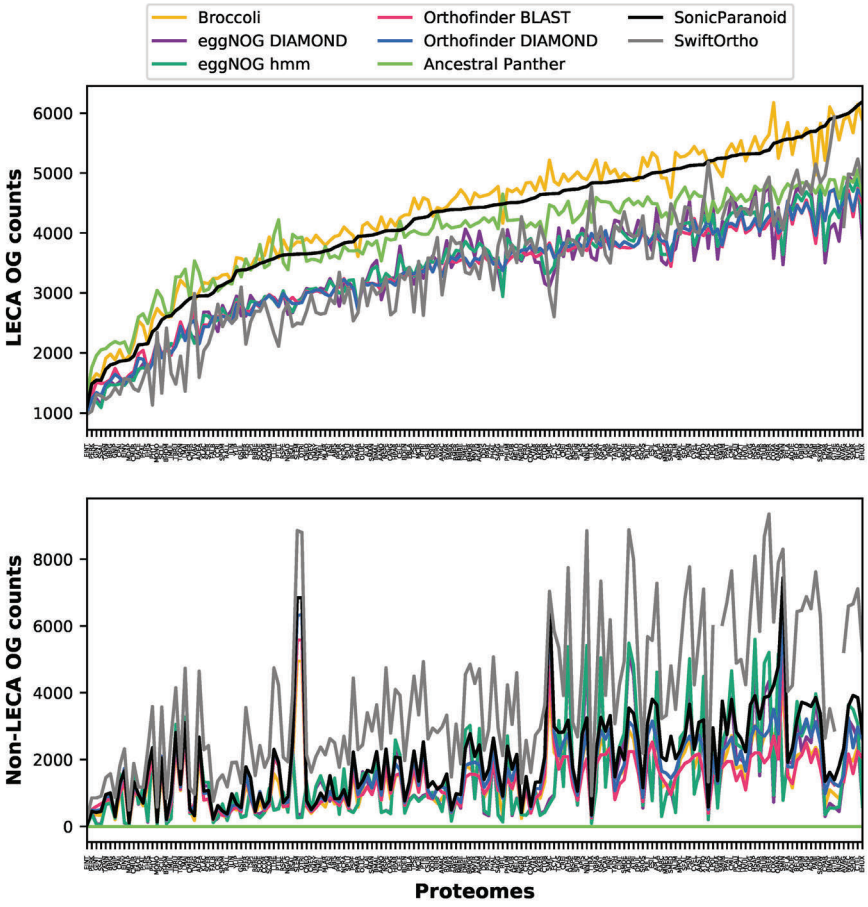
SUPPLEMENTARY FIGURES



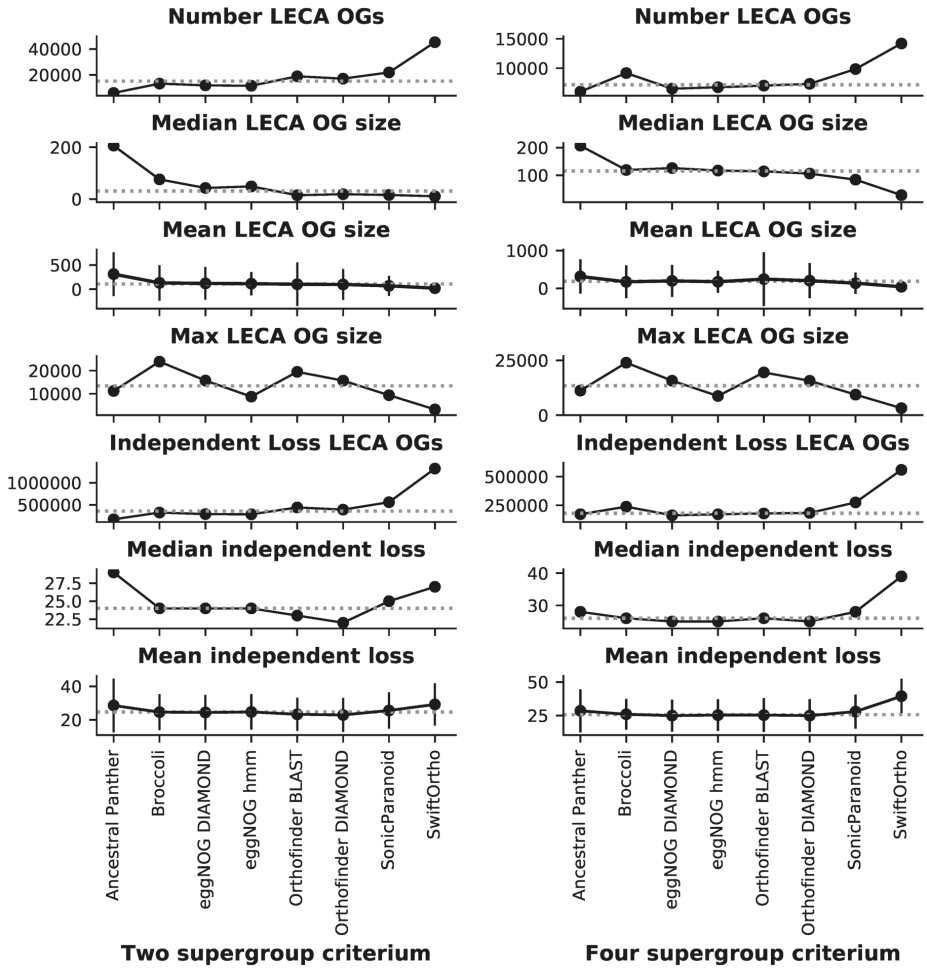
Supplementary Figure 1. Eukaryotic species Tree. The phylogenetic tree of the species used in the analyses and for the Dollo parsimony method. Species names belonging to the IDs in the tree can be found in Supplementary Table 1. The species IDs are coloured according to supergroups and indicated in the legend.



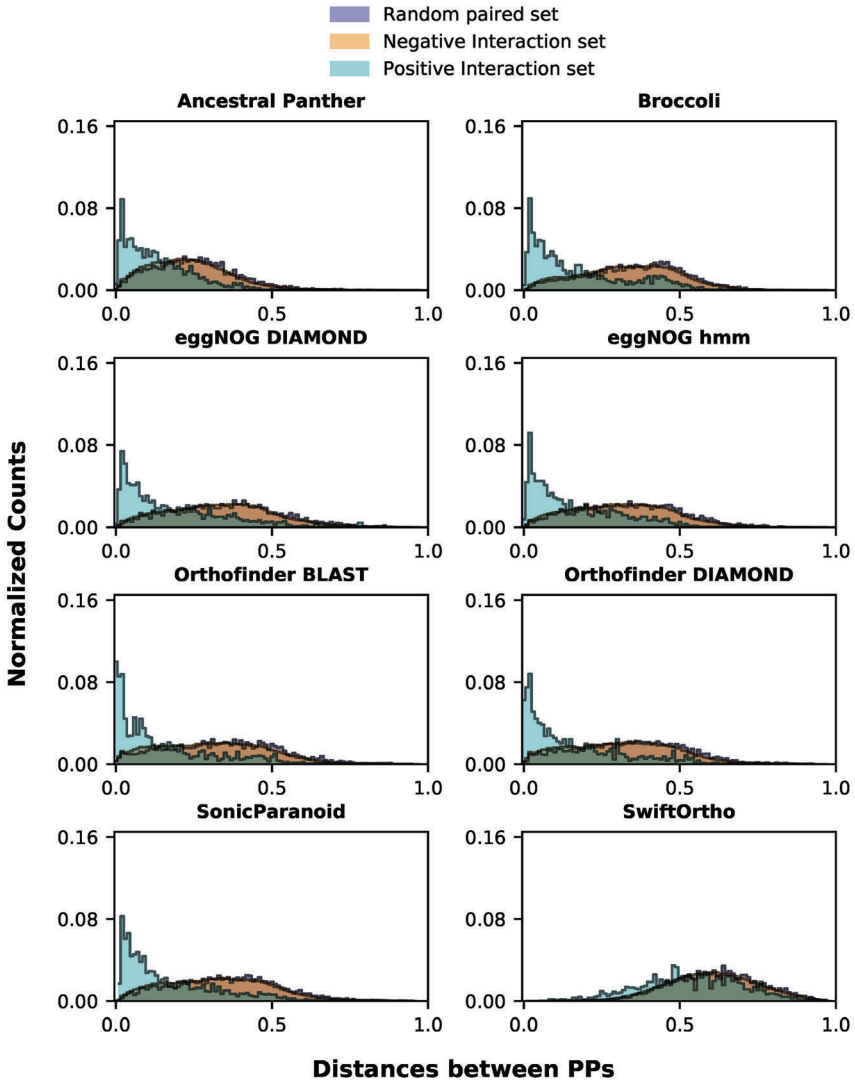
Supplementary Figure 2. Comparison between multiple methods using multiple distances and interaction sets. Calculated ROC Area Under the Curve (AUC) values are shown for all the different methods. Cosine gives the highest AUC value for, except SwiftOrtho, all methods. The Pseudo negative interaction set outperforms the Trabuco et al. (2012) negative interaction set.



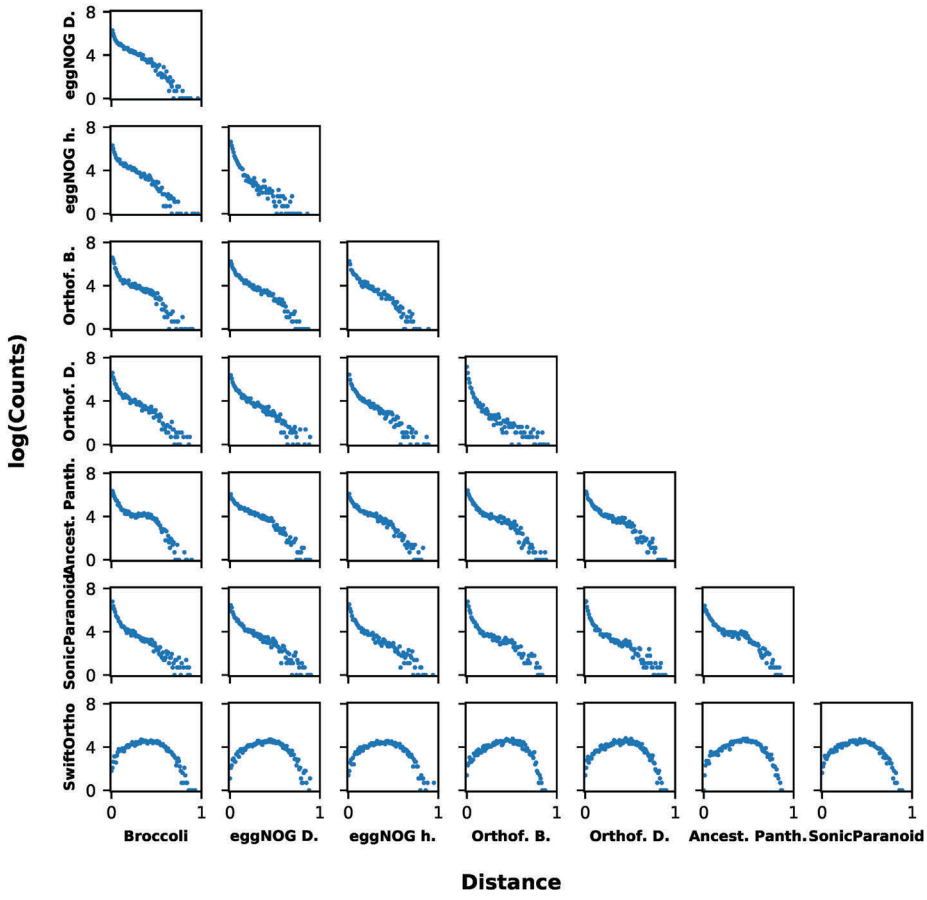
Supplementary Figure 3. LECA OGs and non-LECA OGs per proteome. The values are sorted on the values of SonicParanoid. LECA OGs for every proteome follow similar patterns in all the methods. Note that there are no non-LECA OGs for Ancestral Panther, since these are already ancestral and should in principle only be found in LECA OGs.



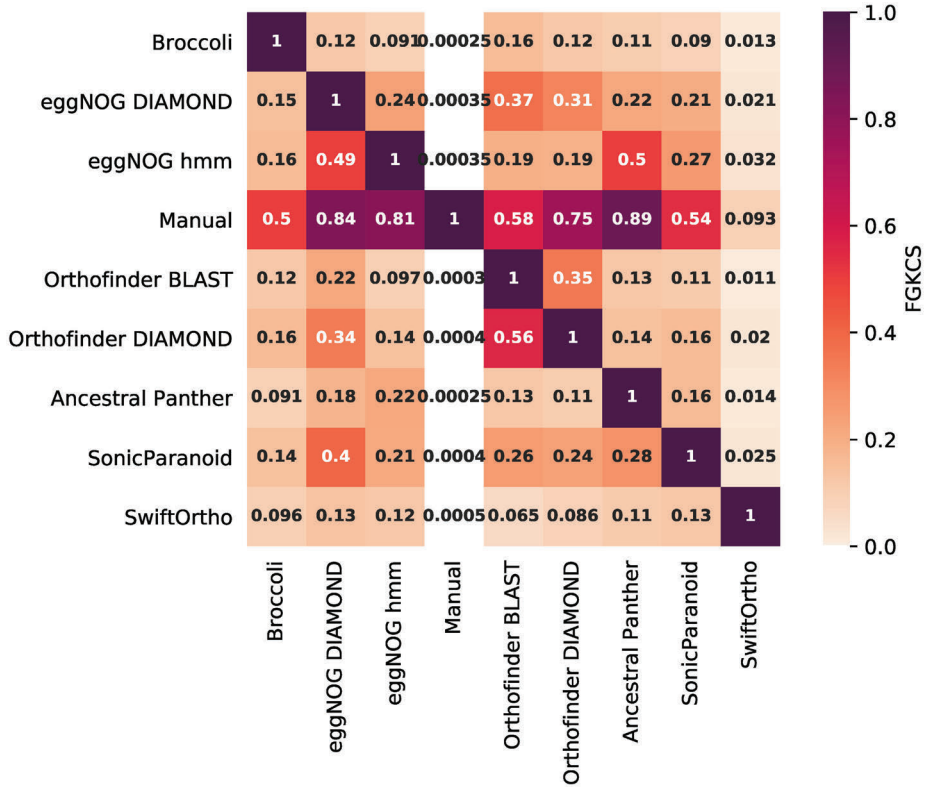
Supplementary Figure 4. Inferred LECA OGs using two and four supergroup criteria. The behaviour between the methods is similar to that of the three supergroup criteria.



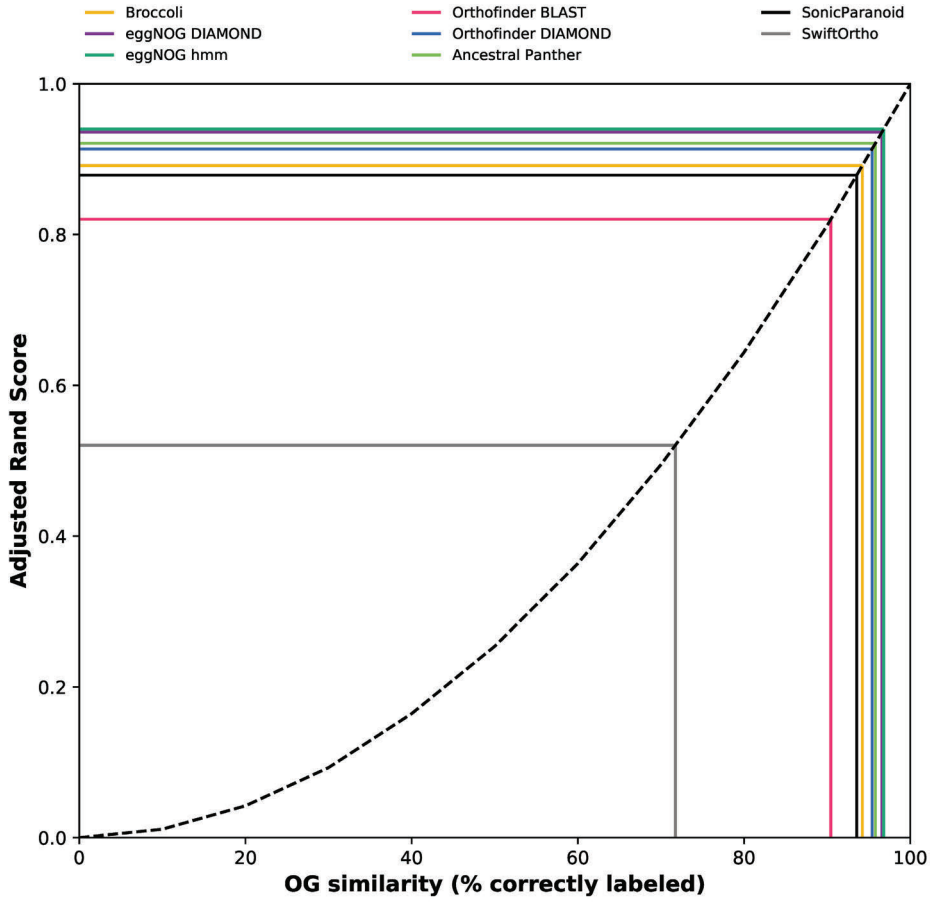
Supplementary Figure 5. Distribution of cosine distances between phylogenetic profiles. Example distributions of the cosine distances between phylogenetic profiles of the positive (blue), negative (orange) and random (purple) protein interaction sets, showing there is a (significantly different) signal in de phylogenetic profile distances for the positive and negative interaction set, compared to the distances of the random protein interaction set (Mann–Whitney U test p-value < 0.001). The counts are normalized by the sum of all counts of the corresponding set.



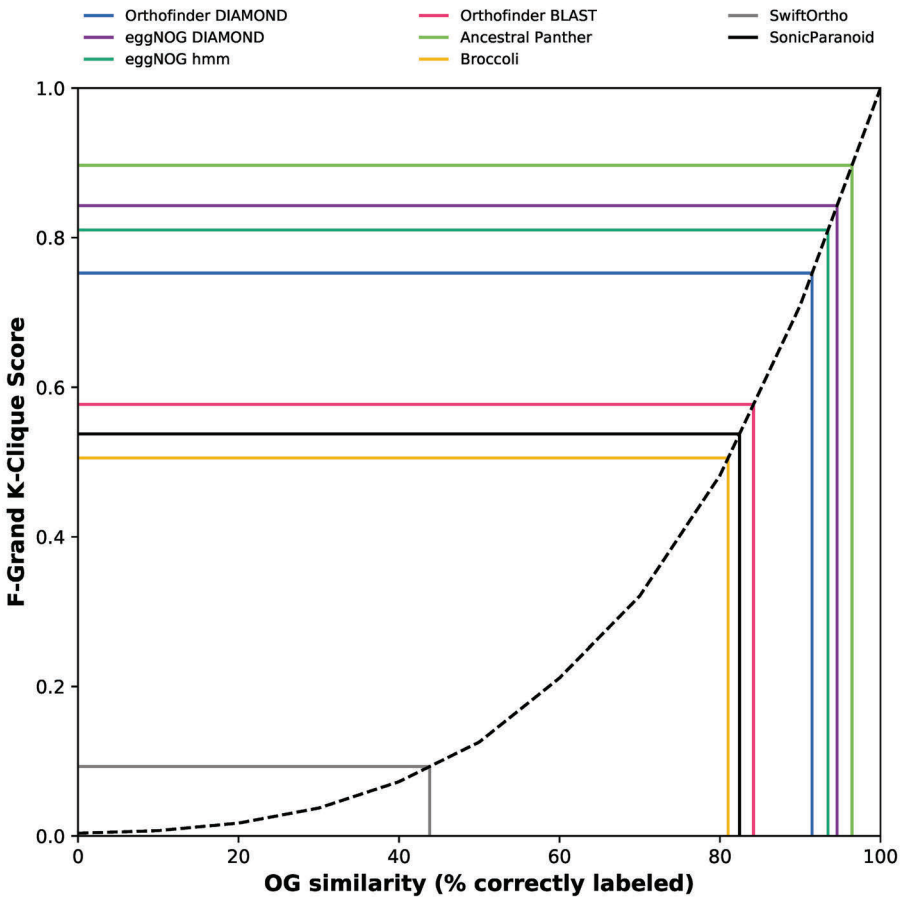
Supplementary Figure 6. The cosine distances between the phylogenetic profiles of OGs from different orthology inference methods. The distances between the phylogenetic profiles of the orthologous groups mapped between the methods should ideally be similar, i.e., a human protein should be found in similar OGs between the different methods. However, this figure shows diversity between the phylogenetic profiles and thus the inferred OGs.



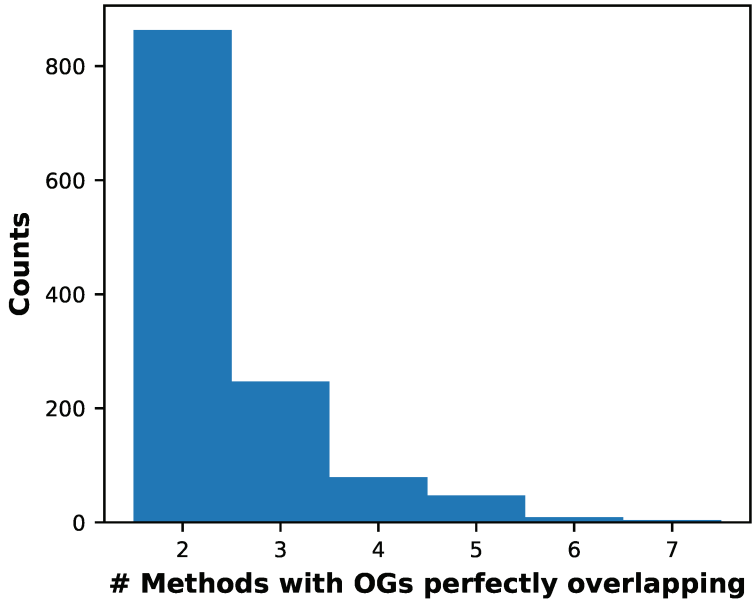
Supplementary Figure 7. The F-Grand K-clique Score (FGKCS) score heatmap. Unlike the ARS (main Figure 4), the FGKCS is not a symmetric score. This way we see how the OGs overlap from method A (vertical axis) to B (horizontal axis), and B to A. For instance, we can make out that the manual OGs (on vertical axis) clusters better with the inferred OGs (on horizontal axis) than the inferred OGs (on vertical axis) cluster to the manual set (on horizontal axis). This has to do with the OG size imbalance between the manual OGs and the inferred OGs (see main text Results).



Supplementary Figure 8. The Adjusted Rand Score (ARS) compared to OG similarity. Shows (black dashed line) the relationship of ARS compared to the OG similarity measured as the percentage of unshuffled members (see Methods and Materials main text). The general shape of this relationship is non-linear. The ARS for all the orthology definitions are plotted (coloured lines) to approximate the percentage of correctly assigned labels from the manually curated OGs by the automated orthology. The ARS shows a clear difference between the low cluster overlap of SwiftOrtho and the (lower) overlap of Orthofinder BLAST.



Supplementary Figure 9. The FGKCS compared to OG similarity. Shows (black dashed line) the relationship of FGKCS compared to the OG similarity measured as the percentage of unshuffled members (see Methods and Materials main text). The general shape of this relationship is non-linear. The FGKCS for all the orthology definitions are plotted (coloured lines) to approximate the percentage of correctly assigned labels from the manually curated OGs by the automated orthology. The FGKCS shows a clear difference between the (lower) cluster overlap of SwiftOrtho, second tier overlap of Broccoli, SonicParanoid and Orthofinder BLAST and the best cluster overlap for the rest of the orthologies.



Supplementary Figure 10. Perfectly overlapping LECA OGs between methods. The bulk of the OGs that are identical in what sequences they contain, are so between only two methods. Only four OGs are identical between seven out of eight methods.

https://github.com/ESDeutekom/ComparingOrthologies/blob/master/Figures/Supplementary/OG_overlap_manual_vs_all_diagonal.pdf

Supplementary Figure 11. The fraction of overlap of the clusters between the manually curated OGs and the inferred OGs from different orthology inference methods.

SUPPLEMENTARY TABLES

Supplementary Table 1. Per species data. Information on the species used in this study, including taxonomic information of each species, counts of absences and found absences, and download locations of genomes/proteomes.

https://github.com/ESDeutekom/ComparingOrthologies/blob/master/Tables/Supplementary%20Table%201_Species%20list.xlsx

Supplementary Table 2. Different methods not used in this study with a short description.

Tool/Dataset	Description
MetaPhOrs	A repository that cannot be easily mapped back to our set [105].
orthAgogue	We were not able to get this tool to complete the analysis, even though we obtained help from the author's [107].
OMA	Smith-Waterman pairwise alignments which would be unrealistic time wise on our set of 2865661 sequences. 900k would take already more than a year [108].
Orthoinspector	Does not infer orthogroups, just one-to-one, one-to-many, many-to-many and in-paralog groups [106].

Supplementary Table 3. Comparison between orthology inference methods and their inferred OGs and derived LECA OGs.

Statistic	Ancestral Panther* N = 167	Broccoli N = 167	EggNOG DIAMOND N = 167	EggNOG hmm N = 167	Orthofinder BLAST (e^{-3}) N = 165	Orthofinder DIAMOND (e^{-3}) N = 167	SonicParanoid (sensitive mode) N = 167	SwiftOrtho (e^{-3}) N = 165
% proteins assiged by orthology[1]	66.2	73.7	67.8	61.9	100.0	100.0	68.2	79.5
Number of OGs	-	57696.0	48183.0	50371.0	77495.0	88849.0	91266.0	351449.0
Median size Ogs	-	4.0	7.0	7.0	4.0	4.0	4.0	3.0
Mean size OGs	-	36.6	37.1	34.8	30.1	24.5	21.4	6.4
% to OG assigned proteins (> 1 sequences in OG)[2]	100.0	100.0	92.0	98.8	82.9	75.9	100.0	100.0
Number of LECA OGs	6061.0	10832.0	8519.0	8554.0	10227.0	10255.0	14237.0	24977.0
Median size LECA OGs	206.0	101.0	86.0	86.0	60.0	59.0	42.0	18.0
Mean size LECA OGs	312.8	153.9	160.6	146.5	178.4	153.8	101.8	30.0
Maximum size LECA OGs	11199.0	23905.0	15708.0	8741.0	19479.0	15670.0	9362.0	3195.0
% to LECA OG assigned proteins[3]	100.0	79.0	76.5	71.4	78.3	72.5	74.2	33.6
% to LECA OG assigned proteins from total[4]	66.2	58.2	47.7	43.7	64.9	55.0	50.6	26.7
Independent loss LECA OGs	173419.0	278632.0	214657.0	218333.0	260980.0	254980.0	393887.0	871018.0
Median independent loss LECA OGs	29.0	26.0	25.0	26.0	26.0	25.0	27.0	33.0

*Ancestral Panther OGs are already LECA OGs.
[1]For N=165: Proteins=2811230. For N=167: Proteins=2865661.
[2]From total assigned proteins.
[3]From total to OGs assigned proteins.
[4]From total assigned proteins.

Supplementary Table 4. Inferred LECA OG counts retained in each species and non LECA OG counts in the different methods. Number included in the column names are corresponding LECA OG and OG counts by each method.

https://github.com/ESDeutekom/ComparingOrthologies/blob/master/Tables/Supplementary%20Table%204__LECA%20per%20proteome.xlsx

Supplementary Table 5. LECA OGs with perfect overlap of all sequences between number of methods

https://github.com/ESDeutekom/ComparingOrthologies/blob/master/Tables/Supplementary%20Table%205__Perfect%20OG%20overlap.xlsx

Supplementary Table 6. Oversplit scores per manual OG in the different orthology methods. The penalty sum is sum times the number of orthologies that show oversplitting.

https://github.com/ESDeutekom/ComparingOrthologies/blob/master/Tables/Supplementary%20Table%206__oversplitting%20manuals.xlsx

CHAPTER 4.

Phylogenetic profiling in eukaryotes: The effect of species, orthologous group, and interactome selection on protein interaction prediction

Eva S. Deutekom, Teunis J.P. van Dam, Berend Snel

Manuscript under review
Preprint available at bioRxiv: https://doi.org/10.1101/2021.05.05.442724

Availability and implementation

All data can be reproduced with the code provided at: https://github.com/ESDeutekom/SelectionPhylogeneticProfiling
All data files can be found at: https://doi.org/10.6084/m9.figshare.14500146.v1

ABSTRACT

Phylogenetic profiling in eukaryotes is of continued interest to study and predict the functional relationships between proteins. This interest is likely driven by the increased number of available diverse genomes and computational methods to infer orthologies. The evaluation of phylogenetic profiles has mainly focussed on reference genome selection in prokaryotes. However, it has been proven to be challenging to obtain high prediction accuracies in eukaryotes. As part of our recent comparison of orthology inference methods for eukaryotic genomes, we observed a surprisingly high performance for predicting interacting orthologous groups. This high performance, in turn, prompted the question of what factors influence the success of phylogenetic profiling when applied to eukaryotic genomes.

Here we analyse the effect of species, orthologous group and interactome selection on protein interaction prediction using phylogenetic profiles. We select species based on the diversity and quality of the genomes and compare this supervised selection with randomly generated genome subsets. We also analyse the effect on the performance of orthologous groups defined to be in the last eukaryotic common ancestor of eukaryotes to that of orthologous groups that are not. Finally, we consider the effects of reference interactome set filtering and reference interactome species.

In agreement with other studies, we find an effect of genome selection based on quality, less of an effect based on genome diversity, but a more notable effect based on the amount of information contained within the genomes. Most importantly, we find it is not merely selecting the correct genomes that is important for high prediction performance. Other choices in meta parameters such as orthologous group selection, the reference species of the interaction set, and the quality of the interaction set have a much larger impact on the performance when predicting protein interactions using phylogenetic profiles. These findings shed light on the differences in reported performance amongst phylogenetic profiles approaches, and reveal on a more fundamental level for which types of protein interactions this method has most promise when applied to eukaryotes.

INTRODUCTION

The post-genomic era has provided us with a wealth of eukaryotic genomes of diverse and underrepresented phyla [25]. Most of the sequences in these new genomes are without precise function assignment and a challenge remains, protein function and interaction discovery [13,14]. Computational approaches that are available for large scale analyses of protein function and interactions include phylogenetic profiling. Phylogenetic profiling uses correlations of the presences and absences of groups of orthologous proteins (orthologous groups) across a set of species [43]. Phylogenetic profiling is a seemingly straightforward method proven to be a valuable alternative resource for studying functional relationships between proteins. Recently the method has played an integral part in identifying the cellular functional role of CENATAC that is a key player in a rare aneuploidy condition in humans [123], identifying eukaryotic reproduction genes [41], and identifying eukaryotic novel recombination repair genes [40].

The information in the phylogenetic profiles given by presence and absence patterns, are shaped by a diverse range of evolutionary forces. These forces include horizontal gene transfer, secondary endosymbiosis and gene loss. The method relies on the principle that proteins with a similar profile indicate that the proteins co-evolved due to them belonging to the same functional pathway or complex. There are countless observations that co-occurring proteins tend to interact [34,36,124]. Phylogenetic profiling can be a powerful tool for function prediction. By comparing, or even clustering, profiles of proteins with unknown function to those with known function enables us to infer to which complexes or functional pathways the uncharacterized proteins likely belong and, in turn, infer their function.

Multiple studies have shown the effectiveness of phylogenetic profiling in large scale analyses of eukaryotes [124,125], which has become possible with the large increase in genomic data and computational methods to (automatically) infer orthologies [126–128] or cluster genes [125]. However, benchmarking and analysing the performance of large-scale phylogenetic profiling has been limited to prokaryotes, for which good performance can be obtained when predicting protein interactions [129–132]. The performance decreases when benchmarking is done solely with eukaryotes or when eukaryotes are combined with prokaryotes [129,130]. Likely, the performance reduction is caused by the different forces driving eukaryotic

genome evolution, compared to the dynamic pan genomes of prokaryotes where the interplay of rampant horizontal gene transfer of operons and loss of genes that create highly informative patterns.

We recently obtained a high protein interaction prediction performance in a large set of eukaryotes in the context of evaluating a diverse set of orthologous group inference methods [133]. The surprisingly high prediction performance only marginally depended on the orthologous group inference methods (which was the focus of the study), suggesting that its cause could be any of the other underlying choices. Therefore, a more elaborate analysis of the choices made for phylogenetic profiling is warranted. Here we evaluate in-depth the meta parameters influencing the performance of phylogenetic profiles in eukaryotes.

Multiple studies have understandably focused their analysis on reference genome selection or the amount of genomes/data needed to increase prediction performance [130–132,134,135]. Besides genome diversity and quality, we analyse orthologous groups and reference interactome selection. Our results demonstrate that an interplay of biological and technical aspects influence phylogenetic profiling. Most importantly, our results show that prediction performance is influenced not only by genome selection but mostly by orthologous and interactome selection.

RESULTS

Each results section describes the analyses of meta parameters encompassing five main concepts: genome quality, genome diversity, performance directed genome selection, orthologous group selection, and reference interactome selection. To rule out any orthology specific issues, we performed the analyses using two orthology inference methods, Sonicparanoid [127] and Broccoli [128]. Sonicparanoid performed the best in our previous study using phylogenetic profiles for protein interaction prediction [133]. We chose Sonicparanoid as the primary method, while broccoli serves to determine to what extent the results are contingent on a specific orthology method. The results for Broccoli can be found in Supplementary figures and are overall in agreement with the results of Sonicparanoid.

1. Lesser quality genomes have more effect on the prediction performance than higher-quality genomes

Phylogenetic profiles can be noisy due to multiple technical reasons, such as gene annotation and genome assembly errors. Consequently, the quality of genomes can be an essential factor, as profiles with a lot of noise would be akin to noisy gene expression or protein interaction measurements. We expect noisy genomes to give much weaker prediction performance. Given this expectation, the first meta parameter assessed was genome quality. We calculated genomes quality using two independent metrics, BUSCO [84] and one of our design (Supplementary figures and Methods and Materials). For clarity, we use only the BUSCO metric in the main text since both metrics generally agree with each other.

The BUSCO metric assesses genome completeness based on the (in our case) absence of single-copy orthologs that are highly conserved among eukaryotic species. The absences of these orthologs can result from incomplete draft genomes or false negatives in gene prediction, which in both cases leads to false absences of orthologs. We selected 50 high-quality genomes with the lowest BUSCO values, i.e., genomes with the least number of unexpected absences. We also selected 50 lower quality genomes with the highest BUSCO values, i.e., genomes with the most number unexpected absences (Figure 1.A.). We compared the quality filtered genome sets with 1000 randomly generated genome sets of 50 genomes each to see if quality-based selection differs from any random sampling of genomes.

The results show that the performance using the highest quality genomes with the least suspect absences falls within the distribution of random genome prediction performance (AUC: 0.765). In contrast, the lower quality genomes fall below the distribution of random prediction performance (AUC: 0.748) (inset Figure 1.B.). This suggests that it is more beneficial to filter out lesser-quality genomes than it is to select for high-quality genomes. This result is consistent between two independent scores of genome quality (Supplementary Figure 1).

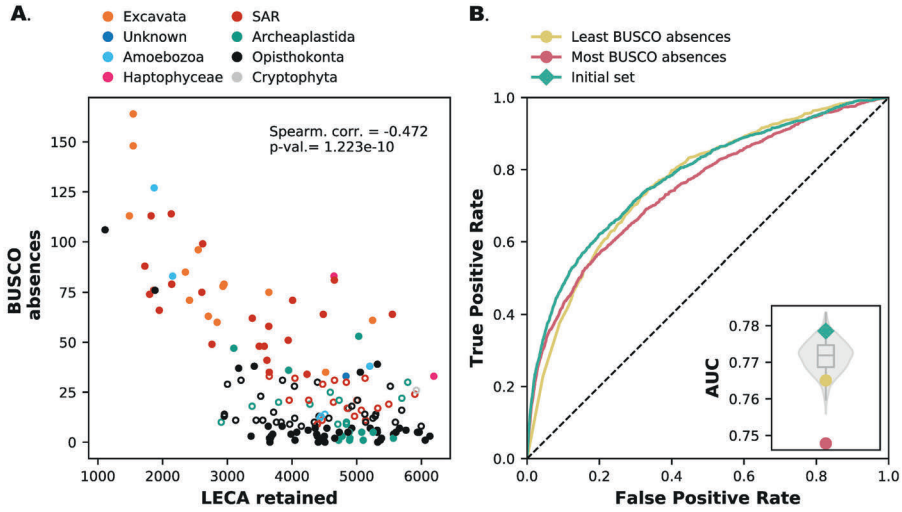


Figure 1. Lesser quality genomes have more impact on protein interaction prediction performance. A. BUSCO absences as a function of retained Last Eukaryotic Common Ancestor (LECA) orthologous groups in different species. Filled data points are the selected genomes for the prediction accuracy calculations. **B.** Receiver-operator Curve of two species sets (n = 50) with the most and least BUSCO absences. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (teal diamond).

With these results, it seems prudent to select genomes only based on quality when applying phylogenetic profiles. However, there is an inherent bias between genome quality and phylogenetic distribution (Figure 1.A). For instance, eukaryotes belonging to the Opisthokonta supergroup have overall lower BUSCO absences, biassing the selection of good genomes towards one eukaryotic supergroup. *A priori*, species diversity seems another meta parameter in genome selection with potential impact. In the next section, we will look at the diversity of species and how that influences phylogenetic profiling.

2. Genome diversity has little effect on prediction performance in eukaryotes

The diversity of species plays a role in the performance of phylogenetic profiles in prokaryotes [130]. We also expect high species diversity to improve how informative profiles are by giving high-resolution information on how genes co-evolve in different organisms. More species diversity allows to maximally discern the effect of evolutionary forces shaping co-evolving proteins, which might not be apparent in, e.g., an animal only

data set. There will be no discernible and informative phylogenetic pattern in a homogeneous species set where most ancestral protein complexes are not frequently lost. A previous study showed that the maximum phylogenetic diversity in Bacteria gives the best predictive performance [132]. Here we want to test how maximal and minimal diversity affects prediction performance in eukaryotes.

We analysed the impact of eukaryotic diversity by selecting two sets of 50 genomes, one containing the most similar species (Figure 2.A.) and the other the most diverse species (Figure 2.B.) from our initial species set. The (dis)similarity was measured using an iterative all-vs-all comparison using the cosine distance between genomes and their orthologous group content. We started with the most diverse or similar species pairs and iteratively added to this set the species with the highest (dis)similarity until we obtained 50 genomes (Materials and Methods). We recalculated the protein-interaction prediction performance for both these sets. The prediction performance is lower than the initial set for both sets, but not worse than any randomly selected genome sets (AUC: 0.760 for the dissimilar set and AUC: 0.764 for the similar set) (Figure 2. C. inset).

Similar species will naturally show more cohesion in profiles, with little separation of protein co-evolution. Highly diverse species will naturally show more discordance, with little information left to see protein co-evolution. In both cases, there will not be a gene-specific signal. A combination of the two should give good separation of actual co-evolving genes. Together with the effects of genome quality, genome diversity can be an important factor for the performance of phylogenetic profiles. However, the interplay between these two factors is complex, and as we previously determined, high genome quality corresponds with lower phylogenetic diversity. To look into this further, we investigate the influence of single genomes on predictive performance.

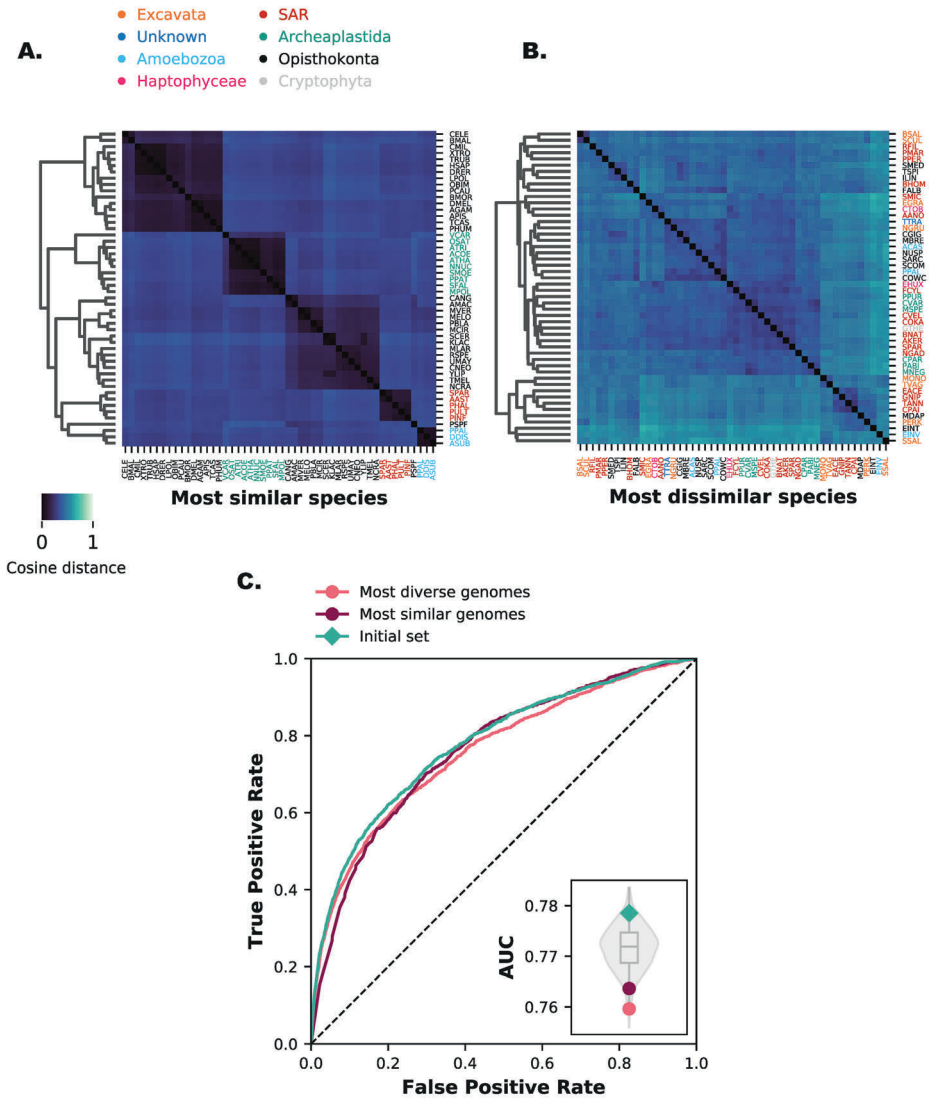


Figure 2. Both high and low diversity sets have little impact on protein interaction prediction performance. A. The most similar species form more clusters and are overall more similar to each other. **B.** The most diverse species show no clustering and are overall less similar to each other. **C.** Receiver-operator Curve of two species sets (n = 50) with the most diverse and most similar species. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (green diamond).

3. Single influential genomes and their combined effect on prediction performance reveal the importance of the type of information in the profiles

Diversity and quality both impact performance and we expect it to have a combined influence on phylogenetic profiling. Instead of a priori selecting genomes based on a measure for each of these two criteria, we can also objectively evaluate the prediction performance by removing genomes from the initial species set one-by-one (Figure 3. A.). Genomes that decrease prediction performance when removed from the initial set we can consider as advantageous to phylogenetic profiling, while genomes that increase prediction performance when removed from the initial set we can consider as disadvantageous to phylogenetic profiling. We selected the top 50 advantageous and top 50 disadvantageous genomes to see whether these genomes together in their respective sets also influence the prediction performance.

For both the advantageous and disadvantageous set we can see a large difference in prediction performance (Figure 3. B.) and a larger difference than the selection based on the measures for either quality or diversity. With the advantageous genome set, the performance increases (AUC: 0.801). In contrast, for the disadvantageous set the performance drops (AUC: 0.730). Both values fall well outside the distribution of 1000 randomly generated genome subsets.

Although a large cumulative effect on performance because we used the genomes' performance to select the genomes, it is still very interesting to see what these genomes share if it is not quality or diversity. We therefore examined the role of different genomes in these genome sets. Comparison of a large number of factors (Supplementary Figure 5) revealed that that the difference in prediction performance of the advantageous and disadvantageous genome sets is related to the (human) interactions retained in the genomes (Figure 3. C. and Supplementary Figure 5. A.). The illogical absence ratios and the complete interactions present (or co-presences) (Figure 3. C. I & II) show intermediate values for the disadvantageous genome set. At the same time, these values are either high or low for the advantageous genome set.

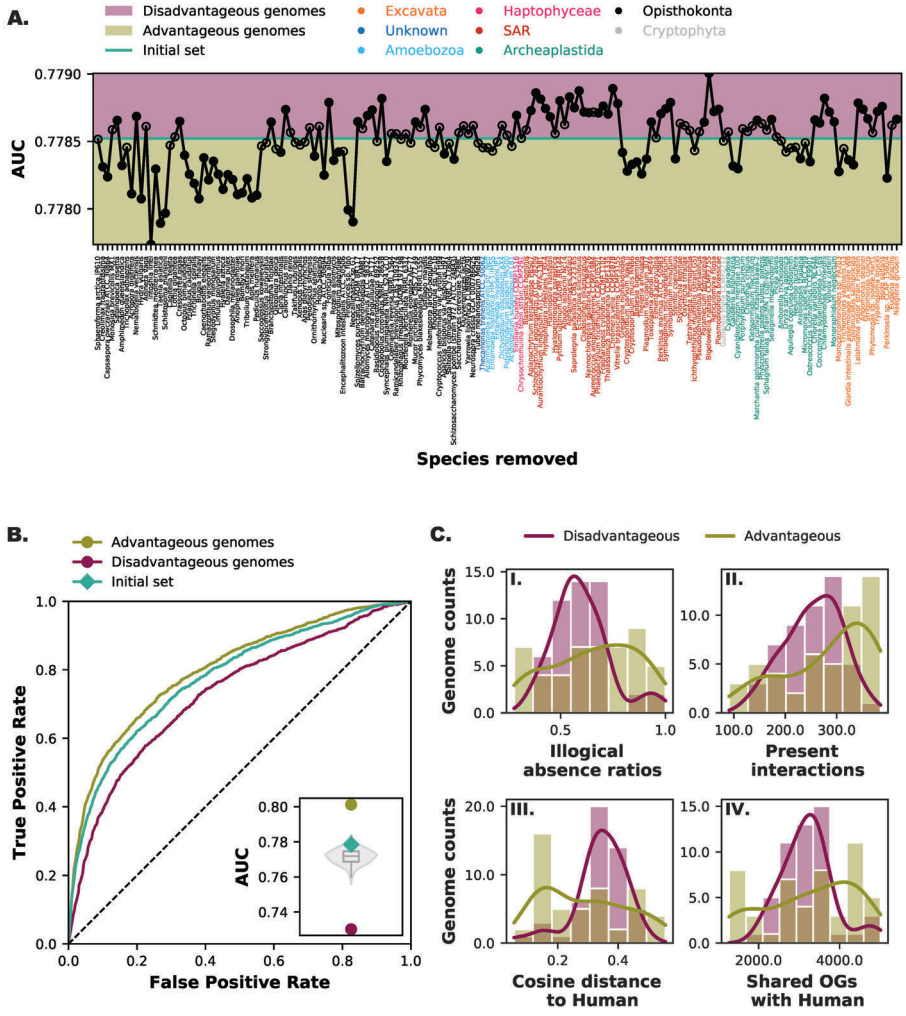


Figure 3. Influential genomes and their combined effect. A. Recalculated Area Under the Curve (AUC) values when a single species is removed from the initial species set. Genomes that increase the AUC value when removed can be considered disadvantages compared to the initial set when predicting protein interactions with phylogenetic profiles. Genomes that decrease the AUC value when removed can be considered advantageous for predicting protein interactions. Top 50 advantageous and top 50 disadvantageous genomes shown with the black fill in the scatter plot. **B.** Receiver-operator Curve of two species sets (n=50) with the most advantageous and disadvantageous genomes. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (green diamond). **C.** Comparison of the counts (histogram) and kernel density estimates (line plot) of (I) illogical absence ratios (illogical absences divided by total interaction absences (co-absences + illogical absences)), (II) present interactions, (III) the cosine distance to human, and (IV) total shared orthologous groups with human.

We can also directly relate the differences between these genome sets to how close the genomes in the sets are to the human genome. The cosine distance and the shared orthologous groups of the genomes with the human genome (Figure 3.C. III & IV) show intermediate values for the disadvantageous set, while the values are either high or low for the advantageous genome set. For the orthologous groups inferred by Broccoli this signal is even more pronounced (Supplementary Figure 5. B. Figure). A surprising finding is that the advantageous set contains numerous parasitic organisms (Supplementary Table 1).

In other words, genomes boosting the performance share either a lot or a little similarity with the reference interactome across a range of dimensions. Thus, phylogenetic profiling in eukaryotes benefits from genomes with a little or a lot of interactions present with regards to the reference interactome. These results reveal the importance of selecting genomes based on the evolutionary information contained within them relative to the query species, and is critical for high performance when predicting interacting proteins.

4. Orthologous group (pre-)selection improves prediction performance by (inadvertently) enriching co-evolving proteins in profiles

Phylogenetic profiling benefits from clear modular co-evolution of proteins and subsets of proteins showing similar evolutionary behaviour [130,136]. A myriad of factors limit the modular co-evolution of interacting protein [137–139]. In previous research [133], which provides the starting meta parameters of this study, we evaluated orthology methods by their ability to recapitulate gene family dynamics in the Last Eukaryotic Common Ancestor (LECA). Consequently, the results so far are based on orthologous groups estimated to be in LECA. To see if this selection criterion was a factor in the strong performance, we performed phylogenetic profiling with other orthologous groups selections: groups estimated to be post-LECA, or groups not filtered on any criteria (post-LECA + LECA), i.e., the raw output of the orthology inference methods. We compared these orthologous group sets with 1000 subsets of randomly selected LECA orthologous groups. The prediction performance was indeed reduced (AUC: 0.691 post-LECA and AUC: 0.734 all orthologous groups) compared to that of LECA orthologous groups or any randomly selected set of orthologous groups (Figure 4.). After some reflection, a myriad of explanations likely factor into this effect.

Profiles of LECA proteins have many losses, and thus a lot of information (entropy) (Supplementary Figure 6 and 8). Profiles of post–LECA proteins have less loss and, by definition, are restricted to specific lineages, and thus contain less information. Combining LECA and post–LECA orthologous groups produce a set of phylogenetic profiles with an overall much lower similarity.

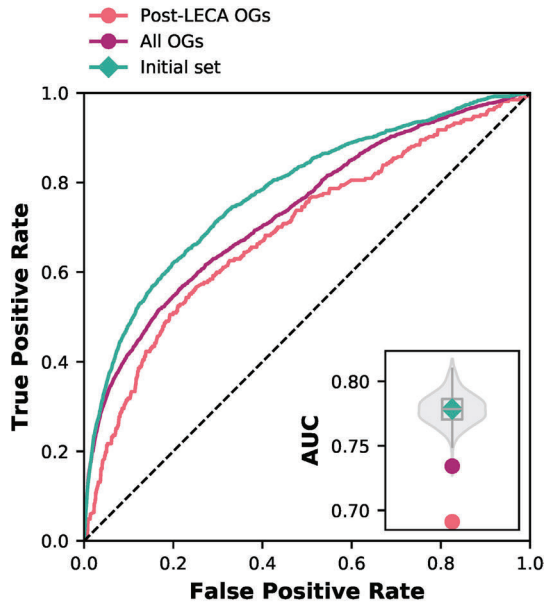


Figure 4. Orthologous group selection has a large impact on prediction performance. Receiver–operator Curve of post–LECA orthologous groups and unfiltered orthologous groups. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of randomly selected LECA OGs (violin plot) and the initial species set (green diamond).

We now have identified a key meta parameter choice explaining why our previous research found such high performance. However, it is unclear what the reason for this effect is and why for specific pairs of proteins, one protein was in LECA and the other not. This separation could be biological reality, i.e., innovations in the evolution from LECA to human, or issues in orthology assignment, i.e., one protein is evolving much more rapidly that causes the protein's predicted orthologous group to give an artifactually lineage-specific distribution in the profile. Consequently, the protein is falsely inferred as a more recent addition or innovation. Manual inspection of this set (Supplementary Figure 9) does not obviously point towards

one of the explanations. It is likely a combination of factors, including orthology prediction errors (e.g., oversplitting) and actual lineage-specific additions/inventions. In any case, the meta parameter of orthologous group selection is perhaps easily overlooked or made implicitly in the OG creation itself. Still, it is highly impactful, and our results show that OG selection improves prediction performance by enriching co-evolving proteins in profiles.

5. Choice of reference interactome and interaction filtering improves prediction performance by increasing the amount of co-evolving proteins and quality of interactions

Phylogenetic profiling attempts to predict which pairs of proteins are part of the same function, pathway, or complex. The performance of phylogenetic profiles can be measured using a data set of proteins that interact or are otherwise functionally linked. For example, we can take KEGG pathways as measuring units, as done in the STRING database [140]. However, these pathways often have an excess of 30 proteins and not all of them are expected to have the mutual functional dependence that results in co-evolution. This unwantedly biases the predictor by having supposedly interacting proteins with little correlation. Similarly, if we would take a very small well-curated set of compact/short linear metabolic pathways as was used to seed the CLIME searches [125], then the choice of what protein pairs to count as false negatives becomes difficult. Hence, our decision in previous work was to parse human interactions from BioGRID to contain only interactions found in at least five independent studies (Methods and Materials). This filtering of interactions has been repeatedly demonstrated to effectively increase the quality and reduce the noise in the interaction data [140,141]. Moreover, the same data set contains a very good indication of which proteins are not functionally related. Proteins that are well studied and repeatedly surface in high throughput assays and are subject to repeated investigations are indeed likely to have no functional relation since these proteins are evidently never identified to interact.

The results in section 3 (Figure 3.C.) reinforce the notion that the reference interaction set plays a role in the performance of predicting interacting proteins. For these reasons, we analysed how the filtering and choice of reference interactome influences protein interaction prediction performance in eukaryotes. Using an unfiltered human protein interaction dataset reduces the prediction performance from an AUC of 0.779 to an

AUC of 0.638 (Figure 5. A.). This performance is also lower than any set of randomly selected LECA orthologous groups (inset). The quality of the interaction data used clearly plays a role in prediction performance, i.e., if we take a noisy "ground truth" it turns out to be difficult to predict this truth. It is difficult to predict interactions with a set littered with false, virtually random, pairs.

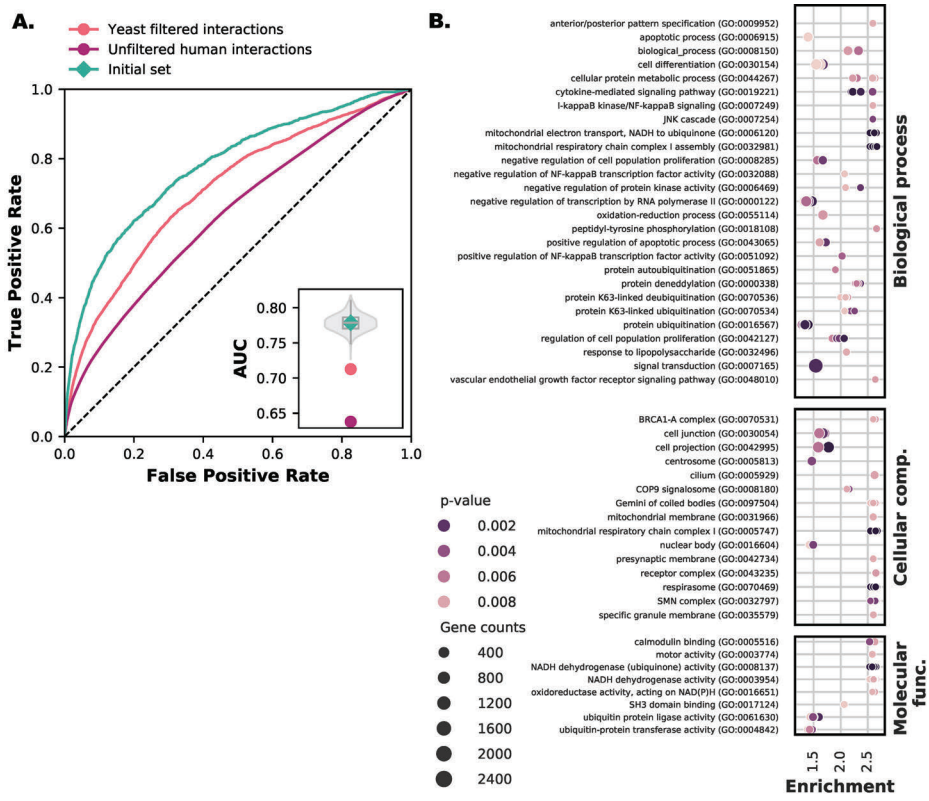


Figure 5. Interactome selection is important for prediction performance. A. Receiver-operator Curve of post-LECA orthologous groups and unfiltered orthologous groups. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of randomly selected LECA orthologous groups (violin plot) and the initial species set (green diamond). **B.** GO-enrichment analysis for genes enriched in interactions present in only human compared to interactions present in human and yeast. Orthologous groups can contain multiple genes. We randomly selected genes from an orthologous group to generate a new sample and population sets ten times and recalculated the enrichment (shown by multiple points in the figure rows).

We further analysed the choice of reference organism for protein interactions. Specifically for eukaryotes, the prediction performance was sensitive to the reference species for protein interactions [135]. Yeast has been the organism of choice as the reference interaction set for eukaryotes. Yeast is a popular model organism that has been extensively researched, and it is with yeast that many protein-protein interaction high throughput methods were pioneered. As a result, we also expect the interaction data of yeast to be of higher quality than that of human and, consequently, interaction predictions to be better.

We used *Saccharomyces cerevisiae* interactions from BioGRID (Materials and Methods) filtered with the same number of publications strictness criterion. Surprisingly, and contrary to for instance [41], the human interaction set performed better with an AUC of 0.779 compared to the yeast interaction set with an AUC of 0.713 (Figure 5. A.). One reason could be that ascomycete fungi and yeast in particular, has lost many co-evolving LECA complexes found in most eukaryotes [142,143]. These losses include Complex I, essential functions in chromatin modification [144], spliceosomal introns and RNAi machinery giving patchy patterns of canonical Dicer and Argonaute [145], ciliary genes [36,146], and the WASH complex [124,125,147]. These observations prompted us to look at the GO term enrichment of interacting LECA orthologous groups that contain only human genes versus interacting LECA orthologous groups that have both human and yeast genes.

We indeed find evidence of multiple genes belonging to ancestral complexes enriched in the human interaction set (Figure 5. B.), including enrichment in more straightforward GO terms related to mitochondria and respiration (e.g., GO:0005747, GO:0006120, GO:0032981 and GO:0070469), cilium (e.g., GO:0005929) and spliceosomal components (e.g., SMN complex GO:0032797). We also find evidence in higher-level GO terms that at lower levels reflect complexes known to be present in human and absent in yeast (Supplementary Table 2), such as chromatin modification (e.g., GO:0042127). For the broccoli inferred orthologous groups, more enriched GO terms reflect at lower-level complexes known to be present in human and not in yeast: Argonaute and Dicer (GO:0010629, GO:0048471 and GO:0030426), WASH (GO:0005814, GO:0005856 and GO:0043005) and chromatin modification (e.g., GO:0007399) (Supplementary Figure 10 and Supplementary Table 3).

Even though there are more yeast than human interactions present in multiple species, the entropies of the profiles participating in yeast interactions are lower (Supplementary Figure 11). This observation and the GO analysis reveal a clear reason why the performance we reported is high relative to others. Namely, we use the human reference interaction set with ancestral complexes that have been frequently lost throughout eukaryotic evolution and are absent in yeast.

DISCUSSION

Phylogenetic profiling is complicated due to many biological and technical issues. These issues include the complex histories of proteins and the choices in the meta parameters for phylogenetic profiling, such as the quantity, quality and diversity of reference genomes and annotations. Meta parameter choices in phylogenetic profiling has been extensively studied in mostly prokaryotes, where generally the focus is on the choice of reference genomes, phylogenetic profile methods, and/or the amount of data. We focus on eukaryotes and investigate qualitative different meta parameters for phylogenetic profiling. We showed that phylogenetic profile performance when predicting protein interactions is influenced by a complex interplay of multiple technical and biological parameters.

Genome diversity plays an important role in prediction performance for prokaryotes [130,132]. In contrast, our measures of eukaryotic diversity did not significantly influence prediction performance. Selecting lesser-quality genomes has a larger effect on prediction performance, while selecting higher-quality genomes does not. Genome selection and the interplay between quality and diversity does matter. However, other meta parameters have a much larger impact on prediction performance, such as the amount of information in the phylogenetic profiles in relation to the reference interaction dataset. This discrepancy suggests that more complex feature selection procedures should be explored for reference genome set selection, especially since (non-linear) interactions between subsets of genomes and combinations of subsets could drastically boost performance.

Other meta parameters, such as orthologous group and reference interactions, have a much larger effect than genome selection. Some results make a lot of sense from a technical point of view. For example, low quality/noisy functional data (unfiltered BioGRID) or mixing phylogenetic

profiles that are at least 50% inconsistent (post-LECA + LECA orthologous groups) have poor performance. A drawback to filtering out post-LECA orthologous groups is that we remove lineage-specific interactions that are still a part of a protein complex and show clear co-evolution. Our analysis shows that we should consider these often hidden choices when encountering large differences in performance between reported studies.

One very counterintuitive finding is that the yeast interaction set showed lower predictive performance. Compared to the human interaction set, yeast should be of equal quality by all accounts, arguably even better. Together with the observation that LECA orthologous groups performed better than post-LECA orthologous groups, this suggests that the performance of phylogenetic profiles in eukaryotes is optimal for modules that fulfil a very particular set of conditions. These modules (i) were present in LECA, (ii) were repeatedly lost in eukaryotic lineages, and (iii) the genes in the module conserved most of their function. This observation fits with notable examples from the WASH complex and cilium [36,124,125], or proteins with great success in predicting its components like the minor spliceosome [123] and RNAi machinery genes including Dicer and Argonaute [145]. These biological patterns should explain the very strong signal found by studies such as [124,125]. Note, both studies show very strong signals for complexes as well as pathways, which we excluded due to the problem of defining a quality negative interaction set.

In conclusion, we find that for eukaryotes more genomes and better-quality genomes are not necessarily better. It is instead the type of information in the genomes. The information in these genomes is not directly related to larger genomes, for instance parasites increase prediction performance. Instead, the information is related to the interactions of the reference species present in a given genome. Genome selection has a minor influence compared to orthologous groups selection and interactome selection, which both greatly improve the performance when predicting protein interactions. Interactome and orthologous group selection is likely the major source for the large variance in reported performances. Ancestral complexes that are repeatedly lost are responsible for the strong performance of phylogenetic profiles in eukaryotes and it is these hidden choices in orthologous group selection that we should consider when we find large differences in performance between studies.

MATERIAL AND METHODS

1.Initial datasets and methods

We started our investigation from the analysis done in our previous work [133], to investigate the influence of different parameters on the performance of predicting protein–protein interactions using phylogenetic profiles. We showed a relatively high prediction performance using a large set of diverse eukaryotes and orthologous groups inferred to be in the Last Eukaryotic Common Ancestor (LECA). This reference set is called the initial set. Any changes that we made are changes in this initial set. In the sections below, we will briefly describe the composition of this initial set and the methods we used to obtain it.

1. 1. Large scale eukaryotic dataset and LECA orthologous groups

We inferred orthologous groups on a diverse genome set of 167 eukaryotes using different orthologous group inference methods in our previous work. For this analysis, we chose the best performing method regarding protein interaction prediction, Sonicparanoid (version 1.3.0) [127]. To rule out any large orthology specific issues during our current analyses, we chose at least one other method: Broccoli (version 1.0) [128].

Ancestral eukaryotic complexes have been lost together multiple times [31]. Phylogenetic profiles should benefit from this clear modular evolution of proteins. Therefore, we selected orthologous groups estimated to be in LECA. Briefly, we inferred LECA orthologous groups using the Dollo parsimony approach [114] with additional strict inclusion criteria [133]. The Dollo parsimony method assumes that genes can be gained only once while minimizing gene loss. Before we assigned an orthologous group to LECA, it must be in at least three supergroups (See Supp. Table X) distributed over the Amorphae and Diaphoretickes (previously known as opimoda and diphoda) [116].

1. 2. Phylogenetic profiling and measuring co-occurrence of proteins

We constructed phylogenetic profiles by determining the presence (1) and absence (0) of orthologous groups in 167 species. To evaluate prediction accuracy, we obtained a higher quality reference interaction set by filtering the human BioGRID interaction database (version 3.5.172 May 2019) [117,148]. BioGRID contains physical interactions between proteins. We filtered this interaction set to keep non redundant interaction pairs found in at least

five independent studies (PubMed IDs). The number of independent studies is a measure of how thoroughly these proteins were investigated and how receptive the proteins are to high-throughput measurements. We mapped the interacting genes to their corresponding orthologous groups.

We used the best performing negative protein interaction set from our previous analyses [133]. We inferred this negative set by taking pairs of interacting proteins that were found to be interacting at least five times, but not with each other. This excludes the possibility that the negative set contains interacting proteins that were not found due to manifold technical reasons.

To calculate the (dis)similarities between phylogenetic profiles we used the from our previous analysis best performing distance measure, the cosine distance.

2. Genome selection procedures

We compared the results of all the genome selection procedures to 1000 sets of genomes randomly selected to exemplify that the differences in prediction accuracies are not due to random variations in genome composition. We calculated the protein interaction prediction performance for each of these random genome sets.

2. 1. Selecting better and worse quality genomes

To measure the quality of the genomes, we used two quality metrics. The first metric is the out of the box BUSCO metric that works by calculating the absences of highly conserved single-copy orthologs [84]. The BUSCO Eukaryota database (odb9) was aligned to the genomes using the hmmsearch alignment tool from the HMMER package 3.1b2 (dated February 2015) [82]. We took the HMM specific quality score given by BUSCO to validate the hits in the alignments.

The second metric is of our own devising. The second quality metric we used was the Illogical absences (IA) metric of our design. We added this second independent metric to remove the dependence of quality on a single measure to establish the completeness of the genomes and gene prediction. The IA metric calculates the number of absences of protein interaction partners, which we termed Illogical absences. Illogical absences follow from the assumption underlying phylogenetic profiling that interacting

proteins are often evolutionary conserved. Therefore, it can be considered suspect when a protein interaction partner is absent. A possible reason could be that the absence is due to gene prediction, genome annotation or even homology detection errors.

We selected the strongest interacting orthologous group pairs by selecting their phylogenetic profiles with the least cosine distance. This selection removes the complex interplay between interacting groups of orthologs. For every interacting orthologous group pair, we calculated the absences of interaction partners in every species. These absences we termed illogical absences or the IA metric.

We can consider the genomes with the most BUSCO absences and illogical absences as lesser quality genomes. In contrast, we can consider the genomes with the least BUSCO absences and illogical absences as higher-quality genomes. We selected 50 genomes of lesser-quality and 50 genomes of higher-quality for each of the metrics and recalculated the protein-protein interaction prediction performance.

2. 2. Selecting highly diverse and similar genomes

We calculated the pairwise cosine distance between all species with the presence and absence profiles of LECA orthologous groups to obtain species sets of maximum diversity and maximum similarity. We then iterated through the resulting pairwise distance matrix and selected the maximally distant pairs for the diverse set or minimally distant pairs for the similar set. Before adding a species of a species pair to a set, we checked to see if the species also had a distance above a certain arbitrary threshold to the other species in the growing set (cosine value ≥ 0.38 for the dissimilar, cosine value ≤ 0.58 for the similar set). We did this until we obtained the desired amount of 50 genomes per set. The maximum diverse and maximum similar genome sets were each used to recalculate the protein-protein interaction prediction performance.

2.3. Selecting single influential genomes and their combined effect on prediction performance

We removed genomes one-by-one from the initial species set of 167 eukaryotes to see how the different genomes influence the performance of protein interaction prediction with phylogenetic profiling. We recalculated the performance for each of these 167 sets. The 50 genomes that increased

the performance compared to the initial species set the most when removed from the initial set were labelled as disadvantageous. The 50 genomes that decreased the performance the most when removed from the initial set were labelled advantageous. For both the disadvantage and advantageous set we recalculated the protein interaction prediction performance.

3. Gene and interactome selection procedures

We compared the results of the orthologous group selection procedures to randomly selected LECA orthologous groups to exemplify that the differences in prediction accuracy is not due to random variations in orthologous group composition. We made a thousand LECA orthologous group sets containing a random selection of 63% of the orthologous groups. We calculated each of these set's protein interaction prediction performance.

3. 1. Selecting orthologous groups

In our initial species set, we used orthologous groups estimated to be in LECA (Methods section 1.1.). We took the raw output of the orthology inference methods and filtered out the LECA orthologous groups to get a set that contains post–LECA orthologous groups. We also recalculated the prediction performance with the raw output of the orthology prediction methods, which is all inferred orthologous groups.

3.2. Selecting different reference interactomes

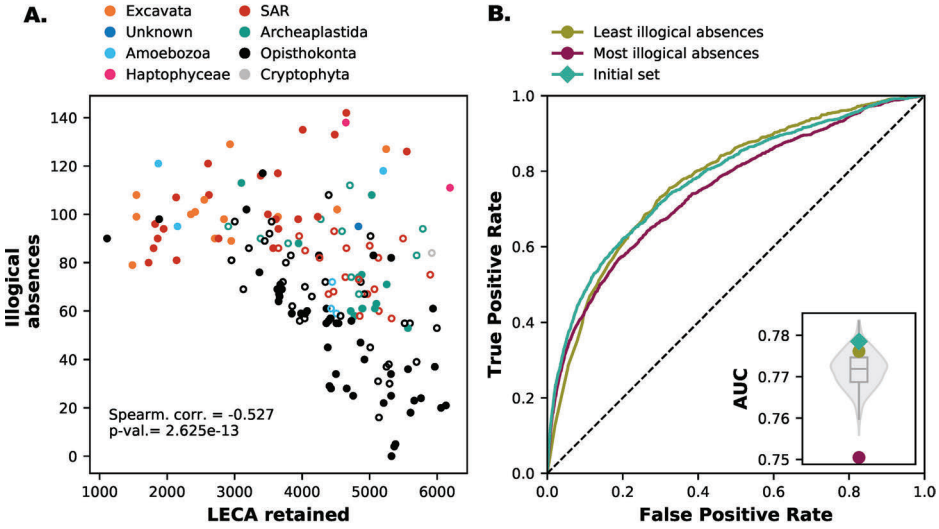
We compared the five PubMed ID filtered human BioGRID set with the unfiltered human BioGRID dataset. Every interaction with less than five PubMed IDs is now included as well. Removing the five PubMed ID filter should indicate how quality filtering of reference interactions influences prediction performance.

We selected next to the human interactions the *Saccharomyces cerevisiae* BioGRID interaction database (version 3.5.175 July 2019) [117] to analyse the influence of the reference interactome. We filtered the interactions to keep only the interaction pairs found in at least five publications (PubMed IDs). We followed the same procedure as with the human interaction set (Methods section 1.2.).

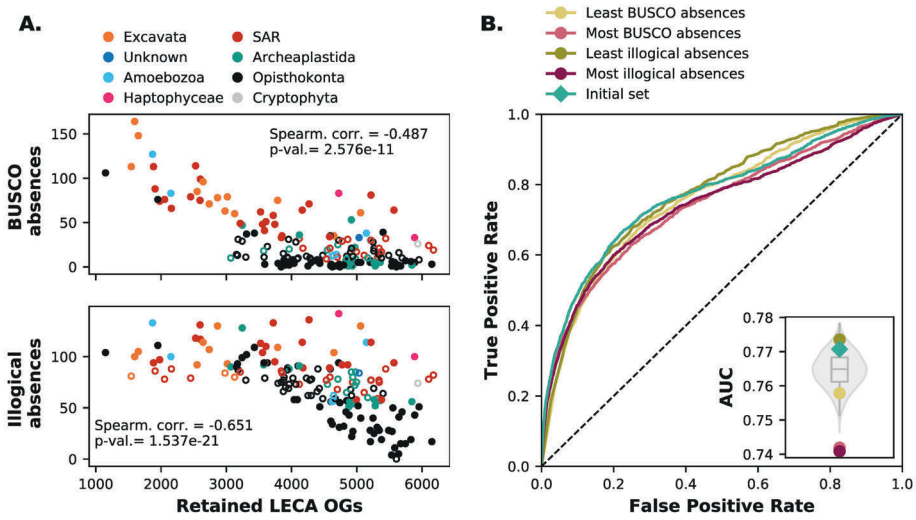
Following this analysis, we hypothesized that the drop in prediction performance for yeast is caused by the loss of ancestral protein complexes

in yeast. To test this, we chose interacting LECA orthologous groups that contained only human genes (sample set) and calculated the enrichment to the set with interacting LECA orthologous groups containing human and yeast genes (population set). We calculated the enrichment using the following equation: $\frac{n}{m} \div \frac{k}{q}$, where n is the total number of genes associated with a GO term (Downloaded GO terms Januari 2021 biomart) in the sample set (overlap), m is the total numbers of genes in the sample set, k is the total number of genes associated with a specific GO term in the population set, and q is the total number of genes in the population set. Since enrichment does not work well for small overlaps, we filtered for a minimum overlap (n) of 3. Enrichment was considered significant for p-values below 0.01. Since orthologous groups can contain multiple genes, we randomly selected genes from an orthologous group to generate a new sample and population sets ten times and recalculated the enrichment.

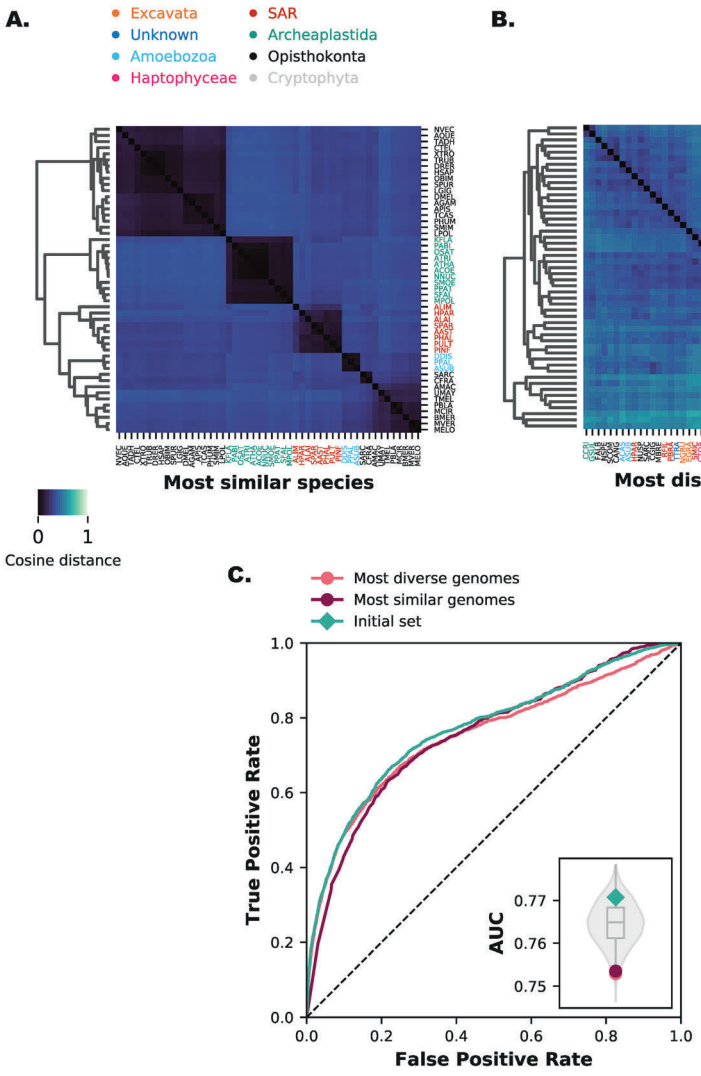
SUPPLEMENTARY FIGURES



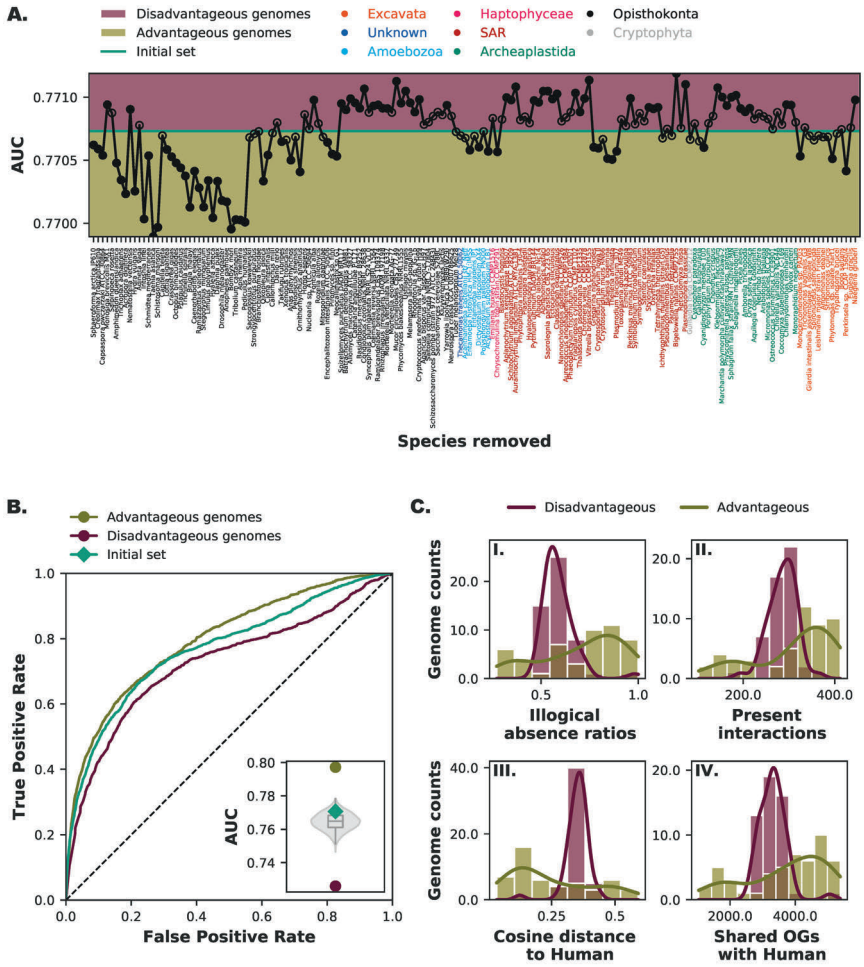
Supplementary Figure 1. Illogical absences and genome quality selection based on illogical absences using Sonicparanoid inferred orthologous groups (OGs). A. Illogical absences as a function of retained LECA OGs in different species. We see that where most Opisthokonta scores similarly low with the BUSCO metric, they score lower with the IA metric indicating a difference between the two metrics. However, the performance of genomes selected with both metrics are similar to each other. Filled data points are the selected genomes for the prediction accuracy calculations. **B.** Receiver-operator Curve of two species sets (n = 50) with the most and least illogical absences. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (teal diamond). Human has a perfect score of 0 illogical absences since the interactions are from the human reference interactome. Therefore, we did not select human for the genome set.



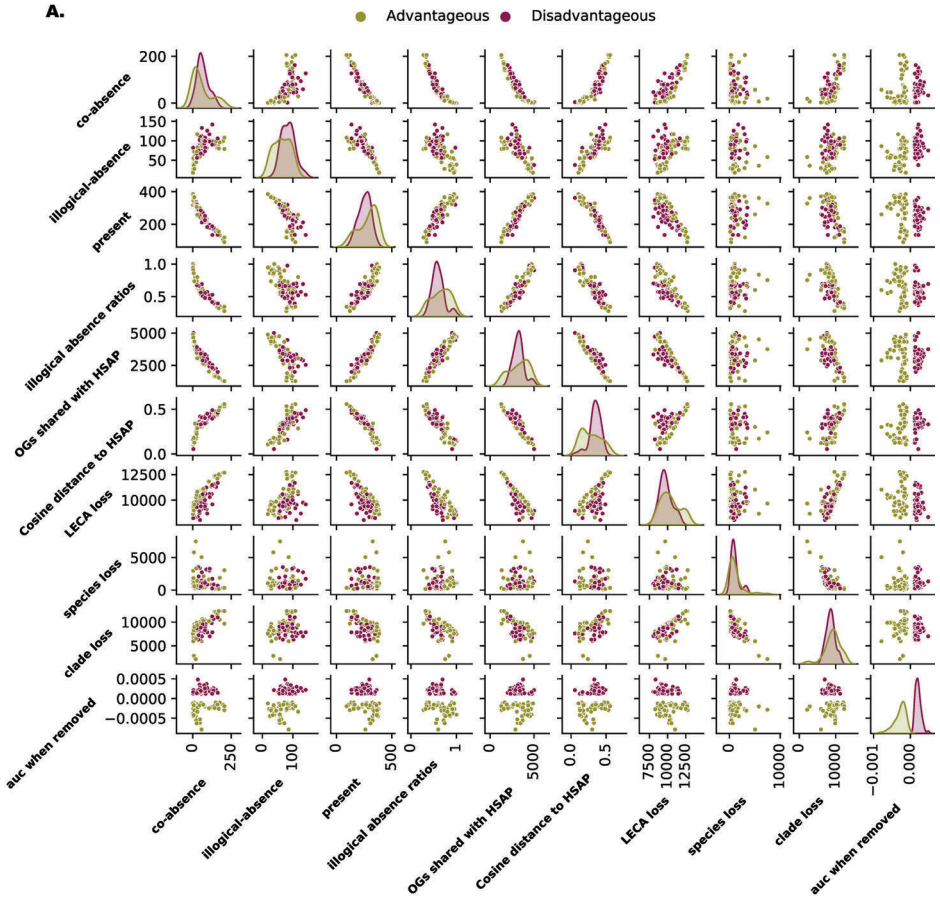
Supplementary Figure 2. Lesser quality genomes have more impact on protein interaction prediction performance also for Broccoli inferred orthologous groups (OGs). A. BUSCO and Illogical absences as a function of retained LECA OGs in different species. Filled data points are the selected genomes for the prediction accuracy calculations. **B.** Receiver-operator Curve of two species sets (n = 50) with the most and least BUSCO and illogical absences. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (teal diamond).



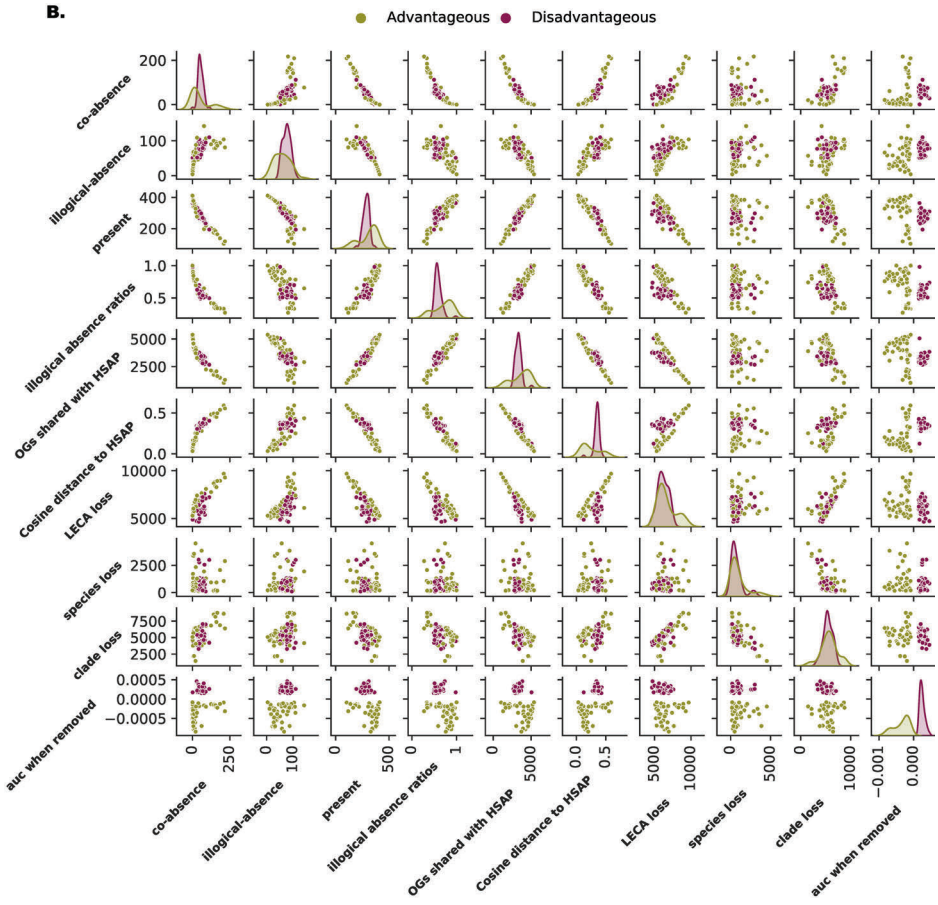
Supplementary Figure 3. Both high and low diversity sets have little impact on protein interaction prediction performance also for Broccoli inferred orthologous groups. A. The most similar species form more clusters and are overall more similar to each other. **B.** The most diverse species show no clustering and are overall less similar to each other. **C.** Receiver–operator Curve of two species sets (n = 50) with the most diverse and most similar species. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (green diamond).



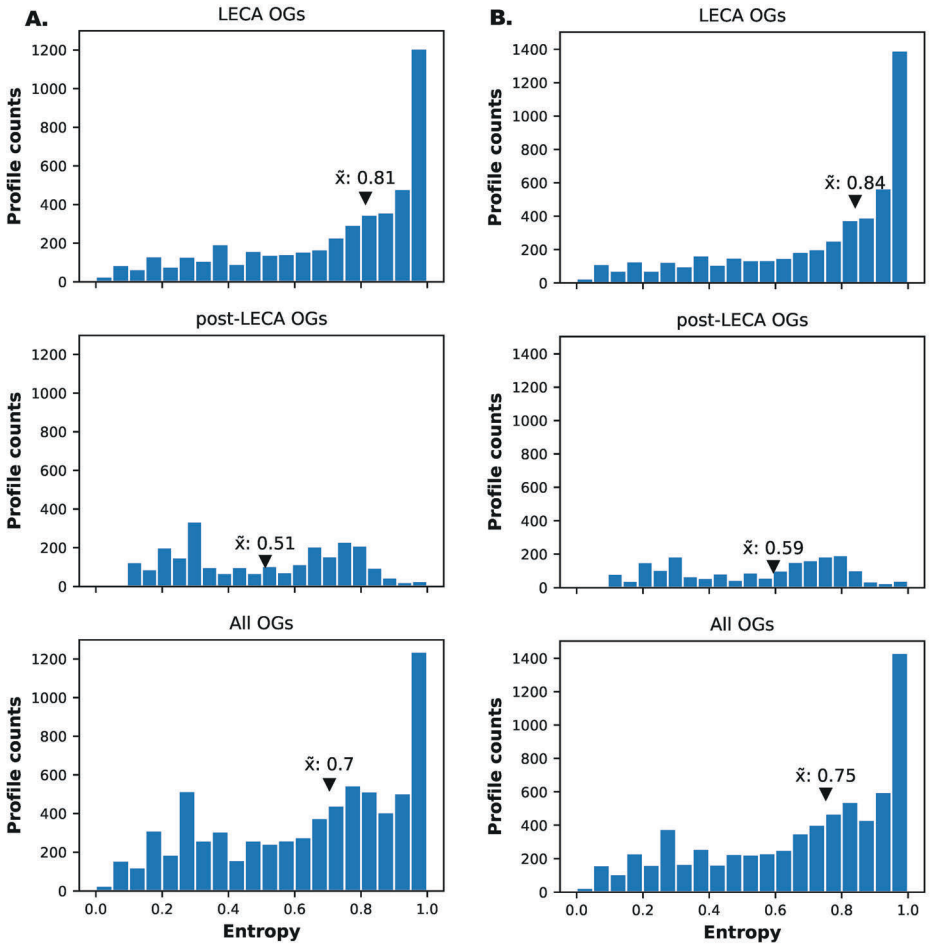
Supplementary Figure 4. Influential genomes and their combined effect for Broccoli inferred OGs. A. Recalculated Area Under the Curve (AUC) values when a single species is removed from the initial species set. Genomes that increase the AUC value when removed can be considered disadvantages compared to the initial set when predicting protein interactions with phylogenetic profiles. Genomes that decrease the AUC value when removed can be considered advantageous for predicting protein interactions. Top 50 advantageous and top 50 disadvantageous genomes shown with the black fill in the scatter plot. **B.** Receiver-operator Curve of two species sets (n=50) with the most advantageous and disadvantageous genomes. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of 1000 random species sets (violin plot) and the initial species set (green diamond). **C.** Comparison of the counts (histogram) and kernel density estimates (line plot) of (I) illogical absence ratios (illogical absences divided by total interaction absences (co-absences + illogical absences)), (II) present interactions, (III) the cosine distance to human, and (IV) total shared orthologous groups with human.



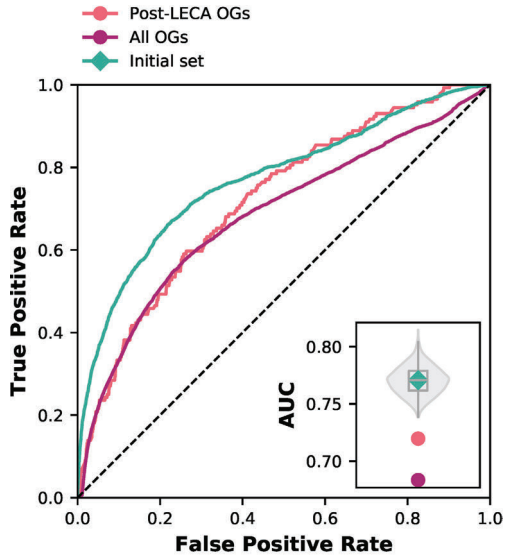
Supplementary Figure 5. Continues on next page.



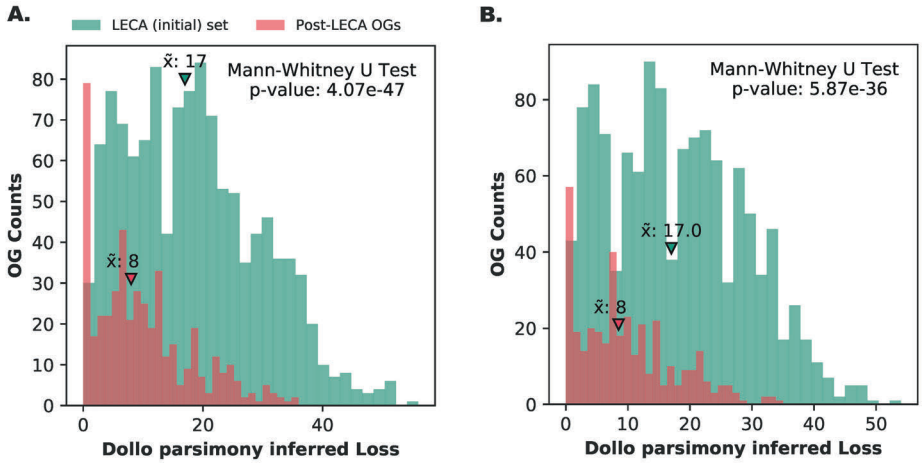
Supplementary Figure 5. Correlations between multiple parameters in the advantageous and disadvantageous genome set. Given for A. Sonicparanoid and **B.** Broccoli inferred orthologous groups (OGs). From top to bottom (or left to right) the interactions that are co-absent; illogically absent; and present; the ratio of illogical absences to total absences; number of OGs shared with the human genome; the cosine distance to the human genome; LECA OGs loss (Dollo parsimony inferred); species (lineage) specific loss; (clade) ancestral loss; and the difference in AUC from the initial set AUC when a genome is removed.



Supplementary Figure 6. Entropy of phylogenetic profiles that have interactions. Given for **A.** Sonicparanoid and **B.** Broccoli inferred orthologous groups (OGs). From top to bottom, the entropy is shown in profiles for LECA, post-LECA and all OGs. Median entropy is presented with a black arrow. Mann-Whitney U test shows significant difference between distributions of LECA, post-LECA and all OGs, p-value < 0.001.



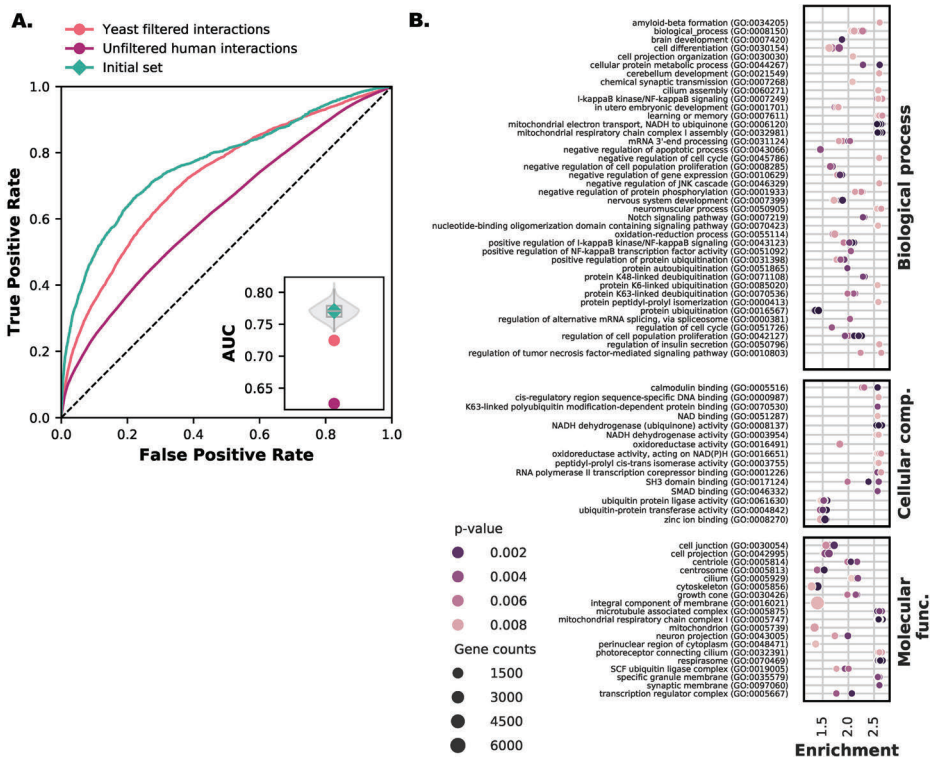
Supplementary Figure 7. Orthologous group selection has a large impact on prediction performance also for Broccoli inferred orthologous groups (OGs). Receiver–operator Curve of post–LECA orthologous groups and unfiltered orthologous groups. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of randomly selected LECA OGs (violin plot) and the initial species set (green diamond).



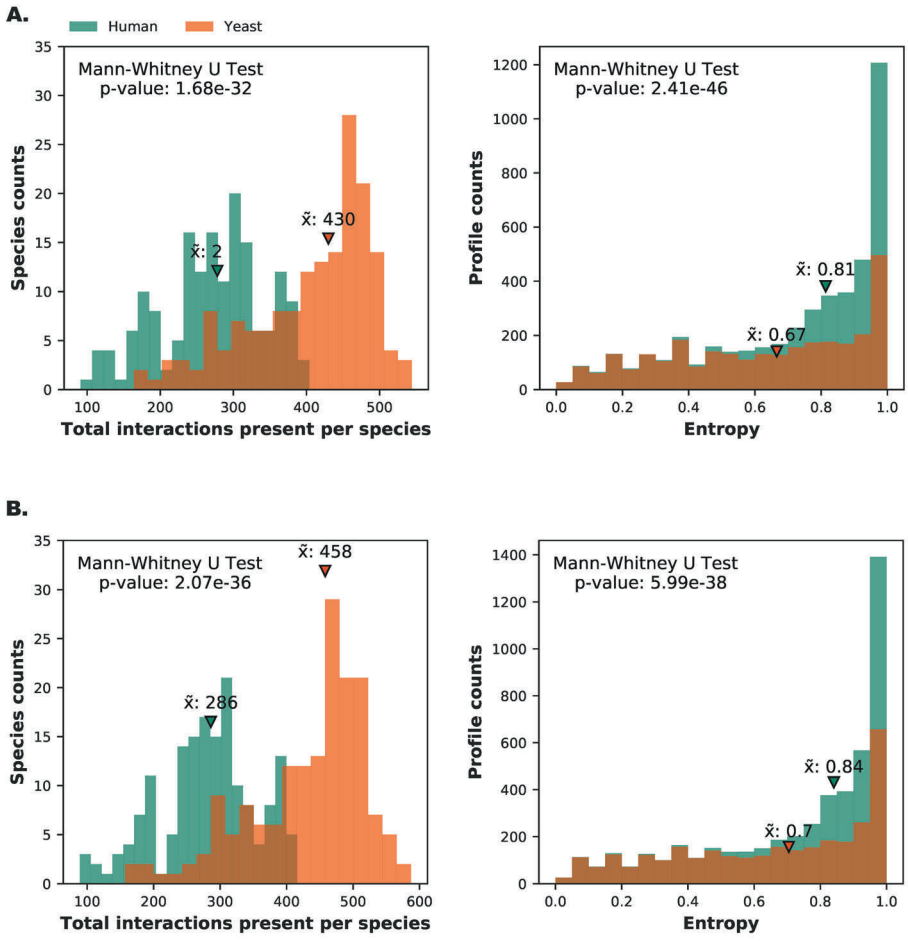
Supplementary Figure 8. Dollo parsimony inferred loss of LECA and post–LECA orthologous groups (OGs). Given for **A.** Sonicparanoid and **B.** Broccoli inferred OGs. Mann–Whitney U test shows significant difference between distributions.

https://github.com/ESDeutekom/SelectionPhylogeneticProfiling/blob/main/Figures/S9_Fig.pdf

Supplementary Figure 9. Groups of interacting orthologous groups (OGs) where one is in LECA (always the last row in a group subplot) and the others are not. The profiles are sorted according to the species tree.



Supplementary Figure 10. Interactome selection is important for prediction performance. A. Receiver-operator Curve of post-LECA and unfiltered orthologous groups (OGs) of Broccoli. The inset gives the Area Under the Curve (AUC) values compared with the random backdrop of randomly selected LECA OGs (violin plot) and the initial species set (green diamond). **B.** GO-enrichment analysis for genes enriched in interactions present in only human vs. interactions present in human and yeast. OGs can contain multiple genes. We randomly selected genes from an OG to generate new sample and population sets 10 times and recalculated the enrichment (shown by multiple points in the figure rows).



Supplementary Figure 11. Interactions of human and yeast interactome present in different species (left) and entropy for LECA profiles that have interactions in human and yeast (right). Given for A. Sonicparanoid inferred and B. Broccoli inferred orthologous groups (OGs). Median values are presented with the arrows. Mann–Whitney U test shows significant difference between distributions.

SUPPLEMENTARY TABLES

Supplementary Table 1. Species table for species used in this study. Green marked species are the species that are in the advantageous set, and red marked species in the disadvantageous set (Sonicparanoid). The measured values are also shown in Supplementary Figure 5.

https://github.com/ESDeutekom/SelectionPhylogeneticProfiling/blob/main/Tables/S1_Table.xlsx

Supplementary Table 2. GO-enrichment table for Sonicparanoid inferred orthologous groups (OGs). Since there can be multiple genes in an OG, we randomly selected one of the genes for the GO-enrichment analysis. We did this ten times, creating ten foreground and background sets (set_num). These values are shown in Figure 5.

https://github.com/ESDeutekom/SelectionPhylogeneticProfiling/blob/main/Tables/S2_Table.tsv

Supplementary Table 3. GO-enrichment table for Broccoli inferred orthologous groups (OGs). Since there can be multiple genes in an OG, we randomly selected one of the genes for the GO-enrichment analysis. We did this ten times, creating ten foreground and background sets (set_num). These values are shown in Supplementary Fiure 10.

https://github.com/ESDeutekom/SelectionPhylogeneticProfiling/blob/main/Tables/S3_Table.tsv

CHAPTER 5.

Exploration of machine learning with phylogenetic profiling for protein interaction prediction in eukaryotes

Eva S. Deutekom, Tristan Schadron, Berend Snel

In preparation

ABSTRACT

Phylogenetic profiling is of continued interest to identify functional relationships of pairs of proteins in eukaryotic species. Getting a good prediction performance in eukaryotes is still challenging due to a myriad of technical and biological reasons. In our previous work we showed how multiple key meta parameters had an unexpected impact on phylogenetic profile performance. The often surprising influence of many choices and their potential non-linear interaction strongly suggest machine learning should be utilized to improve protein interaction prediction in eukaryotes using phylogenetic profiles.

When protein interaction prediction is defined as a classification problem, there exists some difficulties to first overcome. To construct a classifier, we need protein interactions (positive samples), which are relatively easy to obtain (e.g., Chapter 3 and 4 quality filtered BioGRID protein interactions). On the contrary, we also need negative protein interactions (negative samples) that we cannot obtain because for practical and important conceptual reasons, are not annotated in interaction databases. We can choose to define the negative interactions from all the samples that are not positive (Chapter 3 and 4). However, this creates an imbalance class problem, since there is only a small amount of protein interactions that will create a large set of non-interactions. Moreover, the negative interactions might also still contain some false negatives.

Here we explore multiple methods to deal with the difficulties mentioned above and perform machine learning on phylogenetic profiles for protein interaction prediction. Firstly, we assume that our defined negative protein interactions set (Chapter 3 and 4) is not negative, but an unlabelled set. We can then approach this classification problem as a Positive-Unlabelled learning problem. Positive-Unlabelled learning has shown to benefit these kinds of machine learning issues: learning without bona fide or pure negatives, or in other words unlabelled data. We will use a two-step method that first extracts reliable negatives from the large unlabelled data and then applies a classification on this set. Secondly, to tackle the (introduced) class imbalance, we evaluate multiple approaches and algorithms that can be used that handle class-imbalance problems. These approaches include the under sampling of the larger negative interaction

set followed by Random Forest, and using weighted and balanced Random Forest algorithms on the imbalanced sets.

We find that using a two-step Positive-Unlabelled learning approach increases the performance of phylogenetic profiling in eukaryotes considerably, compared to using just quality protein interactions and non-interactions as is. However, we also find that all approaches have a difficulty in predicting protein interactions due to the lack of sufficient amounts of positive data. Nevertheless, we can obtain a good performance to differentiate between positive and negative protein interactions using phylogenetic profiles and improve performance compared to distance based phylogenetic profiling. These results warrant further studies on this topic.

INTRODUCTION

Many genomes still contain proteins of unknown biological role or function, including in depth studied genomes such as human, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. Phylogenetic profiling has proven to be a unique and complementary method to provide function predictions that can be experimentally tested [40,41,60,124,125]. Phylogenetic profiles are the presence and absence of proteins across multiple species. The similarity between profiles can be measured. Similarity between profiles is an indication of co-occurrence of proteins in multiple species. It has been shown that proteins that co-occur tend to interact [ref]. This is referred to as guilt-by-association, where genes with known/equivalent functions are co-present or co-absent.

We were able to get a good performance for protein interaction prediction with phylogenetic profiles for eukaryotes ([133] and Chapter 4), which has been proven difficult in previous studies [129,130,132]. We identified key meta parameters that have a positive effect on performance, including orthologous group and reference interactome selection, and quality filtering of protein interactions. Here we want to improve performance of protein interaction prediction using phylogenetic profiles further by exploring multiple machine learning techniques, such as Random Forest. Random Forest has already been used successfully with phylogenetic profiling in prokaryotes [132].

In our previous study we found that filtering the (non-)interaction set, greatly improved the quality of the dataset and therefore the performance of phylogenetic profiles. We used a negative interaction set defined of protein pairs that were not interacting with each other, but had at least five interactions with other proteins, reducing the chance that a non-interaction is just due to lack of interaction evidence. Even though the negative set will be less noisy when filtered, the absence of a measured interaction does not necessarily mean two proteins never interact [141]. Protein interactions are condition dependent. The proteins taking part in an interaction need to both be expressed. Additionally, other requirements need to be met, such as cofactors that need to be present, correct subcellular localization, or protein modifications necessary for interaction. These signals are under the influence of environmental conditions, cell type or time (cell cycles). Technical difficulties can also cause interacting proteins to be missed,

e.g., posttranslational modifications that are needed to carry out protein functions are unlikely to behave or interact normally in Yeast 2 Hybrid experiments [149]). It is therefore likely not all interactions will or can be found experimentally.

Defining a negative protein interaction set from a positive interaction set like we did in previous chapters, means that there can be positive protein interactions that are not yet found experimentally contaminating the negative set. This point is highlighted in Chapter 3 as well, where we compared a protein negative interaction set inferred by Trabuco et al. [118] and found some of the negative interaction pairs were now deemed interactions in BioGRID. For that same reason, there might still be positives in our own defined negative set. For machine learning this has important consequences: it strongly suggests the need to utilise methods that are able to learn from data that contain both positives and negatives by considering the data to be "unlabelled".

The difficulty of obtaining proper negative samples occurs naturally in many scientific fields, including health care where the absence of disease symptoms is not always the absence of the disease. Problems like this increase the demand for machine learning algorithms that learn from positive and unlabelled samples, or learning without negatives, making it an active field of research that has gained a surge of interest in recent years [150]. A method proposed for the purpose of learning without negatives is the Positive-Unlabelled learning. Positive-Unlabelled learning assumes that the unlabelled sample can belong to either the positive or negative class [151]. One Positive-Unlabelled learning method has been successfully used for protein interaction prediction with gene expression data [152,153]. This Positive-Unlabelled learning is a two-step learning approach, which first infers reliable negatives from the unlabelled set and then builds a classifier with these reliable negatives. In essence, Positive-Unlabelled learning locks on to cases that are most dissimilar to the set of positives which helps the classifier to better distinguish between the negatives and positives while learning. We will explore this method for the use in phylogenetic profiles as well.

Another issue to deal with when we want to use machine learning with phylogenetic profiles for protein interaction prediction in our particular dataset is the class imbalance. In our interaction set we have the "minority

class" of positive interactions that is comparably extremely small to the "majority class" of negative interactions (ratio positive to negative interactions is 1:5899). The degree of imbalance of the classes (i.e., the ratio of the minority class to majority class) and at what degree classification performance deteriorates, depends on multiple factors [154]. These factors include the absolute sample sizes in the classes and separability of the classes. If the minority class is large enough or the classes are really easy to separate, the degree of class imbalance might not be a problem.

With our dataset we might have too little samples in the minority class, i.e., too little positive interactions (1781). If there is a fixed degree of imbalance, the sample size of the positive interactions determines how good a classification model can be. In other words, if there are enough positive interactions to classify with, the class-imbalance does not have to be a problem [155]. With more data, more information about the minority class is available. More information benefits classification by enabling the model to better distinguish between the minority and majority class [154]. Class imbalance often causes problems because of the lack of patterns that belong to the minority class, not because of the ratio between the minority and majority class.

In this chapter we will explore how machine learning approaches can be applied to phylogenetic profiles for predicting protein interactions. This implies analysing the effect of a highly imbalanced set that has very little positive (protein interaction) data. We perform an exploratory analysis on the applicability of multiple readily available solutions and approaches developed for unlabelled and unbalanced learning problems. These approaches include Positive-Unlabelled learning and implementing class-imbalance aware approaches, such as under sampling of the majority class before machine learning with Random Forest, directly implementing class-weighted and class-balanced Random Forests [156]. We analysed how these methods perform specifically for protein interaction prediction using phylogenetic profiles. Based on the outcomes we discuss possible future directions in this field.

RESULTS & DISCUSSION

Our original negative protein interaction set (Chapter 3 and 4) likely still contains positive interactions and can also be seen as a large unlabelled set. This is because there is no actual experimental evidence that the negative protein interactions are not interacting, they might just not have been found yet. To see if we can increase performance from this original negative set, we first used two-step Positive-Unlabelled learning [151,153,158] to pre-process the unlabelled set (step 1) and train the classifiers on this pre-processed set (step 2). We compared the performance to the originally inferred non-interactions set [133,159].

There is also a large imbalance in protein interactions in our data. In other words, a class-imbalance of the positive interacting protein and non-interacting protein classes (1781 interactions and 10505860 non-interactions). Supplementary Table 1 shows this number for all the test and training sets. We used multiple class-imbalance aware approaches to handle the imbalance and see which one provides the best final classifier. We focussed mainly on Random Forest [160] methods implementing some form of class imbalance awareness. Random Forest has been shown to have a good performance for phylogenetic profiling in prokaryotes [132], is less sensitive to noise and generalizes better (i.e., less overfitting) [160], and has the added benefit of being able to predict feature importance that we can use for feature selection later on . The class imbalance aware Random Forest approaches include an under sampling of the larger "majority" class followed by a "classic" Random Forest classification, a weighted Random Forest and a balanced Random Forest classification (Materials and Methods).

We used the different classification approaches for both the pre-processed (from the Positive-Unlabelled learning method) and original negative set for comparison. For the positive unlabelled learning, we are in essence updating the Positive-Unlabelled learning approach [151,153] with classification methods that account for the class imbalance (Materials and Methods).

We finalize this study by doing a feature selection with the best performing approach to see how certain species impact the overall performance of phylogenetic profiling in eukaryotes.

Pre-processing the negative set with positive-unlabelled learning sometimes boosts performance of phylogenetic profiling

In order to improve phylogenetic profiling performance for predicting protein interactions, we explored the effect of pre-processing our original negative set [133,159]. Our positive interaction set should be relatively reliable due to the strict filtering criteria. Namely, in the positive set we include only interactions that have been experimentally verified in five or more independent studies. We inferred our negative interaction set by taking pairs that are not interacting with each other, but have at least five interacting partners. Because of this quality filtering, the negative set will be less noisy than any unlabelled data set. Although we spent quite some effort to increase the quality of both the positive set and negative set, our negative interaction set likely still contains contamination from interacting proteins.

We therefore took a two-step Positive-Unlabelled learning approach to purify the negative interaction set further. Often used Positive-Unlabelled learning approaches are two-step approaches that start with identifying unlabelled samples that can be confidently labelled as negatives (reliable negatives), followed by training the positive and reliable negative set with a standard binary classifier. This two-step Positive-Unlabelled learning can help by making the decision plane between classes clearer. This makes it easier for the classifier to learn from the data.

An appropriate two step Positive-Unlabelled learning method is the Rocchio-Random Forest (Roc-RF) [153] that proved to give good performance for protein interaction prediction when using gene expression data to predict protein interactions. Roc-RF combines the Rocchio technique (or Nearest Centroid classification [161]) to select for strong negatives(Figure 1), followed by Random Forest classification (Materials and Methods). Since our original negative set (or unlabelled set) contains pairs that are likely for the most part non-interacting, Rocchio might prove even more useful, because the Rocchio method works well when the proportion of positives in the set is small and the unlabelled set is large and diverse in character [151]. This is the case for our set if we look at the distances between the profiles in both the negative and positive interaction sets (Figure 1). In other words, there are multiple reasons to indicate that we have negative interacting protein pairs with similar profiles, or positive interacting proteins with dissimilar profiles.

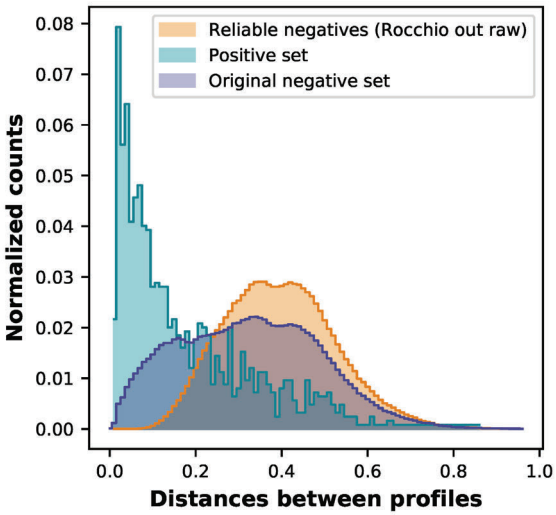


Figure 1. The distance distribution between phylogenetic profiles of the positive, original negative, and Rocchio inferred negative sets.

We continue our analysis by looking at multiple approaches to handle the class imbalance in our data set. The easiest and fastest strategy used for imbalanced classification is random sampling of the initial data. We chose to do a "naive under sampling" approach, which is basically a random subsampling of the larger negative interaction set. The "naïve" in the naïve under sampling comes from the fact that this under sampling does not assume anything about the data and no heuristics are used to select the data. This means that data of the negative interactions will be lost, data that can be important for proper classification. Under sampling of the data is easy and makes classification less time consuming. For this reason, under sampling is often the approach used to handle imbalanced data with a large class, regardless of the implications of throwing away important data. The class under sampling is only applied to the training data. Applying the under sampling only to the training data, allows to test the model on a realistic and thus imbalanced test set.

There is no prior knowledge on the optimal class distribution, or the optimal ratio of classes, to maintain enough information and negate the negative effects of a highly imbalanced set. A ratio of 1:1 positive to negative interactions might perform well, but is not always optimal [162]. Learning is not always optimal with a 1:1 ratio, because another ratio can give better

performance due to, for instance, including more important information for a better classification. Therefore, we resampled the negative set (both the purified and original) with multiple ratios to the positive set to see at which ratio of imbalance the goldilocks zone of performance can be found. We used the following metrics that are suggested to perform well for imbalanced data to compare the different ratios of imbalance (Material and Methods): the macro averaged F-measure (maF1), area under the Receiver Operator Curve (AUC-ROC) and Precision-Recall curve (AUC-PR), balanced accuracy (bAccuracy), and the G-mean.

The results show that until a certain ratio of positives to negatives, the performance as measured by the AUC-ROC value is better with the original negative set compared to the reliable negative set (Figure 2). Random Forest seems to have a hard time identifying positives from the negatives when using the original negative set after a certain positive to negative ratio for training, which can more clearly be seen in confusion matrices shown in Supplementary Figure 1 and 2. The performance as measured by the AUC-ROC value for the reliable negative set remains relatively stable. Compared to training with the original negative set, the training with reliable negative set gives a classifier better capable of identifying positives. The reliable negative set gives a classifier that is overall better capable of identifying positives, as shown by the AUC-PR values (Figure 2). The best performance with both the original and reliable set was with the 1:1 ratio, as measured with (most of) the other performance measures. We therefore took the 1:1 results for further comparison with the other approaches.

A severe limitation with under sampling the data is that possibly important samples are removed that are important or useful in defining a decision boundary. This might be detrimental for inferring protein interactions, since we can be throwing away proteins that are not so clear-cut negative (boundary cases), or negatives that are really far away from the positives that provide useful information for the classifier [154]. For this reason, we wanted to use more sophisticated approaches to handle the data imbalance in our set without having to throw away data.

We first ran a Random Forest classification with class weighting to account for class imbalance between the positive and negative protein interactions. Class weighing is a way to modify the decision tree to take into account the imbalance in the classes by weighing each class. There will be a heavier penalty on misclassifying the minority class [156]. We find that the class weighting performs worse than down sampling the negative set and using the "classic" Random Forest approach, indicating that weighing the protein interaction classes does not increase performance with phylogenetic profiling (Figure 3).

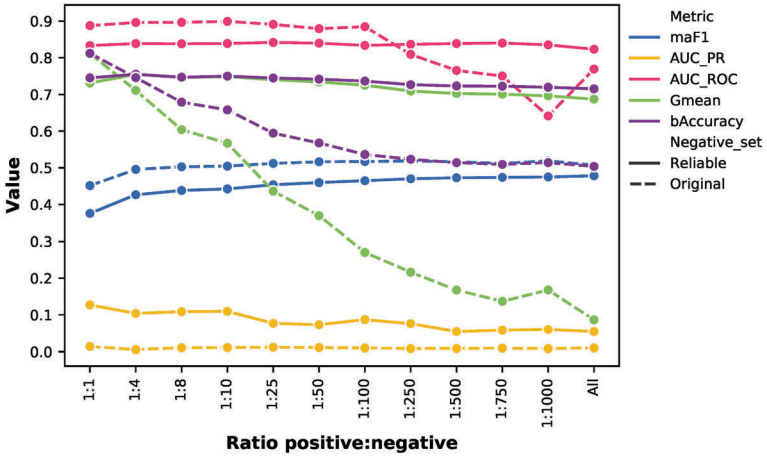


Figure 2. Performance of datasets with different resampling ratios (in training set). We used both the reliable (solid lines) and original negative set (dashed lines). Different performance metrics are shown with the different colours. All ratios are compared to the original (not undersampled) set "All". For the original negative set, this is a positive to negatives ratio of 1:5897. For the reliable negative set, this is a positive to negative ratio of 1:3829).

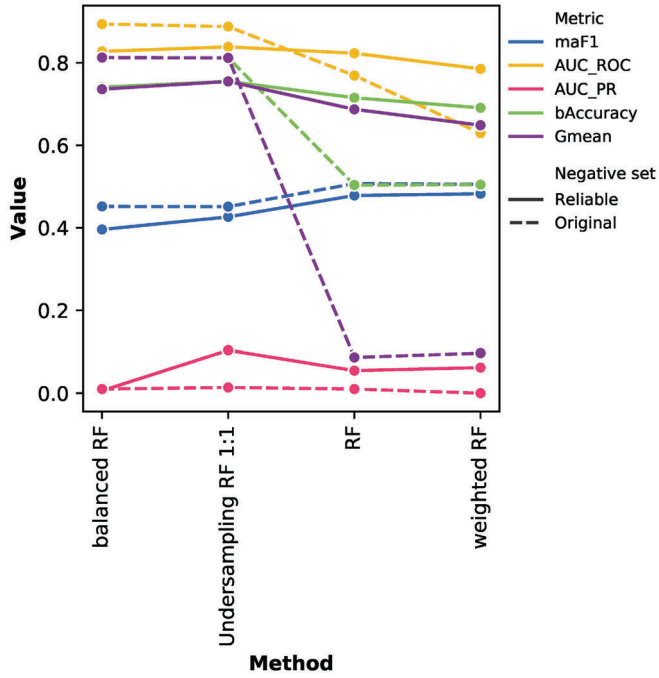


Figure 3. Performance of different methods sorted according to highest AUC-ROC values. We used both the reliable (solid lines) and original negative data set (dashed lines). Different performance metrics are shown with the different colours.

We also tested the use of a balanced Random Forest classifier [156]. The balanced Random Forest performs a more sophisticated data undersampling of the larger non-interacting protein class by balancing the bootstrap samples in order to change the class distributions. Rather than undersampling the whole training data set, all the data is used but the bootstrapped samples are balanced. The balanced Random Forest performs slightly better than the naïve under sampling when using the original negative set (Figure 3, Supplementary Figure 5 and 6). This is reversed when training with the reliable negative set. Although the balancing of the bootstrap samples is a more sophisticated approach to reducing the sample size compared to undersampling the data beforehand, it still performs similarly to the naïve undersampling.

Overall, the best performance is obtained using the original negative set with the balanced Random Forest algorithm (Figure 3). Although, the performance of the balanced Random Forest is only marginally better

than the naïve undersampling. The results also show that depending on which approach you use, the positive unlabelled learning can benefit training. This is particularly the case when we want to also better classify interacting proteins (higher AUC–PR), rather than differentiate between interacting and non-interacting proteins.

The performance increases when using a phylogenetic profiling with machine learning approach (AUC–ROC of 0.89) compared to distance based phylogenetic profiling (AUC–ROC of 0.78) to predict protein interactions (Figure 4). Random Forest has the important benefit that it can reduce the influence of redundant decisions points, in other words the patterns of presences and absences, that would influence distance based phylogenetic profiling [163]. This is because Random Forest uses discission trees that are constructed by searching through random subsets of presences and absences or orthologous groups in a profile when splitting a node in the discission tree. Redundant points are genomes that are not beneficial to profiling. If we are interested in differentiating between positive and negative interactions, we have successfully done so.

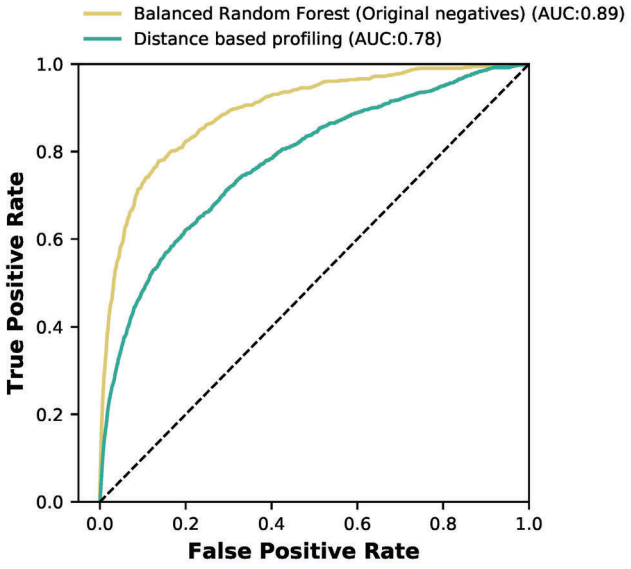


Figure 4. Receiver operator curve for distance based phylogenetic profiling and phylogenetic profiling with machine learning.

Nevertheless, the AUC-PR values (Figure 3) indicate that if the interest lies in solely identifying positive interactions, these methods might not be sufficient. The low AUC-PR is likely due to the lack of positive interaction data that has a large impact on training the models and diminishing performance for predicting positive interactions.

Feature selection does not improve performance of phylogenetic profiling

An important benefit of machine learning for phylogenetic profiling is that non-linear interactions between genomes, or combinations of these subsets, can be inferred. This can possibly improve protein interaction prediction performance. Random Forest grants the opportunity to not only classify, but also select the best performing features in the data. We chose to use the best performing method from our previous section, the balanced Random Forest without PU learning (original negative set), and went further to see if feature selection can improve performance.

We selected the top performing features (Materials and Methods) (60 genomes in total) and re-ran the balanced Random Forest classification. Feature selection did not increase and in fact slightly decreases the performance (Supplementary Figure 9). For instance, the AUC-ROC went from 0.89 for the full feature set, to 0.88 for the selected features. Nevertheless, feature selection might still be useful as it helps to reduce computational costs. We did not check how feature selection influences the other methods and we cannot at the moment say if the other methods benefit (more) from feature selection.

OUTLOOK AND CONCLUSION

We demonstrated that the performance of phylogenetic profiles for protein interaction prediction is improved by utilizing machine learning techniques by increasing the reliability of the negative interaction set, and by addressing class imbalance issues. Our results seem to consistently show that current interaction databases, despite their absolute size, contain too little reliable positive data after filtering to obtain a good classifier for specifically positive interactions.

For very good historical reasons we used a quality filtered positive protein interaction set [133,159], since this set obtained better performance when

used with phylogenetic profiling. However, the quality filtering decreases the number of positive samples available to do machine learning with. It would be interesting to see if removing this quality filter can still give us a good performance, since removing this filter will give us a much larger positive interaction set to train with. On the other hand, this non-filtered interaction set will likely be contaminated with non-interacting proteins [141], which can negatively impact the training of the model.

The difficulty of separating the positive interactions from the negative interactions will influence how well a model can be fitted to the imbalanced data that we have here. In other words, if the interactions are easily separable from the non-interactions, imbalance does not matter [164]. Highly similar phylogenetic profiles for non-interacting proteins can happen for multiple reasons that have little to do with co-evolution, including general genome streamlining (function independent pressure of gene content). If the interacting and non-interacting protein profiles are highly overlapping, the number of correctly classified positive interactions will decrease significantly. Here we did not analyse how well the profiles of the interacting and non-interacting proteins could be separated beforehand. This means that it might not only be the absolute amount of positive protein interactions that are too few to obtain a good classifier, but also the separability of the classes that reduces the performance for predicting specifically positive interactions. Ideally, we would have done a more extensive hyperparameter search. For instance, other parameters besides the number of trees used in the Random Forest (the more trees the more accurate the classification) and the maximum depth of the trees. A hyper parameter sweep can be beneficial when creating a more accurate model for classification. On the other hand, Random Forests should already be good at classification and the default parameters should (initially) be a good start. Running even one algorithm on a large set like ours is very costly and time consuming. Moreover, hyperparameters searches will likely not be able to overcome the major hurdle of lack of positive data.

In this study the phylogenetic profiles of proteins are recoded into profiles for each class, the interacting or non-interacting protein class. The recoding transforms the presence and absences of proteins in multiple species into the presences and absences of (non-) interacting protein pairs in multiple species that can be used by the Random Forest for prediction. The way the recoding is done (Materials and Methods) has a possible

impact on the overall performance when predicting protein interactions. Preliminary results (not shown here) do suggest the current recoding of the profiles compared to other types of recoding do not change performance. However, a more comprehensive analysis into the impact of recoding should be considered.

In the end, experiments are laborious and costly. Computational methods are cost effective and provide a complementary way to infer interactions. Computational methods can also pick up interactions that are hard to detect by certain experimental methods, for instance certain categories of proteins and complexes that are not detected or difficult to purify [von Mering et al., 2002]. Experimental verification can always follow computational predictions. Machine learning and phylogenetic profiling are a way forward for protein interaction prediction and shows promise for further study.

MATERIALS AND METHODS

1. Initial datasets and methods

We started our analysis from the work done in [133,159]. Briefly, we inferred orthologous groups on a diverse set of 167 eukaryotes. We chose the best performing orthology inference method [133], Sonicparanoid (version 1.3.0.) [112]. We showed that filtering for orthologous groups inferred to be in the last common ancestor of eukaryotes (LECA) is beneficial to the performance of protein interaction prediction with phylogenetic profiles in eukaryotes. We therefore inferred LECA orthologous groups using the Dollo parsimony [114] method with a strict inclusion criterion [104]. The Dollo parsimony method assumes genes can only be gained once, while gene loss is minimized. The strict inclusion criteria assures that the orthologous group is in at least three supergroups, distributed over the Amorphae and Diaphoretickes (more or less corresponds to opimoda and diphoda) [116].

We determined the presences (1) and absences (0) of LECA orthologous groups in the 167 species to construct the phylogenetic profiles. We used the human BioGRID interaction database (version 3.5.172 May 2019) [117,148], which contains physical interactions between proteins. We obtained a higher quality non-redundant interaction set by filtering out interactions that were found in less than five independent studies (PubMed ID's). Our previous results showed that this increases prediction accuracy of protein

interactions [159]. We inferred a negative interaction set by taking protein pairs not found to be interacting with each other, but having at least five interactions [133].

2. Recoding of phylogenetic profiles for machine learning

We combine the phylogenetic profiles of interacting and non-interacting protein pairs by summation to get a single profile that is assigned the class "interacting" (1) or "non-interacting" (0). There is a directionality in the presence and absence of (non-)interactions. In other words, one species could have a protein A, but not a protein B. While another species could have a protein B, but not a protein A. We re-code the profiles before combining them to retain this information. For the re-coding a schematic representation is given in Table 1, with an example of each situation given by imaginary species X, Y, Z and Q.

Table 1. Recoding scheme phylogenetic profiles. Summing the profiles does not retain directionality of the presences and absences of the one or other interacting orthologs group (OG). Multiplying and summing retains the directionality. If both OGs are present: 3. If OGA is absent, but OGB is present: 2. If OGA is present, but OGB is absent: 1. If both OGs are absent: 0.

Scheme	**OG pairs**	Species X	Species Y	Species Z	Species Q
Profile1	OGA	1	0	1	0
Profile2	OGB	1	1	0	0
Summed	OGA-OGB	2	1	1	0
Profile 1	OGA	1	0	1	0
Profile 2 multiplied	OGB	2	2	0	0
Multiplied - Summed	OGA-OGB	3	2	1	0

3. Negative set purification using a two-step PU learning approach

The procedure for PU learning is followed from [151,153]. Briefly, the procedure starts with the Rocchio algorithm to extract reliable negatives from our original pseudo negative set. Rocchio performs well when classifying in the absence of a "pure negative" data set [151]. After the creation of reliable negatives, we use Random Forest for classification (RandomForestClassifier from the scikit-learn package version 0.22.2). The procedure is adjusted from [151,153]. We treat the entire unlabelled (original

negative) set as a negative set (*U*). We then use the positive set (*P*) and *U* as the training set to build a Rocchio classifier (NearestCentroid scikit-learn version 0.22.2 with cosine distance). The classifier is then used to classify reliable negatives (*RN*) from *U*. *P* and *RN* are then used to train the Random Forest classifier. In the original approach, the last classification step is done iteratively. The reason to run this iteratively is that a reliable negative set is not large enough to build the best classifier [151,158]. If the reliable negative set *RN* contains mostly negative documents and is sufficiently large, we will be able to build a good classifier without iteration [158]. The latter is true in our case.

To account for class imbalance, we redid the classification also with under sampled the data, the weighted and balanced Random Forest to compare to the "classic" Random Forest.

4. Testing and training sets

To create the test (30% of data) and training set (70% of data), we used a randomized sampling that preserves the percentage of data in each class when splitting the data (stratified sampling) using the stratify parameter in the train_test_split function (scikit-learn package version 0.22.2). This prevents the test set from having almost no positives at all due to the large class imbalance between positives and negatives. Moreover, this stratification will keep the data close to reality.

5. Imbalanced aware classification

For the random resampling of the negative set, we randomly sampled the training set without replacement (RandomUnderSampler from the imbalanced-learn package version 0.7.0). We took the following positive to negative ratios [1:1, 1:4, 1:8, 1:10, 1:25, 1:50, 1:100, 1:250, 1:500, 1:750, 1:1000]. For each of the rebalanced sets we classified with Random Forest. Since these rebalanced sets are considerably smaller, the cost of running multiple Random Forests is negligible.

Additional to the data resampling, to account for class imbalance we ran the weighted Random Forest and the balanced Random Forest. For the weighted Random Forest, we used RandomForestClassifier from the scikit-learn package version 0.22.2) with the class_weight set to "balanced". For the balanced Random Forest (BalancedRandomForestClassifier from the imbalanced-learn package version 0.7.0), we used default parameters.

6. Performance measures

Multiple performance metrics are available to evaluate how capable a model is to distinguish classes. The most common metrics are for instance accuracy, precision, recall, F-measures and AUC-ROC values. When we deal with highly imbalanced sets, the run-of-the-mill metrics might not suffice to give a true view of model performance. It can likely result in the choice of a poor model, or severely mislead about the model performance. The standard metrics treat classes in datasets equally. If we are interested in the smaller class, the smaller class is more important. However, we also are interested in differentiating between positives and negatives if we use phylogenetic profiling in this exploratory analysis. Therefore, the performance metrics need to capture this.

Accuracy is impractical for the use in imbalanced sets, because even if it misses all the smaller class (minority) samples, accuracy can still be very high [155]. This is because accuracy takes the ratio of correct predictions (true positives and true negatives) to the total number of predictions. If true negatives are high, but true positives are very low (or even zero), accuracy is high. The balanced accuracy is better suited for highly imbalanced sets. It takes the average of recall obtained in each class (the average of the true positive rate and true negative rate). However, when a classifier is able to retrieve many negatives, the balanced accuracy will also become higher. If the objective is to identify positives, this measure might also not be the best choice. It will nevertheless give a better indication than accuracy. We will use the balanced accuracy readily available in the scikit-learn package (version 0.22.0).

Precision is not affected by a large negative set and is more focused on the positive set (Precision = true positives / (true positives + false positives)). In other words, it calculates how accurately the smaller class is predicted. Recall (true positive rate) is also more focused on the smaller class and gives an indication of how much of the minority class is "covered" by the model (Recall = true positives / (true positives + false negatives)). Precision and recall can also be plotted in a precision recall curve, giving a diagnostic of the performance for a classifier on the smaller positive interactions class. The scikit-learn package (version 0.22.0) provides all these metrics readily.

Precision and recall can be combined into one measure, the F-measure. If the F-measure is 1, the model has perfect recall and precision. While if the F-measure is 0, either precision or recall are 0. The F-measure is often used in imbalanced classification if only the performance of the positive class is important. If the performance of both classes is important, the G-mean is a better performance measure as it takes both the true positive rate and true negative rate. A high G-mean, means both classes are recognized, while a value of 0 means that at least one class is not recognized [ref].

The receiver operator curve can give us an idea about how well a model is at separating the positives from the negatives. It plots the true positive rate (recall) against the false positive rate. You can see this as a fraction of the correctly predicted positives versus the fraction of incorrectly predicted negatives. With regards to model performance, the receiver operator curve can be too optimistic in a highly unbalanced set. It does give a better indication if we care about both positive and negative classes (or separation thereof) compared to only the precision recall curve.

Lastly, the macro-average F-measure is also suggested for imbalanced data [155], and gives the same importance for all the classes. This measure will be low for models that only perform well on the majority classes, while performing poorly on the minority classes.

7. Cross-validation and hyper parameter search

To reduce possible overfitting while training the models, we did five-fold cross validation together with a hyper parameter sweep. Moreover, it is suggested that hyperparameter tuning can eliminate the effect of class imbalance [155]. We used a stratified five-fold cross validation to maintain the class distributions. This means that each fold (subset) will have and equal class distribution as with the original set. We used stratified sampling since "traditional" cross-validation sampling breaks down in highly unbalanced datasets. Traditional cross-validation sampling uses a random sampling of the classes for each fold (data subset). This random sampling breaks down because there likely will be folds that do not contain samples of the smaller class.

For the hyper parameter search, we varied the number of estimators (decision trees in Random Forests) from [10, 50, 100, 200, 250], and varied the maximum tree depth from [None, 5, 10, 20, 30]. We used

StratifiedShuffleSplit together with GridSearchCV from scikit-learn package version 0.22.2. We then picked the classifiers that showed to be the optimal using a hyperparameter grid search with five-fold stratified cross validation on the training data and selection based on the macro-average F-measure. We use the macro-average F-measure which gives an equal importance to both classes.

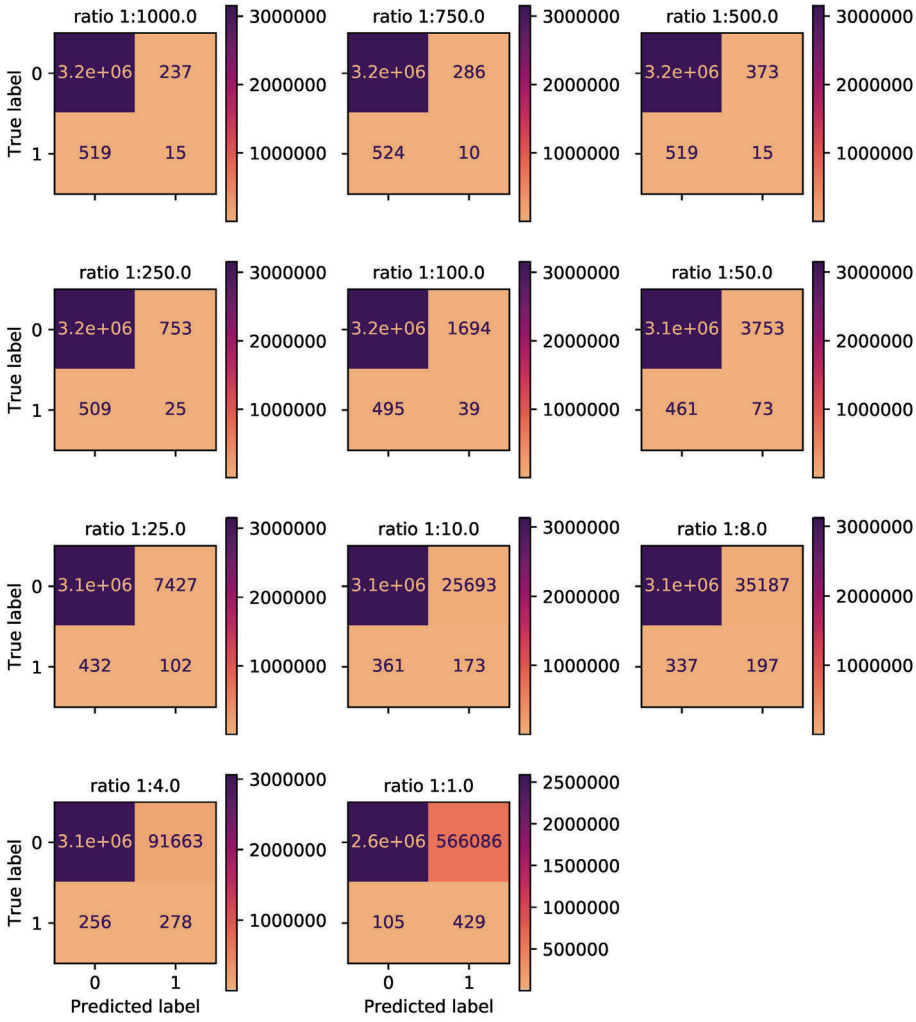
8. Feature selection

Feature selection is done for dimension reduction and therefore computational cost, but can also be done to increase performance. Random Forest can be used together with feature selection (SelectFromModel from the scikit-learn package version 0.22.2). Features are selected if their importance is greater than the mean importance of all the features.

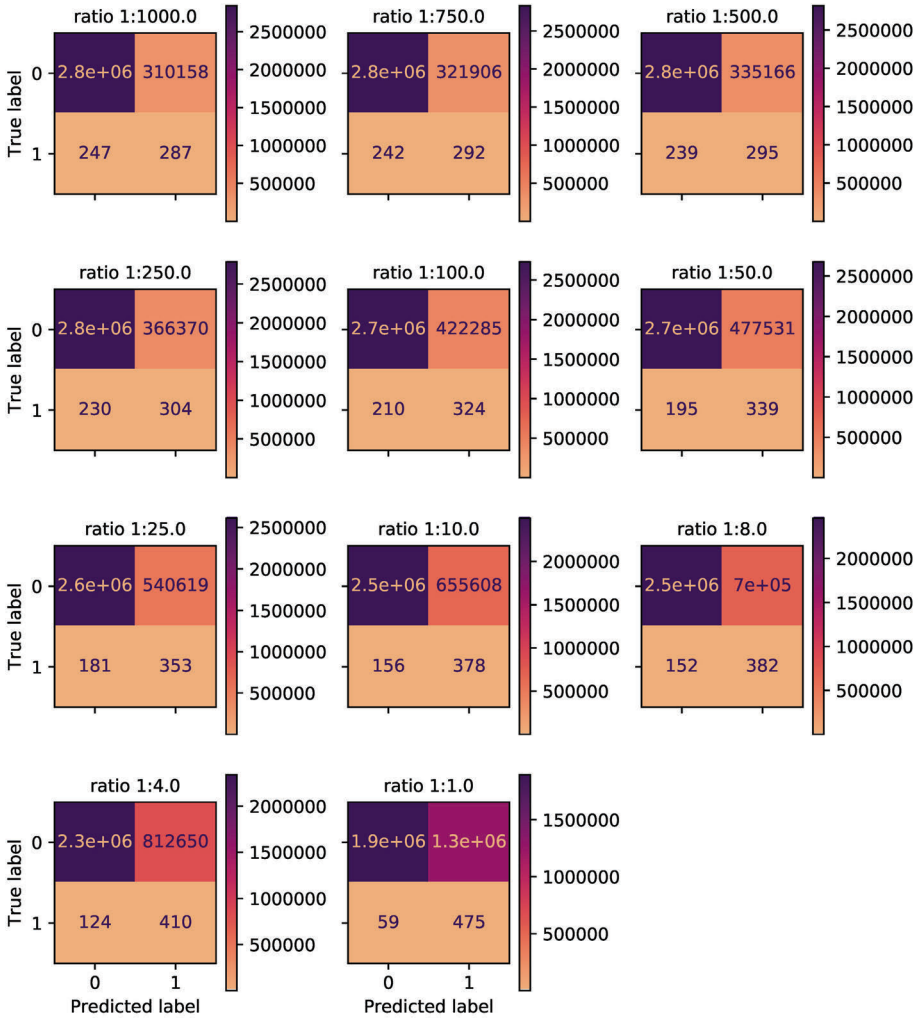
ACKNOWLEDGMENTS

We want to thank Laura van Rooijen for input given during the early stages of this chapter.

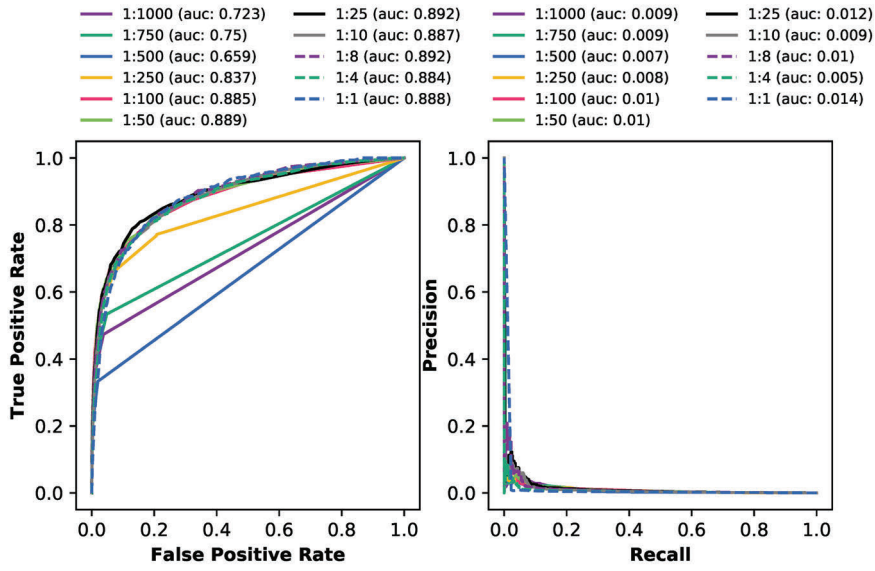
SUPPLEMENTARY FIGURES



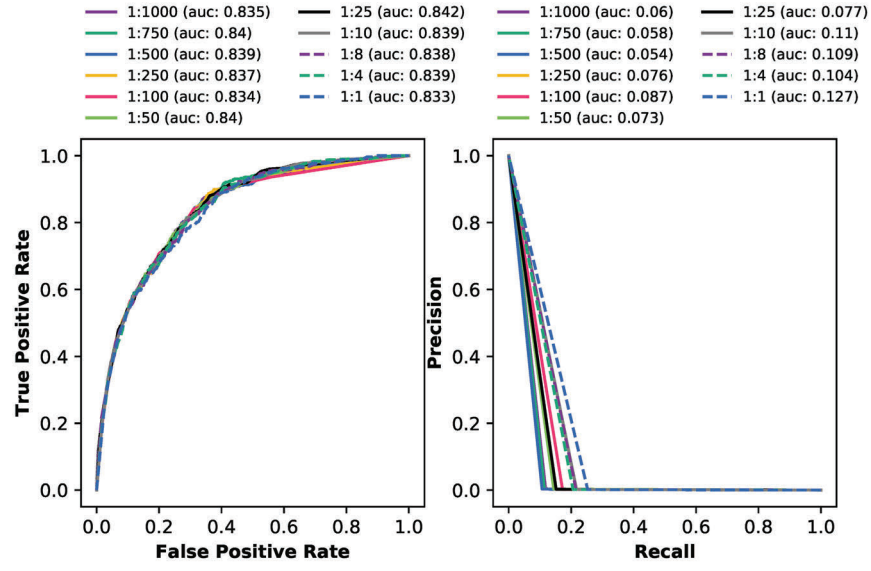
Supplementary Figure 1. Confusion Matrices for Random Forest with different positive to negative ratios and the original negative set.



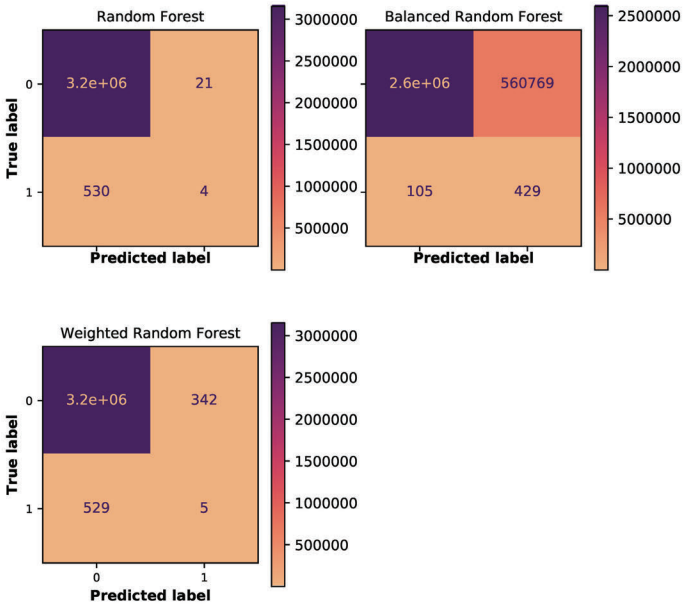
Supplementary Figure 2. Confusion Matrices for Random Forest with different positive to negative ratios and the reliable (Rochhio classified) negative set.



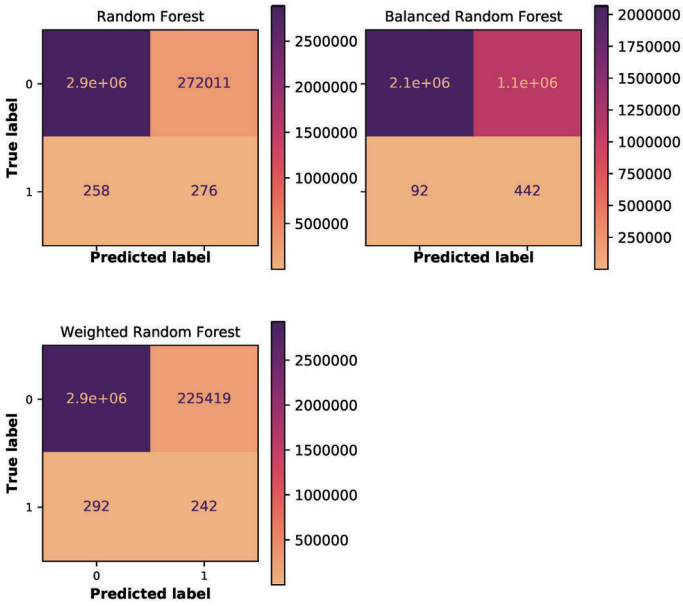
Supplementary Figure 3. Receiver Operator Curve (left) and Precision Recall Curve (right) for the undersampling with Random Forest approach using the original negative set.



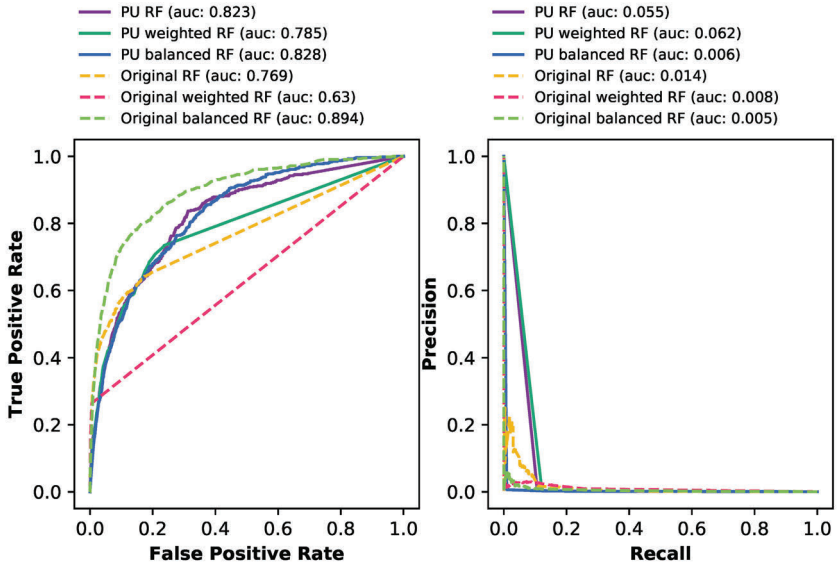
Supplementary Figure 4. Receiver Operator Curve (left) and Precision Recall Curve (right) for the undersampling with Random Forest approach using the reliable negative set.



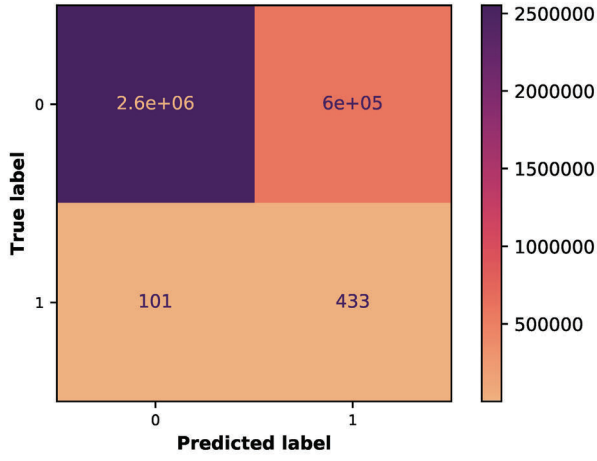
Supplementary Figure 5. Confusion Matrices for Random Forest, weighted Random Forest and balanced random forest the original negative set.



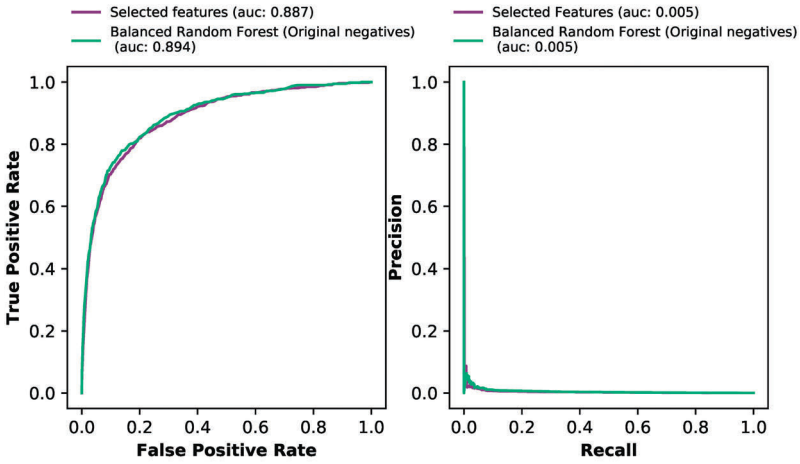
Supplementary Figure 6. Confusion Matrices for Random Forest, weighted Random Forest and balanced random forest the reliable negative set.



Supplementary Figure 7. Receiver Operator Curve (left) and Precision Recall Curve (right) for the "classic" Random Forest, weighted Random Forest and balanced Random Forest with the original (dashed lines) and reliable (solid lines).



Supplementary Figure 8. Confusion matrix for balanced Random Forest with best performing features selected.



Supplementary Figure 9. Receiver Operator curve (left) and Precision Recall Curve for balanced Random Forest with and without best performing features selected.

SUPPLEMENTARY TABLES

Supplementary Table 1. The number of interacting and non-interacting proteins in original, training and test sets.

Classes	Original set	Training set (70%)	Test set (30%)
Original Negatives	10505860	7354101	3151759
Reliable Negatives (Rocchio)	-	4774881	3151759
Positives	1781	1247	534

CHAPTER 6.

General Discussion

THE OLD, THE NEW, AND THE REMAINING

Finding protein-coding genes and other functional elements in a genome is one of the essential parts of genome annotation [17,165]. Errors made in annotation can propagate in downstream analyses of eukaryotic genome evolution. One such downstream analysis we studied in **Chapter 2**. Chapter 2 set out to find falsely inferred absences due to protein prediction errors and how this influences eukaryotic gene loss estimates. We showed that inferring absences from predicted proteomes on average is reliable.

Most importantly, we showed that we should be cautious towards suspicious species-specific absences that are not reinforced by absences in sister species. Species-specific absences show a higher rate of being falsely inferred, leading to increased loss estimates. Our results also showed that referring back to the DNA is a simple but effective approach to see if suspicious absences are really absent. There are multiple reasons why protein prediction fails, and as a consequence, we find false protein absences back in the DNA. Protein prediction is inherently challenging and can be incomplete due to multiple technical and biological reasons, especially in eukaryotes with complex genome architecture. Speculating on and looking into why protein prediction failed in these cases is a paper on itself. Briefly, the reasons protein prediction can fail include (small) sequencing errors, being (too) stringent in inferring which sequences are actual proteins (avoiding false positives that leads to false negatives), lack of RNAseq data, lack of long-read sequencing data, and lack of manual curation.

We are obviously not the only ones running into protein prediction issues and trying to contextualize the problem. A recent case study [51] examined the cause of eukaryotic protein-coding gene prediction errors within primate genomes. The study focused on errors that generally arise when exons are translated in the wrong reading frame, or exon boundaries are poorly predicted and cause non-coding nucleotide sequences to be translated. Almost half of the errors characterised in this study resulted from undetermined regions (N's) in the genome. These undetermined regions cause misprediction of exon-intron structures. Another large part of the errors resulted from abnormal intron lengths in the (other) primates (introns less than 30 nucleotides in length were found). A choice often made with human genes is to regard only introns with more than 30

nucleotides or more. The smaller intron lengths in other primates do not make the cut to be regarded as an intron and are wrongfully not processed as an intron. A smaller number of errors could be attributed to small insertions/deletions causing a frameshift and non-canonical splice sites.

Another noteworthy example, and inverse Chapter 2, is given by a study [166] that looked at falsely inferred duplications. These falsely inferred duplications primarily occur due to increased sequence divergence between paternal and maternal haplotypes, leading to assembly algorithms classifying them as separate genes. A minor source was (accumulated) sequence errors causing under-collapsed sequences that classify as separate genes. Although this study shows the problems regarding genome assembly, it highlights another source of error for genome evolutionary studies and (homology-based) functional annotation. These problems include errors in gene gain and gene family expansion estimates. They also decrease one-to-one orthologs that are key in many comparative genomic and phylogeny studies. Sequence divergence can also cause the wrong inferences of the evolutionary history of genes due to homology detection failures, which was another recent and interesting finding [167]. These detection failures increase the number of lineage-specific genes, or novel genes, due to computational similarity searches failing to pick up distant homologs.

Be it assembly and gene prediction errors, or sequence similarity detection failure, these errors propagate into downstream analyses, through databases and homology-based annotation pipelines. It highlights that with more data comes more responsibility and an urgent need for quality control, re-analysis and error detection. More genomes are sequenced, and sequencing and assembly technologies are improving, but genome annotation has become less accurate [16]. The decrease in annotation accuracy in part has to do with automated annotation needed for large (number of) genomes. The quality of annotations is often affected directly by the input genome quality [165]. Issues such as assembly errors and contamination lead to more errors in annotation. The more draft genomes published, the more errors arise and propagate across species. When such an error is found and corrected, the error often remains in other annotations that relied on the incorrect annotation. We should also take care to not blindly assume the output of any automated annotation method as truth, such as protein interaction prediction or homology detection.

International consortia such as the Vertebrate Genomes Project strive towards complete and error-free genome assemblies for all of ± 70000 vertebrate species and adopt assembly error handling [168]. Once we can minimize errors in sequence assemblies, we can increase the accuracy of annotations and improve existing annotation. Re-annotation is crucial for correcting the misannotations already in public databases [16,168]. Not only will re-annotation correct errors, but it will also open the path to discovering new genes or protein functions, comparing new and existing annotation methods, or asses annotation reproducibility. Automated annotation can save time and resources, but manual annotation is still the gold standard. Comparing multiple automated annotations side-by-side and aggregating them with manual annotation or curation can reduce errors and their propagation.

The PATRIC database and analysis resource is one such integrated database and supports research on bacterial infectious diseases. The repository integrates a variety of data types, such as genomics, transcriptomics, protein interactions, 3D protein structures and sequence data, from a variety of sources [169]. The data integrated into PATRIC is manually curated. PATRIC also provides multiple analysis tools to compare and analyse the data. Initiatives such as PATRIC that shift the focus to database integration and on-site analyses, rather than only provide a repository like structure, could help assist in annotation and improved data accuracy for eukaryotes.

In **Chapter 3,** we evaluated automated orthology inference methods in their ability to retrieve manually curated orthologous groups and how they can recapitulate several observations made regarding eukaryotic genome evolution. More specifically, LECA gene content, gene loss patterns and numbers, and co-occurrence of interacting proteins. There was no clear indication of one method performing better than any other. All the methods are able to find the large scale genomic evolutionary trends that are generally reported in eukaryotic genome evolution. Counterintuitively, while having similar behaviour for general patterns of eukaryotic genome evolution, the methods produce orthologous groups that are vastly different.

That orthologous groups inferred by different methods that are conceptually and practically different might be logical considering the difficulty of

identifying orthologous genes in the first place. The identification of eukaryotic-wide orthologous genes is challenging due to billions of years of dynamic genomes. Particularly for plants or other organisms where whole-genome duplications are rampant, inferring orthologous groups is even more complicated. Issues that likely cause the differences between orthologous groups include the previously mentioned technical and biological intricacies including, wrongfully excluded genes from an orthologous group due to undetected sequence similarity [47,167] caused by (faster) sequence evolution, or a too stringent cut-off value for orthologous group assignment. Orthologous groups can be fragmented into multiple groups due to lineage-specific duplications, or protein domain dynamics such as gains, losses and shuffling (Chapter 1, Box 1).

Chapter 3 shows us that automatic and manual orthology inference should be complementary rather than competitive. To improve comparative genomics further, (orthology) databases should aim to aggregate automatically inferred and manually inferred orthologous group assignments in a convenient way. This aggregation of data can accommodate small scale manual analyses or seed automated orthology methods. Moreover, readily available manually curated orthologues groups can help guide the further development of orthology inference tools.

Another important finding from Chapter 3 was that we obtained a surprisingly high performance when we predicted interacting orthologous groups using phylogenetic profiles. Getting a good performance for predicting protein interactions using phylogenetic profiles is often reserved for prokaryotes [129–132], with only a few success stories in eukaryotes [124,125]. In **Chapter 4,** we wanted to understand why we could get good performance by analysing the choices made in Chapter 3, in other words, choices in meta parameters. We found an effect on performance when using only genomes that are of the worst quality. However, it is unlikely anyone would pick only the worst quality genomes for any phylogenetic profiling analysis. The effect on performance is negligible when using only genomes of the highest quality, compared to using a large and diverse set of eukaryotes with both high and lesser quality genomes. It is instead the information within the genomes that has more of an effect on overall performance. This observation suggests that we should look into more complex species selection procedures to investigate non-linear interactions between subsets of genomes or combinations of subsets, which we can

do with feature selection and machine learning. We explore this later in Chapter 5.

Unexpected but evident in Chapter 4 is that genome selection is not the limiting factor in phylogenetic profiling (anymore). Multiple diverse genomes are available, which leaves room to select between diversity, quality and information in the genomes. We noticed the real difference when we used orthologous groups inferred to be in the last eukaryotic common ancestor. Filtering out profiles that are less dan 50% consistent due to lineage-specific interactions improved phylogenetic profiling in eukaryotes. Also interesting is the effect of filtering out lesser quality interaction data, i.e., lesser quality annotations. This filtering shows again how important good quality annotation is and the downstream impact of lesser quality annotations.

In **Chapter 5**, we wanted to see if we could improve phylogenetic profiling in eukaryotes further and analyse the non-linear effects of genomes and genome sets that we could not capture with "classic" profiling. The initial challenge when we wanted to apply machine learning to our data was the large difference between the number of positive and negative interactions. However, our machine learning explorations suggest that the low abundance of unambiguous positive interactions likely is the biggest problem for our data.

Nevertheless, we showed that machine learning is a promising direction to improve protein interaction prediction with phylogenetic profiles. In Chapter 5, we have shown multiple directions to take for improvement. However, we also saw in Chapter 5 the intricacies and difficulties of applying machine learning to our data, confining the study to preliminary evaluations of techniques and methods that might be suitable for phylogenetic profiling and protein interaction prediction. There are many considerations to take into account.

Although not elaborately compared in Chapter 5, one such consideration is transforming phylogenetic profiles of individual (non-) interacting proteins into class vectors that show the presence and absence of (non-) interacting protein pairs. This recoding can be done in different ways and we have now used only one. Another consideration to take into account is the class imbalance of interacting protein pairs compared to non-interacting protein

pairs. Since there are, and always will be, fewer interacting protein pairs than there are non-interacting protein pairs, dealing with this imbalance is an important issue that is often the subject of study, not only for protein interactions but many biological or real-world problems.

Chapter 5 gave us this first glance into the use of machine learning with phylogenetic profiling for protein interaction prediction in eukaryotes. We think there is more information hidden in the presences and absences of genes across the phylogenetic tree than is captured by simple distances measured between phylogenetic profiles. Machine learning should be able to retrieve this information, much like the deep learning approach AlphaFold has successfully tackled the protein folding problem and retrieved protein structures from sequence alignments [170]. AlphaFold has recently been modified to enable protein complex modelling and already showed good potential for identifying protein interactions and protein modular membership by looking at the co-evolution of amino acid sequences at the protein-protein interaction interface (contact region) [171]. These results suggest that directly inferring protein interactions from sequence alignments might make the use of phylogenetic profiles for that purpose unnecessary. Nevertheless, using AlphaFold for protein interaction prediction is also hampered by the complexity of eukaryotic genome evolution. For instance, the choice of what sequences to use in the case of multiple hits. Particularly for higher eukaryotes, the approach with AlphaFold has still to be proven [171]. For now, both the use of phylogenetic profiling and AlphaFold to infer protein interactions has room left for further development and give an exciting glimpse of possible future advancements in protein-protein interaction prediction.

BACK TO THE FUTURE (PERSPECTIVES)

In this thesis, we have shown that we can improve upon the general inferences made regarding eukaryotic genome evolution when we consider and take into account complex biological and technical challenges. It is clear that multiple approaches and methods in eukaryotic genome evolution and comparative genomics still suffer from very complex issues. It often starts with incomplete or incorrect genome assemblies. The more draft genomes that are not thoroughly quality checked and still published, the more errors we will create that will propagate rather than be removed. Protein prediction directly suffers from assembly errors. Newly developed

methods, like direct RNA sequencing (e.g., nanopore technologies), might provide a possibility to improve genome annotation and help guide re-annotation drastically. Direct RNA sequencing captures RNA directly in long reads without converting RNA first to complementary DNA (cDNA) and amplifying it with PCR that introduces bias [172].

Regardless of the errors in assembly and annotation, the amounts of sequenced genomes have paved the way for (large scale) comparative genomics and allowed an increase in understanding of eukaryotic genome evolution. The knowledge about, for instance, the prevalence of gene loss and a gene-rich eukaryotic ancestor has only been possible because of the increase in available and diverse genomes. In the end, it will be hard to achieve complete error-free assembly and annotation, only awareness and responsibility of anyone embarking on comparative genomics ventures. Arbitrary choices or assumptions that seem innocent with certain analyses often are not innocent at all.

The ultimate issue for automatic orthology inference is that eukaryotic genomes are highly complex, with complex evolutionary dynamics such as horizontal gene transfer, domain dynamics (e.g., duplications, shuffling), and whole-genome duplications spanning over billions of years. Genome evolution may be too intricate to be fully captured by any automatic orthology inference method. Depending on the inferences we want to make, we should decide between automated, manual or a combination of both. Automated methods can inform manual analyses by showing which genes follow expected and clear patterns, and which genes do not. Vice versa, manually checking a few examples from a large-scale study can give a sense of whether automated methods on average produce reasonable results. Nevertheless, there are still ways to improve on these automatic methods. For instance, using the conservation of local gene order (synteny) to help identify orthologs (shortly reviewed here [173]). Using synteny might not be decisive for orthologous group identification in distantly related species because gene order in eukaryotes evolves faster than gene content and sequences. However, synteny can help confidently assign orthologous genes and complement orthology inference methods for closely related species or short evolutionary distances. Other automated methods are also being developed that are protein domain-oriented, to help detect orthologous relationships between proteins that would otherwise be missed due to their domain dynamics [174].

The complexity of orthology prediction trickles down into phylogenetic profiling with eukaryotes. Is automated orthology inference just not "there" yet, and does the lack of high-quality orthology limit the performance of current large scale phylogenetic profile studies by capping the maximum performance obtainable in eukaryotes? Although we looked into the performance of different orthology methods, this metaparameter we could not evaluate since all orthology methods likely fall short of perfect performance. Investigating this question is also hampered by the fact that we likely lack enough good quality interaction data. Therefore, it is hard to decide which factors are limiting at this point and it is likely a combination of all the above. If we make the right choices in hyperparameters beforehand, there is a clear signal when predicting interacting proteins using co-occurrence that we should further explore.

Phylogenetic profiling itself has room left for advancement. For instance, at the heart of phylogenetic profiling is the similarity between profiles. Not only similarity of profiles is the result of evolutionary processes, but dissimilar (discordant) profiles are as well. Similar profiles indicate proteins co-occur or are co-lost. Discordant profiles might indicate functional redundancy (Chapter 1, Figure 2). Functional redundancy arises from lineage-specific duplications (in-paralogs) that changed their function, multifunctional proteins, proteins belonging to multiple protein complexes [175]. Discordant profiles can also indicate analogous proteins that evolved the same function but do not share a common ancestor. Discordant profiles are not taken into account in the studies presented in this thesis, but can in the future provide exciting knowledge of protein (function) evolution. Moreover, despite it being in the name, the phylogenetic profiling method used in the studies presented here do not account for phylogeny. In other words, phylogenetic profiles are not ordered or do not weigh in closely or distantly related species. Nevertheless, the signal that can be found with phylogenetic profiles is strong.

It seems logical that machine learning will be the next step in many biological research areas due to the large amount of data that are becoming available. Nevertheless, as Chapter 5 already touched upon, there are still obstacles to overcome for many real-life situations. Lately, machine learning seems like the holy grail when it comes to complex biological issues. However, to be fully utilised, there still has to be enough quality data. For protein interaction prediction, this is not only the availability of

more protein interaction data. This also includes other types of data that can be used in the feature space, for instance, compartmental and genomic (co-)localization and co-transcription of interacting proteins. However, such aggregated databases where this type of data is easily retrievable do not yet exist for eukaryotic species.

CONCLUDING REMARKS

While it was data we needed in the past, now it is using the data we have and doing that properly. Re-annotation, re-analysis, error correction, quality checks, and data aggregation should become a central subject of this time. This sentiment is also shared by multiple European infrastructure projects that are for instance combining and coordinating recourses and data, and sharing best practices and tools, such as the EMBL-EBI [176] and Elixer [177]. Combining and interconnecting data and resources will enhance biological data discovery, accessibility, and reusability [176]. With the aggregation of, for instance manual and automated methods, we can increase the accuracy and quality of orthology inference and other types of annotation data. Many might ask at this point, what about machine learning? In Chapter 5 I showed that we likely have too little data to get a good predictive model to infer interacting proteins, at least with phylogenetic profiling. Now I am advocating to use the data we have, instead of producing lots and lots of it.

Having more data surely increases the accuracy of any model. Nevertheless, there comes a stage where adding copious amounts of data cannot improve a model or its accuracy. We should also remember the "garbage in, garbage out" principle. If there is no quality data or data is too noisy, any machine learning model will be rendered useless. Nevertheless, so will having too little data. New innovative approaches are popping up to build better models with less or unlabelled data [178], or using neural networks that overall do well with fewer data [179]. Moreover, if more data is aggregated from different resources the feature space for machine learning becomes bigger as well.

In the end, biology is just really complex. We need to stay aware of all the issues with using certain tools and concepts. Many tools and concepts are imperfect, as we have seen, regardless of their usefulness. We also need to realise that producing more data is not always better, nor should

we strive to "just" produce more data. Instead, we would benefit from data production in a smart way to fill the gaps in our knowledge, for reannotation and aggregation. Another issue that fuels the uncontrolled growth of lesser-quality data and annotation errors is publication pressure, or "publish or perish". In the end, we should strive for quality and not just quantity. Adjusting our focus on quality and not quantity (of data) will also alleviate the burdens large data creates, or the data deluge.

The studies described in this thesis showed that on a large scale we can reliably draw large scale conclusions about eukaryotic genome evolution with automated methods. We also saw that we should take care if we draw small scale conclusions with automated methods. Automated and manual inferences will still go hand in hand. With the studies described here, we have also seen that the evolutionary patterns of presences and absences of genes in eukaryotes (phylogenetic profiles) are a reliable and effective technique to infer functional relationships between proteins, warranting further studies into this topic.

Appendices

REFERENCES

1. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87: 4576–4579. doi:10.1073/pnas.87.12.4576
2. Glansdorff N, Xu Y, Labedan B. The Last Universal Common Ancestor: Emergence, constitution and genetic legacy of an elusive forerunner. Biology Direct. 2008. doi:10.1186/1745-6150-3-29
3. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature. 2020;577: 519–525. doi:10.1038/s41586-019-1916-6
4. Baldwin IT, Schultz JC. Rapid Changes in Tree Leaf Chemistry Induced by Damage: Evidence for Communication Between Plants. Science (80-). 1983;221: 277–279. doi:10.1126/science.221.4607.277
5. Simard SW. Mycorrhizal Networks Facilitate Tree Communication, Learning, and Memory. In: Baluska F, Gagliano M, Witzany G, editors. Memory and Learning in Plants Signaling and Communication in Plants. Cham: Springer International Publishing; 2018. pp. 191–213. doi:10.1007/978-3-319-75596-0_10
6. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A. 2011;108: 13624–13629. doi:10.1073/pnas.1110633108
7. O'Malley MA, Leger MM, Wideman JG, Ruiz-Trillo I. Concepts of the last eukaryotic common ancestor. Nat Ecol Evol. 2019;3: 338–344. doi:10.1038/s41559-019-0796-3
8. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. Archaea and the origin of eukaryotes. Nat Rev Microbiol. 2017;15: 711–723. doi:10.1038/nrmicro.2017.133
9. Pillai AS, Chandler SA, Liu Y, Signore A V., Cortez-Romero CR, Benesch JLP, et al. Origin of complexity in haemoglobin evolution. Nature. 2020;581: 480–485. doi:10.1038/s41586-020-2292-y
10. Van Holde KE, Miller KI, Decker H. Hemocyanins and Invertebrate Evolution. J Biol Chem. 2001;276: 15563–15566. doi:10.1074/jbc.R100010200
11. Sidell BD, O'Brien KM. When bad things happen to good fish: The loss of hemoglobin and myoglobin expression in Antarctic icefishes. J Exp Biol. 2006;209: 1791–1802. doi:10.1242/jeb.02091
12. Goffeau AA, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 Genes. 1996;274: 546–567.
13. Nagy LG, Merényi Z, Hegedüs B, Bálint B. Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. Nucleic Acids Res. 2020;48: 2209–2219. doi:10.1093/nar/gkz1241
14. Bork P, Koonin E V. Predicting functions from protein sequences -sequences - where are the bottlenecks ? Nat Genet. 1998;18: 313–318.

15. Brenner SE. Errors in genome annotation. Trends Genet. 1999;15: 132–133. doi:10.1016/S0168-9525(99)01706-0

16. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20: 92. doi:10.1186/s13059-019-1715-2

17. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13: 329–342. doi:10.1038/nrg3174

18. Snel B. Comparative Genome Analysis and Genome Evolution. Utrecht University. 2002. Available: https://www.cell.com/trends/genetics/fulltext/S0168-9525(99)01924-1

19. King JL, Jukes TH. Non-Darwinian evolution. Science. 1969. pp. 788–798. doi:10.1126/science.164.3881.788

20. Gabaldón T, Koonin E V. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013;14: 360–366. doi:10.1038/nrg3456

21. Chen X, Zhang J. The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002784

22. Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Zool. 1970;19: 99. doi:10.2307/2412448

23. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci. 2018;115: 4325 LP – 4333. doi:10.1073/pnas.1720115115

24. Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, et al. Data access for the 1,000 Plants (1KP) project. Gigascience. 2014;3. doi:10.1186/2047-217X-3-17

25. Burki F, Roger AJ, Brown MW, Simpson AGB. The New Tree of Eukaryotes. Trends Ecol Evol. 2020;35: 43–55. doi:10.1016/j.tree.2019.08.008

26. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, et al. Comparative genomics of the eukaryotes. Science (80-). 2000;287: 2204–2215. doi:10.1126/science.287.5461.2204

27. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Res. 2000;10: 950–958. doi:10.1101/gr.10.7.950

28. van Hooff JJ. Origins and divergence of the eukaryotic kinetochore. Utrecht University. Utrecht University. 2018.

29. Van De Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009;10: 725–732. doi:10.1038/nrg2600

30. Wolf YI, Koonin E V. Genome reduction as the dominant mode of evolution. BioEssays. 2013;35: 829–837. doi:10.1002/bies.201300037

31. Albalat R, Cañestro C. Evolution by gene loss. Nat Rev Genet. 2016;17: 379–391. doi:10.1038/nrg.2016.39

32. Sémon M, Wolfe KH. Consequences of genome duplication. Curr Opin Genet Dev. 2007;17: 505–512. doi:10.1016/j.gde.2007.09.007

33. Cuypers TD, Hogeweg P. Virtual genomes in flux: An interplay of neutrality and adaptability explains genome expansion and streamlining. Genome Biol Evol. 2012;4: 212–229. doi:10.1093/gbe/evr141

34. van Hooff JJ, Tromer E, van Wijk LM, Snel B, Kops GJ. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep. 2017;18: 1559–1571. doi:10.15252/embr.201744102

35. Tromer EC, van Hooff JJE, Kops GJPL, Snel B. Mosaic origin of the eukaryotic kinetochore. Proc Natl Acad Sci. 2019;116: 12873–12882. doi:10.1073/PNAS.1821945116

36. van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, et al. Evolution of modular intraflagellar transport from a coatomer-like progenitor. Proc Natl Acad Sci. 2013;110: 6943–6948. doi:10.1073/pnas.1221011110

37. Elias M, Brighouse A, Gabernet-Castello C, Field MC, Dacks JB. Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. J Cell Sci. 2012;125: 2500–2508. doi:10.1242/jcs.101378

38. Zmasek CM, Godzik A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol. 2011;12: R4. doi:10.1186/gb-2011-12-1-r4

39. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al. Endosymbiotic origin and differential loss of eukaryotic genes. Nature. 2015;524: 427–432. doi:10.1038/nature14963

40. Sherill-Rofe D, Rahat D, Findlay S, Mellul A, Guberman I, Braun M, et al. Mapping global and local coevolution across 600 species to identify novel homologous recombination repair genes. Genome Res. 2019;29: 439–448. doi:10.1101/gr.241414.118

41. Moi D, Kilchoer L, Aguilar PS, Dessimoz C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. Ouzounis CA, editor. PLOS Comput Biol. 2020;16: e1007553. doi:10.1371/journal.pcbi.1007553

42. Kensche PR, Van Noort V, Dutilh BE, Huynen MA. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface. 2008;5: 151–170. doi:10.1098/rsif.2007.1047

43. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999;96: 4285–4288. doi:10.1073/pnas.96.8.4285

44. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nat Genet. 1999;21: 108–110. doi:10.1038/5052

45. Korbel JO, Snel B, Huynen MA, Bork P. SHOT: A web server for the construction of genome phylogenies. Trends Genet. 2002;18: 158–162. doi:10.1016/S0168-9525(01)02597-5

46. van Hooff JJE, Tromer E, van Dam TJP, Kops GJPL, Snel B. Inferring the Evolutionary History of Your Favorite Protein: A Guide for Molecular Biologists. BioEssays. 2019;41: 1900006. doi:10.1002/bies.201900006

47. Natsidis P, Kapli P, Schiffer PH, Telford MJ. Systematic errors in orthology inference and their effects on evolutionary analyses. iScience. 2021;24: 102110. doi:10.1016/j.isci.2021.102110

48. Andersson JO, Hirt RP, Foster PG, Roger AJ. Evolution of four gene families with patchy phylogenetic distributions: Influx of genes into protist genomes. BMC Evol Biol. 2006;6: 1–18. doi:10.1186/1471-2148-6-27

49. Martin WF. Too Much Eukaryote LGT. BioEssays. 2017;39: 1–5. doi:10.1002/bies.201700115

50. Leger MM, Eme L, Stairs CW, Roger AJ. Demystifying Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI: 10.1002/bies.201700115). BioEssays. 2018;40: 1700242. doi:10.1002/bies.201700242

51. Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. BMC Bioinformatics. 2020;21: 1–16. doi:10.1186/s12859-020-03855-1

52. Szklarczyk R, Wanschers BFJ, Cuypers TD, Esseling JJ, Riemersma M, Van Den Brand MAM, et al. Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome c oxidase. Genome Biol. 2012;13. doi:10.1186/gb-2012-13-2-r12

53. Altenhoff AM, Glover NM, Dessimoz C. Inferring Orthology and Paralogy. In: Anisimova M, editor. Evolutionary Genomics: Statistical and Computational Methods. New York, NY: Springer New York; 2019. pp. 149–175. doi:10.1007/978-1-4939-9074-0_5

54. Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, et al. Standardized benchmarking in the quest for orthologs. Nat Methods. 2016;13: 425–430. doi:10.1038/nmeth.3830

55. Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernández-Plaza A, et al. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Res. 2020;48: W538–W545. doi:10.1093/nar/gkaa308

56. Emms DM, Kelly S. Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench. Genome Biol Evol. 2020;12: 2258–2266. doi:10.1093/gbe/evaa211

57. Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, et al. TreeFam: 2008 Update. Nucleic Acids Res. 2008;36: D735–40. doi:10.1093/nar/gkm1005

58. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44: D286–D293. doi:10.1093/nar/gkv1248

59. Huang X, Albou LP, Mushayahama T, Muruganujan A, Tang H, Thomas PD. Ancestral Genomes: A resource for reconstructed ancestral genes and genomes across the tree of life. Nucleic Acids Res. 2019;47: D271–D279. doi:10.1093/nar/gky1009

60. de Wolf B, Oghabian A, Akinyi M V, Hanks S, Tromer EC, Hooff JJE, et al. Chromosomal instability by mutations in the novel minor spliceosome component CENATAC. EMBO J. 2021;40: 1–18. doi:10.15252/embj.2020106536

61. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7. doi:10.1371/journal.pcbi.1002073

62. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002514

63. Stamboulian M, Guerrero RF, Hahn MW, Radivojac P. The ortholog conjecture revisited: The value of orthologs and paralogs in function prediction. Bioinformatics. 2020;36: I219–I226. doi:10.1093/BIOINFORMATICS/BTAA468

64. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Comput Biol. 2014;10: e1003998. doi:10.1371/journal.pcbi.1003998

65. Wang Z, Pascual-Anaya J, Zadissa A, Li W, Niimura Y, Huang Z, et al. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. Nat Genet. 2013;45: 701–706. doi:10.1038/ng.2615

66. Lovell P V., Wirthlin M, Wilhelm L, Minx P, Lazar NH, Carbone L, et al. Conserved syntenic clusters of protein coding genes are missing in birds. Genome Biol. 2014;15: 565. doi:10.1186/s13059-014-0565-1

67. Larhammar D, Lagman D. Turtle ghrelin. Nat Genet. 2014;46: 526–526. doi:10.1038/ng.2988

68. Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D. Hidden genes in birds. Genome Biol. 2015;16: 164. doi:10.1186/s13059-015-0724-z

69. Botero-Castro F, Figuet E, Tilak MK, Nabholz B, Galtier N. Avian genomes revisited: Hidden genes uncovered and the rates versus traits paradox in birds. Mol Biol Evol. 2017;34: 3123–3131. doi:10.1093/molbev/msx236

70. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40: 290–301. doi:10.1093/nar/gkr1065

71. Triant DA, Pearson WR. Most partial domains in proteins are alignment and annotation artifacts. Genome Biol. 2015;16: 1–12. doi:10.1186/s13059-015-0656-7

72. Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23: 1061–1067. doi:10.1093/bioinformatics/btm071

73. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

74. Tromer EC. Evolution of the Kinetochore Network in Eukaryotes. Utrecht University. 2017. Available: https://dspace.library.uu.nl/handle/1874/356941

75. Forslund K, Pereira CCC, Capella-Gutierrez S, Da Silva AS, Altenhoff A, Huerta-Cepas J, et al. Gearing up to handle the mosaic nature of life in the quest for orthologs. Kelso J, editor. Bioinformatics. 2018;34: 323–329. doi:10.1093/bioinformatics/btx542

76. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, et al. The genome sequence of Schizosaccharomyces pombe. Nature. 2002;415: 871–880. doi:10.1038/nature724

77. Bitton DA, Wood V, Scutt PJ, Grallert A, Yates T, Smith DL, et al. Augmented annotation of the Schizosaccharomyces pombe genome revealsadditional genes required for growth and viability. Genetics. 2011;187: 1207–1217. doi:10.1534/genetics.110.123497

78. Cavalier-Smith T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett. 2010;6: 342–345. doi:10.1098/rsbl.2009.0948

79. Katz LA, Grant JR, Parfrey LW, Burleigh JG. Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. Syst Biol. 2012;61: 653–660. doi:10.1093/sysbio/sys026

80. He D, Fiz-Palacios O, Fu CJ, Tsai CC, Baldauf SL. An alternative root for the eukaryote tree of life. Curr Biol. 2014;24: 465–470. doi:10.1016/j.cub.2014.01.036

81. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

82. Eddy SR. HMMER. Available: http://hmmer.org

83. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–D285. doi:10.1093/nar/gkv1344

84. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

85. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström Di, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541: 353–358. doi:10.1038/nature21031

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

87. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

88. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

89. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

90. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274. doi:10.1093/molbev/msu300

91. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018;35: 518–522. doi:10.1093/molbev/msx281

92. Huerta-Cepas J, Serra FF, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33: 1635–1638. doi:10.1093/molbev/msw046

93. van Dam TJP, Townsend MJ, Turk M, Schlessinger A, Sali A, Field MC, et al. Evolution of modular intraflagellar transport from a coatomer-like progenitor. Proc Natl Acad Sci. 2013;110: 6943–6948. doi:10.1073/PNAS.1221011110

94. Antonova S V., Boeren J, Timmers HTM, Snel B. Epigenetics and transcription regulation during eukaryotic diversification: the saga of TFIID. Genes Dev. 2019;33: 888–902. doi:10.1101/GAD.300475.117

95. Tromer EC, van Hooff JJE, Kops GJPL, Snel B. Mosaic origin of the eukaryotic kinetochore. Proc Natl Acad Sci U S A. 2019;116: 12873–12882. doi:10.1073/pnas.1821945116

96. Fernández R, Gabaldón T. Gene gain and loss across the metazoan tree of life. Nat Ecol Evol. 2020;4: 524–533. doi:10.1038/s41559-019-1069-x

97. Guijarro-Clarke C, Holland PWH, Paps J. Widespread patterns of gene loss in the evolution of the animal kingdom. Nat Ecol Evol. 2020;4: 519–523. doi:10.1038/s41559-020-1129-2

98. Gabaldón T, Rainey D, Huynen MA. Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (Complex I). J Mol Biol. 2005;348: 857–870. doi:10.1016/j.jmb.2005.02.067

99. Irwin NAT, Keeling PJ. Extensive Reduction of the Nuclear Pore Complex in Nucleomorphs. Archibald J, editor. Genome Biol Evol. 2019;11: 678–687. doi:10.1093/gbe/evz029

100. Koonin E V. The Incredible Expanding Ancestor of Eukaryotes. Cell. 2010. pp. 606–608. doi:10.1016/j.cell.2010.02.022

101. Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C. The Impact of Gene Duplication, Insertion, Deletion, Lateral Gene Transfer and Sequencing Error on Orthology Inference: A Simulation Study. Carmel L, editor. PLoS One. 2013;8: e56925. doi:10.1371/journal.pone.0056925

102. Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldón T, Huerta-Cepas J, et al. Advances and Applications in the Quest for Orthologs. Rogers R, editor. Mol Biol Evol. 2019;36: 2157–2164. doi:10.1093/molbev/msz150

103. Kriventseva E V, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47: D807–D811. doi:10.1093/nar/gky1053

104. Deutekom ES, Vosseberg J, Van Dam TJP, Snel B. Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. PLoS Comput Biol. 2019;15: 1–15. doi:10.1371/journal.pcbi.1007301

105. Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 2011;39: e32. doi:10.1093/nar/gkq953

106. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: Comprehensive orthology analysis and visual exploration. BMC Bioinformatics. 2011;12: 11. doi:10.1186/1471-2105-12-11

107. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics. 2014;30: 734–6. doi:10.1093/bioinformatics/btt582

108. Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Vesztrocy AW, Dalquen DA, et al. OMA standalone: Orthology inference among public and custom genomes and transcriptomes. Genome Res. 2019;29: 1152–1163. doi:10.1101/gr.243212.118

109. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16: 157. doi:10.1186/s13059-015-0721-2

110. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12: 59–60. doi:10.1038/nmeth.3176

111. Derelle R, Philippe H, Colbourne JK. Broccoli: combining phylogenetic and network analyses for orthology assignment. bioRxiv. 2019; 2019.12.13.875831. doi:10.1101/2019.12.13.875831

112. Cosentino S, Iwasaki W. SonicParanoid: fast, accurate and easy orthology inference. Bioinformatics. 2018;35: 149–151. doi:10.1093/bioinformatics/bty631

113. Hu X, Friedberg I. SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. Gigascience. 2019;8: 1–12. doi:10.1093/gigascience/giz118

114. Rogozin IB, Wolf YI, Badenko VN, Koonin E V. Dollo parsimony and the reconstruction of genome evolution. In: Albert VA, editor. Parsimony, Phylogeny and Genomics. 2006. pp. 1–18. doi:10.1093/acprof

115. López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. Philos Trans R Soc B Biol Sci. 2019;374: 20190088. doi:10.1098/rstb.2019.0088

116. Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 2019;66: 4–119. doi:10.1111/jeu.12691

117. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47: D529–D541. doi:10.1093/nar/gky1079

118. Trabuco LG, Betts MJ, Russell RB. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. Methods. 2012;58: 343–348. doi:10.1016/j.ymeth.2012.07.028

119. Drew K, Lee C, Huizar RL, Tu F, Borgeson B, McWhite CD, et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Mol Syst Biol. 2017;13: 932. doi:10.15252/msb.20167490

120. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2: 193–218. doi:10.1007/BF01908075

121. Koumandou VL, Wickstead B, Ginger ML, Van Der Giezen M, Dacks JB, Field MC. Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit Rev Biochem Mol Biol. 2013;48: 373–396. doi:10.3109/10409238.2013.821444

122. Moi D, Kilchoer L, Aguilar PS, Dessimoz C. Scalable Phylogenetic Profiling using MinHash Uncovers Likely Eukaryotic Sexual Reproduction Genes. bioRxiv. 2019; 852491. doi:10.1101/852491

123. de Wolf B, Oghabian A, Akinyi M V, Hanks S, Tromer EC, van Hooff J, et al. Chromosomal instability by mutations in a novel specificity factor of the minor spliceosome. 2020. doi:10.1101/2020.08.06.239418

124. Dey G, Jaimovich A, Collins SR, Seki A, Meyer T. Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. Cell Rep. 2015;10: 993–1006. doi:10.1016/j.celrep.2015.01.025

125. Li Y, Calvo SE, Gutman R, Liu JS, Mootha VK. Expansion of biological pathways based on evolutionary inference. Cell. 2014;158: 213–225. doi:10.1016/j.cell.2014.05.034

126. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20: 238. doi:10.1186/s13059-019-1832-y

127. Cosentino S, Iwasaki W. SonicParanoid: Fast, accurate and easy orthology inference. Bioinformatics. 2018;35: 149–151. doi:10.1093/bioinformatics/bty631

128. Derelle R, Philippe H, Colbourne JK. Broccoli: combining phylogenetic and network analyses for orthology assignment. Falush D, editor. Mol Biol Evol. 2020; 2019.12.13.875831. doi:10.1093/molbev/msaa159

129. Snitkin ES, Gustafson AM, Mellor J, Wu J, Delisi C. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. BMC Bioinformatics. 2006;7: 1–11. doi:10.1186/1471-2105-7-420

130. Jothi R, Przytycka TM, Aravind L. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: A comprehensive assessment. BMC Bioinformatics. 2007;8. doi:10.1186/1471-2105-8-173

131. Muley VY, Ranjan A. Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. PLoS One. 2012;7. doi:10.1371/journal.pone.0042057

132. Škunca N, Dessimoz C. Phylogenetic Profiling: How Much Input Data Is Enough? Escriva H, editor. PLoS One. 2015;10: e0114701. doi:10.1371/journal.pone.0114701

133. Deutekom ES, Snel B, van Dam TJP. Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes. Brief Bioinform. 2020;22: 1–9. doi:10.1093/bib/bbaa206

134. Sun J, Li Y, Zhao Z. Phylogenetic profiles for the prediction of protein-protein interactions: How to select reference organisms? Biochem Biophys Res Commun. 2007;353: 985–991. doi:10.1016/j.bbrc.2006.12.146

135. Simonsen M, Maetschke SR, Ragan MA. Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. Bioinformatics. 2012;28: 851–857. doi:10.1093/bioinformatics/btr720

136. Bloch I, Sherill-Rofe D, Stupp D, Unterman I, Beer H, Sharon E, et al. Optimization of co-evolution analysis through phylogenetic profiling reveals pathway-specific signals. Bioinformatics. 2020;36: 4116–4125. doi:10.1093/bioinformatics/btaa281

137. Snel B, Huynen MA. Quantifying modularity in the evolution of biomolecular systems. Genome Res. 2004;14: 391–397. doi:10.1101/gr.1969504

138. Campillos M, Von Mering C, Jensen LJ, Bork P. Identification and analysis of evolutionarily cohesive functional modules in protein networks. Genome Res. 2006;16: 374–382. doi:10.1101/gr.4336406

139. Fokkens L, Snel B. Cohesive versus flexible evolution of functional modules in eukaryotes. PLoS Comput Biol. 2009;5. doi:10.1371/journal.pcbi.1000276

140. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: A database of predicted functional associations between proteins. Nucleic Acids Res. 2003;31: 258–261. doi:10.1093/nar/gkg034

141. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, et al. Comparative assessment of large-scale data sets of protein-protein interactions. Nature. 2002;417: 399–403. doi:10.1038/nature750

142. Aravind L, Watanabe H, Lipman DJ, Koonin E V. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci U S A. 2000;97: 11319–11324. doi:10.1073/pnas.200346997

143. Münsterkötter M, Steinberg G. The fungus Ustilago maydis and humans share disease-related proteins that are not found in Saccharomyces cerevisiae. BMC Genomics. 2007;8: 1–10. doi:10.1186/1471-2164-8-473

144. Dujon BA, Louis EJ. Genome diversity and evolution in the budding yeasts (Saccharomycotina). Genetics. 2017;206: 717–750. doi:10.1534/genetics.116.199216

145. Tabach Y, Billi AC, Hayes GD, Newman MA, Zuk O, Gabel H, et al. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. Nature. 2013;493: 694–698. doi:10.1038/nature11779

146. Barker AR, Renzaglia KS, Fry K, Dawe HR. Bioinformatic analysis of ciliary transition zone proteins reveals insights into the evolution of ciliopathy networks. BMC Genomics. 2014;15: 1–9. doi:10.1186/1471-2164-15-531

147. Kollmar M, Lbik D, Enge S. Evolution of the eukaryotic ARP2/3 activators of the WASP family: WASP, WAVE, WASH, and WHAMM, and the proposed new family members WAWH and WAML. BMC Res Notes. 2012;5. doi:10.1186/1756-0500-5-88

148. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34: D535–D539. doi:10.1093/nar/gkj109

149. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. Zhou Y, editor. Int J Proteomics. 2014;2014: 147648. doi:10.1155/2014/147648

150. Bekker J, Davis J. Learning from positive and unlabeled data: a survey. Machine Learning. Springer US; 2020. doi:10.1007/s10994-020-05877-5

151. Li X, Liu B. Learning to classify texts using positive and unlabeled data. IJCAI Int Jt Conf Artif Intell. 2003; 587–592.

152. Zhao XM, Wang Y, Chen L, Aihara K. Gene function prediction using labeled and unlabeled data. BMC Bioinformatics. 2008;9: 57. doi:10.1186/1471-2105-9-57

153. Pancaroglu D, Tan M. Improving Positive Unlabeled Learning Algorithms for Protein Interaction Prediction BT - 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014). In: Saez-Rodriguez J, Rocha MP, Fdez-Riverola F, De Paz Santana JF, editors. Cham: Springer International Publishing; 2014. pp. 81–88.

154. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: A review. Int J Pattern Recognit Artif Intell. 2009;23: 687–719. doi:10.1142/S0218001409007326

155. Zheng W, Jin M. The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. SN Comput Sci. 2020;1: 1–13. doi:10.1007/s42979-020-0074-0

156. Chen C, Liaw A, Breiman L. Using Random Forest to Learn Imbalanced Data. Discovery. 2004; 1–12.

157. Liu X-Y, Wu J, Zhi-Hua Z. Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans Syst Man, Cybern Part B. 2009;39: 539–550. doi:10.1109/TSMCB.2008.2007853

158. Kaboutari A, Bagherzadeh J, Kheradmand F. An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification. Int J Comput Appl Technol Res. 2014;3: 592–594. doi:10.7753/ijcatr0309.1012

159. Deutekom ES, van Dam TJP, Snel B. Phylogenetic profiling in eukaryotes: The effect of species, orthologous group, and interactome selection on protein interaction prediction. bioRxiv. 2021; 2021.05.05.442724. doi:https://doi.org/10.1101/2021.05.05.442724

160. Breiman L. Random Forests. Mach Learn. 2001;45: 5–32. doi:10.1023/A:1010933404324

161. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci U S A. 2002;99: 6567–6572. doi:10.1073/pnas.082099299

162. Weiss GM, Provost F. Learning when training data are costly: The effect of class distribution on tree induction. J Artif Intell Res. 2003;19: 315–354. doi:10.1613/jair.1199

163. Škunca N, Altenhoff A, Dessimoz C. Quality of computationally inferred gene ontology annotations. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002533

164. Prati RC, Batista GEAPA, Monard MC. Class imbalances versus class overlapping: An analysis of a learning system behavior. Lect Notes Artif Intell (Subseries Lect Notes Comput Sci. 2004;2972: 312–321. doi:10.1007/978-3-540-24694-7_32

165. Ejigu GF, Jung J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology (Basel). 2020;9: 1–27. doi:10.3390/biology9090295

166. Ko BJ, Lee C, Kim J, Rhie A, Yoo D, Howe K, et al. Widespread false gene gains caused by duplication errors in genome assemblies. bioRxiv. 2021; 2021.04.09.438957. Available: https://doi.org/10.1101/2021.04.09.438957

167. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020;18: 1–24. doi:10.1371/journal.pbio.3000862

168. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592: 737–746. doi:10.1038/s41586-021-03451-0

169. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42: 581–591. doi:10.1093/nar/gkt1099

170. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. doi:10.1038/s41586-021-03819-2

171. Humphreys IR, Pei J, Baek M, Krishnakumar A. Structures of core eukaryotic protein complexes. bioRxiv. 2021.

172. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RN A sequencing on an array of nanopores. Nat Methods. 2018;15: 201–206. doi:10.1038/nmeth.4577

173. Kristensen DM, Wolf YI, Mushegian AR, Koonin E V. Computational methods for Gene Orthology inference. Brief Bioinform. 2011;12: 379–391. doi:10.1093/bib/bbr030

174. Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. Domainoid: Domain-oriented orthology inference. BMC Bioinformatics. 2019;20: 1–12. doi:10.1186/s12859-019-3137-2

175. Schneider A, Seidl MF, Snel B. Shared Protein Complex Subunits Contribute to Explaining Disrupted Co-occurrence. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1003124

176. Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47: W636–W641. doi:10.1093/nar/gkz268

177. Crosswell LC, Thornton JM. ELIXIR: a distributed infrastructure for European biological data. Trends Biotechnol. 2012;30: 241–242.

178. Hoffmann J, Bar-Sinai Y, Lee LM, Andrejevic J, Mishra S, Rubinstein SM, et al. Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. Sci Adv. 2019;5. doi:10.1126/sciadv.aau6792

179. Olson M, Wyner AJ, Berk R. Modern neural networks generalize on small data sets. Adv Neural Inf Process Syst. 2018;2018-Decem: 3619–3628.

FIGURE ATTRIBUTION

The cover of this thesis in part uses premade vector images retrieved and adjusted from Adobe Stock. The credit goes to the following.

© jenesesimre/Adobe Stock
© mirifadapt/Adobe Stock
© natality/Adobe Stock
© Sergj/Adobe Stock
© Ekaterina Glazkova /Adobe Stock
© aksol /Adobe Stock
© Marina Gorskaya/Adobe Stock

Figure 1 in Chapter 1 in part uses vector images downloaded from Phylopic.org or from Adobe Stock that were combined into one larger figure. The credit goes to the following.

Matt Crook, Gareth Monger, Yan Wong from photo by Gyik Toma, under the following license: http://creativecommons.org/licenses/by-sa/3.0/

© pingebat/Adobe Stock

SAMENVATTING

Overal om ons heen, en zelfs in ons, leven talloze (micro-)organismen. Sommige organismen kunnen we makkelijk zien, zoals bomen, zeewier en insecten. Andere organismen zien we alleen met een microscoop, zoals bacteriën, gist of een pantoffeldiertje. Het leven op aarde is verbonden door een gezamenlijke eigenschap, namelijk het genetische materiaal dat zit opgeslagen in het DNA. Het complete DNA wat zich in een organisme bevindt heet ook wel het genoom van een organisme. Het DNA is een lange streng van vier bouwstenen die we kunnen representeren met vier letters. Deze vier letters staan in bepaalde volgordes en vormen een code. Deze code vormt net als de letters in een alfabet woorden in een zin. De stukjes code, or woorden, zijn genen en deze genen zijn een blauwdruk voor de belangrijke componenten die een organisme nodig heeft om te kunnen functioneren, de eiwitten.

Met de recente opkomst van snelle DNA sequencing methoden, wat "next-generation sequencing" of "high-throughput sequencing" heet, is het mogelijk om steeds accurater, makkelijker en sneller DNA uit organismen te extraheren en af te lezen. Door de vele genomen die tegenwoordig gesequenced zijn kunnen we nu de overeenkomsten en de verschillen tussen organismen en hun genen beter vergelijken. Dit laatste wordt ook wel vergelijkende genoom analyse genoemd. De route die het leven heeft afgelegd gedurende de evolutie van verschillende organismen zit verscholen in hun DNA. Met vergelijkende genoom analyse kunnen we de overeenkomsten en verschillen onderzoeken tussen gedeelde genen van organismen en genen die aanwezig of afwezig zijn in deze organismen. Op deze manier kunnen we een reconstructie maken van welke organismen evolutionair aan elkaar gerelateerd zijn. Met het DNA kunnen we daarom de stamboom van het leven veel accurater maken.

Alle organismen zijn uiteindelijk (verre) familie van elkaar. De familieboom bestaat uit Bacteria, Archaea en Eukaryota. Bacteriën en archaea noemen we ook wel prokaryoten. Uit vele jaren onderzoek weten we dat de eukaryoten afstammen van de prokaryoten, waarbij de eukaryoten zijn ontstaan na een aantal grote evolutionaire gebeurtenissen met betrekking tot hun cellulaire structuur. Zo onderscheiden eukaryoten zich van de prokaryoten doordat hun cellulaire structuur complexer is. De cellen van eukaryoten hebben bijvoorbeeld een celkern waarin hun genetisch materiaal zit

verpakt. Ook hebben de cellen van eukaryoten andere compartimenten met gespecialiseerde functies. Bijna alle (grotere) organismen die we met het blote oog kunnen zien zijn meercellige eukaryoten. Mensen zijn dus eukaryoten, net zoals bomen, zeewier, insecten en schimmels.

De technische en conceptuele ontwikkeling rondom vergelijkende genoom analyse en daarnaast de verkregen kennis dat alle organismen van elkaar afstammen heeft ook geleid tot de wetenschap dat genen die gedeeld worden tussen organismen ook van elkaar afstammen. Door te kijken naar de code van genen en deze naast elkaar te leggen kunnen we zien of deze genen aan elkaar verwant zijn. Deze verwantschap tussen genen noemen we homologie. Er zijn verschillende soorten homologie tussen genen. Als verschillende organismen hetzelfde gen hebben, betekend het dat de voorouder van deze organismen het gen ook had moeten hebben. Dit soort homologie tussen genen noemen we orthologie. Een groep van orthologe genen van een set organismen komt dus voort uit een gen in de gemeenschappelijke voorouder van die set organismen die de genen hebben. Deze genen zijn functioneel ook vergelijkbaar. De kennis dat orthologe genen functioneel vergelijkbaar zijn wordt daarom gebruikt om de functie van onbekende genen af te leiden in (nieuw) geëxtraheerde genomen. Naast een gen die organismen delen vanwege een gemeenschappelijk voorouder, zijn er ook andere relaties tussen genen. Als een gen in een bepaald organisme gedupliceerd (gekopieerd) is, zijn de twee kopieën ook homoloog. Maar in dit geval veranderd de functie van één van de kopieën. Dit soort homologie tussen genen noemen we paralogie.

Uiteraard zijn er ook ingewikkeldere constructies hoe genen aan elkaar verwant zijn, bijvoorbeeld als in een voorouder een gen is gedupliceerd en de kopieën (de paralogen) vervolgens terecht komen in de afstammelingen. De afstammelingen hebben dan paraloge genen waarvan beide genen een ortholoog gen hebben in zuster organismen. Naast het dupliceren van individuele genen, dupliceren hele genomen in sommige organismen. Dit gebeurt heel vaak in planten en we weten nu dat dit een reden is voor de enorme diversiteit in het plantenrijk. Organismen wisselen ook (geslachtloos) genen uit door middel van horizontale gen overdracht. Horizontale gen overdracht lijkt in eukaryoten echter een kleine(re) bijdrage te hebben in genoom evolutie. De ingewikkelde gen relaties en vele andere complexe evolutionaire gebeurtenissen die plaatsvinden tijdens

de evolutie van eukaryote genomen maakt vergelijkende genoom analyse vaak heel uitdagend.

Er is nu ook veel onderzoek gedaan naar de aan- en afwezigheden van genen in verschillende eukaryoten en welke genen worden gedeeld tussen deze eukaryoten. Doordat veel verschillende genen in meerdere verschillende eukaryoten voorkomen is tegen alle verwachtingen in aangetoond dat de voorouder van de eukaryoten al veel genen had en daarom al vrij complex was. Waar men voorheen dacht dat eukaryoten door de miljoenen jaren heen complexer zijn geworden doordat ze steeds meer nieuwe genen kregen, blijkt de voorouder van alle eukaryoten al heel veel genen te bezitten die zijn doorgegeven aan de afstammelingen. Deze genen zijn in de loop van generaties vaak gekopieerd (gedupliceerd). Daarnaast is er nog een interessante uitkomst van vele jaren onderzoek dat meer genen verloren zijn gegaan in eukaryoten dan dat er nieuwe genen zijn ontstaan of "uitgevonden". Uiteindelijk blijkt dus dat de evolutie van eukaryote genomen helemaal niet gedreven wordt door het ontstaan van nieuwe genen, maar juist door de duplicatie van bestaande genen en het verlies van genen.

Om de vele duizenden genomen en genen van vele eukaryoten tegelijkertijd te vergelijken hebben we computers nodig. Door bioinformatische analyses uit te voeren zien we welke genen aan- en afwezig zijn in eukaryoten. Door te kijken naar de genen die aan- en afwezig zijn brengen we in kaart waar genen verdwijnen, dupliceren of ontstaan. Om aan- of afwezige genen in eukaryoten te onderzoeken, kijken we eerst naar welke genen ortholoog zijn (welke genen uit hetzelfde voorouderlijk gen voortkomen). Dit doen we met behulp van gen sequentie vergelijkingen, oftewel we kijken hoeveel de sequenties van genen op elkaar lijken. Sequenties kunnen gaan verschillen van elkaar door mutaties die optreden. Sommige sequenties muteren sneller en hierdoor evolueren de genen sneller dan hun orthologe tegenhangers in andere organismen. Door de snellere evolutie van bepaalde genen lijken de genen minder op elkaar en worden orthologe genen minder snel herkent. Hierdoor missen we vaak genen die belangrijk zijn voor vergelijkende genoom analyse.

Naast de biologische processen die hierboven staan beschreven, zoals gen duplicatie, verlies, en evolutie snelheid, maken technische factoren vergelijkende genoom analyses ook complex. Het verkrijgen van

betrouwbare genoom en gen sequenties hangt van veel dingen af. Zo heeft het onder andere te maken met de techniek die we gebruiken om het DNA te extraheren en te vertalen naar de letters die wij kunnen lezen. De staat van het geëxtraheerde DNA beïnvloed ook de betrouwbaarheid van het uiteindelijke genoom, veel onderzoekers publiceren het genoom al voordat het volledig is. Dit kan mede door publicatie druk zijn of puur om andere wetenschappers de kans te geven al naar het genoom te laten kijken.

Als het genoom uit een organisme is gehaald, weten we niet meteen welke genen erin zitten. Niet het hele DNA bestaat namelijk uit genen die voor iets functioneels coderen en we zullen de genen eerst nog moeten voorspellen. Het voorspellen noemen we genoom annotatie. Genoom annotatie betekend dat we de functionele componenten van het DNA in kaart brengen, oftewel de genen die coderen voor eiwitten. Bij het voorspellen van genen gebruiken we veelal geautomatiseerde gen voorspellingsmethoden. Genen voorspellen is niet altijd makkelijk omdat eukaryoten nogal een complexe genoom structuur hebben. Genen worden herkend door naar specifieke stukjes sequentie kijken, maar deze stukjes sequentie zijn niet altijd precies hetzelfde tussen alle eukaryoten. In dit geval kan het geautomatiseerde gen voorspeller de genen niet herkennen, omdat we vaak uitgaan van een sequentie die we al kennen uit een ander organisme, bijvoorbeeld die van de mens. Soms wordt het voorspellen van genen bemoeilijkt doordat er stukjes DNA missen. De stukjes die we missen kunnen juist belangrijk zijn voor het herkennen van de genen. In beide gevallen missen we genen en wordt er foutief afgeleid dat de genen afwezig zijn in een organisme.

Om met al deze inwikkelde zaken nog steeds goed onderzoek te doen naar genoom evolutie in eukaryoten zullen we continue ons bioinformatische gereedschap moeten verbeteren en de conclusies die we ermee trekken bijschaven. Om grootschalig analyses uit te voeren en naar de patronen te kijken in de evolutie van eukaryoten en hun genomen zullen we ons moeten richten op geautomatiseerde methoden die op grote schaal genen kunnen voorspellen, vergelijken en verwantschap af kunnen lijden. In dit proefschrift vergelijk ik verschillende geautomatiseerde methoden en kijk ik hoe betrouwbaar deze methoden zijn in het benaderen van een aantal patronen in de evolutie van eukaryote genomen.

Samenvatting van de hoofdstukken

Uit onderzoek van het laatste decennium weten we dat eukaryoten meer genen lijken te verliezen dan dat ze nieuwe genen verkrijgen. Tegelijkertijd weten we ook dat genoom data niet altijd compleet is. De annotatie van genomen verloopt ook niet altijd perfect. **Hoofdstuk 2** is een studie die kijkt naar de hoeveelheid genen waarvan we foutief aannemen dat ze afwezig zijn in eukaryoten door gen predictie problemen. Daarbij is een belangrijke vraag hoeveel van de genen waarvan we denken dat ze verloren zijn gegaan gedurende de evolutie van eukaryoten, eigenlijk niet verloren zijn gegaan maar foutief afgeleid zijn als afwezig. Ook kijken we of we van begin af aan bepaalde afwezige genen in eukaryoten moeten wantrouwen. Met andere woorden, we willen weten welke factoren een aanwijzing zijn dat een afwezig gen niet echt afwezig is. Dit soort "verdachte afwezige" genen zijn bijvoorbeeld genen die afwezig zijn in één organisme, terwijl we de genen in een zuster organismen nog wel vinden. We noemen dit soort verdachte afwezigheden ook wel organisme-specifieke gen afwezigheden.

In hoofdstuk 2 beginnen we onze analyse door te kijken naar eiwitten die voorspeld zijn in een set genomen van 209 diverse eukaryoten. De eukaryoten bevinden zich op belangrijke evolutionaire posities in de eukaryote evolutionaire boom. Dit betekent dat we representatieve eukaryoten hebben gekozen die zich in belangrijke taxonomische groepen bevinden binnen de eukaryoten. Vervolgens benaderen we welke eiwitten in de voorouder zat van deze eukaryoten. Dit doen we met een algoritme die kiest tussen eiwitten aan de hand van bepaalde criteria, onder andere, dat de eiwitten in een minimaal aantal eukaryoten moeten voorkomen in bepaalde taxonomische groepen. Zodra we deze vooroudelijke eiwitten hebben van de eukaryoten in onze set, kunnen we traceren waar en in welke eukaryoten de vooroudelijke eiwitten zijn verloren. Hiermee brengen we in kaart welke vooroudelijke eiwitten aanwezig of afwezig zijn in de verschillende organismen in onze set. De vooroudelijke eiwitten die afwezig zijn zoeken we vervolgens terug in het originele DNA. Hierbij kijken we ook of de eiwitten niet van pseudogenen afkomstig zijn, oftewel gebroken genen die geen functionele eiwitten meer kunnen maken en daarom niet zijn voorspeld. Alle eiwitten die we hierna toch terugvinden in het DNA achten we al foutief afgeleid als afwezig.

Onze resultaten tonen dat gemiddeld gezien genen correct worden afgeleid als afwezig. Er verandert daarom weinig aan het geschatte totaal aantal

genen die zijn verloren in eukaryoten. Daarom blijft de conclusie behouden dat eukaryoten meer genen verliezen dan dat ze verkrijgen. Wel zien we dat "verdachte" organisme-specifieke afwezige genen met een veel hoger percentage foutief afgeleid worden als afwezig. Voorzichtigheid is daarom geboden met deze verdachte organisme-specifieke gen afwezigheden. Met deze studie hebben we het bekende probleem van gen predictie gekwantificeerd en laten zien welke gen afwezigheden alarmbellen moeten doen rinkelen.

Onderzoek naar de evolutie van een klein aantal genen kan met de hand gedaan worden en is vaak het meest betrouwbaar. Bij dit kleinschalig onderzoek wordt door een individuele wetenschapper grotendeels maatwerk verricht, gebruikmakend van specifieke software. Uiteindelijk wordt door een persoon bekeken welke genen samen in een orthologe groep behoren, zonder dat de beslissing goed te automatiseren is. Om grootschalige analyses te doen naar genoom evolutie en de evolutie van genen in honderden eukaryoten tegelijk zullen we geautomatiseerd orthologe genen moeten voorspellen. In **hoofdstuk 3** gaan we dieper in op automatische orthologie voorspellingsmethoden. We zullen deze methoden vergelijken door te kijken naar hoe goed ze zijn in het benaderen van bepaalde waarneming en trends in eukaryote genoom evolutie, zoals: de gen inhoud van de laatste eukaryote voorouder, gen verlies, eiwit-eiwit interactie voorspellingen, en het benaderen van manueel voorspelde orthologe groepen. We vergelijken ook de verkregen orthologe groepen van de verschillende methoden met elkaar.

Uit de resultaten van deze analyse kunnen we opmaken dat alle methoden de waarnemingen en algemene trends in eukaryote genoom evolutie redelijk benaderen. Deze trends zijn bijvoorbeeld een groot gen repertoire van de laatste eukaryote voorouder, veel gen verlies en het correct voorspellen van eiwit-eiwit interacties. Daarentegen is de overlap met de manueel voorspelde orthologe groepen onvolmaakt. Geen van de automatische methode heeft een duidelijke voorsprong op de andere, ze presteren grotendeels hetzelfde. Contra-intuïtief is dat ondanks dat alle methoden gemiddeld even goed presteren in het voorspellen van de grootschalige trends in eukaryote genoom evolutie, de voorspelde orthologe groepen tussen de methoden enorm verschillen van elkaar.

Deze studie laat zien dat geautomatiseerde methoden algemene trends kunnen vinden in eukaryote genoom evolutie. Daarentegen, de studie bewijst ook dat het met de hand voorspellen nog steeds de voorkeur moet hebben als de interesse ligt in een specifieke (groep) genen en hun orthologe groepen. Automatische methoden kunnen namelijk de kwaliteit van hand werk nog niet benaderen. Deze bevindingen benadrukken dat hand werk en geautomatiseerd voorspellen van orthologe groepen complementair zijn. Deze complementariteit kunnen we juist gebruiken om vergelijkende genoom analyse te versterken en verbeteren.

In **hoofdstuk 4** en **hoofdstuk 5** focussen we op eiwit-eiwit interacties binnen eukaryoten. Eiwit-eiwit interacties hebben een belangrijke rol in cellulaire processen. Het voorspellen van deze interacties is daarom een belangrijk uitgangspunt in het voorspellen van functionele relaties van eiwitten. Het voorspellen van de functie van eiwitten met een onbekende functie kan ook worden gedaan aan de hand van het voorspellen van hun interacties met eiwitten met een bekende functie. In hoofdstuk 3 lukte het ons met bepaalde keuzes en aannamen op grote schaal goed eiwit-eiwit interacties te voorspellen in eukaryoten.

In hoofdstuk 3 hebben we gebruik gemaakt van fylogenetische profielen om eiwit-eiwit interacties te voorspellen. Fylogenetische profielen zijn niets meer dan de aan- en afwezigheid van dezelfde (orthologe) eiwitten in meerdere organismen. Deze aan- en afwezigheden geven een evolutionair signaal of patroon. Dit signaal weergeeft in welke (groep) organismen bepaalde eiwitten samen steeds voorkomen of niet. Het signaal kan gebruikt worden om af te leiden welke eiwitten samen evolueren. Als twee eiwitten altijd samen aan- of afwezig zijn (co-evolueren) in bepaalde organismen dan duidt dit op een functionele connectie tussen de eiwitten, zoals een interactie tussen de eiwitten. Door de wiskundige afstanden tussen fylogenetische profielen te berekenen kunnen we zien of profielen op elkaar lijken. Als de profielen van de eiwitten een kleine wiskundige afstand hebben tot elkaar, lijken ze op elkaar en betekent dit dat de eiwitten vaak samen voorkomen. De eiwitten co-evolueren dus.

Fylogenetische profielen zijn volop in de belangstelling om gebruikt te worden bij het voorspellen van eiwit-eiwit interacties of functionele relaties van eiwitten. Er zijn nog heel veel eiwitten waarvan we niet weten wat ze doen en welke nog een onbekende functie hebben in eukaryoten.

Veel onderzoek is gaande naar goedkopere en efficiëntere manieren om functies van eiwitten te achterhalen met behulp van bioinformatische analyses. In bacteriën is al veel onderzoek gedaan naar het gebruik van fylogenetische profielen voor eiwit functie voorspellingen. Hierbij werd voornamelijk gekeken naar welk soort bacteriële organismen je het beste kon gebruiken in de fylogenetische profielen.

Met eukaryoten blijkt het uitdagender te zijn om fylogenetische profielen te gebruiken voor het voorspellen van functionele relaties tussen eiwitten. Dit heeft te maken met de complexiteit van genoom evolutie in eukaryoten in vergelijking met bacteriën. Wij laten in hoofdstuk 3 echter zien dat fylogenetische profielen wel degelijk succesvol gebruikt kunnen worden. In **hoofdstuk 4** bouwen we voort op deze bevindingen. We onderzoeken waarom wij wel een goede presentie krijgen bij het voorspellen van eiwit interacties. Hierbij kijken we naar de invloed van de eukaryote organismen en de genomen waarmee we de profielen opbouwen, de (pre-)selectie van orthologe groepen waarvan we de profielen hebben, en de gebruikte interacterende referentie eiwitten om de voorspellingen mee te doen.

In hoofdstuk 4 laten we zien welke keuzes van invloed zijn bij het gebruik van fylogenetische profielen voor eiwit-eiwit interactie voorspellingen in eukaryoten. Van invloed is onder andere bij het selecteren van genomen de kwaliteit van deze genomen, in mindere mate de diversiteit van de organismen van welk de genomen zijn, maar het meest de (evolutionaire) informatie in de genomen. Een nog belangrijkere bevinding is dat het niet de keus is van de genomen die de meeste invloed heeft. Het is de selectie van de orthologe groepen, de interacterende referentie eiwitten en de datakwaliteit van deze interacterende referentie eiwitten. Deze resultaten laten mogelijk zien waarom veel studies niet of nauwelijks erin zijn geslaagd eiwit interacties te voorspellen in eukaryoten met behulp van fylogenetische profielen. Met deze studie hebben we op een fundamenteel niveau laten zien voor welke type eiwitten fylogenetisch profielen in eukaryoten wel werken om eiwit interacties te voorspellen.

In hoofdstuk 4 hebben we nu aangetoond hoe fylogenetische profielen het beste werken om eiwit-eiwit interacties te voorspellen in eukaryoten. In **hoofdstuk 5** willen we het gebruik van fylogenetische profielen nog verder verbeteren. Machine learning biedt hier de mogelijkheid. Machine learning leert van patronen die we met het menselijk oog niet kunnen waarnemen.

In dit geval zijn we geïnteresseerd in de patronen van gezamenlijke aan- en afwezigheden van eiwitten die kunnen duiden op een eiwit-eiwit interactie. De signalen die we kunnen oppikken uit de patronen versterken of verzwakken door bepaalde genomen die in de fylogenetische profielen worden meegenomen. Dit is al duidelijk uit de resultaten van hoofdstuk 4, waarbij we zien dat sommige genomen een positieve of negatieve invloed hebben op hoe effectief we eiwit-eiwit interacties kunnen voorspellen met fylogenetische profielen. Het is ook mogelijk dat bepaalde combinaties van genomen, oftewel sub groepen van genomen, beter presteren bij het voorspellen van interacties. Dit laatste hebben we in hoofdstuk 4 niet kunnen bekijken omdat we niet met de hand alle mogelijk combinaties van genomen kunnen maken om te testen welke groep genomen het beste werken in een fylogenetisch profiel. Ook daar biedt machine learning een uitkomst.

In hoofdstuk 5 gebruiken we het machine learning algoritme Random Forest. Dit algoritme is uitermate geschikt voor het gebruik met fylogenetische profielen en eiwit-eiwit interactie voorspellingen. Dit komt omdat dit algoritme goed bestand is tegen ruis in de data. Fylogenetische profielen hebben last van ruis door de complexe dynamiek in eukaryote genoom evolutie en de verschillende technische complicaties bij het afleiden van aan- en afwezigheden van genen en orthologe groepen. Het is ook al aangetoond dat Random Forest effectief is met fylogenetische profielen.

Om Random Forest (of machine learning in het algemeen) toe te passen op onze data zullen we eerst met verschillende dingen rekening moeten houden met betrekking tot hoe onze data eruitziet. Om goed te kunnen leren van onze data, zal het Random Forest algoritme genoeg fylogenetische profielen moeten kunnen zien van interacterende en niet interacterende eiwitten. Dit is nodig om een onderscheid tussen interacterende en niet interacteren groepen eiwitten te kunnen maken. Ook moeten we zorgen dat het Random Forest algoritme niet te veel "blijft hangen" op een bepaalde groep eiwitten. Het algoritme kan bijvoorbeeld blijven hangen als er van de ene groep eiwitten (interacterend of niet interacterend) heel weinig of juist heel veel aanwezig zijn. Het Random Forest algoritme leert dan het verkeerde patroon. Als het algoritme van het verkeerde patroon leert, zal het geen goed onderscheid kunnen maken en niet goed kunnen zien bij welke groep bepaalde eiwitten horen.

In hoofdstuk 5 hebben we gekeken hoe we het beste de Random Forest kunnen toepassen op fylogenetische profielen om zo de beste prestatie te krijgen bij het voorspellen van eiwit–eiwit interacties. We laten zien dat we in vergelijking met de klassieke manier die wiskundige aftanden tussen de profielen berekend (hoofdstuk 4), we inderdaad een verbeterde prestatie krijgen bij het voorspellen van eiwit interacties met machine learning. We zien dat we redelijk goed onderscheid kunnen maken tussen interacteren en niet interacterende eiwitten.

De studies die staan beschreven in dit proefschrift geven weer hoe we met geautomatiseerde technieken en grootschalige analyses betrouwbaar conclusies kunnen trekken over eukaryote genoom evolutie. Ze benadrukken ook dat we op moeten passen als we met deze geautomatiseerde technieken kleinschalige analyses doen, bijvoorbeeld de evolutie van individuele genen. Automatische technieken en hand werk zijn complementair aan elkaar en zullen elkaar blijven aanvullen en niet uitsluiten. Daarnaast kunnen we met de resultaten die in het proefschrift staan beschreven, bevestigen dat de patronen van gen aan- en afwezigheden in eukaryoten (fylogenetische profielen) een sterk genoeg (evolutionair) signaal geven om effectief functionele afleidingen te maken van eiwitten en hoe machine learning dit nog (verder) kan verbeteren. Met het in acht nemen van deze waarnemingen kunnen we onze bioinformatische gereedschapskist verder uitbreiden en daarmee eukaryote genoom evolutie steeds beter onderzoeken en begrijpen.

CURRICULUM VITAE

Eva Stephanie Deutekom was born on the island Curaçao, the Netherlands Antilles, on 23rd of March 1988. She moved to the Netherlands in 2008 to study at the University of Amsterdam where in 2012 she finished the bachelor Bio-Exact that included subjects such as, data analysis, thermodynamics, (bio)chemistry, (bio)physics, mathematics, genomics, ecology and (micro)biology. In 2014 she graduated from the master Systems Biology and Bioinformatics. For 2.5 years she worked in the Computational Science Lab at the University of Amsterdam on simulating the physiology of stony coral calcification. In 2017 she started her Ph.D. in the Theoretical Biology and Bioinformatics group at Utrecht University under the supervision of prof. Berend Snel and dr. Teunis J.P. van Dam. The resulting work is described in this thesis.

LIST OF PUBLICATIONS

1. **Deutekom ES**, Vosseberg J, Van Dam TJP, Snel B. Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. PLoS Comput Biol. 2019;15: 1–15. doi:10.1371/journal.pcbi.1007301

2. van Belzen IAEM, **Deutekom ES**, Snel B. PhyRepID: A comparative phylogenomics approach for large-scale quantification of protein repeat evolution. bioRxiv. 2020; 0–28. doi:10.1101/2020.02.14.947036

3. **Deutekom ES**, Snel B, van Dam TJP. Benchmarking orthology methods using phylogenetic patterns defined at the base of Eukaryotes. Brief Bioinform. 2020;22: 1–9. doi:10.1093/bib/bbaa206

4. **Deutekom ES**, van Dam TJP, Snel B. Phylogenetic profiling in eukaryotes: The effect of species, orthologous group, and interactome selection on protein interaction prediction. bioRxiv. 2021; 2021.05.05.442724. doi:https://doi.org/10.1101/2021.05.05.442724

ACKNOWLEDGEMENTS

I have read many acknowledgements with the thought that if my time came, I would be able to easily write my own. The very last pages that are certainly not the most important ones in a thesis, but were everyone most of the time leafs to the moment they receive it. I thought the words of thanks would easily flow from my hand, because it is obvious who played an important role during my PhD. But when the days neared for me to write these last pages, the harder it got for me to start writing. Obtaining a PhD is not only these four+ years you spent on the project, but also the road that led to it. And there are so many people on that road. Writing the acknowledgments is a moment of looking back and reflection, of the people that played their parts and of the work you have now finished. It is the nearing of an end, and an end to something can sometimes be heavy and emotional. Luckily, ends also mark beginnings. Talking about beginnings, let me start with the one from my PhD.

In particular, I want to begin with the person that years ago saw something in me and trusted me enough to join the group as a PhD candidate, my promotor Berend. Maybe it started out a little difficult, because I was someone that liked to crawl into her shell way too often and you were more the person that at certain times burst out of his. That started out a little awkward, but with time I became to see it as "echt Berend" and you hauled that shell that I lugged around right off of me. Now I look at it with nostalgia. I came to know it as your passion for science and bioinformatics. You often showed your humanity and it made you approachable. Those moments you walked into the office to then walk out again without saying a word, leaving behind your students in a dazed stupor, is one such example that will stay with me (and I suspect many others) and make me laugh. I assume you just don't want to bother us and walk away again. The time you give your students and your work is admirable. Your insightful knowledge and feedback have sharpened my thinking and pulled our work up to a higher level. Even though it was sometimes hard to keep up with the "Sneltrein" of knowledge during our meetings and I had to write really fast. I will always want to take your approach to science and supervision as an example and hope to come close to your workstyle and skills, even by a little. There are still so many things I could still say, but at the risk of it becoming trite I will quickly finish with this last thing. I have learned so much from you. You are a great inspiration. Thank you.

To my co-promoter John, your help and feedback has also helped me immensely throughout the years. We have worked a lot together on a lot of beautiful data. I could always turn to you with questions, but also for reassurance. You are calmness itself! Thank you for the trust you have placed in me during our collaboration and for showing it. You showed me that the experiences I went through as a PhD were normal and you often shared your experience with me. This has given me a lot of peace. I am certain your new group is extremely happy with a leader so understanding and kind as you.

My gratitude goes to the reading and assessment committee for their valuable input on my thesis: prof. van den Ackerveken, prof. Dessimoz, prof. de Ridder, prof. Huijnen, and dr. Boekhorst.

I would also like to thank my paranymphs, Joleen and Juliane. Joleen, we have taken many walks, walks that always do me good. We can always share and compare our experiences as PhDs and the life during and after it. I am glad you are now *my* paranymph. Juliane, we started our PhD in the TBB on the same day. We went through eerily similar events that were partly PhD related, but also in large part not related to our work at all. Even though I want to kick your *** ever so often because you doubt yourself too much, I am proud of what you accomplished and how you fought through all of it. I believe in you!

I want to show my gratitude to the whole TBB as well. TBB is a place where science and integrity come first and communal effort to increasing the quality of its science and scientists are at the forefront. The input from the weekly lab meetings and group meetings have been essential in many of the chapters in this thesis. Even though the TBB is an entity with all its Binfs important cogs in a well-oiled science machine, there are many (staff) members of the TBB I want to thank individually. Each and every one of you left a mark in my life, during our meetings or otherwise.

Rob, even though you were never my supervisor, it is clear that you watch over your students and the TBB with a fatherly love. That has always made me very happy to see. Paulien, I have enormous respect for you as a scientist and power woman. I aspire to accomplish only a fraction of what you have accomplished. Can, I will never forget your kindness and interest for your group members. Bas, even if you go to Jena, I think you

are and will always remain a real TBB member. You say it like it is and that can be scary at times, but I respect it. Basten, thanks for all the help, you were always quick to help out with questions and problems (especially in R). Michael, from day one you were extremely kind and fun to be around! Rutger, if anyone has a question about statistics, you are the man with the answers! I was able to come with my questions to you, and you found the time to help out. In the end I understood only half of it, but that certainly was not your fault. Kirsten, I cannot match your love for Coke Zero, and I really love Coke Zero.

In the beginning of my PhD, I was told one staff member would get his own paragraph in the thesis of every PhD candidate. Now I completely understand why. Jan Kees, without you a lot of work would likely have crashed and burned, and multiple UPSs would likely have exploded. I think I will never be fully satisfied with a computational infrastructure other than the one you created at TBB. Thank you for all your hard work, keeping it all running, and saving us after deleting our home directories. Speaking of home directories, I beg you to never look into the gargantuan gargoyle that mine has become until I can finish with it and I finally can ask you kindly (but urgently) kill it with fire.

Tina, stay awesome and thanks for the fun times! You are just a great person and I hope to have many more Disney/BBQ nights. To my roomy Laura, thank you for the friendship. We had some great walks and hope to have more. I wish you all the best in your project and I am proud of your resilience. I respect your positive outlook and perseverance. To my other roomy Xin, good luck with everything and trust in yourself! Julian, good luck with you political ventures and I hope you find your way towards becoming a supervisor. Sam, stop eating noodles for breakfast. And also, thanks to the other PhDs and group members I had great times with: Peter, Jeroen, Bas, Arpit, Thea, Jaap, Laurens, Sarah, Petros, Daniel, David, Erdem, Lingyi, and Nikos!

Then there were some people that started and finished at the TBB before me and I looked up to as my seniors, or I often had questions for during the first months when everything was still scary and new. Jolien, thank you for all the help when we were still roommates, but also after you left. I really like our conversations about future roads to take. Also, thanks to

Eelco, Leny, Hilje, Bram, and Bastiaan. You all made it look easy. Wish you all the best, I see cool things happening in your scientific careers.

Two (then) students I want to thank for their great work during their internships, Ianthe and Tristan. May we meet again!

Thank you all for making the experience at TBB awesome!

To my good friends that all were examples for me during my PhD. Johanna and Fredrik, thanks for all the input, wisdom and great game nights. Fredrik, I always try to stop, think, "hmm", and then speak, as per your example. Also, thank you for showing me ± six years ago that pressing *enter* after typing *ls* is not going to break the computer! Amir, I always hear your voice yelling at me about design. I did my best to think about alignments, fonts, and well-designed figures. Roland and Kim, your eye for detail and good work are an inspiration. Yuki and Zoli, I hope to see you and your beautiful kids again soon! Brecht, thank you for helping me see that art is not so strict and that I should let loose a little. It really helped with making my cover.

Mijn allerliefste vriendinnen die ik al vanaf mijn jeugd in Curaçao naast me heb staan. Christel, dank je voor al je onuitputtelijke interesse in mijn werk en hartstikke leuke en lange gesprekken die we hebben over de wetenschappelijke onderwerpen van mijn thesis, koralen, maar ook over het leven, liefde en nog veel meer. Dominique, die eerste tijd in Amsterdam was heel fijn om liefde en leed te delen in de grote boze stad waar we moesten zien te overleven in onze studies. Kijk waar we nu zijn, zoveel jaar later! Ik heb veel lol in mijn gesprekken met jou en Rick die heel vaak gaan over, o.a., biologie en dino's.

Dan mijn familie die ervoor heeft gezorgd dat ik een zachte landing had in het koude kikkerland. Zonder jullie had ik hier waarschijnlijk niet gestaan. Jullie vingen me op toen ik mijn thuis moest achterlaten en accepteerde me als een van jullie. Ik herinner mijn aankomst op Schiphol en Churchilllaan nog als de dag van gisteren. Lieve familie Henny en Mulder, dank voor jullie warmte en een thuis weg van thuis. Ook dank aan de familie Wammes, Deenik, Deutekom, van de Linden en Zahradnik die er ook altijd voor me zijn.

Lieve pap en mam, de afstand is soms verschrikkelijk. Maar ik hoop dat ik jullie trots heb gemaakt met dit boekje. Jullie hebben me alle liefde gegeven die jullie konden geven en me geleerd hoe ik moet liefhebben, hard werken en doorzetten. Mede daardoor ligt dit boekje nu voor jullie.

Wes, je was er al bij tijdens mijn bachelor, toen mijn master en nu mijn promotie. Ik denk dat ik de laatste jaar niet had overleefd als je niet zoveel begrip had voor de vele dagen en uren die ik achter de computer besteed. Ik ben bang dat onze Drakensteyn, inclusief alle inwonende draken, in puin had gelegen. Dank je voor al je geduld bij de duizenden verschillende versies van de voorkant van dit boekje die je hebt moeten bekijken. Dank je voor al je steun. Dank je dat je me altijd laat lachen. Ik hoop dat we nog heel veel jaren samen zullen blijven leren.