Transposable Element Evolution in Stramenopiles and Choanoflagellates

Chloe Demaine

School of Applied Sciences, University of Huddersfield

MSc Biological Sciences

I confirm that, unless indicated otherwise, I undertook all of the work presented in this report. Any work performed by other parties is fully acknowledged.

Abstract

Transposable elements are parasitic mobile elements that can transpose within host genomes and may have a large impact on the host evolution. Despite this, very few studies have looked into how transposable elements themselves have evolved within unicellular organisms, which make up the vast majority of organisms in existence. In this study, evidence is presented for selection for transposable element codon usage in both the closest relatives to Metaoza – choanoflagellates, and stramenopiles. Both choanoflagellates studied show evidence for selective accuracy whilst neither of the stramenopiles do. Selection appears to be stronger within the choanoflagellates than the stramenopiles. *Phaeodactylum tricorntum* transposable elements are likely to be evolving under mutation pressure rather than selection pressure but there is evidence to suggest *Thalassiosira pseudonana* transposable element's codon usage is evolving via selection. The most likely explanation for these findings is that the choanoflagellates have a larger effective population size than the two stramenopiles in this study.

Contents

Introduction

1.1 Background

Transposable Elements (TEs) are DNA sequences that have the ability to transpose around in the genome. They are ubiquitous (Flavell et al., 2007) and are usually detrimental to their host genome but they can occasionally convey a benefit to it. TEs are capable of making any host gene non-functional if they transpose into a key part of the host gene, and can therefore be very harmful to the host. However, TEs have been responsible for creation of novel host genes due via molecular domestication, which involves the host genome incorporating the TE (Miller et al., 1999). They have been shown to have an impact on evolution (Zhang et al., 2006; Jiang et al., 2004). There have been many studies investigating TE activity within multicellular organisms, but very few have focused upon TEs within eukaryotic unicellular organisms. Unicellular organisms make up the vast majority of the eukaryotes (Baldauf et al., 2013) and therefore only a small portion of eukaryote's TEs have been studied. There is little evidence to suggest that multicellular TEs are evolving under direct selection and published studies suggest TEs in multicellular organisms have evolved via mutation pressure (Rouzic et al., 2007). However, Jiang & Govers (2006) made a remarkable discovery in finding that TEs within unicellular *Phytophthora infestans* are evolving under selection pressure. Selection pressure may also be driving the evolution of TE codon usage within other unicellular organisms, and this study aims to discover whether this is the case.
Many studies have investigated TEs of *Drosophila melanogaster*, *Saccharomyces cerevisiae, Homo sapiens*, *Arabidopsis thaliana* and *Caenorhabditis elegans* (Carr et al., 2012). *S. cerevisiae* is unicellular but has evolved from a multicellular ancestor (Ratcliff et al., 2012; Ito & Kakutani, 2014) so its TEs have gone through a period in which the host species was multicellular. Therefore, TEs of a unicellular species that has not evolved from a multicellular ancestor have not been studied in great detail. Additionally, many of the species that have been studied are either Opisthokonta or Archaeplastida, which are only two of seven eukaryotic super groups. This study incorporates a third super group's TEs – Heterokonta. The group Heterkonta includes stramenopiles, Opisthokonta includes Metazoa, choanoflagellates and fungi and Archaeplastida include land plants. This study also incorporates TEs from choanoflagellates as choanoflagellates are the sister group to Metazoa (Carr et al., 2008) and are therefore the closest related unicellular organisms. This study will enable comparison of unicellular TE evolution in both species distantly related to and closely related to those already studied.

Fig. 1. Maximum likelihood phylogenetic tree of eukaryotes
This phylogeny was retrieved from Brown et al. (2018). Branch numbers indicate the bootstrap support value followed by the posterior probabilities for two sets of converged chains using the Maximum Likelihood model.

1.2 Transposable Elements

Transposable elements, which are parasitic mobile elements within host genomes, can be divided into two classes. Class 1 TEs are retrotransposons. These transpose in the genome via a RNA-intermediate. Reverse transcriptase then converts the RNA into DNA and the resulting DNA sequence is integrated into the genome (Finnegan, 2012). Their method of transposing is nicknamed copy and paste. Retrotransposons are further sub-divided into two divisions. Retrotransposons that are flanked by long terminal repeats (LTRs) are LTR-retrotransposons. The LTR is double stranded DNA that is repeated at each side of the retrotransposon, and is usually 250 – 600 base pairs long. Retroviruses also have LTRs (Hine & Martin, 2016). Other retrotransposons that are not flanked by LTRs are subsequently called non-LTR-retrotransposons. Class II elements are referred to as transposons and can be flanked by inverted terminal repeats (ITRs). They transpose without an RNA intermediate, and unlike class I elements, use the enzyme Transposase to facilitate their movement around the genome. This enzyme is responsible for both excising the element from the genome and integrating it at a different genome site. This method of transposing is often called cut and paste. When a TE, either class I or class II, is inserted into the genome, a target site

duplication (TSD) is generated. This arises from a short sequence of host DNA being duplicated, and each TE family shares a TSD of a similar length (Sinzelle et al., 2009). LTR elements can undergo recombination with one another which leaves a solo LTR. The recombination event is often between the two LTRs of the same element (Shirasu et al., 2000). If the LTR recombination was between the LTRs of the same element, the TSD will be identical on the 5 prime and 3 prime side (Zhang et al., 2015). LTR sequences aren't always completely conserved between different elements. There may also be alternative reasons why LTRs are divergent from one another, other than single substitutions. One reason could be the occurrence of biased gene conversion, which involves the sequence from one LTR replacing the sequence from another LTR. This biased gene conversion could change an LTR to look more like an LTR from a different element. If the 5' looks more similar to an LTR from a different element it may be due to biased gene conversion (Lesecque et al., 2012).

The life cycle of a TE involves transposing and increasing copy number until they become inactive and become TE fossils (Wicker, 2004). Inactivity can result from accumulation of mutations that prevent the TE from transposing, or it can result from host defence mechanisms. Cells have evolved defences against TEs and can inactivate the TEs via DNA methylation or small RNA pathways (Casacuberta & Santiago, 2003).

Within each class of TE, there are many variations of TE. Some LTR-retrotransposons of particular interest in this study are *gypsy*-like and *copia*-like elements. *Copia*-like elements have several highly conserved domains, in the typical order of Protease, Integrase, Reverse transcriptase and Ribonuclease H (Berg & Howe, 1989). *Gypsy*-like elements are similar to *copia*-like elements in terms of their conserved domains, except that they have the Integrase at the 3' end of their Ribonuclease H domain. The Ribonucleoase H protein is responsible for degrading the original RNA strand and for removing the primers on the newly synthesised strands (Cerritelli & Crouch, 2009). This domain has also being studied using phylogenetic analysis. By dating the TEs based on their Ribonuclease H domain, it appears that LTR retrotransposons are younger than the youngest non-LTR retrotransposons, suggesting non-LTR retrotransposons evolved first (Malik, 2005). *Chromoviruses* have been found to be a lineage of *gypsy*-like TEs and have a chromodomain found at the C-terminal of the sequence. It is thought that the chromodomain may enhance insertion site specificity (Gao et al., 2008). *Chromoviruses* are abundant in fungi and have been found in choanoflagellates and *C. owczarzaki*, which is a filasterean (Carr & Suga, 2014). Fliastereans are unicellular opisthokonts that are the sister group to choanoflagellates (Shalchian-Tabriz et al., 2008). *Chromoviruses* have also been found within amphibians, reptiles and fish but there is no evidence for presence of *chromoviruses* within any other metozoan lineage (Gorinsek et al., 2004). It has been theorised that, as *chromoviruses* are present in choanoflagellates, *chromoviruses* may have been present in the genomes since the last common ancestor of holozoans (a group of organisms that includes Metazoa, Filasterea and choanoflagellates) or even opisthokonts. There would therefore have been loss of *chromoviruses* across many metazoans (Carr et al., 2008).
Major groups of non-LTR retrotransposons include Long interspersed nuclear element-1 (LINE-1) and Short interspersed nuclear elements (SINEs). Whilst it has been estimated that two thirds of the human genome is TE derived (de Koning et al., 2011), LINE-1 is the only active autonomous TE in humans, and its activity has been found to cause many various genetic diseases (Hancks et al., 2016). TEs can also be autonomous or non-autonomous (McClintock, 1950). Autonomous TEs are mobile and encode a Transposase enzyme, whereas non-autonomous TEs do not. These can only transpose if an autonomous TE is

present (Wang et al., 2015). L1 elements have been found within *C. owczarzaki*, plants and animals (Suga et al., 2013; Schwarz-Sommer et al., 1987; Yamashita & Tahara, 2006; Papenfuss et al., 2007) and the first evidence that L1 can undergo horizontal transfer has been recently found within the cattle genome (Ivancevic et al., 2018). It is likely that L1 has played a large role in the evolution of mammals, including humans, due to its abundance. L1 expresses two proteins: ORF1 and ORF2. ORF2 protein contains Endonuclease and Reverse transcriptase, and ORF1 protein has a chaperone activity (Moran et al., 1996; Martin et al., 2005). The SINE family Alu is also present within the human genome but is not autonomous and requires L1 ORF2 protein in order to transpose (Wallace et al., 2008).
Mutator-like elements (MULEs) are major transposons and were first discovered in maize (Robertson, 1978). There are both autonomous (MuDR) and non-autonomous (Mu) MULEs. MULEs are widespread across Opisthokonts and Archaeplasdida (Neuvéglise et al., 2005; Marquez & Pritham, 2010), including being present in *C. owczarzaki* (Suga et al., 2013) and the choanoflagellate *S. rosetta* (Carr & Suga, 2014). MULEs have also been identified in eukaryotic supergroup Excavata – in the Parabasalid *Trichomonas vaginalis* (Lopes et al., 2009). This is evidence for MULEs existing in species other than plants and opisthokonts, suggesting that MULEs may also exist in other eukaryotic super groups such as Heterokonta.
PiggyBac is another transposon and was first extracted from the metazoan *Trichoplusia ni* in order to be used as a vector (Fraser et al., 1983). *Tigger* elements are also transposons and inactive *Tigger* elements have been identified to be present in the human genome (Arian et al., 1996). *Tigger* elements have also been identified in the horse genome (Paulis et al., 2004) and in the choanoflagellate *Salpingoeca rosetta* (M. Carr, Personal Communication, 18 November, 2017). This means that *Tigger* elements may have been present in opisthokonts since before the divergence of choanoflagellates and Metazoa but it does not provide any evidence for the origin of *Tigger* elements.

1.3 Transposable Element Evolution

Within both multicellular and unicellular organisms, there is evidence of selection on the host against TEs due to their deleterious effect upon their hosts (Pereira, 2004). When there is a lack of purifying selection, TEs can transpose rapidly (Roberston, 2002). This may occur when a TE enters a new genome via horizontal transfer as the host will not yet have evolved genome defences against the TE. An example of this is the TE *rider*. It is likely to have transferred into the tomato genome from *Arabidopsis* and rapid transposition lead to there being around 2000 copies (Cheng et al., 2009).
Whilst there is evidence for selection on TEs, few studies have investigated TE codon usage selection within multicellular organisms. Codons are made up from three bases, and multiple combinations of nucleotide bases can make up the same codon. The third base in the codon can vary and still form the same codon in eighteen amino acids. For example, tyrosine can be made up from the bases UAU or UAC. This is the degeneracy of the genetic code (Barnett et al., 1967). Organisms often preferentially use certain combinations over others, which is codon bias (Komar, 2016). Codons that a species uses preferentially are called optimal codons. Genomes are often either AT rich or GC rich. This is possible because the genomes may preferentially use codon combinations that often end in AT or in GC (Quax et al., 2015). In multicellular organisms, TE codon usage may not be driven by selection pressure because multicellular organisms have relatively small effective populations compared to unicellular organisms, and genetic drift overcomes selection in small effective populations (Lynch, 2006). However, TE's codon usage have been found to evolve via selection in some unicellular genomes such as in the stramenopile *P. infestens* (Jiang & Govers, 2006). In

species with large effective populations, it may be necessary for TE families to evolve under selection in order to continue transposing. Investigating codon bias of TEs will indicate if the TE's codon usage is evolving under selection, mutation or a combination of the two pressures.

Selective evolution of TE codon bias could be derived from selective efficiency, selective accuracy or both. Selective efficiency is only seen in genes that are highly expressed. During translation, tRNAs carry anti codons which bind to the codons on mRNA. Some tRNAs may be more abundant than others. For example, the tRNA carrying the anticodon to UAU to form tyrosine may be more abundant that the tRNA carrying the anticodon UAC to form tyrosine. Therefore proteins would be produced faster, and more efficiently, if the species contained more UAU codons than UAC codons. It confers a benefit to the organism to contain more of these codons, and so the genome evolves codon bias under selection (Ehrenberg & Kurland, 1984). TEs may have evolved under selection efficiency as using the codons that bind to the anticodons of the host's most abundant tRNA would also allow the TE proteins to be produced faster and more efficiently. Selective accuracy involves improved translational accuracy. Selective accuracy comprises of selection for correct protein folding and function, as oppose to selection for quicker protein production, though the two phenomenon's can coincide and are not mutually exclusive. This causes functionally and structurally important regions to gain non-synonymous substitutions at a slower rate than regions that are less important structurally and functionally (Akashi, 1994). Therefore, it is likely that functionally important regions will also have a higher proportion of optimal codons than regions with less functional importance as non-optimal codons are more likely to be mistranslated than optimal codons (Drummond & Wilke, 2008). It could therefore be predicted that TE families that use a higher proportion of optimal codons will also show more evidence for selective efficiency and selective accuracy. It would also stand to reason if TE that are expressed more would have more evidence for selective efficiency, as this would improve the speed of translation and therefore the TE protein would be produced more. However, as gene expression levels were unavailable for this study, this hypothesis cannot be tested here. Another prediction based on gene expression levels would be that the more abundantly expressed, the more tRNA molecules would be present for translation of the TE. This is because the more abundant levels of tRNA are, the faster the TE protein can be translated and the more it will be expressed. Again, without gene expression levels, this cannot be tested within this study.

If TE codon bias have evolved via selective accuracy, it would be expected that more functionally important sections of the TE will have higher codon usage bias than sections that are less functionally important. Therefore, regions of the TE that make up domain sequences would be expected to use a higher proportion of host optimal codons than regions of the TE that do not make up domain sequences. As oppose to this, it would be expected that if the TE codon usage evolution was driven more by selective efficiency than accuracy, then a similar level of host optimal codons would be used by both functionally important and less functionally important regions of the TE. This is because selective efficiency acts to speed up the process of translation regardless of how functionally important the sequences are, whereas selective accuracy is more important within sequences that must be translated correctly in order for the protein to function properly. Although it is expected that there will be little difference between domain regions and non-domain region's codon bias for TEs whose codon usage evolution is driven by selective efficiency, it would be expected that there would be a large proportion of host optimal codons used throughout the TE sequence.

Whilst TE codon usage may evolve via selection pressure or mutation pressure, TEs themselves can evolve in a variety of way, one of which is horizontal transfer. This is the

transfer of genetic material from one organism to another, through any method other than from parent to descendants, which is vertical transfer. There can be a variety of ways in which horizontal transfer takes place. For example, TEs can undergo host-parasite transfer. This involves a TE transferring from a parasite to the parasite's host and has been witnessed in metazoans (Pace II et al., 2010). This helps the TE avoid extinction as remaining in a single host increases the chances of selection acting upon the host, providing it with the defences it requires to rid itself of the TEs (Schaack et al., 2010). TEs can also use predator-prey transfer as a form of horizontal transfer. TEs may be transferred from an organism to another, from an organism that the receiving organism eats. Without horizontal transfer, TEs are likely to become extinct through mutational changes, host selection and genetic drift (Hartl et al., 1997). However, LINE-1 TEs have survived in mammals for over 100 million years with what was thought to be vertical transfer only (Khan et al., 2006; 2005). This could be because TEs can avoid extinction in ways other than horizontal transfer. The most obvious method is by eliciting minimal harm to the host. This reduces the host's selection effects upon the TEs (Schaack et al., 2010), though LINE-1 TEs have caused human genetic disease suggesting this mechanism to prevent extinction may not be prevalent in LINE-1 TEs. Additionally, as recent evidence has surfaced that there has been horizontal transfer of LINE-1 within mammals (Ivancevic et al., 2018), there may be more horizontal transfer yet to be observed that could explain how the TE has survived for so long within mammals. LINE-1 elements could be jumping between species in order to survive within mammals. However, the evidence Ivancevic et al. (2018) found of horizontal transfer of LINE-1 elements did not appear to be recent. If the rate of transposition outweighs the selective disadvantage of the TE deleterious effects upon the host, this will also help the TE survive.
TEs can also become domesticated within their host, becoming a host gene after gaining a function that benefits the host. This process is called molecular domestication (Miller et al., 1992), and TE derived proteins often provide functions such as transcription regulation (Sinzelle et al., 2009) or domesticated TEs may beome transcription factors (Feschotte, 2008). These genes no longer possess ITRs, can usually no longer transpose and are present in the host as only single copies, as oppose to TEs which usually have multiple copies within a host (Sinzelle et al., 2009). An example of a TE derived domesticated protein is the Jerky protein, which is found in vertebrates, including humans. It has evolved from a *pogo* element and has a gene regulatory function, controlling the availability of mRNAs in neurons (Liu et al., 2003) as well as playing an important cellular function (Donovan et al., 1995) which may be linked to the cellular stress response system (Liu et al., 2003).

1.4 Stramenopiles

Stramenopiles are a particular focus of this study as the one organism in which there is evidence for selection on TE codon usage is the stramenopile *Phytophthora infestans* (Jiang & Govers, 2006). *P. infestans* is a plant pathogen that causes potato blight and was one of the factors causing the Irish potato famine in the mid 1800's (Bourke, 1964). There are many other *Phytophthrora* species that are plant pathogens with great consequence for humans (Haas et al., 2009). However, no other stramenopile TE codon usage has been investigated for evolving under selection. Therefore, it is currently unknown if TEs within stramenopiles universally show selection for codon usage or if *P. infestans* is unique in doing so.
Stramenopiles have three particular features that define them. At some point in their life, each stramenopile has two flagella. One is smooth and one is a tinsel flagella, which has distinctive flagella hairs (Walker & van West, 2007). Another feature is that stramenopiles have chlorophyll a and c but lack chlorophyll b and many have a double layer of endoplasmic

reticulum around their chloroplast (Hine & Martin, 2016). Stramenopiles can be either photosynthetic or heterotrophic. The photosynthetic stramenopiles have been shown to cluster together under phylogenetic studies, whilst the heterotrophic stramenopiles are in basal lineages, but further phylogenetic analysis is required to identify the correct phylogenetic position of stramenopiles (Fletcher et al., 2016).
Stramenopiles can be divided into the clades Bigyra and Gyrista. *P. infestans* and the two stramenopiles investigated within this study fall into the clade Gyrista. Gyrista can then be further subdivided into Oomycota (pseudofungi), which is the clade into which *P. infestans* falls, and Ochrophyta. Within the many subclasses of Ochrophyta, the clade Diatomeae lies. This is the clade in which both *P. tricornutum* and *T. pseudonana* reside (Derelle et al., 2016).

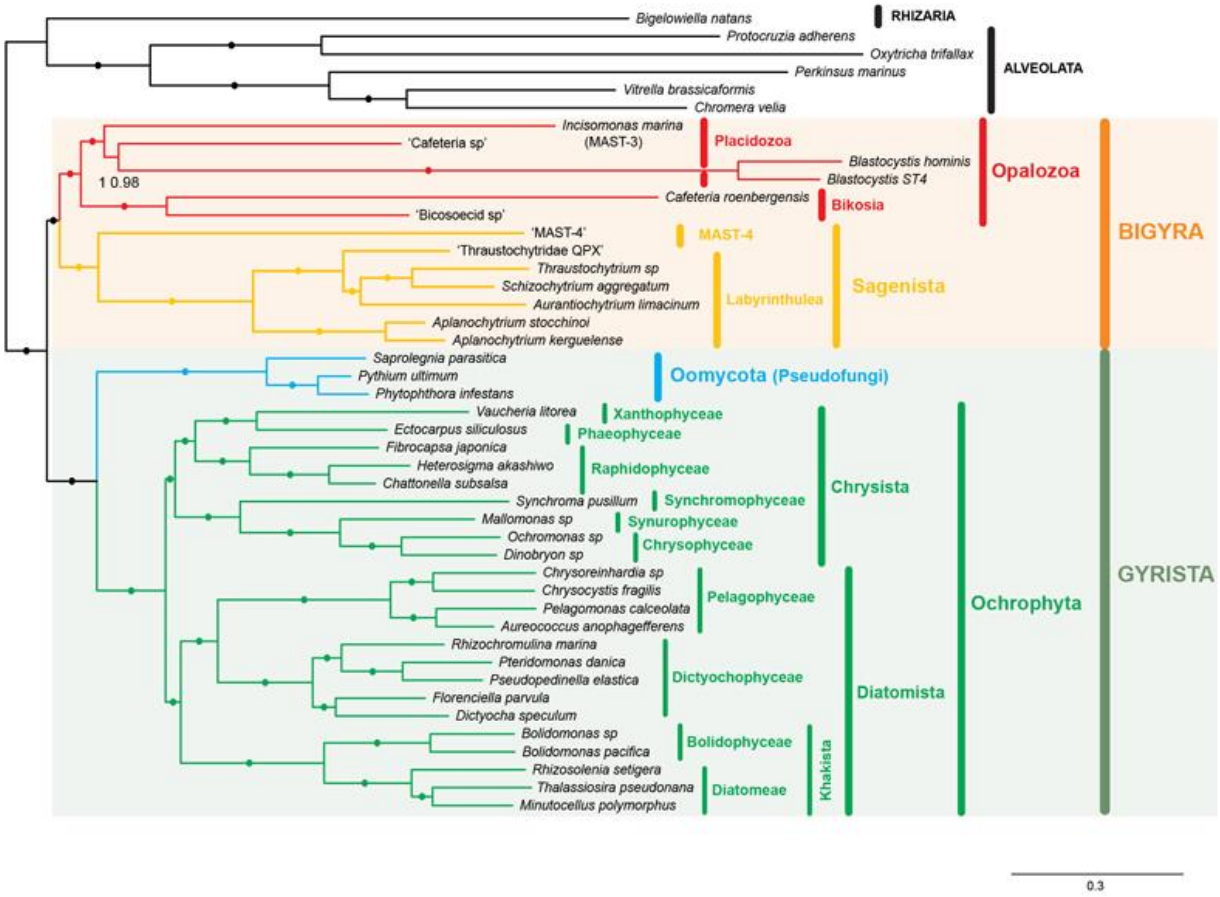


Fig. 2. Bayesian phylogenetic tree of stramenopiles
This phylogeny was obtained from (Derelle et al., 2016) and is rooted on Rhizaria and Alveolata. Bayesian inference posterior probability is denoted to the right of diverging branches and support of 1.00 is denoted by a bullet.

P. tricornutum and *T. pseudonana* are both free-living marine diatoms. There are between 10,000 to 100,000 diatom species (Norton et al., 1996), and diatoms may be responsible for around 20% of global productivity due to their carbon fixating abilities (Falkowski et al., 1998). The divergence rate of diatoms is very high, relative to that of yeast and metazoa. Wyrwicz et al. (2008) suggest that TE evolution could be a contributing factor to this, along with selective gene family expansions. Diatoms can be further subdivided into two classes: pennate and centric diatoms. *P. tricornutum* is a pennate diatom (Wyrwicz et al., 2008) whereas *T. pseudonana* is a centric diatom (Armbrust et al., 2004). These two classes have

very different genomes and centric diatoms are the older of the two, aging at least 180 million years, with pennates being at least 90 million years old (Wyrwicz et al., 2008). Centric diatoms are typically radially symmetrical (symmetrical about a central axis), whereas pennate diatoms are usually elongated and have striae that are parallel to one another, arranged so that the diatom is symmetrical along the long axis. Striae are rows of holes in between the lines of silica in the diatom cell wall, and the lines of silica are termed the costae. The cell wall is called the frustule and consists of two halves that come together to enclose the cell within. Each half is called a theca. Frustules are very diverse and have been utilised in many different ways (Parkinson and Gordan, 1999) such as in drug delivery, enzyme immobilisation and separation of heavy metals (Yuan et al., 2013; Poulsen et al., 2007; Bariana et al., 2013).

1.5 Choanoflagellates

Choanoflagellates are the sister group to Metazoa (Carr et al., 2008). Studying them conveys a greater insight into the evolution of both the choanoflagellates themselves and Metazoa, as well as the origins of Metazoa. Choanoflagellates live in water, both marine and freshwater, are unicellular and have a collar of microvilli and a flagellum. This flagellum stirs the water, creating currents which cause bacteria to move towards the choanoflagellates and into the choanoflagellate's collar for phagocytosis. In free living choanoflagellates, the flagellum is also used to propel the organism through the water (King, 2005).

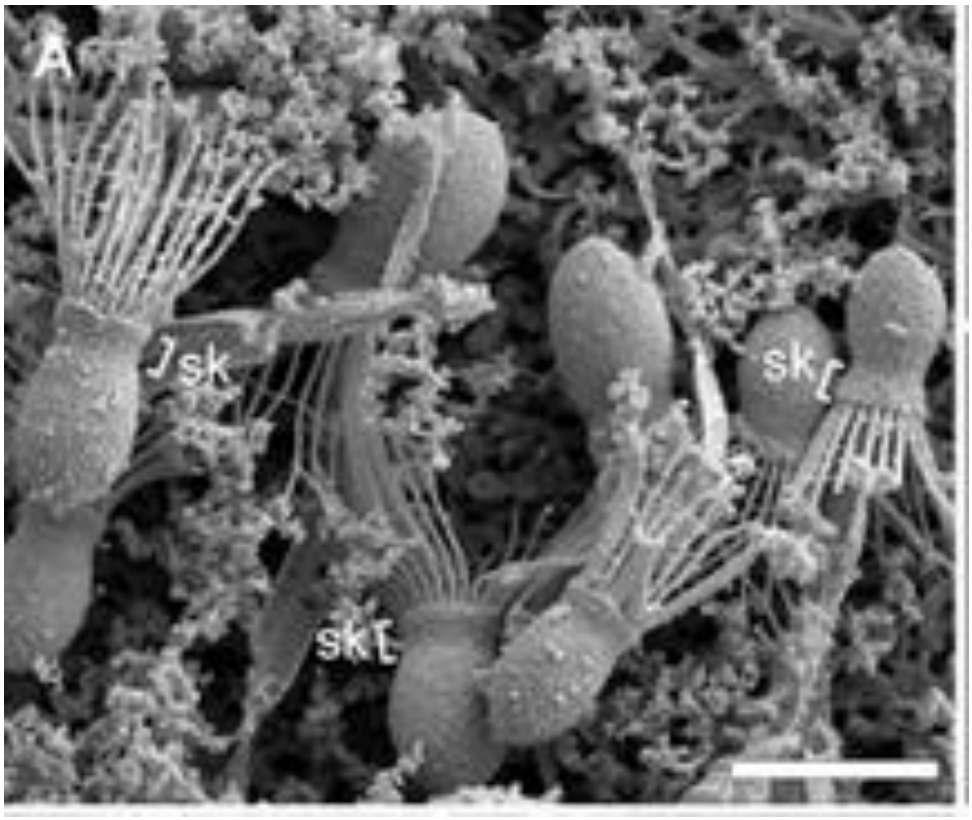


Fig. 3. Collar morphology of *Monosiga brevicollis*
This image was obtained from Mah et al. (2014) and scanning electron microscopy was used. Here there are several *M. brevicollis* organisms where it is visible that the collar of microvilli emerges from the skirt (sk).

In the past, choanoflagellates have been divided up into three classes, based on morphological characteristics. These are Codosigidae, Salpingoecidae, and Acanthoecidae. Acanthoecidae choanoflagellates have a lorica made up of silica strips into a basket like formation (Norris, 1965). The function of the lorica is still unknown (Asadzadeh et al., 2019). The Salpingoecidae choanoflagellates were described as having theca, which is a rigid membrane that is found in a variety of ways. Different morphologies of the theca include the flask, the cup and the tube (Carr et al., 2017). Codosigidae choanoflagellates are often referred to as 'naked cells' although they are covered in a fine membrane called the glycocalyx. However, multiple studies (Cavalier-Smith, 2002; Medina et al., 2003) found that neither Codosigidae nor Salpingoecidae were monophyletic in phylogenetic studies. Instead, choanoflagellates are now split into two classes. These are Craspedida and Acanthoecida. Craspedida choanoflagellates are split into three clades (Carr et al., 2008).

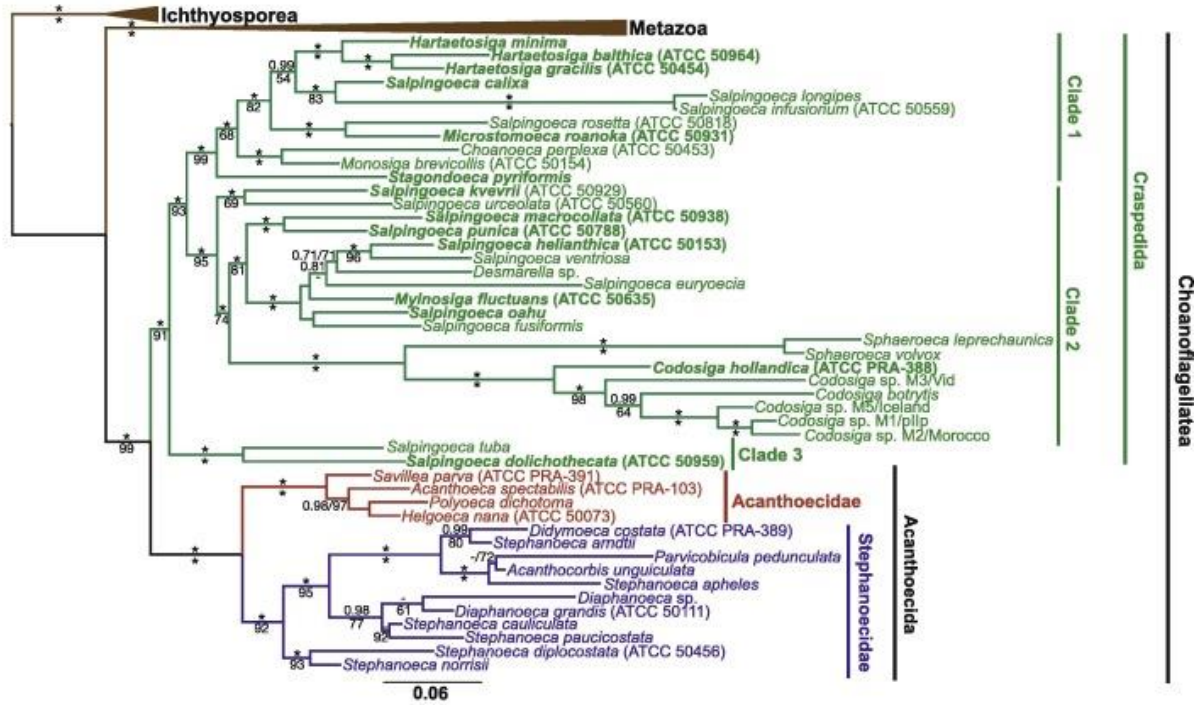


Fig. 4. Maximum likelihood phylogeny of choanoflagellates
This phylogenetic tree was obtained from Carr et al. (2017) and support values of 1.00bbPP and 100% maximum likelihood bootstrap percentage are denoted by an *. Branches with a support value of less than 0.7bbPP and 50% maximum likelihood bootstrap percentage were omitted.

Choanoflagellates have many morphological similarities to choanocytes (collar cells) of sponges, despite there being at least 600 million years of evolution separating them (Mah et al., 2014). Both are of a similar size, with choanoflagellates being 1.2 to 10µm in length and choanocytes being 2 to 10µm in length. Both have a microvilli collar, a flagellum and both have glycocalyx on the surface of their cells (Mah et al., 2014). Other metazoans have also been identified as having similar collar cells now, including echinoderms (Crawford & Campbell, 1993). It has been suggested that this is the missing link between metazoans and their unicellular ancestors (Cavalier-Smith, 1993).
The TEs of two choanoflagellates in particular have been investigated in this study. One is *Mlynosiga fluctuans*, which was once given the name *Monosiga ovata*, and the other is *Diaphanoeca grandis*. Carr et al. (2017) found that *Monosiga ovata* and *Mlynosiga flucutans* were two different species as the species described as *Monosiga ovata* by Saville-Kent (1880) was a marine organism whereas *Mylnosiga fluctuans* is found in freshwater. The morphology

is different too. *Monosiga ovata* has a short peduncle whereas *M. fluctuans* does not. *M. fluctuans* is in clade 2 of the Craspedida (Carr et al., 2017). *D. grandis* has a lorica (Manton et al., 1981) and is an Acanthoecida choanoflagellate (Carr et al., 2017).
Whilst choanoflagellates are the closest related holozoans to Metazoa, Filasterea are another clade of holozoans and these are the sister group to choanoflagellates and Metozoa. Many genes previously thought to be present in metazoans alone have been identified within both choanoflagellates and filastereans (Hehenberger et al., 2017; Sebé-Pedrós et al., 2016). Genes found within choanoflagellates and filastereans are likely to be ancestral to both clades, and therefore to Metazoa. This may even be the case if such genes are absent from metazoan genomes (Southworth et al., 2018). Investigating choanoflagellates, and TEs within choanoflagellates, may reveal more about the unicellular ancestry of Metazoa and differences between TEs residing in relatively closely related unicellular and multicellular organisms.

2.0 Study Aims:

The aim of this study is to identify whether TE codon usage evolves via selection in any unicellular organism other than *P. infestens*. If there is evidence for selection for TE codon usage, I would like to investigate whether it is selection accuracy, selection efficiency or both. This study also aims to gain a greater insight into the evolution of the TE themselves, within choanoflagellates.

3.0 Methods:

3.1 Stramenopile Codon Usage

The entire genome for *P. tricornutum* and *T. pseudonana* were downloaded from the EnsemblProtists database. The forms used were:
Thalassiosira_pseudonana.ASM14940v2.cds and
Phaeodactylum_tricornutum.ASM15095v2.cds. The contigs and proteins for *M. fluctuans* and *D. grandis* were obtained from Daniel Richter (Richter et al. 2018). Codon usage statistics, including GC3s, effective number of codons (Nc) and frequency of optimal codons (Fop) were generated using CodonW (Peden, 2000). GC3s is the fraction of codons that have guanine or cytosine in the third position. Nc shows the general level of codon bias and generates a number from 20-61. This is the number of combinations of codons that have been used to make the amino acids, with 20 being the minimum possible combinations of codons used and 61 the maximum. Fop is the ratio of optimal codons to synonymous codons and requires an input of the species optimal codons. This was obtained from M. Carr (personal communication, June 22, 2018).
Relative synonymous codon usage was found within the genomes using CodonW, which generated a Fop.coa file. Optimal codons of the stramenopiles were found using the hilo.coa files generated by CodonW following finding the Fop values and Nc values for each TE. The hilo file identified the most commonly used codons for each amino acid for the stramenopiles, which were assumed to be the optimal codons. Optimal codons of the choanoflagellates was found by Juan José Ginés (Personal communication, June 24, 2018), also using CodonW.

3.2 tRNA Gene Screening

The software tRNAscan-SE 2.0 (Lowe & Eddy, 1997) was used to identify tRNA genes within the choanoflagellate contigs. Default settings were used as was the online server (Lowe & Chan, 2016).

3.3 Optimal Codon Usage in Domain regions and Non-domain regions

Each stramenopile TE family sequence was divided into functional domain regions and non-domain regions. Domains were identified by using Blastp with default settings and the database set to Non-redundant protein sequences (nr) (Altschul et al., 1990). The TE protein sequences were inputted into Blastp to identify any domains. These domains were then selected and the beginning and end of these sections were identified. Everything within these sections were classed as a domain region and sections of protein sequence that were not included within these domain sections were classed as non-domain regions. As Blastp is very reliable for characterising sequences it can be assumed that the domain regions and non-domain regions identified this way are likely to be accurate. CodonW was used to generate Fop values and Nc values for each TE's functional domain regions and non-domain regions.

3.4 GC Content in Non-coding and Synonymous Sites

Each stramenopile TE family sequence was divided into translated regions and untranslated regions. The translated regions included all coding regions which are translated into TEs. LTR sequences were not included within either section as they may contain sequences that are under selective constraint. NCBI was used to identify these sequences as NCBI identifies sections of the TEs that are long terminal repeats, gag, pol and sections in between these that are untranslated. CodonW was used to generate Fop values and Nc values for each TE's translated regions and untranslated regions. GC content was not compared in choanoflagellate non-coding and synonymous sites as non-coding DNA could not be identified.

3.5 Copy Number

The software Blast trace (Altschul et al., 1990) was used to identify other copies of TEs within each stramenopile. The database used was *Phaeodactylum tricornutum*-WGS and *Thalassiosira pseudonana* - WGS depending on which stramenopile the TE family was within. The TE LTR sequence was used as the input. The outputs were then downloaded and the LTR within each Blast hit was identified manually, and the target site duplication (TSD) found manually. These were listed and copy number calculated based on the number of different TSDs within each TE family. It is likely that the copy number has been underestimated rather than overestimated as when there was any doubt that a TSD was genuine then it was excluded. This was done because including LTRs that were not genuine would falsely increase the copy number, whereas choosing to not include LTRs where it was uncertain that they were genuine may not always decrease the estimated copy number if that LTR was also identified elsewhere in the TE. Another reason the copy number could have been underestimated is that the Blast hits with the least similarity to the query sequence may have be partial sequences due to a deletion mutation, and therefore the LTRs identified may be genuine. However, there may be partial sequences due to some of the nucleotides not being in the sequencing read and the TSD identified may not be genuine. Due to this

ambiguity, the majority of partial sequences that could not be classified as 5 prime, 3 prime or solo with certainty were not considered as possible copies. Also, the first thirty to forty base pairs of the sequencing reads and around the last hundred base pairs are often poor quality. Therefore, if the TSD is located at the very beginning or near the end of the sequence, it is often difficult to ascertain what the true TSD is. Looking at the trace helped deduct if the nucleotides were genuine. Often only the 5' or the 3' LTRs were found. This could be because their corresponding LTR may have been in a Blast hit which was excluded due to the above reasons, or it could be that the genome sequencing didn't cover it. When only one LTR was found, its Blast Trace pattern was analysed to decide whether it was a genuine LTR sequence or not. If it was found likely to be so, it was including in the copy number count.
There is no trace archive for either *D. grandis* or *M. fluctuans* so their copy number could not be investigated.

3.6 Identifying Choanoflagellate TEs

The software Repeat Masker (Version open-4.0.6, http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) was used to identify potential TEs within both *D. grandis* and *M. fluctuans.* The choanoflagellate's contigs were used as the input. A list of potential TEs was generated and TEs with an E-value of more than e-5 was excluded. E-5 was used as TEs with an E-value more than this were not likely to be genuine TEs and their sequences only similar to that of a TE by chance. Blastp was used to examine the remaining protein sequences, with the database Non-redundant protein sequences (nr), and to identify whether they were genuine TEs. If they had domains of a TE such as reverse transcriptase and integrase, or were closely related to another sequence that had previously been identified as a TE then they were deemed as being likely TEs.

3.7 Phylogenetics

MAFFT (version 7: improvements in performance and usability, Katoh & Standley, 2013) was used to generate alignments of each stramenopile TE's copy's sequences. The output format was Pearson/FASTA with default settings. RAxML -HPC2 on XSEDE (Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies) was used to construct the phylogenetic trees. The parameters used included nucleotide as the data type, the maximum hours to run was 96, rapid bootstrap replicates were used, the Bootstrap Iterations were set to 1000 and the model for the bootstrapping phase was GTRCAT. The remaining parameters were default parameters. FigTree (Rambaut, 2007) was then used to view the trees. Sequences for the choanoflagellate phylogenetic trees were obtained by using Blastp with Non-redundant protein sequences (nr) as the database, using already identified TE's from the choanoflagellates as the input. tBlastn was also used to find sequences similar to the choanoflagellate TEs. The database used for this was Whole-genome shotgun contigs (wgs) and different taxonomic groups were specified. These were Alveolate, Amoebozoa, Apusozoa, Cryptophyta, Euglenozoa, Fornicata, Glaucocystophyceae, Haptophyceae, Heterolobosea, Jakobida, Oxymonadida, Parabasalia, Rhizaria, Rhodophyta and Stramenopiles. *Chromovirus,* MULE and L1 outgroups were obtained from Carr & Suga (2014). MAFFT was again used to generate alignments with the same settings and RAxML was used to construct the phylogenetic trees with the same settings. MrBayes on XSEDE was also used to construct phylogenetic trees. A mixed model was used as the matrix for amino

acids and the parameters were set to 5,000,000 generations with a sampling frequency of 1000. The first 25% of the sample values were set to be discarded as burnin, resulting in the consensus trees being generated from a total of 3750 phylogenetic trees. The type of consensus tree created was all compatible groups. Figtree was also used again to view the trees.

4 Results

4.1 Stramenopile Codon Usage

For each Stramenopile TE, codon usage bias was investigated using the frequency of optimal codons (Fop) and effective number of codons (Nc). Table 1. shows codon usage bias within *P. tricornutum* TEs and Table 2. shows codon usage bias within *T. pseudonana* TEs.

TE Family	Frequency of Optimal Codon (Fop)	Effective Number of Codons (Nc)
4.3	0.319	47.10
6.5	0.395	51.96
6.7	0.408	56.82
4.1	0.443	57.90
4.4	0.450	52.88
2.3	0.455	55.32
7.1	0.490	58.08
6.4	0.500	57.70
5.3	0.506	52.32
3.1	0.508	54.64
5.2	0.510	52.76
6.6	0.515	49.36
5.1	0.585	46.44
Average	0.468	53.33

Table 1. Codon usage statistics for the TE families of *P. tricornutum*

Fop in *P.tricornutum* TEs ranges from 0.319 in TE 4.3 to 0.585 in TE 5.1. Nc ranges from 46.44 in TE 5.1 to 58.08 in TE 7.1.

TE Family	Frequency of Optimal Codon (Fop)	Effective Number of Codons (Nc)
6.3	0.430	51.56
6.2	0.444	52.90
5.5	0.447	54.22
5.6	0.467	51.21
4.5	0.510	58.90
5.4	0.569	50.02
6.1	0.597	53.60
Average	0.495	53.20

Table 2. Codon usage statistics for the TE families of *T. pseudonana*

Fop in *T.pseudonana* TEs ranges from 0.430 in TE 6.3 to 0.597 in TE 6.1. Nc ranges from 50.01 in TE 5.4 to 58.90 in TE 4.5.

Fop values and Nc show *P. tricornutum* TEs have a wider range of codon usage bias than TEs within *T. pseudonana*. Average Nc values for the two stramenopile's TEs are very

similar, and whilst average Fop value for *T. pseudonana* TEs is higher than that of *P. tricornutum's* TEs, it is still similar.

The most frequently used codon for each amino acid within each TE was identified to see how often the most frequently used codon matches with the host optimal codon. This will indicate each TEs codon bias.

Amino Acid	4.1	5.1	5.2	5.3	3.1	4.4	4.3	2.4	7.1	6.6	6.7	6.4	6.5
Phe	UUU	UUU	**UUC**	UUU	**UUC**	UUU	UUU	UUU	UUU	UUU	UUC	UUU	UUU
Leu	UUG	CUC	CUC	CUU	CUC	UUG	UUG	CUU	CUU	CUC	CUU and **CUA**	CUC	UUG
Ile	AUU	AUU	**AUC**	AUU	AUU	AUU	AUU	**AUC**	AUU and **AUC**	AUU	**AUC**	**AUC**	AUU
Val	GUU	GUC	GUC	GUC	GUC	GUU	GUU	GUC	GUU	GUU	GUU	GUU	GUU
Tyr	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	UAU	**UAC**	**UAC**	**UAC**	**UAC**	**UAC**	UAU
His	CAU and **CAC**	**CAC**	**CAC**	**CAC**	**CAC**	**CAC**	CAU	CAU and **CAC**	**CAC**	**CAC**	CAU	CAU	CAU
Gln	CAG	**CAA** and CAG	**CAA**	**CAA**	**CAA**	CAG	**CAA**	**CAA**	**CAA**	**CAA**	**CAA**	CAG	**CAA** and CAG
Asn	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	AAU	AAU	AAU	**AAC**	AAU	**AAC**	**AAC**	AAU
Lys	**AAA**	**AAG**	**AAA**	**AAG**	**AAA**	**AAA and AAG**	**AAA**	**AAG**	**AAG**	**AAA**	**AAG**	**AAG**	**AAA**
Asp	GAC	GAC	GAC	GAC	GAU and GAC	GAU	GAU	GAU	GAC	GAC	GAC	GAC	GAU
Glu	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	GAG	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**	**GAA**
Ser	UCC	UCC	UCC and UCU	UCC	UCU	UCG	UCG	AGU	UCG	UCC	**UCA**	AGC	UCG
Pro	**CCA**	CCC	CCC	CCU	CCC	CCG	**CCA**	CCU	CCG	CCC	**CCA**	**CCA**	CCG
Thr	**ACA**	ACC	ACC	ACC	ACC	ACG	**ACA**	ACC	ACG	ACC	**ACA**	ACG	ACA and ACG
Ala	**GCA**	GCC	GCC	GCC	GCC	**GCA**	GCU	GCC	GCC	GCC	**GCA**	**GCA**	**GCA**
Cys	UGU	UGU	**UGC**	UGU	**UGC**	UGU	UGU	UGU	**UGC**	UGU	**UGC**	**UGC**	UGU and **UGC**
Arg	CGA	CGC	CGC	CGC	CGC	**CGG**	CGA	CGU	AGA	CGC	CGA and AGA	CGC	CGA
Gly	**GGA**	**GGC**	**GGC**	**GGC**	GGU	**GGA**	GGU	GGU	**GGA**	**GGC**	**GGA**	**GGA**	**GGA**
	7/18	15/18	16/18	12/18	13/18	8/18	2/18	9/18	12/18	11/18	8/18	11/18	6/18

Table 3. Host optimal codon usage for TEs of *P. tricornutum*

TEs within *P. tricornutum* use host optimal codons for each amino acid most frequently between two times for TE 4.3, to sixteen times for TE 5.2. Grey boxes indicate when the TE is using the host's optimal codon. Codons in bold indicate when a TE is using the host's most abundant tRNA codon.

Amino Acid	5.4	5.5	5.6	4.5	6.1	6.3	6.2
Phe	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**	**UUC**
Leu	CUC	CUC	CUC	CUC	CUU	**UUG**	**UUG**
Ile	AUC	AUC	AUC	AUC	AUC	**AUU**	AUC

Val	GUC	GUG	GUC	GUC	GUC	GUG	GUU
Tyr	**UAC**	**UAC**	**UAC**	UAU	**UAC**	UAU	UAU
His	**CAC**	**CAC**	**CAC**	CAU	**CAC**	CAU	CAU
Gln	CAA	CAA	CAA	CAA and **CAG**	CAA	CAA	CAA
Asn	**AAC**	**AAC**	**AAC**	**AAC**	**AAC**	AAU	AAU
Lys	AAA	AAA	AAA	AAG	AAG	AAG	AAA
Asp	**GAC**	**GAC**	**GAC**	**GAC**	**GAC**	GAU	GAU
Glu	GAA	GAA	GAA	**GAG**	**GAG**	GAA	GAA
Ser	UCC	UCA	**AGC**	AGU	UCU	AGU	UCA
Pro	CCA	CCA	CCA	CCA	**CCU**	CCA	CCA
Thr	ACC	ACA	ACA	ACU	ACC	ACA	ACU
Ala	GCC	GCA	GCA	**GCU**	**GCU**	GCA	**GCU**
Cys	UGC	UGU	UGU and UGC	UGC	UGC	UGC	UGU
Arg	CGA	CGA	CGA	AGA	CGU	CGA	CGA
Gly	GGC	**GGA**	**GGA**	GGU	GGU	GGU	**GGA**
	12/18	8/18	10/18	12/18	16/18	5/18	6/18

Table 4. Host optimal codon usage for TEs of *T. pseudonana*
TEs within *T. pseudonana* use optimal codons for each amino acid most frequently between five times, in TE 6.3, and sixteen times, in TE 6.1. Grey boxes indicate when the TE is using the host's optimal codon. Codons in bold indicate when a TE is using the host's most abundant tRNA codon.
In *P. tricornutum* there are more TEs with very low host optimal codon use (TE 4.3). Both stramenopiles have TEs with very high host optimal codon use: TE 5.2 in *P. tricornutum* and TE 6.1 in *T. pseudonana*.

4.2 tRNA Gene Screening

Codon W identified optimal codons for both stramenopiles but may have identified false optimal codons. Genuine optimal codons match the most abundant tRNA gene within each stramenopile so identifying the most abundant tRNA gene ensures that the codons identified as optimal are genuine.

Amino Acid	Most Abundant tRNA Gene	Number of tRNA Genes	Optimal Codon	Codons Not Present in TEs
Phe	TTC/GAA	4	UUC	TTT
Leu	CTA/UAG	9	CUC and CUG	CTC
Ile	ATC/GAU	3	AUC	ATA
Val	GTT/AAC and GTA/UAC	5	GUC and GUG	GTC
Tyr	TAC/GUA	2	UAC	TAT

His	CAG/CUG and CAC/GUG	2	CAC	-
Gln	CAA/UUG	3	CAG	-
Asn	AAC/GUU	3	AAC	AAT
Lys	AAG/CUU and AAA/UUU	4	AAG	-
Asp	GAC/GUC	3	GAC	GAT
Glu	GAA/UUC	3	GAA	-
Ser	AGC/GCU and TCA/UGA	6	UCC and UCG	TCC and AGT
Pro	CCA/UGG	5	CCC and CCG	CCC
Thr	ACA/UGU	4	ACC and ACG	ACC
Ala	GCA/UGC	6	GCC and GCG	GCC
Cys	TGC/GCA	1	UGC	TGT
Arg	CGG/CCG	6	CGU, CGC and CGG	CGC
Gly	GGA/UCC and GGC/GCC	6	GGC	GGT and GGG

Table 5. Most abundant tRNA genes and optimal codons for *P.tricornutum*
The most abundant tRNA within *P. tricornutum* for each amino acid most often matches the optimal codon. In the instances where this is not the case, the first two nucleotides always match.

Amino Acid	Most Abundant tRNA Gene in TE (Codon/Anticodon)	Number of tRNA Genes	Optimal Codon	Codons Not Present in TEs
Phe	TTC/GAA	3	UUC	TTT
Leu	TTG/CAA	6	UUG, CUU and CUC	CTC
Ile	ATT/AAU	3	AUU and AUC	ATC and ATA
Val[1]	GTT/AAC and GTA/CAU	5	GUC	GTC
Tyr	TAC/GUA	3	UAC	TAT
His	CAC/GUG	2	CAC	CAT
Gln	CAG/CUG	3	CAG	-
Asn	AAC/GUU	4	AAC	AAT
Lys	AAA/UUU and AAG/CUU	6	AAG	-
Asp	GAC/GUC	4	GAC	GAT

Glu[3]	GAG/CUC	7	GAG	-
Ser	AGC/GCU	8	UCU and UCC	TCC and AGT
Pro[2]	CCT/AGG	4	CCU and CCC	CCC
Thr[2]	ACT/AGU and ACA/UGU	5	ACU and ACC	ACC
Ala[2]	GCT/AGC	6	GCU and GCC	GCC
Cys	-	0	UGC	TGC and TGT
Trp	TGG/CCA	2	-	-
Arg[2,3]	AGG/CGU, CGA/UCG, AGA/UCU and CGG/CCG	4	CGU, CGC and AGG	CGT and CGC
Gly	GGA/UCC	7	GGA	GGT and GGG

Table 6. Most abundant tRNA genes and optimal codons for *T. pseudonana*
The most abundant tRNA within *T. pseudonana* for each amino acid most often matches the optimal codon. In the instances where this is not the case, the first two nucleotides of the codon always match.

4.3 Codon Usage in Domain regions and Non-domain regions

Fop values were generated for both domain regions and non-domain regions of each TE and compared. Differences in Fop between the regions suggest the codon usage is different in

them. Nc values were also calculated and compared in order to reduce any bias from misidentified optimal codons.

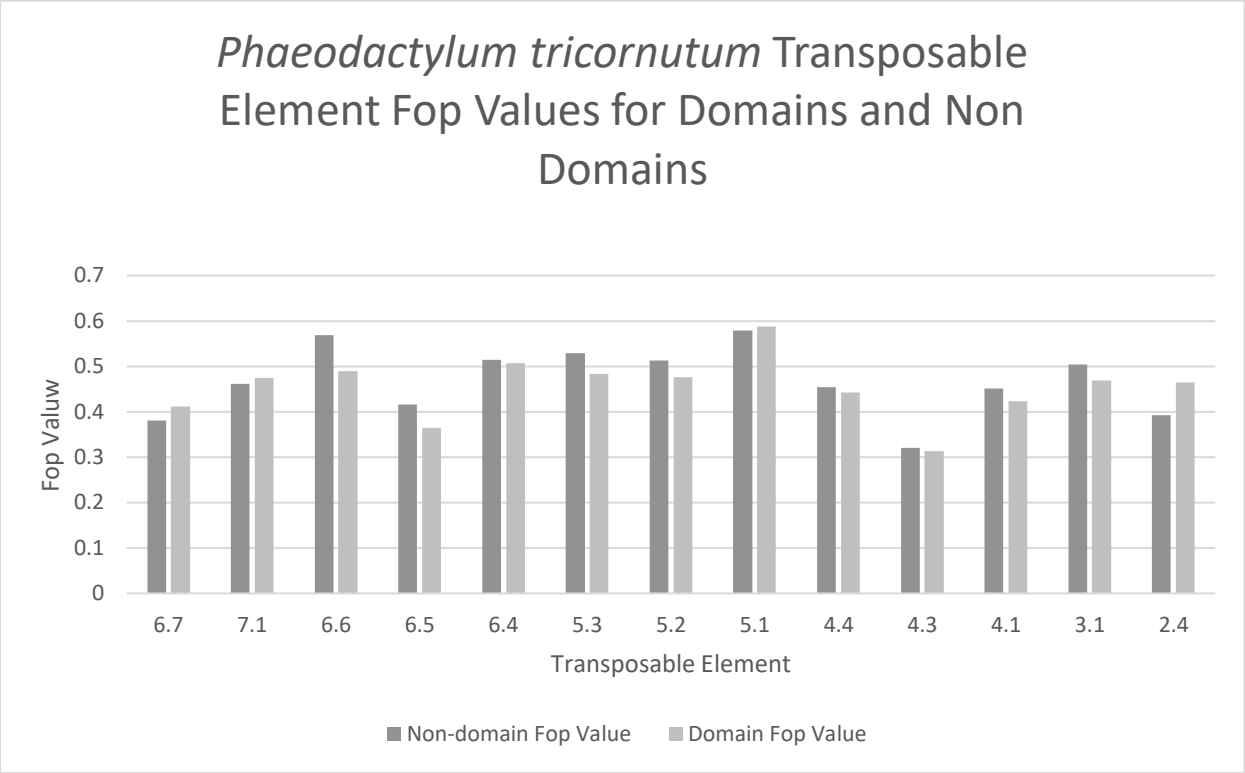


Fig. 5. TE Fop values in domain regions and non-domain regions of *P. tricornutum*
The Fop values of the domain regions and the non-domain regions were analysed using Fisher's exact test and it was found that none of them were significantly different as the p-value was larger than 0.05%. The Fop value in domain regions was higher than the Fop value of the non-domain regions in three TEs out of the thirteen in *P. tricornutum*.

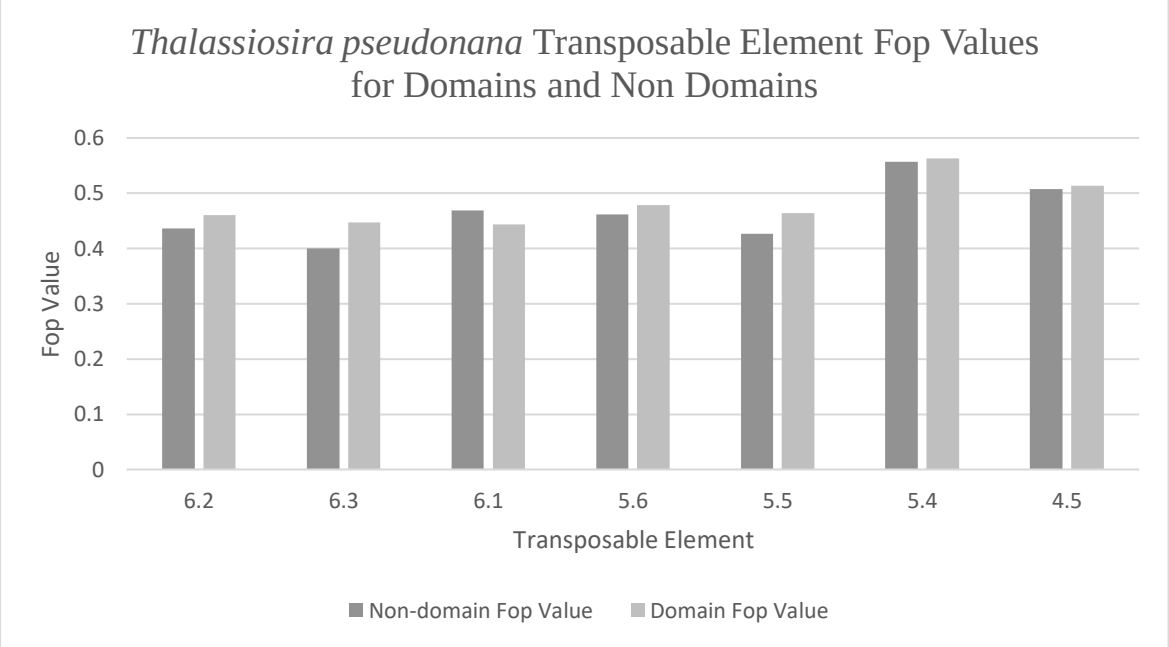


Fig. 6. TE Fop values in domain regions and non-domain regions of *T. pseudonana*
The Fop values of the domain regions and the non-domain regions were analysed using Fisher's exact test and it was found that none of them were significantly different as the p-

value was larger than 0.05%. The Fop value in domain regions was higher than the Fop value of the non-domain regions in five of the seven TEs within *T. pseudonana*.

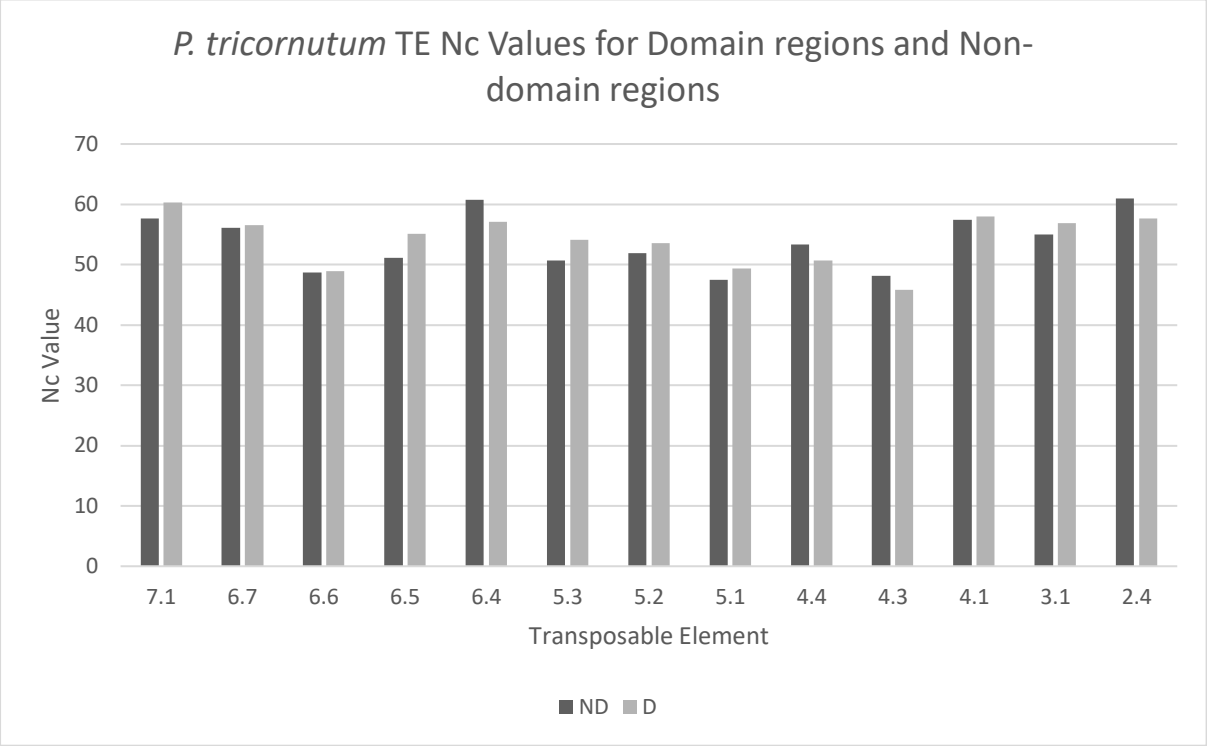


Fig. 7. TE Nc values in domain regions and non-domain regions of *P. tricornutum*
None of the Nc values of the domain regions were significantly different from the Nc values of the non-domain regions. This was calculated using Fisher's exact test and the p-values were higher than 0.05%. The Nc value was higher in domain regions in nine out of the thirteen TEs.

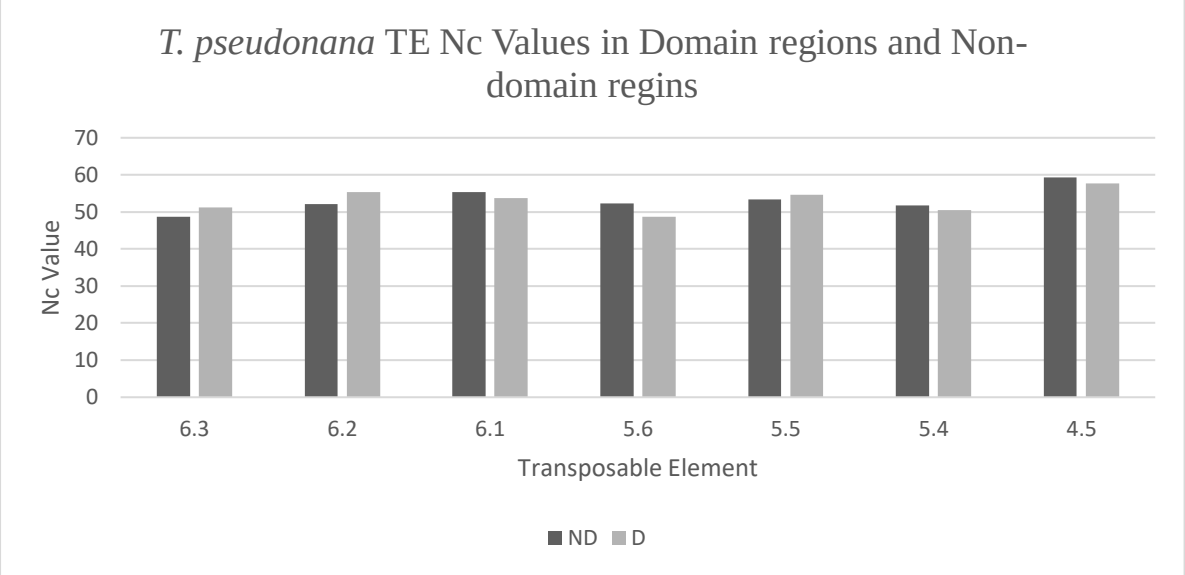


Fig. 8. TE Nc values in domain regions and non-domain regions of *T. pseudonana*
None of the Nc values of the domain regions were significantly different from the Nc values of the non-domain regions. This was calculated using Fisher's exact test and the p-values were higher than 0.05%. The Nc value was higher in domain regions in three out of the seven TEs.

4.4 GC Content in Non-synonymous and Synonymous Sites

Total GC content and GC3s were compared within non-coding regions of the TEs. An association between the two is consistent with mutation pressure contributing to codon usage bias. This is because if selection pressure is not the main driver of codon usage evolution then both GC and GC3s will have similar mutation pressures acting upon them.

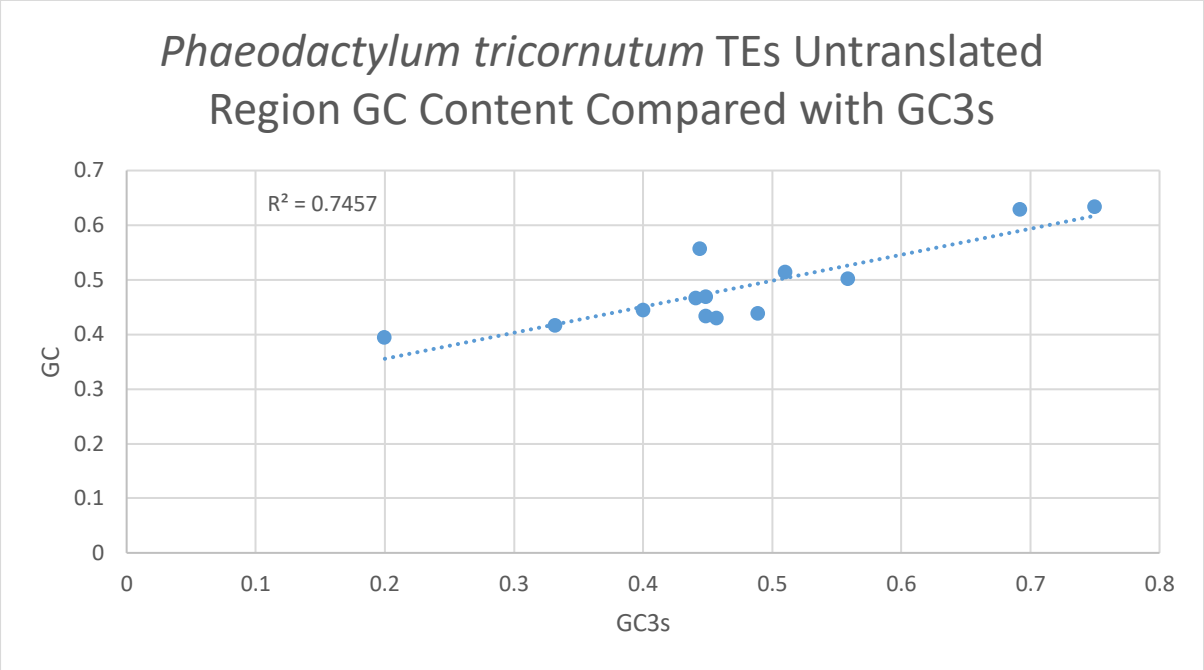


Fig. 9. Correlation of GC content and GC3s of *P. tricornutum's* TE's untranslated regions

As GC content of TE non-coding regions increases, the GC3s also increases.

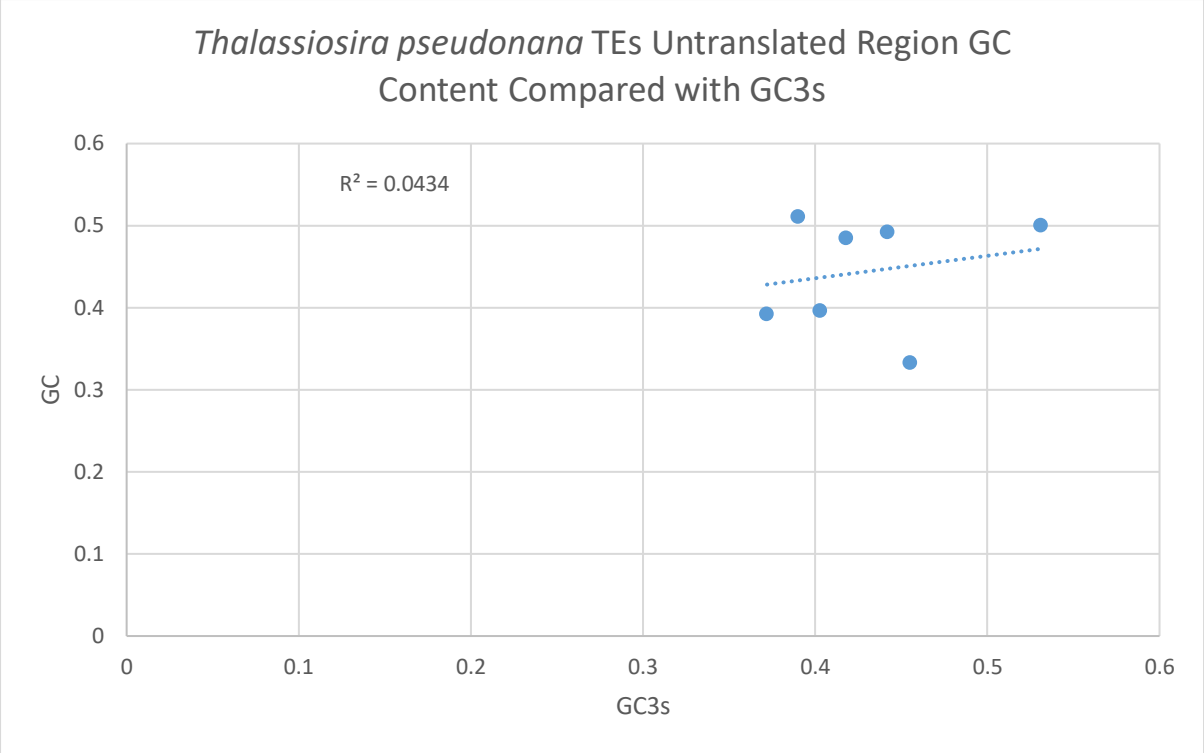


Fig. 10. Correlation of GC content and GC3s of *T. pseudonana's* TE's untranslated regions

There is no strong relationship between the two variables within *T. pseudonana* TEs.

The positive correlation is stronger within the TEs of *P. tricornutum* than within *T. pseudonana* TEs. Mutation pressure appears to make a larger contribution to *P. tricornutum* TEs than to *T. pseudonana* TEs.

4.5 Copy Number Compared to Host Optimal Codon Usage

Copy number for each TE was calculated and compared to the proportion of major tRNA genes that match the host optimal codon for each TE. This indicates if selection is contributing to the evolution of the TEs as TEs with more copies may have been more successful in transposing. If this correlates with the proportion of major tRNA genes that match the host preferred codon then it may indicate that their transposing success may be linked with selection pressures.

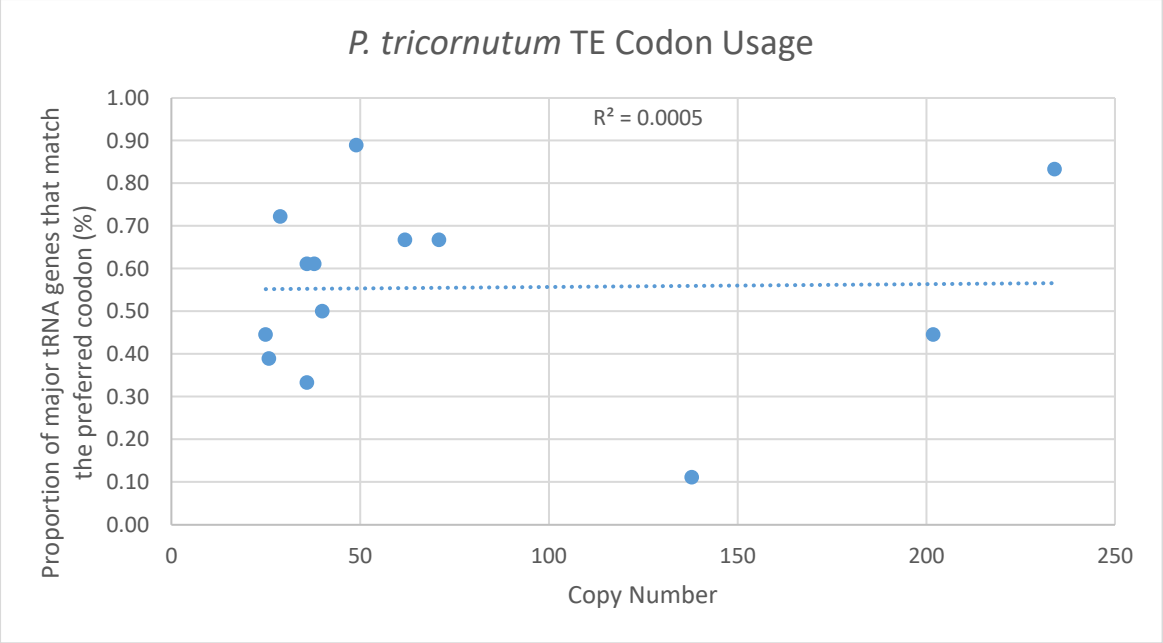


Fig. 11. Correlation of *P. tricornutum* TE copy number and their respective host optimal codon usage

There is little relationship between the two variables and the R^2 value is very low.

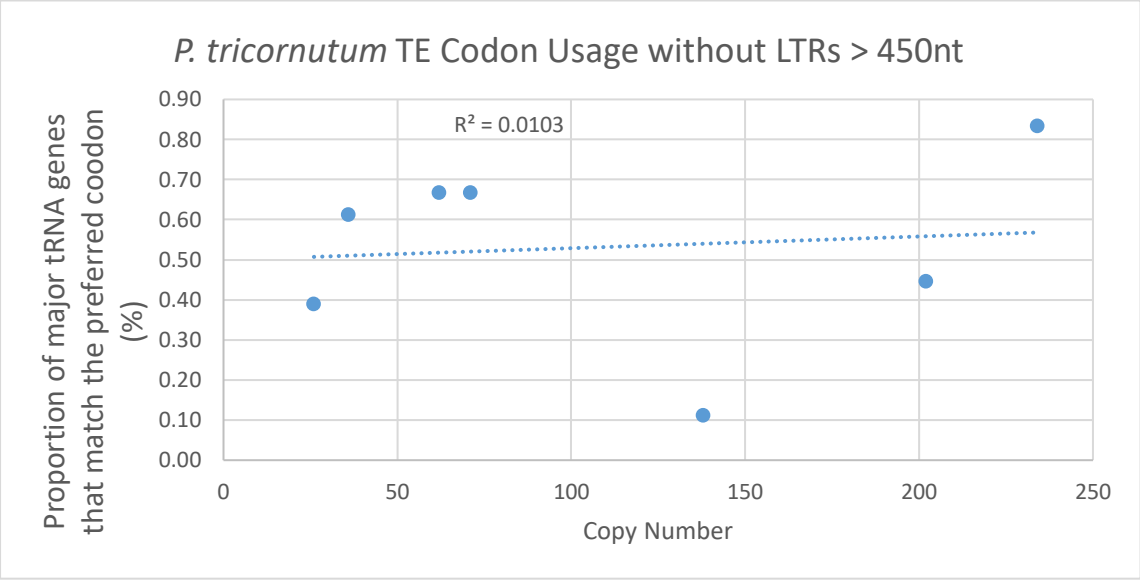


Fig. 12. Correlation of *P. tricornutum* TE copy number and their respective host optimal codon usage, including only TEs with an LTR of less than 450 nucleotides

Many of the TEs had large LTRs which may have made the copy number count lower than the genuine count for those particular TEs. Therefore this graph only includes TEs with LTRs smaller than 450 nucleotides as the copy number for these will be more accurate. There is positive correlation, although the R^2 value is still low.

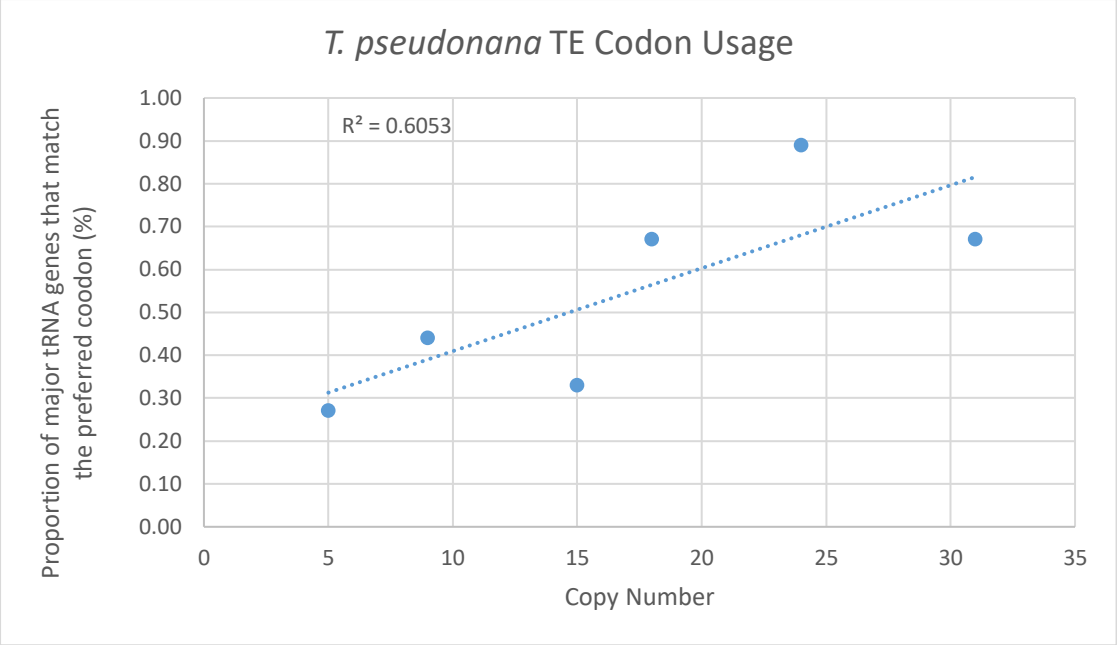


Fig. 13. Correlation of *T. pseudonana* TE copy number and their respective host optimal codon usage
There is strong positive correlation between the two variables.

Correlation is strongest in *T. pseudonana* TEs. The *P. tricornutum* TEs had much weaker correlation, but it was stronger within Fig. 10. where only TEs with a LTR sequence of less than 450 nucleotides were included. However, the R^2 value was still low and the correlation still weak meaning there may not be an association between the two variables. The association was much clearer in *T. pseudonana* TEs.

4.6 Stramenopile Phylogenetics

Phylogenetic trees of each TE were created in order to gain a greater insight into the transposing activity of each of the TEs. Phylogenetic trees show how many nucleotides each copy differs from one another, indicating how recent the transposition was, how old the element is and how active the element is. For the stramenopile phylogenetic trees (Figure 14, Figure 15 and Appendix 1-17) the branch lengths are proportional to nucleotide substitutions per site.

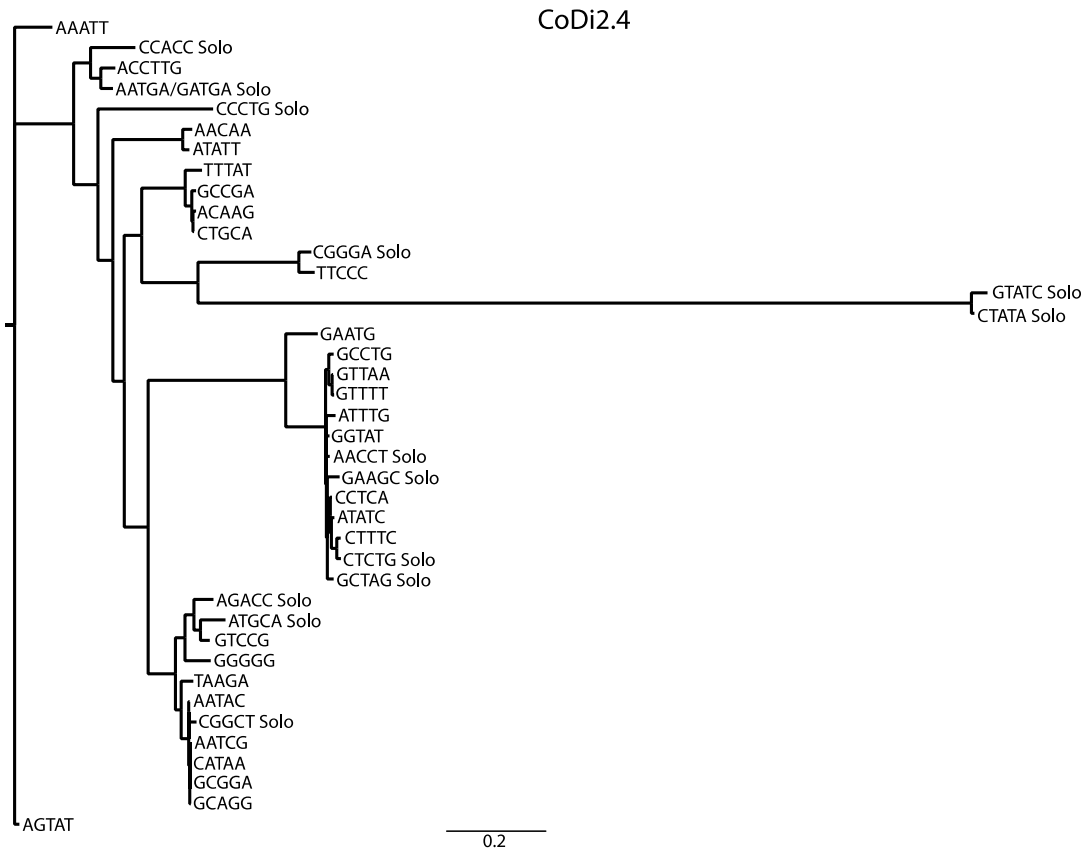


Fig. 14. Phylogenetic Tree of *P. tricornutum* TE CoDi2.4
Some of the copies have identical sequences, meaning that there is evidence of recent transposition. Thfere are two sets of 2 copies with identical sequences. There are 13 solo elements and two appear to be much older than the other elements. There is a long internal branch of 1.561 substitutions/ site. There are 40 copies in this family.

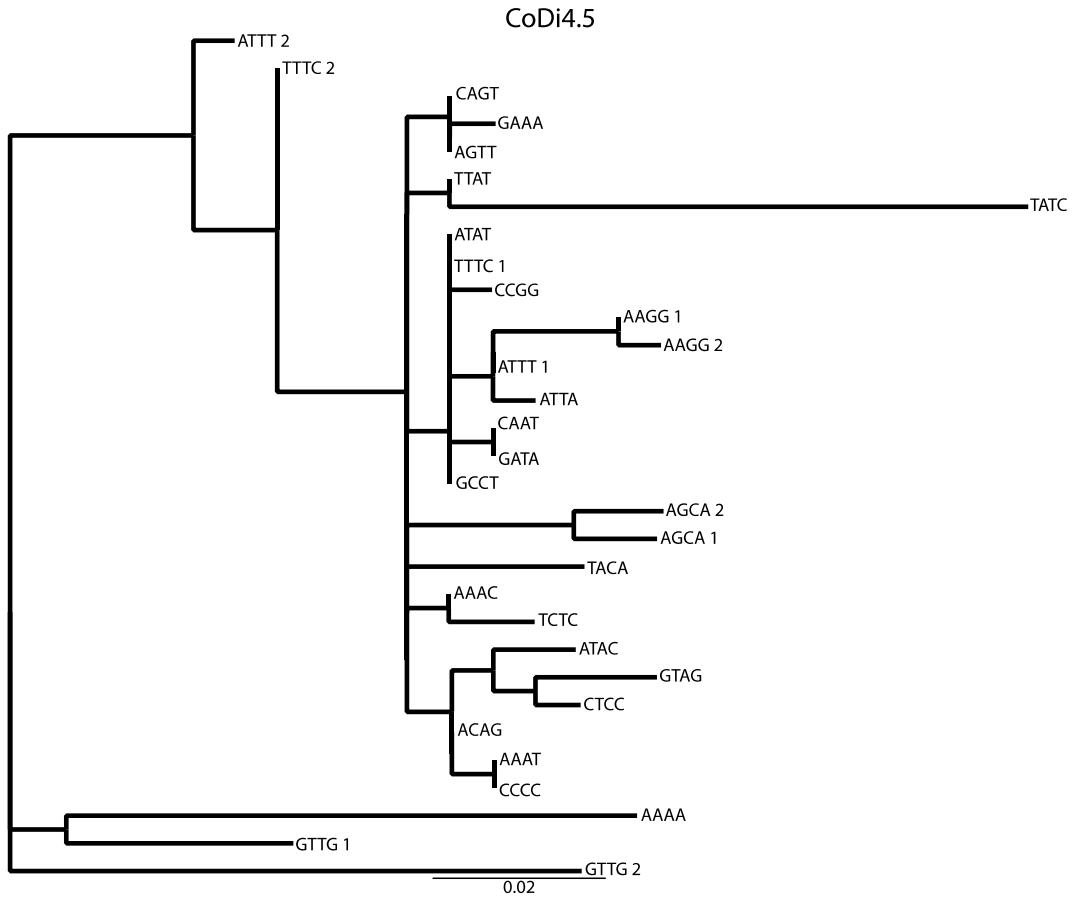


Fig. 15. Phylogenetic tree of *T. pseudonana* TE CoDi4.5
Some of the copies have identical sequences, meaning that there is evidence of recent transposition. There are three sets of 2 copies with identical sequences. There are no solo copies. This family has a TSD of 4 nucleotides and there are no copies with very short branch lengths. The longest branch length is 0.0672 substitutions per site. There are 31 copies within this family.

Mean branch lengths within *T. pseudonana* were lower than in *P. tricornutum* and had a lower average copy number, with an average of 17 copies and 75 respectively.

4.7 Choanoflagellate Phylogenetics

Phylogenetic analysis of *M. fluctuans* and *D. grandis* provides a further insight into the evolution of TEs found within the choanoflagellates and may provide evidence for the origin of the TEs. The phylogenetic trees will show which species' TEs the choanoflagellate TEs are most closely related to.
For the following phylogenetic trees (Figure 16 – Figure 23), the scale bar shows branches are proportional to nucleotide substitutions per site. 1.00 bayesian inference posterior probability (bbPP) support is indicated by an * and other values of bbPP are given above branches. A – indicates low bbPP support. Choanoflagellates and filastereans are written in red, fungi in brown, excavates in black, stramenopiles in purple, metazoans in blue, plants in green and amoebozoans in orange.

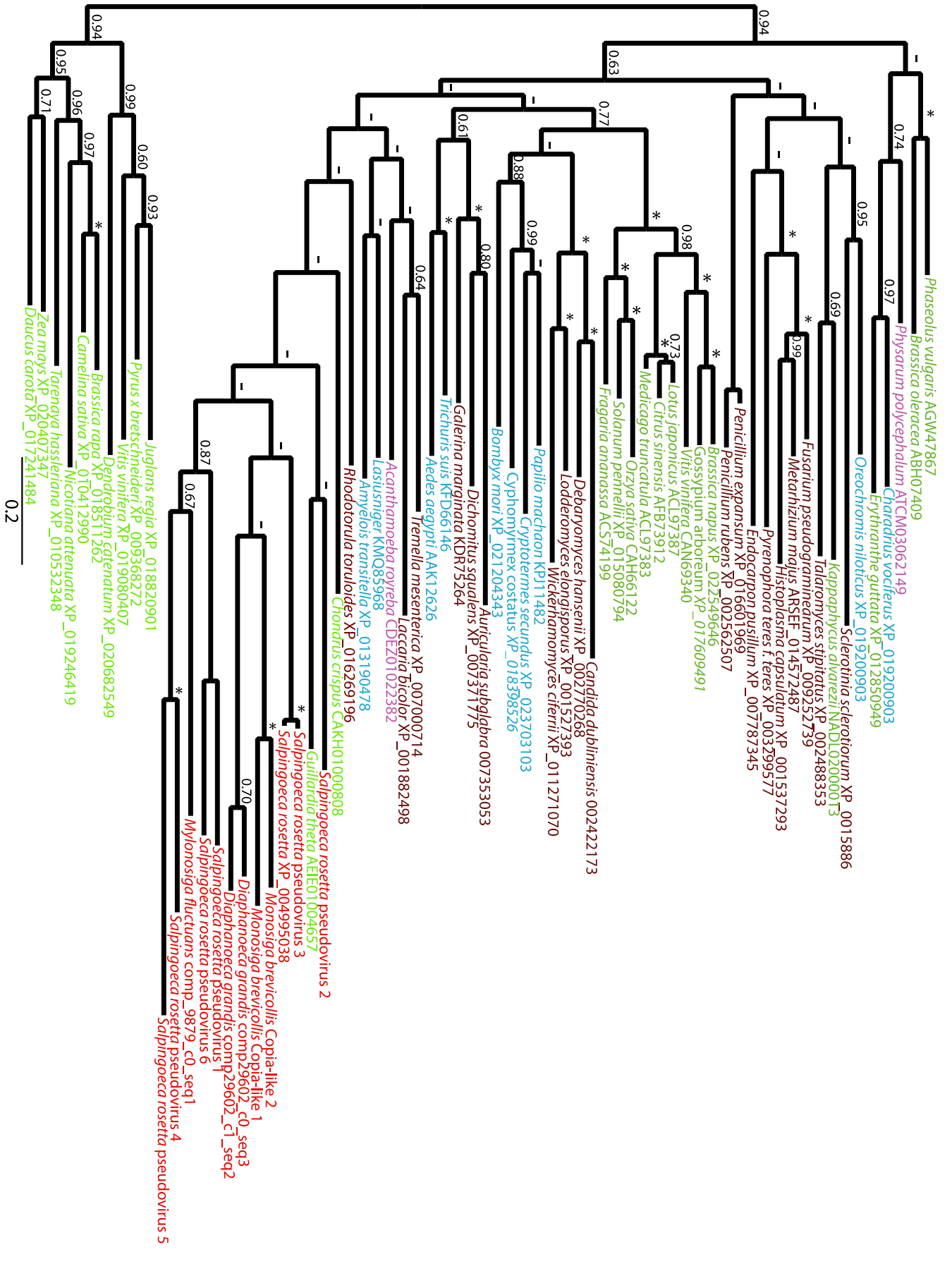

Phaseolus vulgaris AGW47867
Brassica oleracea ABH07409
Physarum polycephalum ATCM03062149
Charadrius vociferus XP_011920903
Erythranthe guttata XP_012850949
Oreochromis niloticus XP_019200903
Sclerotinia sclerotiorum XP_0015886
Kappaphycus alvarezii NADL02000073
Talaromyces stipitatus XP_002488353
Fusarium pseudograminearum XP_009252739
Metarhizium majus ARSEF_014572487
Histoplasma capsulatum XP_001537293
Pyrenophora teres f. teres XP_003299577
Endocarpon pusillum XP_007787345
Penicillium rubens XP_002562507
Penicillium expansum XP_016601969
Brassica napus XP_022549646
Gossypium arboreum XP_017609491
Vitis vinifera CAN69340
Lotus japonicus ACL97387
Citrus sinensis AFB73912
Medicago truncatula ACL97383
Oryza sativa CAH66122
Solanum pennellii XP_015080794
Fragaria ananassa ACS74199
Candida dubliniensis 002422173
Debaryomyces hansenii XP_002770268
Lodderomyces elongisporus XP_001527393
Wickerhamomyces ciferrii XP_011271070
Papilio machaon KPJ11482
Cryptotermes secundus XP_023703103
Cyphomyrmex costatus XP_018398526
Bombyx mori XP_021204343
Auricularia subglabra 007353053
Dichomitus squalens XP_007371775
Galerina marginata KDR75264
Trichuris suis KFD66146
Aedes aegypti AAK12626
Tremella mesenterica XP_007000714
Laccaria bicolor XP_001882498
Acanthamoeba royreba CDE201022382
Lasiusniger KMQ85968
Amyelois transitella XP_013190478
Rhodotorula toruloides XP_016269196
Chondrus crispus CAKH01000808
Salpingoeca rosetta pseudovirus 2
Guillardia theta AEE01004657
Salpingoeca rosetta pseudovirus 3
Salpingoeca rosetta XP_004995038
Monosiga brevicollis Copia-like 2
Monosiga brevicollis Copia-like 1
Diaphanoeca grandis comp29602_c0_seq3
Diaphanoeca grandis comp29602_c1_seq2
Salpingoeca rosetta pseudovirus 1
Salpingoeca rosetta pseudovirus 6
Mylonosiga fluctuans comp_9879_c0_seq1
Salpingoeca rosetta pseudovirus 4
Salpingoeca rosetta pseudovirus 5
Juglans regia XP_018820901
Pyrus x bretschneideri XP_009368272
Vitis vinifera XP_019080407
Dendrobium catenatum XP_020682549
Brassica rapa XP_018511262
Camelina sativa XP_010412990
Nicotiana attenuata XP_019246419
Tarenaya hassleriana XP_010532348
Zea mays XP_020407347
Daucus carota XP_017241484
0.2

Fig. 16. MrBayes phylogenetic tree of *copia*-like elements, including *G. theta*
The phylogeny is based on 68 aligned sequences either with similarity to the *copia*-like element found within *M. fluctuans*, with similarity to the *copia*-like elements found within *D. grandis* or sequences already confirmed to be *copia*-like elements.

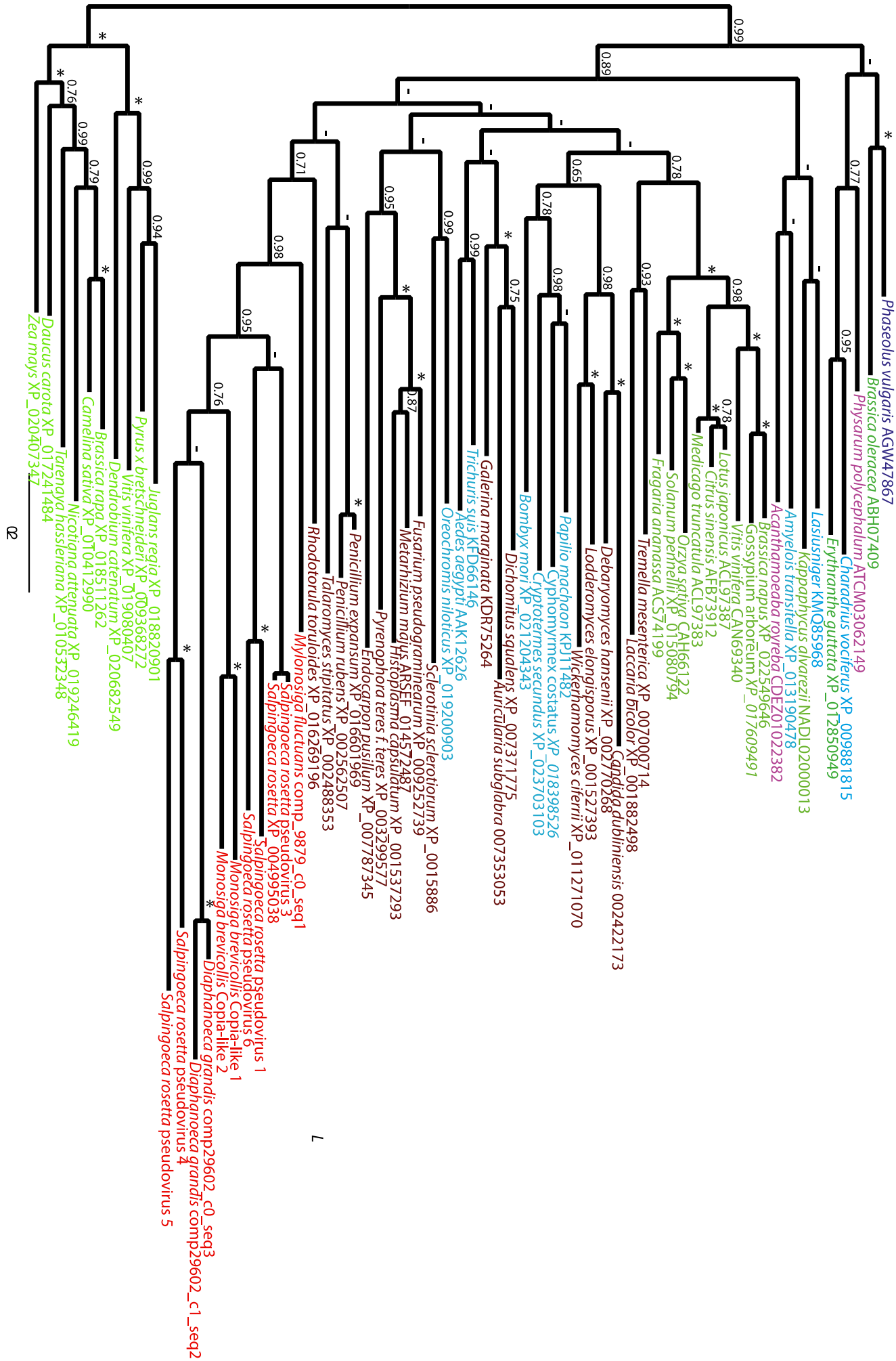

Phaseolus vulgaris AGW47867
Brassica oleracea ABH07409
Physarum polycephalum ATCM03062149
Charadrius vociferus XP_009881815
Erythranthe guttata XP_012850949
Lasiusniger KMQ85968
Kappaphycus alvarezii NADL02000013
Amyelois transitella XP_013190478
Acanthamoeaba royreba CDEZ01022382
Brassica napus XP_022549646
Gossypium arboreum XP_017609491
Vitis vinifera CAN69340
Lotus japonicus ACL97387
Citrus sinensis AFB73912
Medicago truncatula ACL97383
Oryza sativa CAH66122
Solanum pennellii XP_015080794
Fragaria ananassa ACS74199
Tremella mesenterica XP_007000714
Laccaria bicolor XP_001882498
Candida dubliniensis 002422173
Debaryomyces hansenii XP_002770268
Lodderomyces elongisporus XP_001527393
Wickerhamomyces ciferrii XP_011271070
Papilio machaon KPJ11482
Cyphomyrmex costatus XP_018398526
Cryptotermes secundus XP_023703103
Bombyx mori XP_021204343
Dichomitus squalens XP_007371775
Auricularia subglabra 007353053
Galerina marginata KDR75264
Trichuris suis KFD66146
Aedes aegypti AAK12626
Oreochromis niloticus XP_01920903
Sclerotinia sclerotiorum XP_0015886
Fusarium pseudograminearum XP_009252739
Metarhizium majus ARSEF_014572487
Histoplasma capsulatum XP_001537293
Pyrenophora teres f. teres XP_003299577
Endocarpon pusillum XP_016601969
Penicillium rubens XP_002562507
Talaromyces stipitatus XP_002488353
Penicillium expansum XP_007787345
Rhodotorula toruloides XP_016269196
Mylonosiga fluctuans comp_9879_c0_seq1
Salpingoeca rosetta pseudovirus 3
Salpingoeca rosetta XP_004995038
Salpingoeca rosetta pseudovirus 1
Salpingoeca rosetta pseudovirus 6
Monosiga brevicollis Copia-like 1
Monosiga brevicollis Copia-like 2
Diaphanoeca grandis comp29602_c0_seq3
Diaphanoeca grandis comp29602_c1_seq2
Salpingoeca rosetta pseudovirus 4
Salpingoeca rosetta pseudovirus 5
Juglans regia XP_018820901
Pyrus x bretschneideri XP_009368272
Vitis vinifera XP_019080407
Dendrobium catenatum XP_020682549
Brassica rapa XP_018511262
Camelina sativa XP_010412990
Nicotiana attenuata XP_019246419
Tarenaya hassleriana XP_010532348
Daucus carota XP_017241484
Zea mays XP_020407347
20
L

Fig. 17. MrBayes phylogenetic tree of *copia*-like elements, excluding *G. theta*
The phylogeny is based on 65 aligned sequences either with similarity to the *copia*-like element found within *M. fluctuans*, with similarity to the *copia*-like elements found within *D. grandis* or sequences already confirmed to be *copia*-like elements. *G. theta* was removed as a process of refinement.

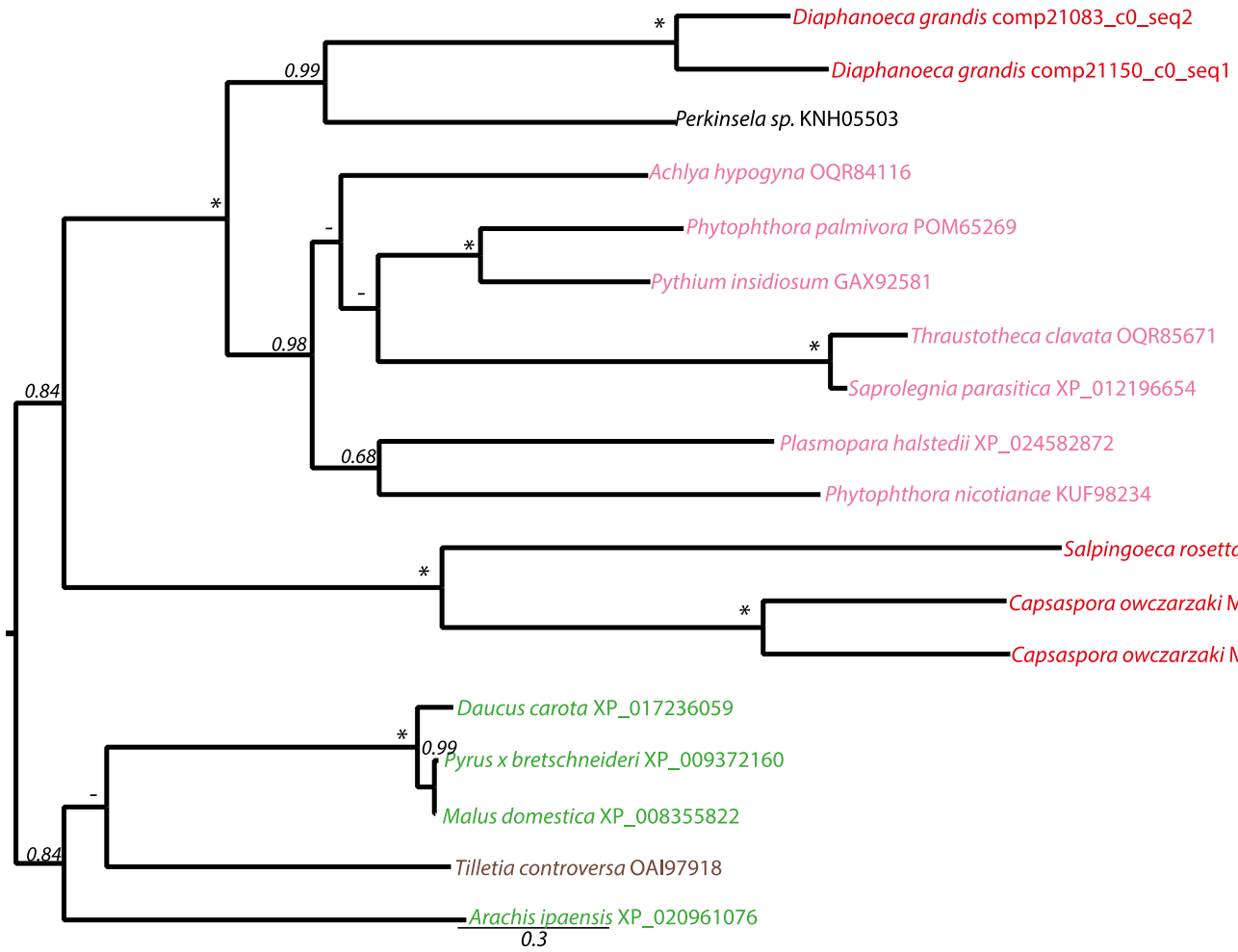


Fig. 18. MrBayes phylogenetic tree of MULE-like elements, including *Perkinsela sp.*
This phylogeny is based on 18 aligned sequences with either similarity to the MULE-like elements found within *D. grandis* or sequences already confirmed to be MULE-like elements.

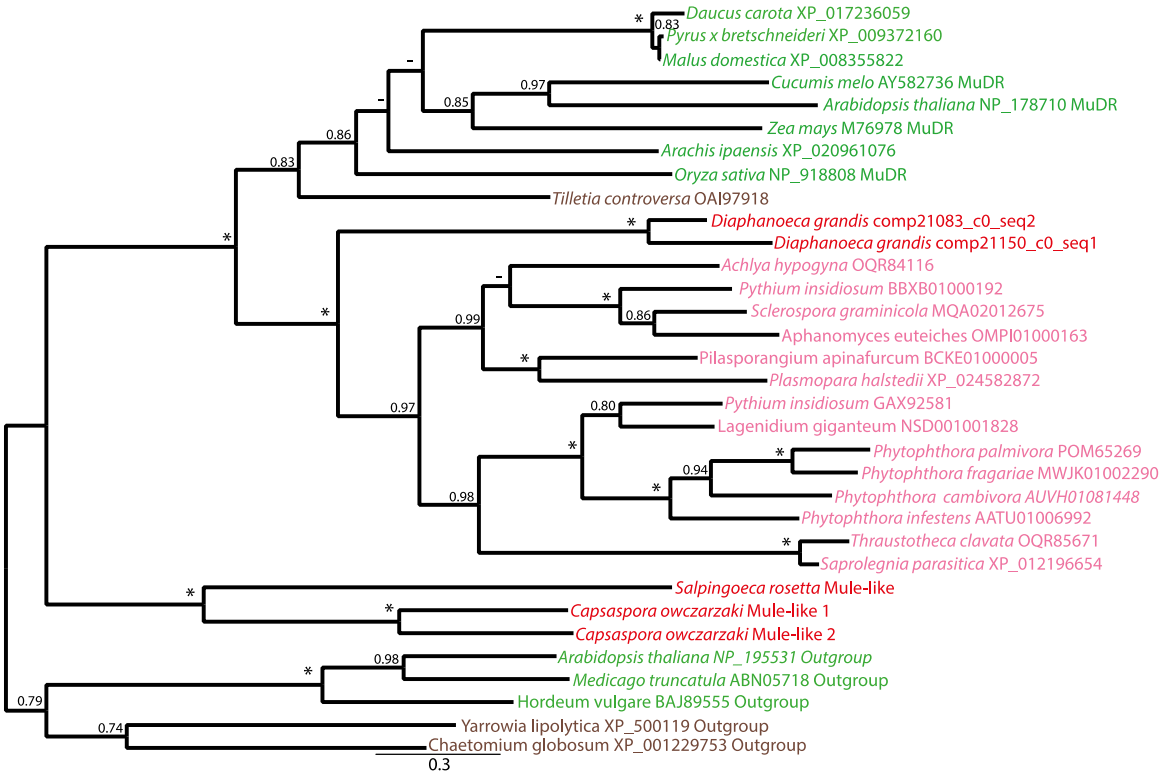


Fig. 19. MrBayes phylogenetic tree of MULE-like elements, excluding *Perkinsela sp.* This phylogeny is based on 33 aligned sequences with either similarity to the MULE-like elements found within *D. grandis*, sequences already confirmed to be MULE-like elements and outgroup sequences.

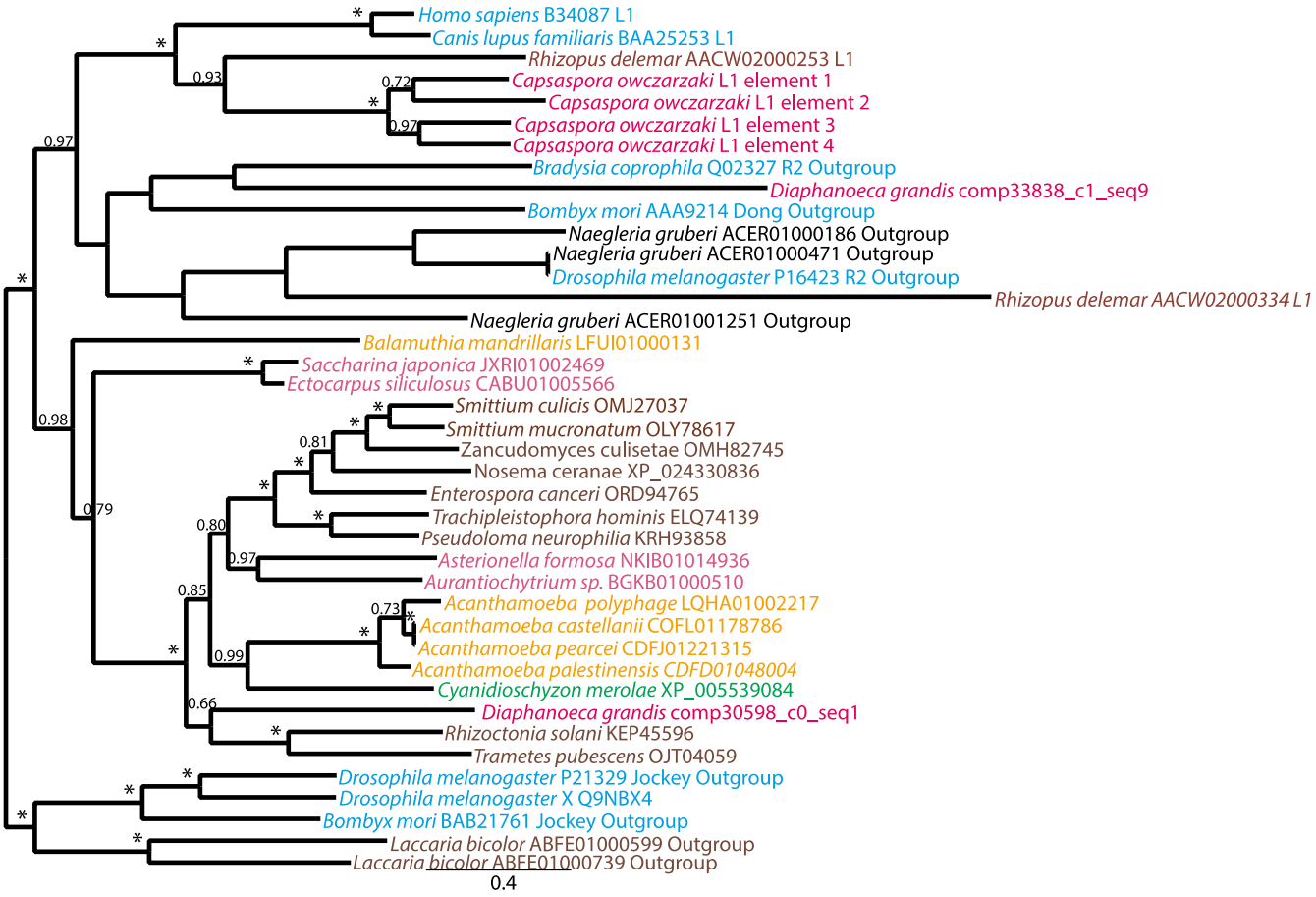


Fig. 20. Mrbayes phylogenetic tree of LINE-1 like elements
This phylogeny is based on 40 aligned sequences that were found to have similarity with the LINE-1 like elements within *D. grandis*, sequences that were already identified as LINE-1 like elements and sequences that were used as outgroups and are not LINE-1 like sequences.

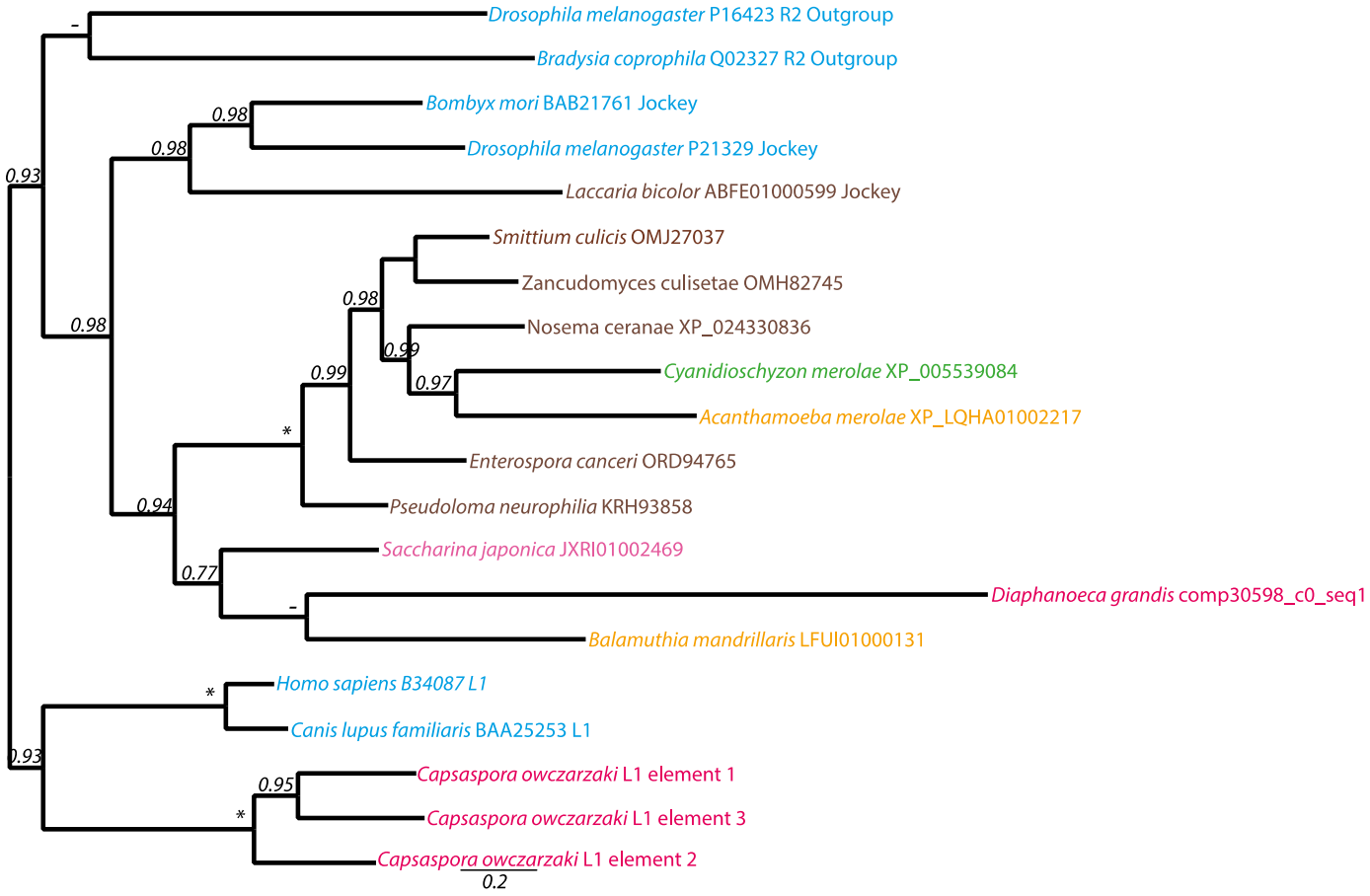


Fig. 21. MrBayes phylogenetic tree of *jockey*-like elements
This phylogeny is based upon 20 aligned sequences that either had similarities to the *D. grandis* element comp30598_c0_seq1, sequences that were already identified as *jockey*-like elements and sequences used as outgroups that were not *jockey*-like sequences. LINE-1 like elements were used as the outgroups.

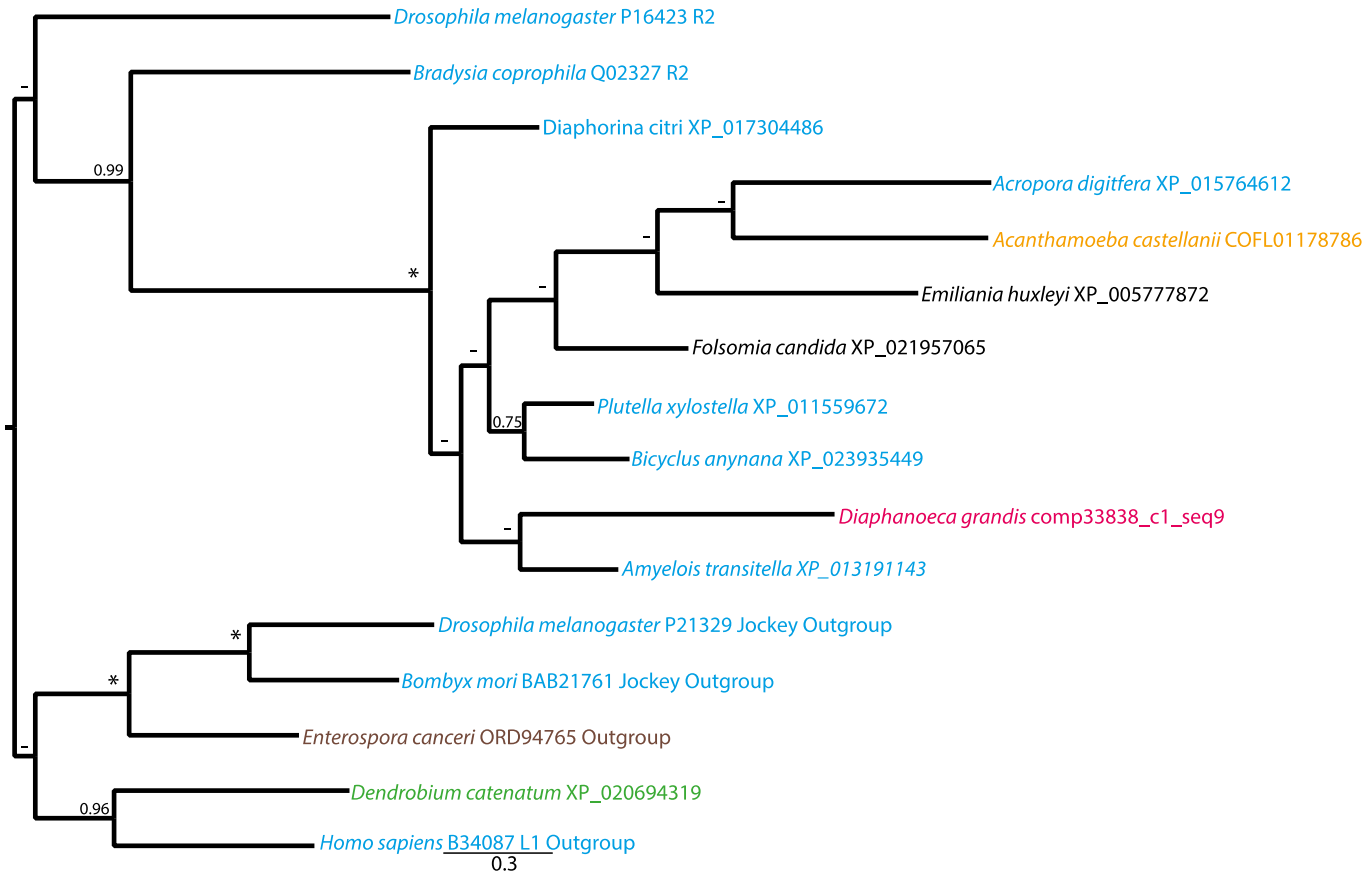


Fig. 22. MrBayes Phylogenetic tree of *R2*-like elements
This phylogeny is based upon 16 aligned sequences that were either found to have similarity with *D. grandis* TE comp33838_c1_seq9, sequences that were already identified as *R2*-like elements and sequences that were not *R2*-like elements as outgroups.

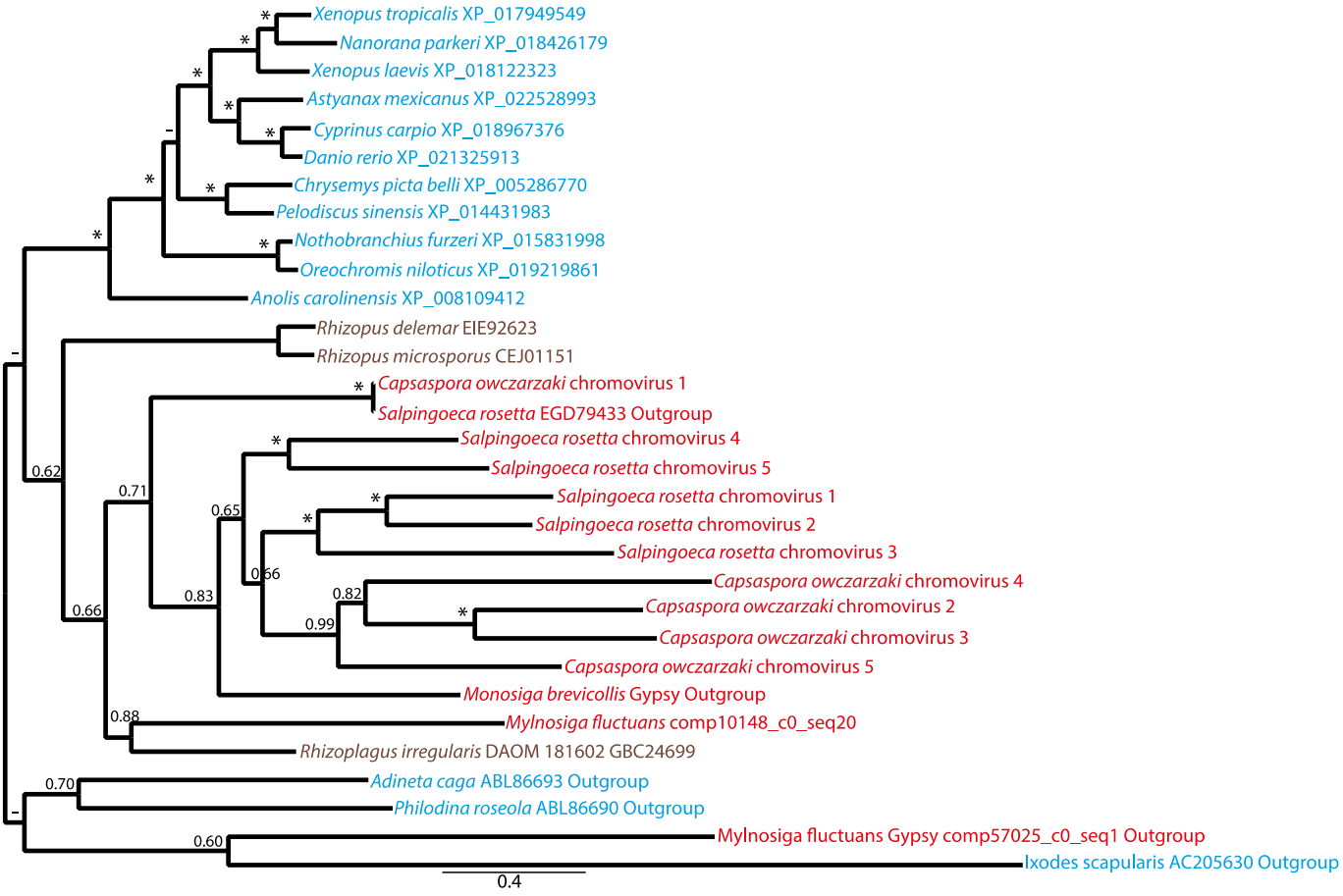


Fig. 23. MrBayes phylogenetic tree of *chromovirus* like elements
This phylogeny is based upon 31 aligned sequences that were either found to have similarity to *M. fluctuans chromovirus* like element (comp10148_c0_seq20), sequences that have already been identified as *chromovirus* elements and sequences that were not *chromovirus* elements as outgroups.

4.8 Choanoflagellate Codon Usage

Differences in Fop value between domain regions and non-domain regions in the TEs indicate whether selection accuracy is driving TE codon usage evolution.

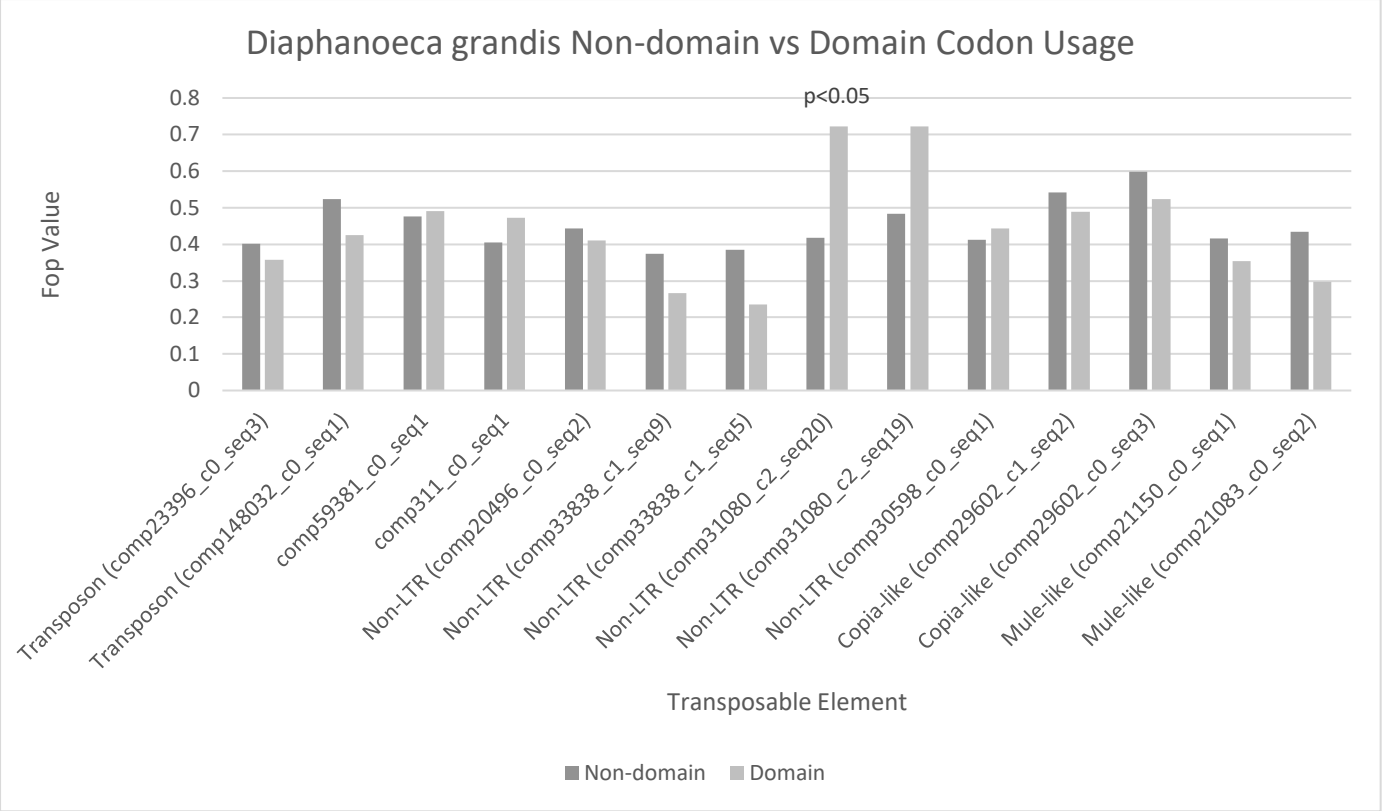


Fig. 24. TE Fop values in domain regions and non-domain regions in *Diaphanoeca grandis*

The Fop values of the domain regions and the non-domain regions were analysed using Fisher's exact test and it was found that the only significant difference was between the regions of the non-LTR retrotransposon comp31080_c2_seq20. The p-value was smaller than 0.05%. The Fop value in domain regions was higher than the Fop value of the non-domain regions in five TEs out of the fourteen in *D. grandis*.

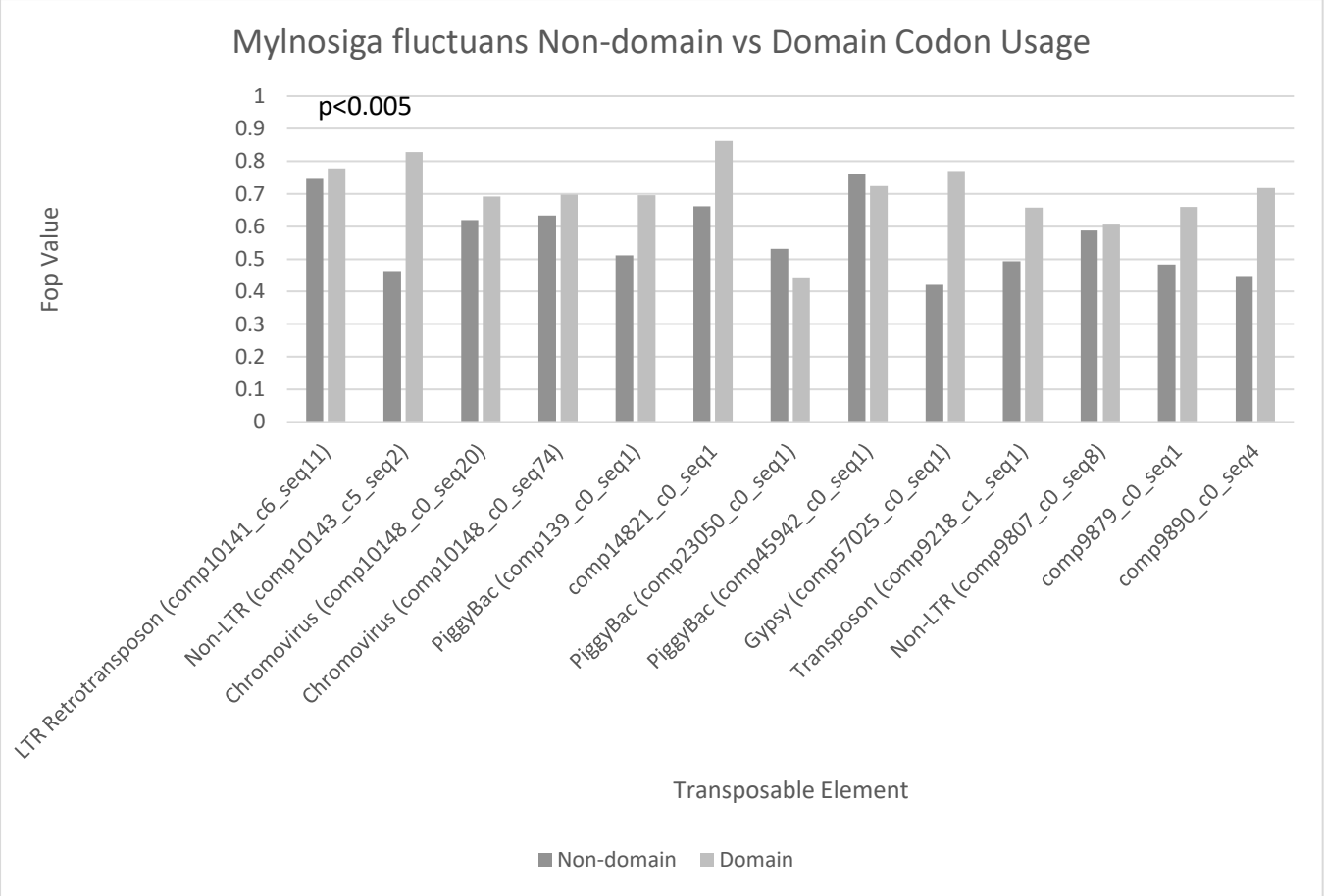


Fig. 25. TE Fop values in domain regions and non-domain regions in *Mylnosiga fluctuans*

The Fop values of the domain regions and the non-domain regions were analysed using Fisher's exact test and it was found that four TEs had significant difference between the regions. The p-value was smaller than 0.05% for *piggyBac*-like TE comp139_01_seq1 and non-LTR retrotransposon comp9890_c0_seq4. The p-value was smaller than 0.01% for non-LTR retrotransposon comp10143_c5_seq2 and *gypsy*-like TE comp57025_c0_seq1. The Fop value in domain regions was higher than the Fop value of the non-domain regions in ten TEs out of the twelve in *M. fluctuans*.

More TEs showed evidence for selection accuracy within *M. fluctuans* than *D. grandis*.

5 Discussion

Most of the previous work on TEs has focused upon multicellular organisms. Most diversity within eukaryotes is within unicellular organisms and unicellular organisms are at least 1.2 billion years old (Rasmussen et al., 2008), whereas multicellular organisms are estimated to be only around 600/700 million years as a result of fossil evidence (Ye et al., 2015). As multicellular organisms have been the primary focus of the most studies, only the minority of eukaryotic TE activity has been investigated, as well as only the recently evolved species. Here I investigate TEs of two distantly related eukaryotic groups in order to gain a better insight into TE evolution. This includes stramenopiles and choanoflagellates.

5.01 Selection vs Mutation in Stramenopiles

The two stramenopiles included within this study are distantly related to *P. infestans* within the stramenopile phylogeny (Derelle et al., 2016). Other than the one study on *P. infestans*, TE codon usage within stramenopiles has not been investigated before so there is a vast potential for significant discoveries.

P. tricornutum and *T. pseudonana* TEs are using optimal codons more than would be expected if their codon usage was entirely down to mutation pressure. Using host optimal codons as the most frequently used codon conveys a benefit to the TEs because the host's replicating machinery will recognise the TE codons as being similar to its own, potentially allowing the TE to transpose more (Ehrenberg & Kurland, 1984), or allows the TE protein to be translated more accurately, reducing mistranslations (Akashi, 1994). Therefore, TEs that use host optimal codons as the most frequent codon are likely to be evolving under selection. However, some TEs may be under stronger selection than others. This appears to be the case as TE 5.2 uses host optimal codons as the most frequently used codon for sixteen out of eighteen amino acids with degeneracy whereas, at the other end of the scale, TE 4.3 uses host optimal codons as the most frequently used codon two out of eighteen amino acids. This suggests that overall, the TEs in *P. tricornutum* may appear to be under moderate selection but in reality, there are TEs under strong selection and TEs under very weak selection. It could also suggest that mutation is acting against any selection in TE that appear to be under weak selection.

TE 6.7 is using host optimal codons for eight out of eighteen amino acids with degeneracy but has a low Fop value relative to the other *P. tricornutum* TE elements – 0.408. This suggests that this TE is under weak selection pressure because weak selection may allow optimal codons to be the most frequently used codon, more than would be expected by chance, but still at low frequencies. Alternatively, the low Fop value could be a result of chance and mutation, with no selection acting upon it at all.

In *P. tricornutum*, the most frequently used codon in the TE often complements the most abundant tRNA but is not the host optimal codon. This could be because of a variety of reasons. It is possible that the host is simply not using their tRNAs very efficiently, or that CodonW has misidentified optimal codons. When mutation is a strong factor in codon usage, CodonW may identify a codon as an optimal codon but it may simply be a high frequency codon due to mutation. Peden (2000) described how CodonW can only be certain of identifying correct optimal codons if translational efficiency is the main driver of codon usage for a species. In order to confirm whether optimal codons identified by CodonW are genuine or not, gene expression data of the stramenopiles and choanoflagellates would be necessary and that was not available for this study.

In this study, selection as a driver of evolution has been estimated by investigating codon usage and looking at which TEs use host optimal codons as their most frequently used codon. If the theory that TEs have evolved to use the host's most abundant tRNAs regardless of whether they are an optimal codon or not is correct, then the level of selection involved in the TEs appear weaker than it actually is. Alternatively, the TEs of *P. tricornutum* may be favouring the codons which bind to major tRNA molecules because they are transitionally optimal or because the TE mutation pressure by chance has driven codon usage towards major tRNA genes. It could be a combination of both selection for translational efficiency and mutation that has led to TEs most frequently using codons complementary to the most abundant host tRNAs. This phenomenon occurred thirteen times in *T. pseudonana* across the seven TEs (an average of 1.86 times per TE, Table 3), compared to twenty-seven times in *P. tricornutum* (an average of 2.07 times per TE, Table 4). This difference could suggest that *P. tricornutum* TEs are more translationally optimal than *T. pseudonana* TEs as they are using

the major tRNA more often than *T. pseudonana*. However, all other evidence points toward *T. pseudonana* TE codon usage being under stronger selection that *P. tricornutum* TE codon usage. Therefore, this theory is unlikely.

Another explanation for the TEs to use codons complementary to the most abundant host tRNA even if they are not host optimal codons is that mutation may be more important than selection if increased variation within the population is important. This is because increased variation allows selection within some individuals while allowing the species to survive.

As TEs can jump between species (Daniels et al., 1990), optimisation between tRNAs and TE codon usage may not be expected but we can expect optimisation between host genes and the tRNA, if there is selection acting on codon usage, because host genes stay within the species. However, TE optimisation may be present if the donor species has similar codon usage as the recipient species. Horizontal transfer of TEs is more likely to occur within closely related species (Bolotin & Hershberg, 2017), so it is possible that host and donor species are likely to have similar codon usage.

TEs in a broad range of eukaryotic organisms tend to prefer AT ending codons (Jia & Xue, 2009; Lerat et al., 2002). Some of the TE families within *P. tricornutum* and *T. pseudonana* are strongly preferring GC ending codons. Eight out of thirteen TEs within *P. tricornutum* are using GC ending codons as their most frequently used codons 50% of the time or more (Table 3), as are four out of seven TEs within *T. pseudonana* (Table 4). These GC ending codons match the host tRNA genes. The expectation was that if the TEs are evolving by mutation pressure alone, like those of multicellular organisms, the codon third position would most likely be A or T as that has what been seen in other eukaryotic organisms. As that is most often not the case within the stramenopile TEs, this is more evidence that the TEs are evolving under selection pressure.

Comparing GC content of non-coding regions of TEs with their GC3s will indicate if there is mutation bias within their codon usage. The LTRs of the TEs are involved in replication (Boeke, 1989), meaning that LTRs may be under selection constraints and may not be showing the true mutation pattern of the TEs.

GC content and GC3s would be expected to have similar mutation pressures acting upon them if selection is not a driver of codon usage evolution, and the two variables would correlate if mutation pressure is stronger than selection pressure upon the codon usage. This is seen within the TE GC content of non-coding regions and GC3s increase together in *P. tricornutum* (Figure 9). This suggests that mutation is an important driver in the TE's evolution. Although there is positive correlation between GC and GC3s of *T. pseudonana's* TEs, the R value is very low and there is therefore little relationship between the two variables (Figure 10). The results are consistent with the theory that *T. pseudonana* TEs are under stronger selection and that mutation is more important in the evolution of *P. tricornutum* TEs than *T. pseudonana* TEs. It is possible that the effective population size of *T. pseudonana* is larger than that of *P. tricornutum* and therefore under stronger selection. This would make the environment for the TEs in *T. pseudonana* harsher and they may have to be under selection in order to survive, whereas the *P. tricornutum* TEs may be in a host that is under weaker selection.

5.02 Selection Efficiency vs Selection Accuracy in Stramenopiles

Fop values were also generated to compare the codon usage between domain regions and non-domain regions of the TE in the stramenopiles. The expectation was that, as the domain regions are usually enzymes, they are more important to the TE and are therefore more likely to be using the optimal codons if they are under selection for translational accuracy

(Drummond & Wilke, 2008). Non-optimal codons are more likely to be mistranslated than optimal codons (Precup & Parker, 1987) so if a TE is evolving under translational accuracy selection then there will be more optimal host codon usage within domain regions than in non-domain regions, if the domain regions are more functionally important to the TE than the non-domain regions. If incorrect amino acids were placed in the domain regions it would result in a mistranslated protein. Selection accuracy works to prevent this occurring as oppose to speeding up translation (translational efficiency). If selection efficiency is driving the evolution of the TEs then the likely outcome is that optimal codons are included in both non-domain regions and domain regions in order to speed up the process of translation. In this instance, there would be less significant difference in Fop values between non-domain regions and domain regions. Translational accuracy and efficiency could both be drivers for selection within the stramenopile TEs.

The results show that Fop values for both domains and non-domains in *P. tricornutum* TEs were very similar (Figure 5). This means that the results are not consistent with the prediction for translational accuracy. Whilst there was slightly more difference in Fop values between the domain regions and non-domains regions in *T. pseudonana* TEs (Figure 6), these results were also not statistically significant. Here there is no evidence of selection accuracy in *T. pseudonana* as nearly all the domain Fop values are higher than non-domain Fop values. The TE of *T. pseudonana* are more consistent with translational accuracy than those of *P. tricornutum*, but there is not significant evidence for either species.

Despite the results not being consistent with selection for translational accuracy in these stramenopiles, there could be a few alternative explanations. It is possible that there were very few sites within the domain regions that are under selection for translational accuracy and that the majority of them are not, leading to non-significant differences between domain region and non-domain region Fop values. Another plausible explanation is that there may be many sites within the non-domain region that are under selection for translational accuracy, as well as many sites within the domain regions. These non-domain sites could be structurally important, and ensuring the correct amino acid is in place via translational accuracy could allow the TE protein to form the correct bonds and structure. Translational efficiency would also lead to optimal codons in non-domain regions so identifying optimal codons in non-domain regions could suggest either, or both types of selection were acting upon codon usage.

In order to remove any bias from incorrectly identified optimal codons, Nc values were calculated. It would be expected that TE domain regions would have a lower value of Nc, showing a bias towards using specific codons. Non-domains would have a higher value of Nc as they would be expected to have more randomised codon usage. Although not significantly different, nine out of the thirteen *P. tricornutum* TEs had a higher Nc value in their domain regions (Figure 7). This is consistent with the theory that there is no selection bias driving TE evolution within *P. tricornutum* TEs. Four out of the seven TEs within *T. pseudonana* showed a higher Nc value in non-domain regions (Figure 8), although this was also not significantly different. This is also consistent with the theory that *T. pseudonana* TEs are under stronger selection than *P. tricornutum* TEs.

5.04 Stramenopile Phylogenetics

Phylogenetic analysis of TEs indicates how successful the TEs have been at transposing over time, in terms of copy number. It allows a clearer picture of each individual TE family and will help identify any TE family whose codon usage is evolving under selection. This is

because different types of selection upon codon usage may produce different patterns within TE copies which can be identified when looking at TE copy phylogenies.

Within *P. tricornutum* there appears to be twelve TEs that show signs of recent transposition (Figure 14; Appendix 1-12). There are copies with zero branch lengths which means they are identical sequences. The sequences have not had time to mutate since one copy transposed from the other. Some families have copies on very long branch lengths, indicating that these copies are old. For example, GTATC solo and CTATA solo within CoDi2.4 TE (Figure 14) branch lengths are 1.9636 and 1.9095 substitutions per site respectively. These copies can no longer transpose but have not been removed via natural selection because solo LTRs are less harmful to the host (Carr et al., 2008). There are eight other TEs within *P. tricornutum* that appear to be particularly old (Appendix 3-9; Appendix 11). It has been theorised that solo LTRs have fewer regulatory regions than 5 prime or 3 prime LTRs, leading to a less harmful effect upon the expression of adjacent genes. Solo LTRs also do not produce transposable element RNA or proteins so they use less of the host's resources in terms of tNRAs than a full-length element would (Carr et al., 2002).

Some TE families have a shorter mean branch length than other families, suggesting that they are younger families. For example, the mean branch length in CoDi3.1 (Appendix 1) is much shorter than that of CoDi2.4 (Figure 14). There are also a smaller number of copies within this family compared to CoDi2.4, which is consistent with the idea that CoDi3.1 may be a younger family within *P. tricornutum*, if there are similar rates of transposition and similar deleterious selection within the two families. Although a smaller number of copies of a TE may indicate that it is a younger family, there is an alternative reason why a TE may have fewer copies. It is possible that the TE is more deleterious to the host and that selection against the family is stronger and that more copies of these TEs have been removed due to purifying selection.

Purifying selection could also lead to the age of copies being estimated incorrectly. Purifying selection removes copies that are deleterious to the host so many old copies are removed which may make the families appear younger than they really are. However, some old copies remain in families and can be identified because they are on long branches and are distantly related to the other copies within the family. For example, *P. tricornutum* CoDi5.3 (Appendix 7) has a solo element distantly related to the oldest group of copies. The solo copy is likely to not have been removed via purifying selection with the other copies due to solo copies not being as deleterious to the host as previously stated. This solo LTR also has a different 5 prime TSD from the 3 prime TSD. As they differ from one another by one nucleotide, it is likely that a substitution mutation took place within the TSD. Solo LTR's may also have two different TSDs due to recombination. As there are identical sequences involved in TE copies, recombination may occur, causing the 5 prime TSD from one element to end up with the 3 prime TSD from another element.

Some TEs within *P. tricornutum* have two groups of closely related copies, whilst the groups themselves are distantly related, such as in CoDi5.2 (Appendix 7). A possible explanation for this is that the family is very old and many of the copies were removed from the host genome through purifying selection. However, recently, events have occurred which allowed the family to undergo rapid transposition and the few old copies that remained within the host have also transposed. It is possible that as the copy number declined, the family was less deleterious to the host and therefore there was less selection pressure from the host upon the family. This, in turn, enabled the family to transpose more, although as soon as copy number increases, the family will become increasingly deleterious. This family is also using the highest number of host optimal codons as its most frequently used codon – 16 out of 18

amino acids with degeneracy - which supports this theory (Table 3). However this theory would be impossible to prove.
There are families that have similar longest branch lengths to other families but fewer copies, such as *P. tricornutum's* CoDi4.1 TE (Appendix 2) compared to CoDi3.1 (Appendix 1). This could mean that the two families are of a similar age but has transposed less which has resulted in fewer copies. Looking at the host optimal codon usage also reveals that in this case, CoDi4.1 uses less host optimal codons than CoDi3.1 which may result in less efficient translation, contributing to the less successful transposition in terms of copy number.
Poor sequencing may also lead to a families age being overestimated. It may lead to a copy appearing as if it has many more substitutions/ site than it genuinely does and making the family appear much older than it really is. For example, *P. tricornutum* TE CoDi6.4 copy with the TSD TGACCC is on a very long branch length of 2.655 substitutions/ site which is much longer than any of the other branches (Appendix 8). If this is due to poor sequencing, then the branch looks deceptively old and the family may be younger than it appears.

5.05 Stramenopile Copy Number and Selection

P. tricornutum's TE CoDi4.3 has a much higher copy number than CoDi2.4, CoDi3.1 or CoDi4.1 (Appendix 3). The longest branch length is 0.4696 substitutions/site, between that of CoDi2.4 and CoDi4.1 but the frequency of host optimal codons used is the lowest out of all the TEs studied, including those from *T. pseudonana*. CoDi4.3 also has a relatively low Nc value of 47.10 and the lowest Fop value. It is possible that there are no true optimal codons for *P. tricornutum* and that the CodonW estimated optimal codons are false positives. The Fop value for 4.3 may therefore, be incorrect. However, if the results are genuine, there could be a low Fop value due to a high mutation rate. Mutation is more likely to be changing codons to host non-optimal codons from either host optimal codons or non-optimal codons rather than from host non-optimal codons to optimal codons as long as mutation patterns are all equal, due to chance. Therefore, random mutation may be reducing codon bias by chance. Unless there is selection for codon usage, the host optimal codons will be lost. Alternatively, or additionally, recent horizontal transfer may have taken place. This would mean that the family would not have had time to adapt to its new host and would have a low proportion of codon usage. This would also mean that there would be a lack of host defences and the family would be able to transpose rapidly. However, as the longest branch length is 0.4696 substitutions/ site, this family appears old and recent horizontal transfer into *P. tricornutum* is not likely.
A TE that may have undergone recent horizontal transfer is *P. tricornutum* TE CoDi6.7. This has no copies on zero branch lengths so it cannot be ascertained whether this family is still transposing or not (Appendix 11). The copies are all on relatively short branch lengths suggesting that this family is relatively young, except from two copies, which are distantly related to the other copies. It is possible that recent horizontal transfer has occurred and because the family is not well adapted to the host's codon usage compared to other copies, it has not been able to survive. However, it is unlikely that codon bias will lead to the family's extinction. It is also possible that if the family has undergone recent horizontal transfer and that *P. tricornutum* is a new host to CoDi6.7, then the TE may not be able to recognise the host proteins and will therefore be unable to transpose. The copies that have been identified could be fossils that are no longer transposing. In addition, this family has 25 copies, which is the lowest number of copies out of all the TEs studied within *P. tricornutum*, as well as having a low Fop value of 0.408 and a high Nc value 56.82 (Table 1). These results are consistent with the theory of recent horizontal transfer.

T. pseudonana TE CoDi6.3 has only five copies, which are on very short branch lengths (Appendix 17). It is still transposing because two of the copies have identical sequences. These two pieces of evidence suggest that it may have undergone recent horizontal transfer as that would explain why the element appears young with its short branch lengths, and may be able to transpose because the host has not had time to evolve defences against the family. As it is only using host optimal codons in five out of eighteen amino acids with degeneracy, it appears that this family's codon usage is not under selection (Table 4). However, if it has recently entered the *T. pseudonana* genome, it will still be adapting to the host codon usage. Alternatively, the family may not be evolving under selection and have very few copies because other families with better adapted codon usage are out-competing it for host tRNA use.

CoDi4.5 is the TE with the highest copy number out of all TEs found within *T. pseudonana*, with 31 copies (Figure 15). It is one of the *T. pseudonana* TEs that appears old because there are three copies with relatively long branch lengths. *T. pseudonana* TEs CoDi5.4, CoDi5.5 and CoDi6.2 only have copies on shorter branch lengths, suggesting that they are young families (Appendix 13, 14 & 16). CoDi6.1 has only copies on short terminal branches but some of the copies are very distantly related from the others (Appendix 15). CoDi6.1 has the highest Fop value out of all the stramenopile TEs within this study with a value of 0.597 (Table 2), and it also has the second highest copy number with 24 copies. This suggests that codon usage is under selection within this family. This finding is almost unique within TEs, as codon usage selection has not been studied extensively. The exception to this is within some stramenopiles, notably the study on *P. infestans* TEs where codon usage was found to be under selection (Jiang & Govers, 2006).

All TE families within *T. pseudonana* have very few copies compared to *P. tricornutum* (Figure 14; Figure 15; Appendix 1-17). Copy number in *T. pseudonana* TEs ranges from 5 to 31 copies, and *P. tricornutum* TE copy number ranges from 25 to 234. This could indicate that the *T. pseudonana* TEs may be under stronger selection from the host, removing more older copies from *T. pseudonana* than from *P. tricornutum*. However, it could also suggest that the TEs within *T. pseudonana* are younger than those within *P. tricornutum* as the mean branch lengths for *T. pseudonana* were lower than those for *P.ticorntutum*. These shorter branch lengths could also be the result of higher transposition rates within *T. pseudonana*. This suggests that each TE has had less time to evolve and fewer mutations have taken place. There is also a much stronger positive correlation between copy number and TE codon usage within *T. pseudonana* than *P. tricornutum* (Figure 11 – Figure 13). These findings together suggest that *T. pseudonana* TEs are evolving under stronger selection for codon usage than TEs within *P. tricornutum*. Transposing TE families may be more likely to be under selection, whereas other families may be no longer transposing due to the lack of selection driving their evolution or due to selection against individual copies due to their deleterious effects on the host. The fact that there are far more copies in the majority of the *P. tricornutum* TEs may suggest that the majority of TEs in *P. tricornutum* are under weaker purifying selection, which allowed the families to proliferate. *P. tricornutum* TE CoDi6.7 has no identical sequences and therefore there is no evidence that the copies are still transposing. It is possible that selection pressure upon this family have changed and this TE may be under weaker selection or the host may have developed defences through selection that protect them from the deleterious effects of the TE more. However, if copies are more deleterious then they are likely to be under stronger selection pressure too in order to continue being active despite the host evolving defences against them.

P. tricornutum TE CoDi5.1 has the highest copy number of all stramenopile TEs in this study, with 234 copies. It also appears to be an old element due to having relatively long branch lengths (Appendix 5) and it uses host optimal codons for fifteen out of eighteen amino

acids with degeneracy (Table 5). It also has a very high Fop value of 0.585 and low Nc value of 46.44 (Table 1), which all are consistent with the theory that CoDi5.1 codon usage may be under selection. It is possible that because there is a high number of copies, there will be more transcripts and this family will out-compete other families for the host tRNAs, leading to more efficient translation and therefore more copies. Additionally, there may be strong competition between individual copies within CoDi5.1. Copies with the best-adapted codon usage may out-compete other copies from the same family, resulting in the family having a high Fop value and proportion of host optimal codons. However, if the effective population size of *P. tricornutum* is small, there will not be a drive to have higher codon bias within the TEs as there will be no selection for codon usage within the host.

Within the majority of the TE families, there are multiple copies with the same TSD. As they have different sequences from one another, they are unlikely to be the same copy. The most feasible explanation is that the TSDs are the same by chance, particularly when the TSD is four nucleotides long. In a family with TSDs of four nucleotides long, such as CoDi4.5, there is one in sixteen chance of two TSDs being the same. For a family with TSDs of 5 nucleotides long, such as CoDi5.4, there is one in twenty chance of two TSDs being the same and for a family with TSDs of 6 nucleotides long, such as CoDi6.2, there is one in 24 chance of two TSDs being the same. However, if the copy is old then the same insertion may have different sequences in different individuals. This means that copies with the same TSD but different sequences may actually be the same copy. This may have occurred in elements such as *P. tricornutum* TE CoDi4.4 (Appendix 4). CoDi4.4 also has copies with a five base pair TSD as well as having the same copy being a solo LTR that has a four base pair TSD. This could possibly occur if a mutation in the *integrase* gene occurred, affecting the Integrase protein and leading to the shift of the LTR from a solo LTR to the LTR with either a 5' or 3' 4 base pair TSD.

5.05 Identified Choanoflagellate TEs

Exploring how TEs may have evolved within choanoflagellates would be enhanced by investigating the origin of the choanoflagellate TEs. Therefore, I identified 16 potential TEs within *M. fluctuans* and 14 potential TEs within *D. grandis* to undergo phylogenetic analysis. It was difficult to discern if comp14821_c0_seq1 is a genuine TE. It has a relatively short protein that, when inputted into Blast, did not appear to have any conserved domains. In Repeat Masker, it had a low E-value of $4.5e^{-07}$ and an E-value of $2e^{-17}$ in Blast. It may be that the sequence is partial and that the domains are in the missing part of the sequence. Another possible theory is that the TE has become a host gene and is no longer a TE. This can happen via natural selection, and happens when the TE ends up providing a function for the host (Sinzelle et al., 2009). Alternatively, the sequence could be a genuine *Tigger* element. If it had been possible to assess *M. fluctuans* in the lab within the time constraints, then it would be attempted to find the full sequence of comp14821_c0_seq1. As pure bioinformatics cannot conclusively say whether this element is a genuine TE or not, laboratory work would have to take place. If multiple copies of the sequence were found, this would suggest that the sequence is a TE and if there was only one copy then it is probably a host gene. However, TEs can be found as single copies when either many copies have been removed from the genome or a TE is new to a genome. A TE could be identified if it has TSDs. Host genes do not have TSDs as they do not transpose. Additionally, if ITRs were found, that would also indicate that the sequence is a genuine TE as host genes do not have these.

Blast results of this sequence show that it has similarities to *jerky*-like domesticated proteins. However, *jerky*-like domesticated proteins have, as of yet, only been found in vertebrates (Sinzelle et al., 2009). For it to be the same domesticated protein in both vertebrates and choanoflagellates, the protein would have to have been lost in all the other animals or there may have been an independent domestication event. It is therefore more likely that comp14821_c0_seq1 has more similarity to *jerky*-like domesticated proteins than to other *Tigger* TEs by chance, and that it is a genuine TE.

5.06 Choanoflagellate Codon Usage

As well as analysing the TEs from *T. pseudonana* and *P. tricornutum*, TEs from the choanoflagellates *Mylonsiga flucutans* and *Diaphanoeca grandis* were investigated. Very little study into TE activity of the eukaryotic group of Stramenopiles has taken place yet, whereas the majority of TE studies have focused on Opisthokonts, which is the eukaryotic group in which choanoflagellates belong. Despite this, there have been no studies which investigate codon usage evolution of TEs within choanoflagellates. In this study, selection for TE codon usage bias has been identified within the unicellular stramenopile *T. pseudonana* and the following sections discuss possibilities for selection for TE codon bias within the unicellular choanoflagellates.

The genome of *M. fluctuans* was found to house twelve potential TEs. Two of the three non-LTR retrotransposons were found to be using significantly more host optimal codons in TE domain regions than non-domain regions. This finding is evidence for the theory that these TEs have evolved via selection accuracy as the domain regions are more functionally important to the TE. The third non-LTR retrotransposon, which did not have significant difference in host optimal codon usage between domain regions and non-domain regions, may have done so for a variety of reasons. It may be a non-functional family, or a family that only transposes rarely, meaning that it is not under selection as much as the other two non-LTR retrotransposons. Copy number could not be determined as only the choanoflagellate transcriptomes were available. Within *C. owczarzaki*, non-LTRs have been found to be under strong selection (M. Carr, personal communication, May 31, 2018). As *C. owczarzaki* are in the sister group to the clade containing both choanoflagellates and metozoa, and are unicellular like choanoflagellates, it could be expected for their TEs to have similar evolution to choanoflagellates. However, the two stramenopiles within this study have shown different trends in their codon usage, so TEs in closely related hosts may not share similar codon usage.

The *copia*-like element also did not have significant difference in using host optimal codons in domain regions compared to non-domain regions. A piwi RNA pathway is an organism's usual way of silencing TEs. When TEs transpose into piwi RNA generating loci, the host begins producing piRNA, which inactivates the TEs (Pélisson et al., 2007). However, *copia* elements have been found to be inactivated via a post-transcriptional method instead (Nuzhdin et al., 1998). It is possible that this *copia* family could be old and have stopped transposing. This is unlikely though as the family was found within a transcriptome, meaning the family must be expressed.

Looking at the transposons within *M. fluctuans*, one of the *piggyBac*-like TEs had significantly more optimal codons in the domain regions than non-domain regions. Another study underway has found similar results in choanoflagellate *Salpingoeca rosetta*, with another transposon - the *Tigger* family - also having significantly more host optimal codons in the domain than non-domain regions (M. Carr, personal communication, May 11, 2018). A

possible theory for this is that the significant *piggyBac*-like TE may be a young family. A young transposon family may be under strong selection because there will not be many autonomous elements to drive the transposition of the non-autonomous elements. The majority of the elements will need their own functional transposases and will be under selection. An older family may have considerably less copies with transposase. These few copies will possess transposase that other copies can use and there will be many transposing copies. As long as the copies have an intact ITR, they can survive with non-functional transposase and there will be no selection. The results that are significant are consistent with translational accuracy.

D. grandis has some amino acids with no optimal codons. (J. J. Ginés, personal communication, June 06, 2018). This suggests that *D. grandis* codon usage is weak, as a species that lacks optimal codons may be under less selection compared to other choanoflagellates and be less effective at translation. Investigating the codon usage of *D. grandis* TEs gives more insight into whether they also have evolved more or less through selection than TEs of other choanoflagellates.
In *D. grandis* only one TE shows significant difference in Fop value between the domain regions and non-domain regions, which is non-LTR retrotransposon comp31080_c2_seq20. This is consistent with the theory that non-LTR retrotransposon codon usage is under selection across choanoflagellates and *C. owczarzaki*, although only a few species have been investigated. The results do indicate that in *M. fluctuans* and *D. grandis*, *copia*-like elements have not strongly evolved through selective accuracy. Based on the fact that there were more elements with significantly more optimal codons in domain regions than non-domain regions in *M. fluctuans* than *D. grandis*, this suggests that whilst selective accuracy may not play the largest part in *M. fluctuans* TEs evolution, it may play a bigger part in *M. fluctuans* TE evolution than in *D. grandis* TE evolution. A possible reason for this is that *D. grandis* may have weaker codon usage than *M. fluctuans*. This may mean that selection upon *D. grandis* TE codon usage may be weaker as purifying selection from *D. grandis* is weaker so the TEs can transpose without being as well adapted to use host translational machinery as *M. fluctuans* TEs must be.
Both choanoflagellates had more significant results comparing host optimal codon usage in domain regions and non-domain regions than in either *P. tricornutum* or *T. pseudonana* TEs. This suggests that both *M. fluctuans* and *D. grandis* TEs are under stronger selection than the TEs of the stramenopiles in this study. It is possible that the choanoflagellate effective population sizes are larger than the stramenopile effective population sizes as this would mean that the choanoflagellate's evolution would be driven by stronger selection. This might result in the TEs codon usage selection being stronger in order for the TEs to survive in the harsher environment.

5.07 *Copia*-like TEs Found Within Choanoflagellates

Copia elements are abundant within plants, fungi, Metozoa (Flavell et al., 1997) and have been identified in other choanoflagellates (Carr et al., 2008). Therefore, it comes as no surprise that *copia*-like elements were found in both *D. grandis* and *M. fluctuans*. Both sequences were incorporated into one *copia* phylogenetic tree. In a eukaryotic tree, plants are on one side of the presumed root and choanoflagellates and metazoans are on the opposing side (Baldauf et al., 2013). That is mostly reflected on the *copia*-like tree generated as part of this study (Figure 17). The majority of the TE phylogeny matches the host phylogeny, except from within the choanoflagellates, where the TE phylogeny does not match the host

phylogeny. As *D. grandis* is the most distantly related to the other choanoflagellates (Carr et al., 2017), it could be expected that the *D. grandis* TE would be the outgroup of the choanoflagellate TEs if purely vertical inheritance had taken place. This is not the case, so the TEs phylogeny does not directly mirror that of the host's. However *D. grandis* TEs are on a long branch. It therefore shares very few amino acid variants with other sequences. Of the variants present, some may be shared due to convergent evolution - rather than common ancestry - with distant relatives, resulting in an erroneous phylogenetic placement. As the erroneous clustering is due to a small number of convergent sites, the support values are often weak. This phenomenon is known as long-branch attraction (Felsenstein, 1978). The branch *D. grandis* is on is poorly supported, supporting the theory that the sequence may be misplaced within the choanoflagellates of the phylogenetic tree. Another aspect of this tree that does not support the theory of only simple vertical inheritance taking place is that whilst choanoflagellates are the sister group to Metazoa, the majority of the Metazoa *copia* TEs are not closely related to the choanoflagellate TEs. This is another example of the TE phylogeny not matching the host phylogeny. Instead of clustering with Metazoa, the choanoflagellate's TEs in this *copia* tree cluster with Fungi TEs. As the metazoans are nestled within the fungi, it is possible that *copia* elements from fungi have radiated and undergone horizontal transfer into both Metazoa and choanoflagellates. However, the support values of the branches separating Metazoa from choanoflagellates and clustering choanoflagellates with fungi are not supported (Figure 17).

The *M. brevicollis* TEs do cluster together and the *M. fluctuans* TE clusters with the *S. rosetta* TEs, which is expected based on the host species phylogeny. These choanoflagellate positions do support the theory of vertical inheritance of the *copia* elements. *D. grandis* TEs cluster together, suggesting that their TEs divergence occurred within *D. grandis* or a choanoflagellate closely related, such as another acanthoecidae. *M. brevicollis* TEs also cluster together so their TEs are also likely to have diverged within *M. brevicollis* or a choanoflagellate closely related to *M. brevicollis*, such as *Choanoaca perplexa* - another clade 1 Craspedida (Carr et al., 2017). Carr & Suga (2014) also found that the two *copia* elements within *M. brevicollis* were closely related to one another and most likely diverged within the choanoflagellate lineage. These choanoflagellate TE positions do support the theory of vertical inheritance of the *copia* elements. It could also be consistent with the theory that a single *copia* element within the ancestor of all these choanoflagellates existed, and that as these choanoflagellates have radiated, so have the *copia* elements. These positions do not rule out the possibility of horizontal transfer between choanoflagellates. *S. rosetta* TEs do not cluster together. It could be that *D. grandis* and some *S. rosetta* TEs are clustering together because an element has jumped from *D. grandis* into *S. rosetta.* Another possibility is that the ancestor of the choanoflagellate had a number of different *copia* elements, such as three or four, and some of the choanoflagellate lineages in this study have kept some of these *copia* elements and other choanoflagellates have kept other *copia* elements. This is complex vertical inheritance. It could be a combination of some of these theories, such as there been both horizontal transfer and ancestral complexity. There are many potential options that this tree does not rule out, but is merely consistent with each of the many theories. As the tree is not strongly supported, the positions may also reflect genuine relationships. In order to develop a closer insight into how these *copia* elements have evolved within choanoflagellates, it would be necessary to involve a greater number of choanoflagellates *copia* elements. In this tree there are only four different choanoflagellate species' *copia* elements included due to time constraints but a more diverse selection incorporated could reveal which of the many possible theories of choanoflagellate *copia* element evolution is most likely. This may also be helped by including additional metazoan and fungal *copia* Pols.

One theory that this tree is not consistent with is that a *copia* element in a choanoflagellate was acquired from another source as the choanoflagellate *copia*-like elements in this study are monophyletic. An ongoing study has found that horizontal transfer of a transposon has occurred from a stramenopile to a choanoflagellate and it has been theorised that this was caused by predator/prey interactions (M. Carr & J. Southworth, personal communication, July 12, 2018). Whilst Tucker (2013) showed evidence that horizontal transfer with choanoflagellates often involve prey species, there is no evidence from other studies to suggest that this might have happened with *copia* elements within choanoflagellates. However, there is only one published paper on TEs within choanoflagellates, which contains data from only one species: *M. brevicollis* (Carr et al., 2008).

In early trees, *Guillardia theta copia* element clustered with plant TEs with moderate to low support. After including sequences similar to *M. fluctuans copia* element, *G. theta* TE moved and clustered with the choanoflagellates. The support values of the choanoflagellate's position was very low. As *G. theta* is single celled alga, it is a possibility that horizontal transfer of a TE could have occurred from *G. theta* to a choanoflagellate, but it could also be a long-branch artefact. Horizontal transfer is unlikely as *G. theta* is a photosynthetic alga and its opportunities for acquiring a TE from a choanoflagellate are likely to be limited. It is possible that a virus could have facilitated the transfer (Filée et al., 2006;2007) but still unlikely as it is improbable that a virus would affect both *G. theta* and choanoflagellates. Once *G. theta* was removed, the support values of the choanoflagellate positions increased, which is consistent with long-branch attraction for the *G. theta* TE. This could mean that the positions that the choanoflagellates are in is genuine and that the *S. rosetta* TEs are genuinely not clustered together, increasing the probability of the theory that an element jumped from *D. grandis* into *S. rosetta*. This is likely because once *G. theta* TE was removed, *D. grandis* TEs and *S. rosetta* TEs clustering together was strongly supported, and the *S. rosetta* TEs not all clustering together became significantly supported. Alternatively, the *S. rosetta* TEs may be diversifying within the same genome. They could be undergoing speciation and diversification within the genome, leading to distinct TE groups in *S. rosetta*. Removing the *G. theta* TE also changed the position of the *M. fluctuans* TE. It was originally clustered with *S.rosetta pseudovirus-4* and *S.* rosetta pseudovirus-5. *M. fluctuans* and *S. rosetta* are not closely related to one another (Carr et al., 2017), and the position had poor support, suggesting it was an artefact and not in its genuine position. *M. fluctuans* TE's new position is not clustering with *S. rosetta* TEs directly but still clusters with the choanoflagellates, suggesting that the *M. fluctuans copia*-like TE is the most distantly related to the other choanoflagellate *copia*-like TEs. Its new position within the choanoflagellates is also significantly supported. Based upon species phylogeny, this is not expected because *M. fluctuans* is a clade 2 Craspedida and is closely related to *M. brevicollis* and *S. rosetta* as they are both clade 1 Craspedidas. This is further support for horizontal transfer of the copia element among the choanoflagellates as the species phylogeny and the TE phylogeny do not mirror one another.

Another thing that this tree does not provide evidence for is the origin of the choanoflagellate *copia*-like elements. It shows that the *copia* elements have been in the choanoflagellates since choanoflagellates evolved as this tree includes TEs from species which fall on either side of the choanoflagellate root. *D. grandis* is an Acanthoecidae compared to the other three Craspedidas (Carr et al., 2017). Despite this, the origin of the *copia* elements in choanoflagellates cannot be discerned with this phylogenetic tree. In early *copia*-like phylogenetic trees within this study, Metazoa *copia*-like elements were found to be closely related to those of the choanoflagellates. This suggests that the *copia* elements have been

present at least since the origin of choanoflagellates and Metazoa. However, the most refined phylogenetic tree has metazoan *copia*-like elements spread around the tree. Within this tree, the support values are very low for the placement of the metazoan TEs, which means that the theory that the *copia*-like elements being present at least since the origin of both choanoflagellates and Metazoa is not disproved.
The *M. fluctuans copia*-like element appears to be split into two open reading frames between the *gag* and *pol* ORFs. There may have been a sequencing error that has caused a frameshift as some sequences in the transcriptome have single A nucleotides incorporated (M. Carr, Personal communication, August 28, 2018). It is likely to be a genuine frameshift as many gag-pol open reading frames require a frameshift for the ribosome to translate Pol.

5.08 MULE-like TEs Found Within Choanoflagellates

MULE TEs have been found in a choanoflagellate (Carr & Suga, 2014) and are abundant in fungi, plants and metazoans (Lisch, 2015; Chalvet et al., 2003) so the finding that *D. grandis* contained a MULE-like element was not surprising. In the MULE phylogenetic tree, *D. grandis* TEs do not cluster with the *S. rosetta* MULE TEs, which was the only other choanoflagellate to contain MULE TEs (Figure 19). The *D. grandis* MULE TEs are distantly related to those of *S. rosetta* and *C. owczarzaki*. *S. rosetta* and *C. owczarzaki* MULE TEs are clustering together with strong support with a Bayesian inference posterior probability of 1.00. *S. rosetta* is a closer relation of *D. grandis* than of *C. owczarzaki*, as *C. owczarzaki* is not a choanoflagellate but a close relative, which suggests that the MULE TEs have not evolved purely through vertical inheritance. Choanoflagellates are more closely related to animals than they are to *C. owczarzaki* but unfortunately, no metazoan MULE TEs were found for this phylogenetic tree to compare. A potential explanation as to why *D. grandis* does not cluster with *S. rosetta* is that there may have been a horizontal transfer of a MULE TE from a stramenopile to *D. grandis*. However, the stramenopile MULE TEs are monophyletic, with significant support, suggesting that the horizontal transfer was not recent, and therefore not a transfer of a TE from the stramenopiles to a choanoflagellate. Instead, a transfer from a eukaryotic group related to stramenopiles, such as alveolates or rhizarians, may have taken place, to a choanoflagellate related to *D. grandis* or *D. grandis* itself. However, all stramenopiles TEs within this tree are oomycota stramenopiles and therefore the species are closely related. A donor species outside this range of sampled oomycota taxa could produce a sister-relationship, even if that donor was a stramenopile. TEs from a greater variety of stramenopiles need to be included in this phylogenetic tree to ascertain from what eukaryotic group a MULE element may have been transferred from, to the choanoflagellate. Also, if an alveolate or rhizarian was the donor group it would most likely have been recovered by Blast. As this wasn't the case, the most likely donor is a stramenopile that is not an oomycota.

The length of the branch that *D. grandis* is on suggests that the transfer may have been to a choanoflagellate related to *D. grandis*, which then was transferred to *D. grandis* via vertical inheritance. Because of the branch length, it is impossible to discern at which point the MULE element was transferred into a choanoflagellate without further investigation. Other choanoflagellates would have to be included in this study to gain a better insight into which choanoflagellate was the recipient of MULE horizontal transfer. As *D. grandis* is an acanthoecid, if only other acanthoecid's MULE elements were found to be closer related to that of stramenopiles than *S. rosetta*, then it would be a plausible theory that the horizontal transfer had occurred in an ancestor of acanthoecids. However, if choanoflagellates other than

acanthoecids were found to have a MULE TE closer related to the MULE in stramenopiles, it would suggest that the horizontal transfer occurred earlier in the evolution of choanoflagellates. Evidence for horizontal transfer of TEs from stramenopiles to choanoflagellates has also been found in other studies (Tucker, 2013), which supports this theory as it has been seen to happen in other choanoflagellates.
Tilletia controversa is a smut fungus and clusters with plants with strong support. It is likely that horizontal transfer of the MULE TE took place from a plant to *T. controversa* as it is a plant pathogen. Host-parasite transfer may have taken place. This type of transfer has been observed within a variety of metazoans (Pace II et al., 2010), but there is a lack of TE host-parasite transfer outside of metazoans.

Before it was removed as a process of refinement, the *Perkinsela sp.* TE clustered with the *D. grandis* TE with significant support and was on the same side of the presumed root of the tree as the choanoflagellate TEs and stramenopile TEs. However, *Perkinsela sp*. is distantly related from choanoflagellates and stramenopiles. *Perkinsela sp*. is an excavate, which is another eukaryotic supergroup (Baldauf et al., 2013) and a plausible theory would be that *D. grandis* has fed on *Perkinsela sp*. and taken up a *Perkinsela sp*. TE. However, *Perkinsela sp*. is an intracellular organism and in order to ingest the *Perkinsela sp.*, the choanoflagellate would have had to also consume *Perkinsela sp*.'s host, which is still plausible. The TE has not come directly from *Perkinsela sp.* – it has come from something related to *Perkinsela sp*. as *Perkinsela sp*. and *D. grandis* are separated from one another by long branches. The TE may have been transferred to *D. grandis* by a free-living excavate relation of *Perkinsela sp*., which would be more likely as *D. grandis* would then not have needed to also consume the excavate's host. This is a predator –prey type transfer. This could be vertical transfer between the TE hosts as a parasite or it may be the case that dispersal cells have found new hosts in *D. grandis*. Predator – prey horizontal transfer appears to be particularly common within choanoflagellates. Prey is caught in the choanoflagellate's mucous net between microvilli at the base of the flagellum (Tucker, 2013). There is also evidence for TE transferring to new species via predator-prey transfer in other studies (Kuraku, 2012). As genes from the prey have been found to be incorporated into the choanoflagellate's genome, it is certainly a possibility that TEs from the prey may also have been incorporated. *Perkinsela sp.* reside in an amoebozoan called *Neoparamoeba*, which are parasitic. This is also, therefore, another potential occurrence of host-parasite transfer, although there is no evidence that *Neoparamoeba* are parasites within choanoflagellates (Tanifuji et al., 2011).
Other than the exceptions of *D. grandis* and *T. contraversa*, the MULE elements phylogeny within this tree matches that of the host's phylogeny. On one side of the tree there are choanoflagellates and fungi, and the other side of the tree consists of stramenopiles and plants. This suggests that for the majority of MULE elements, including those in *S. rosetta*, vertical inheritance has been responsible for the evolution. The fact that there is a variety of host species on both sides of the eukaryotic tree also suggests that MULE has been vertically inherited since the origin of eukaryotes.

5.09 *LINE-1* TEs Found Within Choanoflagellates

In the *Line-1* tree, refining led to the conclusion that the TEs in *D. grandis* that had sequence similarity to *Line-1* elements were not actually *L1* elements (Figure 20). The *D. grandis* TEs did not cluster with the other *L1* elements, but instead clustered more with the *R2* and *jockey* outgroups, though with poor support. This prompted further phylogenetic study into what the TEs actually were. An *R2* tree was generated to attempt to discover what TE the sequence

comp33838_c1_seq9 was. This sequence had clustered with *R2* and *L1* elements within the *L1* tree but in the *R2* tree it was found to be unlikely to be an *R2* element as it did not cluster with the other *R2* elements. It could be distantly related to both types of elements. Additionally, this sequence was on a long branch. The weak support could result from long-branch attraction.
Similar efforts were made to find out what TE the sequence comp30598_c0_seq1 was. In the *L1* tree, this sequence had clustered with *jockey* elements. When a *jockey* tree was generated, it was found that comp30598_c0_seq1 did not cluster with *jockey, R2* or *L1* elements. Investigation of sequences found to be similar to this sequence with the use of Blast found that the closest named TE to comp30598_c0_seq1 was *jockey*. This finding suggests that comp30598_c0_seq1 is similar to a *jockey* element. Future study would involve further examination of these two sequences, which are Non-LTR retrotransposons that have similarities to *L1, R2* and *jockey* TEs. Comp30598_c0_seq1 is split into two open reading frames, so further study would involve using PCR to confirm if the annotation for this sequence is genuine or not. It may be that there is a frameshift between the *gag* and *pol* ORFs, or the split could be due to poor sequencing.
Other *D. grandis* sequences were all found to have similarity to *L1* elements by using Blastp. These sequences are: comp33838_c1_seq9; comp33838_c1_seq5; comp31080_c2_seq19 and comp20496_c0_seq2. However these sequences were too short to align and it was not possible to make a phylogenetic tree including these sequences. The proteins were 82 to 279 amino acids long. It is possible, therefore, that *D. grandis* has *L1* elements. These TEs could also be non-LTR retrotransposons that have similarities to *L1*, and not actually be *L1* elements.
Amoebozoans and opisthokonts are on one side of the presumed root of the eukaryotic tree and, other than a few stramenopiles, that is what is mainly present in the tree. The tree is therefore consistent with the theory that these elements have been vertically inherited within the Amorphea. Amorphea is the group which consists of Amoebozoa and Obazoa. Opisthokonts are within the group Obazoa (Cavalier-Smith, 2002).

5.10 *Chromoviruses* Found Within Choanoflagellates

Chromoviruses are not present in many metazoans (Kordiš, 2005). Either they have been lost many times or they have been transferred into the metazoans from another organism. It has been proposed that the donor species may have been a fungi (Carr & Suga, 2014). *Chromoviruses* are present in amphibians, reptiles and fish but no other metazoans, which suggests that regardless of how the elements were inherited, they have been lost again as they were present in the ancestor of vertebrates. Many close relatives of metazoans have *chromoviruses*, including choanoflagellates, fungi and *C. owczarzaki* (Carr & Suga, 2014; Gorinsek et al., 2004). The *gypsy*-like TE found in M. fluctuans was included in the *chromovirus* tree to gain further insight into what element it is. Its position is not supported. Blast searches suggest that it is a gypsy element but as the chromodomain is at the end of the pol, it may be the case that the sequence is actually a *chromovirus* and is missing the end of the pol ORF.
The *chromovirus* tree shows the choanoflagellate *chromovirus* TEs clustering together, with the *C. owczarzaki chromoviruses* and the fungal *chromoviruses*. Metazoa *chromoviruses* cluster together forming an outgroup. This is consistent with vertical inheritance since the last common ancestor of fungi and metazoa, as well as with horizontal transfer from fungi into the group of vertebrates included in the phylogenetic tree. However, as the metazoa *chromoviruses* are in a separate clade in this study, these findings do not support either

theory. A very limited selection of vertebrates and fungi have been included in this study due to time constraints and the fact that very few vertebrates with *chromoviruses* have had their genomes sequences. Whilst the groups are well supported, a larger variety of *chromoviruses* from other organisms needs to be included for the tree to not be consistent with the theories.

6.0 Conclusion

Transposable element activity and evolution within unicellular eukaryotes is an area of biology where there is much more to be discovered. TE codon usage has been found to be under selection in unicellular organisms, with previous studies found TE codon usage to be under selection in the host *P. infestans* (Jiang & Govers, 2006) and within this current study, TE codon usage has been found to be under selection in the hosts *T. pseudonana*, *D. grandis* and *M. fluctuans*. As there was little evidence for selection in the TE codon usage of *P. tricornutum*, it appears that not all stramenopiles have strong selection for codon usage, suggesting selection for TE codon usage is not universal within unicellular eukaryotes. Therefore, if TEs from another stramenopile were studied, it would be impossible to correctly predict what evolutionary pressures would be acting upon them. There is great potential for future work covering TE evolution in stramenopiles. Equally, there are still many eukaryotic groups where very little investigation into their TE evolution has taken place. Rhizaria, Alveolata and Amoebozoa species may also host TEs with strong selection for codon usage. Other future work as an extension of this study could involve a greater selection of stramenopile and choanoflagellate TE's codon usage compared to one another to see if there is a pattern of choanoflagellate TEs having stronger codon usage than stramenopile TEs, or if that was just the case for the species in this study. Differences between evolution of Class I and class II TEs could be investigated, or differences between various TE families such as *gypsy* and *copia* families. Within gypsy families, evolution and codon usage of *chromovisuses* and non-chromoviral *gypsy*-like families could be compared. Another area of codon usage that could also be explored is the relationship between the level of host codon usage bias and TE codon usage bias. If a host does not have translational machinery adapted for selection on codon usage, it is not expected the TEs to show selection. However, as so few unicellular eukaryotic TE's codon usage have been studied, it could be the case that TEs do show codon usage bias regardless of the host's codon bias. There may also be a pattern in that as the codon bias of a host is higher, the TE codon usage bias may also be higher. Additionally, laboratory work could be implemented to find out if the TEs identified within *M. fluctuans* and *D. grandis* were genuine, in order to further explore the origins of these TEs. This would involve using primers from the upstream and downstream sections of the TE sequences, finding out the true sequence in between these sections. The primers could then be reverse complemented to perform inverse PCR to obtain the rest of the TE sequence. This would result in fully annotated TEs. Laboratory work could also identify any links between TE codon usage via selection and gene expression levels.

Though it is unlikely that either stramenopile in this study showed evidence for TE codon usage selection as strong as that in *P. infestans*, it appears that *T. pseudonana* TE's codon usage are also evolving under selection. Mutation pressure is the largest driver of codon usage evolution in the TEs of *P. tricornutum*. *P. tricornutum* and *T. pseudonana* are much closer related to one another than they are to *P. infestans*. The fact that there is evidence that both *T. pseudonana* and *P. infestans* TE's codon usage in under selection suggests that there is not one clade of stramenopiles that show stronger selection on TE codon usage than

another, but that selection on TE codon usage may vary between each individual stramenopile's TEs.
It is possible that *P. infestans* has the largest effective population size, hence stronger codon usage selection which is driving the selective evolution of TEs. Following this theory, it would stand to reason that *P. tricornutum* had the smallest effective population out of these three stramenopiles. This is supported by the fact that *P. infestans* is the only stramenopile of these three to definitely undergo sexual reproduction (Yuen & Andersson, 2013; Moore et al., 2017; Falciatore & Bowler, 2002). Sexual reproduction increases a species effective population and lack of it may supress effective population size. However, *P. infestans* is also a parasite whereas *P. tricornutum* and *T. pseudonana* are free living. A parasitic organism's effective population size is dependent on the host, meaning that it would be expected for a free-living organism to have a higher effective population size than a parasitic organism. When *P. infestans* organisms transfer from one host to another, their effective population will be bottlenecked, as only a small proportion will survive. Bottlenecking reduces the effective population size, making the host selection less efficient.
It is very difficult to estimate an effective population size in these protists as there is no population data. Protist effective population sizes differ massively (Snoke et al., 2006; Watts et al., 2013). Additionally, effective population size has been shown to effect the codon usage of the species, but its effects upon TE codon usage are unknown. It is likely that it has effects but these may manifest themselves in ways hitherto undiscovered. A logical theory would be that the larger the host effective population size, the larger the TE effective population size is.

As both choanoflagellates presented evidence for TE codon usage selection and no other studies have investigated TE codon usage evolution within choanoflagellates, it is unknown whether selection for TE codon usage is universal within choanoflagellates. It may be that just the two choanoflagellates within this study show evidence for selection in TE codon usage. However, this study does show that selection for TE codon usage exists in choanoflagellates - the closest unicellular relative to Metazoa. Additionally, choanoflagellates have been shown to undergo sexual reproduction (Levin & King, 2013). Selection appeared to be stronger within the choanoflagellates compared to the stramenopiles within this study, which could be because the choanoflagellate's effective population size is larger than that of the stramenopiles.
TEs in a broad range of unrelated organisms show a weak bias towards AT-ending codons, whereas this study, along with Jiang & Gover's (2006) paper on *P. infestens*, show that at least in some TE families within unicellular eukaryotes do have the signature of selection on codon usage.

7.0 References:

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., . . . Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology, 59*(5), 429-514. doi:10.1111/j.1550-7408.2012.00644.x

Akashi, H. (1994). Synonymous codon usage in drosophila melanogaster: Natural selection and translational accuracy. *Genetics, 136*(3), 927-935. Retrieved from http://library.hud.ac.uk/summon

Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1006/jmbi.1990.9999

Arian F. A. Smit, & Riggs, A. D. (1996). Tiggers and other DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the United States of America, 93*(4), 1443-1448. doi:10.1073/pnas.93.4.1443

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., . . . Rokhsar, D. S. (2004). The genome of the diatom thalassiosira pseudonana: Ecology, evolution, and metabolism. *Science, 306*(5693), 79-86. doi:10.1126/science.1101156

Asadzadeh, S. S., Nielsen, L. T., Andersen, A., Dölger, J., Kiørboe, T., Larsen, P. S., & Walther, J. H. (2019). Hydrodynamic functionality of the lorica in choanoflagellates. *Journal of the Royal Society Interface, 16*(150), 20180478. doi:10.1098/rsif.2018.0478

Austin Bourke, P. M. (1964). Emergence of potato blight, 1843-46. *Nature, 203*(4947), 805-808. doi:10.1038/203805a0

Azam, F., & Bidle, K. D. (1999). Accelerated dissolution of diatom silica by marine bacterial assemblages. *Nature, 397*(6719), 508-512. doi:10.1038/17351

Baldauf, S. L., Romeralo, M., & Carr, M. (2013). The evolutionary origin of animals and fungi. In *Evolution from the Galapagos: Two Centuries after Darwin* (pp. 73-106). Springer.

Bariana, M., Aw, M. S., Kurkuri, M., & Losic, D. (2013). Tuning drug loading and release properties of diatom silica microparticles by surface modifications. *International Journal of Pharmaceutics, 443*(1-2), 230-241. doi:10.1016/j.ijpharm.2012.12.012

Barnett, L., Brenner, S., F. H. C. Crick, Shulman, R. G., & Watts-Tobin, R. J. (1967). Phase-shift and other mutants in the first part of the rII B cistron of bacteriophage T4.*Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 252*(780), 487. doi:10.1098/rstb.1967.0030

Bartual, A., Villazán, B., & Brun, F. G. (2011). Monitoring the long-term stability of pelagic morphotypes in the model diatom phaeodactylum tricornutum. *Diatom Research, 26*(2), 243. doi:10.1080/0269249X.2011.619365

Belshaw, R., Anna L. A. Dawson, Woolven-Allen, J., Redding, J., Burt, A., & Tristem, M. (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *Journal of Virology, 79*(19), 12507-12514. doi:10.1128/JVI.79.19.12507-12514.2005

Berg, D.E. & Howe, M.M. (Eds.) (1989). *Mobile DNA*. Washington, D.C: American Society for Microbiology.

Boeke, J.D. (1989). Transposable Elements in Saccharomyces cerevisiae. In D . E. Berg & M .M. Howe (Eds.) *Mobile DNA* (pp. 336-356). Washington, D.C: American Society for Microbiology.

Bolotin, E., & Hershberg, R. (2017). Horizontally acquired genes are often shared between closely related bacterial species. *Frontiers in Microbiology, 8*, 1536. doi:10.3389/fmicb.2017.01536

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., . . . Roger, A. J. (2018). Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biology and Evolution*, 10(2), 427-433. doi:10.1093/gbe/evy014

Carr, M., & Suga, H. (2014). The holozoan capsaspora owczarzaki possesses a diverse complement of active transposable element families. *Genome Biology and Evolution, 6*(4), 949-963. doi:10.1093/gbe/evu068

Carr, M., Bensasson, D., & Bergman, C. M. (2012). Evolutionary genomics of transposable elements in saccharomyces cerevisiae. *PloS One, 7*(11), e50978. doi:10.1371/journal.pone.0050978

Carr, M., Leadbeater, B. S. C., Hassan, R., Nelson, M., Baldauf, S. L. (2008). Molecular phylogeny of choanoflagellates, the sister group to metazoa. *Proceedings of the National Academy of Sciences of the United States of America, 105*(43), 16641-16646. doi:10.1073/pnas.0801667105

Carr, M., Nelson, M., Leadbeater, B. S. C., Baldauf, S. L. (2008). Three families of LTR retrotransposons are present in the genome of the choanoflagellate monosiga brevicollis. *Protist, 159*(4), 579-590. doi:10.1016/j.protis.2008.05.001

Carr, M., Richter, D. J., Fozouni, P., Smith, T. J., Jeuck, A., Leadbeater, B. S. C., & Nitsche, F. (2017). A six-gene phylogeny provides new insights into choanoflagellate evolution. *Molecular Phylogenetics and Evolution, 107*, 166-178. doi:10.1016/j.ympev.2016.10.011

Carr, M., Soloway, J. R., Robinson, T. E., & Brookfield, J. F. (2002). Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of drosophila melanogaster. *Chromosoma, 110*(8), 511-518. doi:10.1007/s00412-001-0174-0

Casacuberta, J. M., & Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: Control of transposition and impact on the evolution of plant genes and genomes. *Gene, 311*, 1-11. doi:10.1016/S0378-1119(03)00557-2

Cavalier-Smith, T. (1993). Kingdom protozoa and its 18 phyla. *Microbiological Reviews, 57*(4), 953-994. Retrieved from http://library.hud.ac.uk/summon

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *International Journal of Systematic and Evolutionary Microbiology, 52*(2), 297-354. doi:10.1099/00207713-52-2-297

Cerritelli, S. M., & Crouch, R. J. (2009). Ribonuclease H: The enzymes in eukaryotes. *FEBS Journal, 276*(6), 1494-1505. doi:10.1111/j.1742-4658.2009.06908.x

Chalvet, F., Grimaldi, C., Kaper, F., Langin, T., & Daboussi, M. (2003). Hop, an active mutator-like element in the genome of the fungus fusarium oxysporum. *Molecular Biology and Evolution, 20*(8), 1362-1375. doi:10.1093/molbev/msg155

Cheng, X., Zhang, D., Cheng, Z., Keller, B., & Ling, H. (2009). A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. *Genetics, 181*(4), 1183-1193. doi:10.1534/genetics.108.099150

Crawford, B. J., & Campbell, S. S. (1993). The microvilli and hyaline layer of embryonic asteroid epithelial collar cells: A sensory structure to determine the position of locomotory cilia?. *The Anatomical record, 236*(4), 697-709. doi:10.1002/ar.1092360414

Daniels, S. B., Peterson, K. R., Strausbaugh, L. D., Kidwell, M. G., & Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between drosophila species. *Genetics, 124*(2), 339-355. Retrieved from http://library.hud.ac.uk/summon

de Koning, A P Jason, Gu, W., Castoe, T. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genetics, 7*(12), e1002384. doi:10.1371/journal.pgen.1002384

Derelle, R., López-García, P., Timpano, H., & Moreira, D. (2016). A phylogenomic framework to study the diversity and evolution of stramenopiles (=Heterokonts). *Molecular Biology and Evolution, 33*(11), 2890-2898. doi:10.1093/molbev/msw168

Diao, X., Freeling, M., & Lisch, D. (2006;2005). Horizontal transfer of a plant transposon.*PLoS Biology, 4*(1), e5-e5. doi:10.1371/journal.pbio.0040005

Donovan, G. P., Buzsaki, G., Grimsby, J., & Toth, M. (1995). Epileptic seizures caused by inactivation of a novel gene, jerky, related to centromere binding protein-B in transgenic mice. *Nature Genetics, 11*(1), 71-75. doi:10.1038/ng0995-71

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell, 134*(2), 341-352. doi:10.1016/j.cell.2008.05.042

Ehrenberg, M., & Kurland, C. G. (1984). Costs of accuracy determined by a maximal growth rate constraint. *Quarterly Reviews of Biophysics, 17*(1), 45-82. doi:10.1017/S0033583500005254

Falciatore, A., & Bowler, C. (2002). Revealing the molecular secrets of marine diatoms. *Annual Review of Plant Biology, 53*, 109. Retrieved from http://library.hud.ac.uk/summon

Falkowski, P. G., Barber, R. T., & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science, 281*(5374), 200-206. doi: 10.1126/science.281.5374.200

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology, 27*(4), 401-410. doi:10.2307/2412923

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics, 9*(5), 397-405. doi:10.1038/nrg2337

Filée, J., Siguier, P., & Chandler, M. (2006;2007). I am what I eat and I eat what I am: Acquisition of bacterial genes by giant viruses. *Trends in Genetics, 23*(1), 10-15. doi:10.1016/j.tig.2006.11.002

Finnegan, D. J. (2012). retrotransposons. *Current Biology, 22*(11), R432-R437. doi:10.1016/j.cub.2012.04.025

Flavell, A. J., Pearce, S. R., Heslop-Harrison, P., & Kumar, A. (1997). The evolution of Ty1-copia group retrotransposons in eukaryote genomes. *Genetica, 100*(1), 185-195. doi:10.1023/A:1018385713293

Flavell, A., Chalhoub, B., Sabot, F., Hua-Van, A., Leroy, P., Panaud, O., . . . Morgante, M. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics, 8*(12), 973-982. doi:10.1038/nrg2165

Fletcher, K. I. G., van West, P., & Gachon, C. M. M. (2016). Nonagonal cadherins: A new protein family found within the stramenopiles. *Gene, 593*(1), 64-75. doi:10.1016/j.gene.2016.08.003

Fraser, M. J., Smith, G. E., & Summers, M. D. (1983). Acquisition of host cell DNA sequences by baculoviruses: Relationship between host DNA insertions and FP mutants of autographa californica and galleria mellonella nuclear polyhedrosis viruses. *Journal of Virology, 47*(2), 287-300. Retrieved from http://library.hud.ac.uk/summon

Gao, X., Hou, Y., Ebina, H., Levin, H. L., & Voytas, D. F. (2008). Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Research, 18*(3), 359-369. doi:10.1101/gr.7146408

Gorinsek, B., Gubensek, F., & Kordis, D. (2004). Evolutionary genomics of chromoviruses in eukaryotes. *Molecular Biology and Evolution, 21*(5), 781-798. doi:10.1093/molbev/msh057

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., . . . Glykovetenskap. (2009). Genome sequence and analysis of the irish potato famine pathogen phytophthora infestans. *Nature, 461*(7262), 393-398. doi:10.1038/nature08358

Hancks, D. C., & Kazazian, J., Haig H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA, 7*, 9. doi:10.1186/s13100-016-0065-9

Hartl, D. L., Lohe, A. R., & Lozovskaya, E. R. (1997). MODERN THOUGHTS ON AN ANCYENT MARINERE:Function, evolution, regulation. *Annual Review of Genetics, 31*(1), 337-358. doi:10.1146/annurev.genet.31.1.337

Hehenberger, E., Tikhonenkov, D. V., Kolisko, M., del Campo, J., Esaulov, A. S., Mylnikov, A. P., & Keeling, P. J. (2017). Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Current Biology, 27*(13), 2043-2050.e6. doi:10.1016/j.cub.2017.06.006

Hine, R. & Martin, E. (Eds). (2016). *long terminal repeat*. (7th ed.) Oxford University Press.

Inada, E., Saitoh, I., Watanabe, S., Aoki, R., Miura, H., Ohtsuka, M., . . . Sato, M. (2015). Piggy bac transposon-mediated gene delivery efficiently generates stable transfectants derived from cultured primary human deciduous tooth dental pulp cells (HDDPCs) and HDDPC-derived iPS cells. *International Journal of Oral Science, 7*(3), 144-154. doi:10.1038/ijos.2015.18

Ito, H., & Kakutani, T. (2014). Control of transposable elements in arabidopsis thaliana. *Chromosome Research, 22*(2), 217-223. doi:10.1007/s10577-014-9417-9

Ivancevic, A. M., Kortschak, R. D., Bertozzi, T., & Adelson, D. L. (2018). Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. *Genome Biology, 19*(1), 1-13. doi:10.1186/s13059-018-1456-7

Jackman, J. E., & Alfonzo, J. D. (2013). Transfer RNA modifications: Nature's combinatorial chemistry playground. *Wiley Interdisciplinary Reviews: RNA, 4*(1), 35-48. doi:10.1002/wrna.1144

Jia, J., & Xue, Q. (2009). Codon usage biases of transposable elements and host nuclear genes in arabidopsis thaliana and oryza sativa. *Genomics, Proteomics & Bioinformatics, 7*(4), 175. doi: 10.1016/S1672-0229(08)60047-9

Jiang, N., Eddy, S. R., Bao, Z., Zhang, X., & Wessler, S. R. (2004). Pack-MULE transposable elements mediate gene evolution in plants. *Nature, 431*(7008), 569-573. doi:10.1038/nature02953

Jiang, N., Feschotte, C., & Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. *Nature Reviews Genetics, 3*(5), 329-341. doi:10.1038/nrg793

Jiang, R. H. Y., & Govers, F. (2006). Nonneutral GC3 and retroelement codon mimicry in phytophthora. *Journal of Molecular Evolution, 63*(4), 458-472. doi:10.1007/s00239-005-0211-3

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772-780. doi:10.1093/molbev/mst010

Khan, H., Smit, A., & Boissinot, S. (2006;2005). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research, 16*(1), 78-87. doi:10.1101/gr.4001406

King, N. (2005). choanoflagellates. *Current Biology, 15*(4), R113-R114. 10.1016/j.cub.2005.02.004

Komar, A. A. (2016). The yin and yang of codon usage. *Human Molecular Genetics, 25*(R2), R77-R85. doi:10.1093/hmg/ddw207

Kordiš, D. (2005). A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. *Gene, 347*(2), 161-173. doi:10.1016/j.gene.2004.12.017

Kuraku, S., Qiu, H., & Meyer, A. (2012). Horizontal transfers of Tc1 elements between teleost fishes and their vertebrate parasites, lampreys. *Genome Biology and Evolution, 4*(9), 929-936. doi:10.1093/gbe/evs069

Lambert, C., Ivanov, P., & Sockett, R. E. (2010). A transcriptional "scream" early response of E. coli prey to predatory invasion by bdellovibrio. *Current Microbiology, 60*(6), 419. doi:10.1007/s00284-009-9559-8

Lerat, E., Capy, P., & Biémont, C. (2002). Codon usage by transposable elements and their host genes in five species. *Journal of Molecular Evolution, 54*(5), 625-637. doi:10.1007/s00239-001-0059-0

Lesecque, Y., Keightley, P. D., & Eyre-Walker, A. (2012). A resolution of the mutation load paradox in humans. *Genetics, 191*(4), 1321-1330. doi:10.1534/genetics.112.140343

Levin, T., & King, N. (2013). Evidence for sex and recombination in the choanoflagellate salpingoeca rosetta. *Current Biology, 23*(21), 2176-2180. doi:10.1016/j.cub.2013.08.061

Lim, G. W., Lim, J. K., Ahmad, A. L., & Chan, D. J. C. (2015). Influences of diatom frustule morphologies on protein adsorption behavior. *Journal of Applied Phycology, 27*(2), 763-775. doi:10.1007/s10811-014-0356-9

Lisch, D. (2015). Mutator and MULE transposons. *Microbiology Spectrum, 3*(2), MDNA3. doi:10.1128/microbiolspec.MDNA3-0032-2014

Liu, W., Seto, J., Sibille, E., & Toth, M. (2003). The RNA binding domain of jerky consists of tandemly arranged helix-turn-Helix/Homeodomain-like motifs and binds specific sets of mRNAs. *Molecular and Cellular Biology, 23*(12), 4083-4093. doi:10.1128/MCB.23.12.4083-4093.2003

Longnecker, K., Kido Soule, M. C., & Kujawinski, E. B. (2015). Dissolved organic matter produced by thalassiosira pseudonana. *Marine Chemistry, 168*, 114-123. 10.1016/j.marchem.2014.11.003

Lopes, F. R., Silva, J. C., Benchimol, M., Costa, G. G. L., Pereira, G. A. G., & Cararetо, C. M. A. (2009). The protist trichomonas vaginalis harbors multiple lineages of transcriptionally active mutator-like elements. *BMC Genomics, 10*(1), 330-330. doi:10.1186/1471-2164-10-330

Lowe, T. M., & Chan, P. P. (2016). tRNAscan-SE on-line: Integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research, 44*(W1), W54-W57. doi:10.1093/nar/gkw413

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*(5), 955-964. doi:10.1093/nar/25.5.0955

Lynch, M. (2006). The origins of eukaryotic gene structure. *Molecular Biology and Evolution, 23*(2), 450-468. doi:10.1093/molbev/msj050

Mah, J. L., Christensen-Dalsgaard, K. K., & Leys, S. P. (2014). Choanoflagellate and choanocyte collar-flagellar systems and the assumption of homology. *Evolution & Development, 16*(1), 25-37. doi:10.1111/ede.12060

Malik, H. S. (2005). Ribonuclease H evolution in retrotransposable elements. *Cytogenetic and Genome Research, 110*(1-4), 392-401. doi:10.1159/000084971

Manton, I., Bremer, G., & Oates, K. (1981). Problems of structure and biology in a large collared flagellate (diaphanoeca grandis ellis) from arctic seas. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 213*(1190), 15-26. doi:10.1098/rspb.1981.0050

Marquez, C. P., & Pritham, E. J. (2010). Phantom, a new subclass of mutator DNA transposons found in insect viruses and widely distributed in animals. *Genetics, 185*(4), 1507-1517. doi:10.1534/genetics.110.116673

Martin, S. L., Cruceanu, M., Branciforte, D., Wai-lun Li, P., Kwok, S. C., Hodges, R. S., & Williams, M. C. (2005). LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. *Journal of Molecular Biology, 348*(3), 549-561. doi:10.1016/j.jmb.2005.03.003

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America, 36*(6), 344-355. doi:10.1073/pnas.36.6.344

Medina, M., Collins, A. G., Taylor, J. W., Valentine, J. W., Lipps, J. H., Amaral-Zettler, L., & Sogin, M. L. (2003). Phylogeny of opisthokonta and the evolution of multicellularity and complexity in fungi and metazoa. *Intenrational Journal of Astrobiology, 2*(3), 203-211. doi: 10.1017/S1473550403001551

Miller, W. J., Hagemann, S., Reiter, E., & Pinsker, W. (1992). P-element homologous sequences are tandemly repeated in the genome of drosophila guanche. *Proceedings of the National Academy of Sciences of the United States of America, 89*(9), 4018-4022. doi:10.1073/pnas.89.9.4018

Miller, W. J., McDonald, J. F., Nouaud, D., & Anxolabéhère, D. (1999). Molecular domestication – more than a sporadic episode in evolution. *Genetica, 107*(1), 197-207. doi:10.1023/A:1004070603792

Moore, E. R., Bullington, B. S., Weisberg, A. J., Jiang, Y., Chang, J., & Halsey, K. H. (2017). Morphological and transcriptomic evidence for ammonium induction of sexual reproduction in thalassiosira pseudonana and other centric diatoms. *PloS One, 12*(7), e0181098. doi:10.1371/journal.pone.0181098

Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., & Kazazian, H. H. (1996). High frequency retrotransposition in cultured mammalian cells. *Cell, 87*(5), 917-927. doi:10.1016/S0092-8674(00)81998-4

Neuvéglise, C., Chalvet, F., Wincker, P., Gaillardin, C., & Casaregola, S. (2005). Mutator-like element in the yeast yarrowia lipolytica displays multiple alternative splicings. *Eukaryotic Cell, 4*(3), 615-624. doi:10.1128/EC.4.3.615-624.2005

Nitsche, F., Carr, M., Arndt, H., & Leadbeater, B. S. C. (2011). Higher level taxonomy and molecular phylogenetics of the choanoflagellatea. *Journal of Eukaryotic Microbiology, 58*(5), 452-462. doi:10.1111/j.1550-7408.2011.00572.x

Norris, R. E. (1965). Neustonic marine craspedomonadales (choanoflagellates) from washington and california. *Deep Sea Research and Oceanographic Abstracts, 13*(4), 780. doi:10.1016/0011-7471(66)90666-8

Norton, T. A., Melkonian, M., & Andersen, R. A. (1996). Algal biodiversity. *Phycologia, 35*(4), 308-326. doi:10.2216/i0031-8884-35-4-308.1

Nuzhdin, S. V., Pasyukova, E. G., Morozova, E. A., & Flavell, A. J. (1998). Quantitative genetic analysis of copia retrotransposon activity in inbred drosophila melanogaster lines. *Genetics, 150*(2), 755-766. Retrieved from http://library.hud.ac.uk/summon

Pace II, J. K., Brindley, P. J., Gilbert, C., Schaack, S., & Feschotte, C. (2010). A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature, 464*(7293), 1347-1350. doi:10.1038/nature08939

Papenfuss, A. T., Gu, W., Kamal, M., Aken, B., Mauceli, E., Sharpe, T., . . . Uppsala universitet. (2007). Genome of the marsupial monodelphis domestica reveals innovation in non-coding sequences. *Nature, 447*(7141), 167-177. doi:10.1038/nature05805

Park, C. & Allerby, M. (Eds). (2017). *photic zone*. (3rd ed.) Oxford University Press.

Parkinson, J., & Gordon, R. (1999). Beyond micromachining: The potential of diatoms. *Trends in Biotechnology, 17*(5), 190-196. doi:10.1016/S0167-7799(99)01321-9

Paulis, M., Moralli, D., Bensi, M., De Carli, L., & Raimondi, E. (2004). Isolation from the horse genome of a new DNA transposon belonging to the tigger family. *Mammalian Genome, 15*(5), 399-403. doi:10.1007/s00335-004-3040-6

Peden, J. F. (2000). *Analysis of codon usage.* Retrieved from http://library.hud.ac.uk/summon

Pélisson, A., Sarot, E., Payen-Groschêne, G., & Bucheton, A. (2007). A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the drosophila ovary. *Journal of Virology, 81*(4), 1951-1960. doi:10.1128/JVI.01980-06

Pereira, V. (2004). Insertion bias and purifying selection of retrotransposons in the arabidopsis thaliana genome. *Genome Biology, 5*(10), R79-R79. doi: 10.1186/gb-2004-5-10-r79

Poulsen, N., Berne, C., Spain, J., & Kröger, N. (2007). Silica immobilization of an enzyme through genetic engineering of the diatom thalassiosira pseudonana. *Angewandte Chemie - International Edition, 46*(11), 1843-1846. doi:10.1002/anie.200603928

Precup, J., & Parker, J. (1987). Missense misreading of asparagine codons as a function of codon identity and context. *Journal of Biological Chemistry, 262*(23), 11351-11355. Retrieved from http://library.hud.ac.uk/summon

Quax, T.E.F., Claassens, N.J., Soll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell, 59* (2), 149-161. doi: 10.1016/j.molcel.2015.05.035.

Rambaut, A. (2007). *Figtree.* Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Rasmussen, B., Kilburn, M. R., Brocks, J. J., & Fletcher, I. R. (2008). Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature, 455*(7216), 1101-1104. doi:10.1038/nature07381

Ratcliff, W., Denison, R. F., Borrello, M., & Travisano, M. (2012). Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences of the United States of America, 109*(5), 1595-1600. doi:10.1073/pnas.1115323109

Richter, D. J., Fozouni, P., Eisen, M. B., & King, N. (2018). Gene family innovation, conservation and loss on the animal stem lineage. *Elife, 7* doi:10.7554/eLife.34226f

Roberston, H.M. (2002) Evolution of DNA transposons in eukaryotes. (pp. 1093-1110) ASM Press. doi:10.1128/9781555817954.ch48

Robertson, D.S. (1978). Characterization of a mutator system in maize. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis, 51*(1), 21-28. doi: 10.1016/0027-5107(78)90004-0.

Rouzic, A. L., Boutin, T. S., & Capy, P. (2007). Long-term evolution of transposable elements.*Proceedings of the National Academy of Sciences of the United States of America, 104*(49), 19375-19380. doi:10.1073/pnas.0705238104

Salamov, A., Gruber, A., Mock, T., Robison, M., Rynearson, T. A., Maumus, F., . . . Rychlewski, L. (2008). The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature, 456*(7219), 239-244. doi:10.1038/nature07410

Saville-Kent, W. (1880). *A manual of the infusoria*. England:United Kingdom

Schaack, S., Gilbert, C., & Feschotte, C. (2010). Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution, 25*(9), 537-546. doi:10.1016/j.tree.2010.06.001

Schwarz-Sommer, Z., Leclercq, L., Göbel, E., & Saedler, H. (1987). Cin4, an insert altering the structure of the A1 gene in zea mays, exhibits properties of nonviral retrotransposons. *The EMBO Journal, 6*(13), 3873-3880. doi:10.1002/j.1460-2075.1987.tb02727.x

Sebé-Pedrós, A., Peña, M., Capella-Gutiérrez, S., Antó, M., Gabaldón, T., Ruiz-Trillo, I., & Sabidó, E. (2016). High-throughput proteomics reveals the unicellular roots of animal

phosphosignaling and cell differentiation. *Developmental Cell, 39*(2), 186-197. doi:10.1016/j.devcel.2016.09.019

Shalchian-Tabrizi, K., Minge, M. A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K. S., & Cavalier-Smith, T. (2008). Multigene phylogeny of choanozoa and the origin of animals. *PloS One, 3*(5), e2098. doi:10.1371/journal.pone.0002098

Shirasu, K., Schulman, A. H., Lahaye, T., & Schulze-Lefert, P. (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research, 10*(7), 908-915. doi:10.1101/gr.10.7.908

Shiratori, T., Thakur, R., & Ishida, K. (2017). Pseudophyllomitus vesiculosus (larsen and patterson 1990) lee, 2002, a poorly studied phagotrophic biflagellate is the first characterized member of stramenopile environmental clade MAST-6. *Protist, 168*(4), 439-451. doi:10.1016/j.protis.2017.06.004

Sinzelle, L., Izsvák, Z., & Ivics, Z. (2009). Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences, 66*(6), 1073-1093. doi:10.1007/s00018-009-8376-3

Smit, A.F.A., Hubley, R., & Green, P. *RepeatMasker Web Server*. Retrieved from http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.

Snoke, M. S., Berendonk, T. U., Barth, D., & Lynch, M. (2006). Large global effective population sizes in paramecium. *Molecular Biology and Evolution, 23*(12), 2474-2479. doi:10.1093/molbev/msl128

Southworth, J., Armitage, P., Fallon, B., Dawson, H., Bryk, J., & Carr, M. (2018). Patterns of ancestral animal codon usage bias revealed through holozoan protists. *Molecular Biology and Evolution, 35*(10), 2499-2511. doi: 10.1093/molbev/msy157

Suga, H., Chen, Z., de Mendoza, A., Sebé-Pedrós, A., Brown, M. W., Kramer, E., . . . Ruiz-Trillo, I. (2013). The capsaspora genome reveals a complex unicellular prehistory of animals.*Nature Communications, 4*(1), 2325. doi:10.1038/ncomms3325

Takahashi, F. (2016). Blue-light-regulated transcription factor, aureochrome, in photosynthetic stramenopiles. *Journal of Plant Research, 129*(2), 189-197. doi:10.1007/s10265-016-0784-5

Tanifuji, G., Kim, E., Onodera, N. T., Gibeault, R., Dlutek, M., Cawthorn, R. J., . . . Archibald, J. M. (2011). Genomic characterization of neoparamoeba pemaquidensis (amoebozoa) and its kinetoplastid endosymbiont. *Eukaryotic Cell, 10*(8), 1143-1146. doi:10.1128/EC.05027-11

Tucker, R. P. (2013). Horizontal gene transfer in choanoflagellates. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 320*(1), 1-9. doi:10.1002/jez.b.22480

Walker, C. A., & van West, P. (2007). Zoospore development in the oomycetes. *Fungal Biology Reviews, 21*(1), 10-18. doi:10.1016/j.fbr.2k007.02.001

Wallace, N., Wagstaff, B. J., Deininger, P. L., & Roy-Engel, A. M. (2008). LINE-1 ORF1 protein enhances alu SINE retrotransposition. *Gene, 419*(1), 1-6. doi:10.1016/j.gene.2008.04.007

WANG Xue-feng CHEN Jiao-yue TAN Jin DUAN Suo DENG Xiao-ling CHEN Jian-chi ZHOU Chang-yong. (2015). High genetic variation and recombination events in the vicinity of non-autonomous transposable elements. *农业科学学报：英文版, 14*(10), 2002-2010. doi:10.1016/S2095-3119(14)60979-5

Watts, P. C., Lundholm, N., Ribeiro, S., & Ellegaard, M. (2013). A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species. *Biology Letters, 9*(6), 20130849-20130849. doi:10.1098/rsbl.2013.0849

Wicker, T. (2004). The repetitive landscape of the chicken genome. *Genome Research, 15*(1), 126-136. doi:10.1101/gr.2438004

Wilson, D. P. (1946). The triradiate and other forms of nitzschia closterium (Ehrenberg) WM.Smith,forma minutissima of Allen and Nelson. *Journal of the Marine Biological Association of the United Kingdom, 26*(3), 235–270. doi:10.1017/S002531540001211X

Woodard, L. E., Cheng, J., Welch, R. C., Williams, F. M., Luo, W., Gewin, L. S., & Wilson, M. H. (2017). Kidney-specific transposon-mediated gene transfer in vivo. *Scientific Reports, 7*(1), 44904. doi:10.1038/srep44904

Wyrwicz, L. S., Mangogna, M., Rayko, E., Green, B. R., Lommer, M., Van de Peer, Y., . . . Schmutz, J. (2008). The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature, 456*(7219), 239-244. doi:10.1038/nature07410

Yamashita, H., & Tahara, M. (2006). A LINE-type retrotransposon active in meristem stem cells causes heritable transpositions in the sweet potato genome. *Plant Molecular Biology, 61*(1), 79-84. doi:10.1007/s11103-005-6002-9

Ye, Q., Tong, J., Xiao, S., Zhu, S., An, Z., Tian, L., & Hu, J. (2015). The survival of benthic macroscopic phototrophs on a neoproterozoic snowball earth. *Geology, 43*(6), 507-510. doi:10.1130/G36640.1

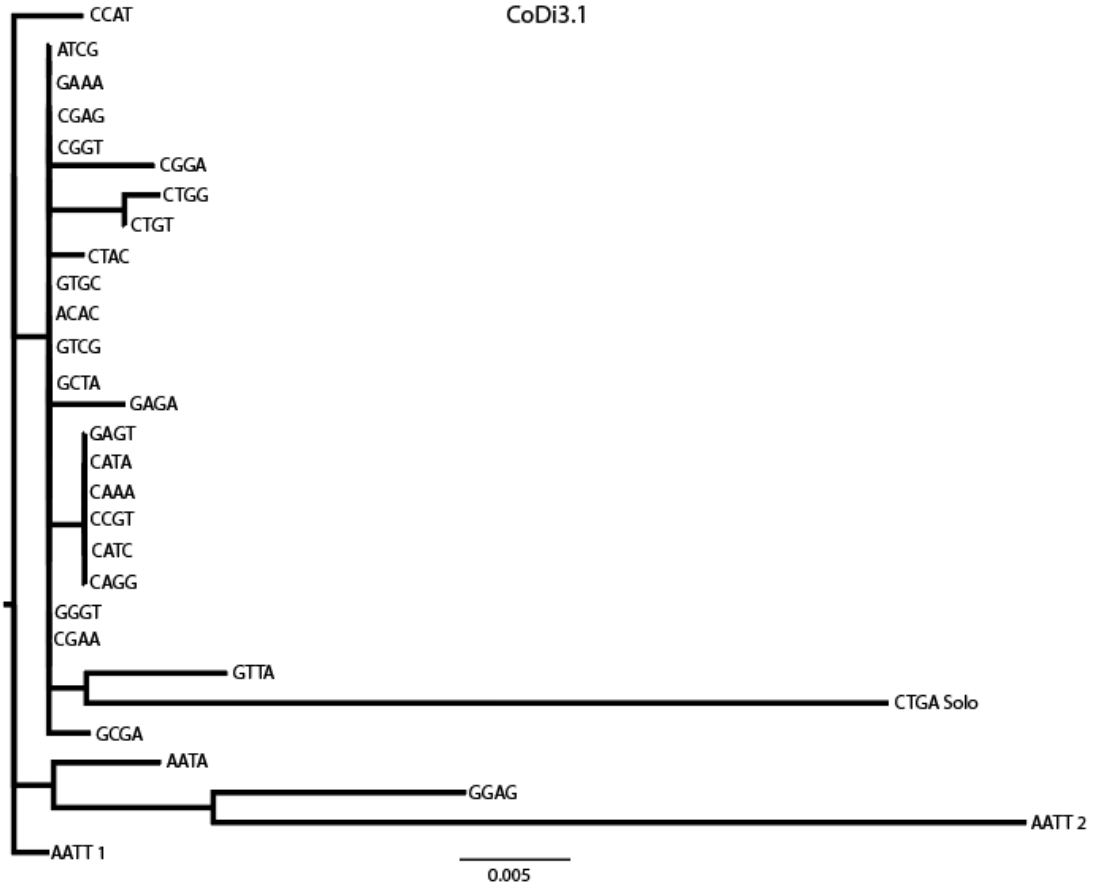
Yuan, P., Yuan, A., Liu, D., Liu, K., Tan, D., Yu, H., . . . He, H. (2013). Surface silylation of mesoporous/macroporous diatomite (diatomaceous earth) and its function in cu(II) adsorption: The effects of heating pretreatment. *Microporous and Mesoporous Materials, 170*, 9-19. doi:10.1016/j.micromeso.2012.11.030

Yuen, J. E., Andersson, B. (2013). What is the evidence for sexual reproduction of phytophthora infestans in europe? *Plant Pathology, 62*(3), 485-491. doi:10.1111/j.1365-3059.2012.02685.x

Zhang, J., Zhang, F., & Peterson, T. (2006). Transposition of reversed ac element ends generates novel chimeric genes in maize. *PLoS Genetics, 2*(10), e164-e164. doi:10.1371/journal.pgen.0020164

Zhang, L., Mao, D., Xing, Y., Xing, F., Bai, X., Zhao, H., . . . Xie, W. (2015). Loss of function of OsMADS3 via the insertion of a novel retrotransposon leads to recessive male sterility in rice (oryza sativa). *Plant Science, 238*, 188-197. doi:10.1016/j.plantsci.2015.06.007

Appendix 1

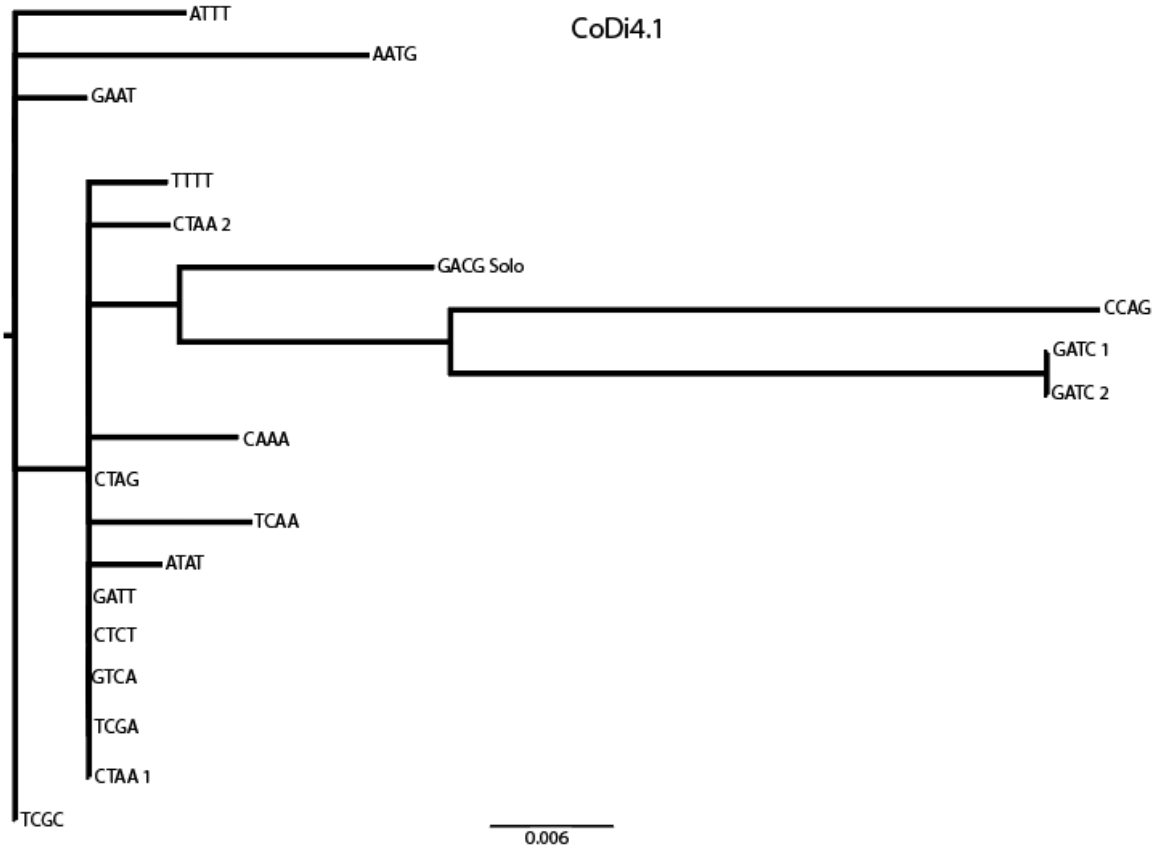


Phylogenetic Tree of *P. tricornutum* TE CoDi3.1

Many of the copies have identical sequences, meaning that this is an active element. There are two sets of 4 copies that are identical, a set of 6 copies that are identical and two more copies that have identical sequences. This family has a TSD of 4 nucleotides. There are two copies on long branch lengths compared to the other copies, but the lengths are only 0.0395 for CTGA Solo and 0.0548 for AATT 2. There is only one solo element. There are 29 copies in this family.

\

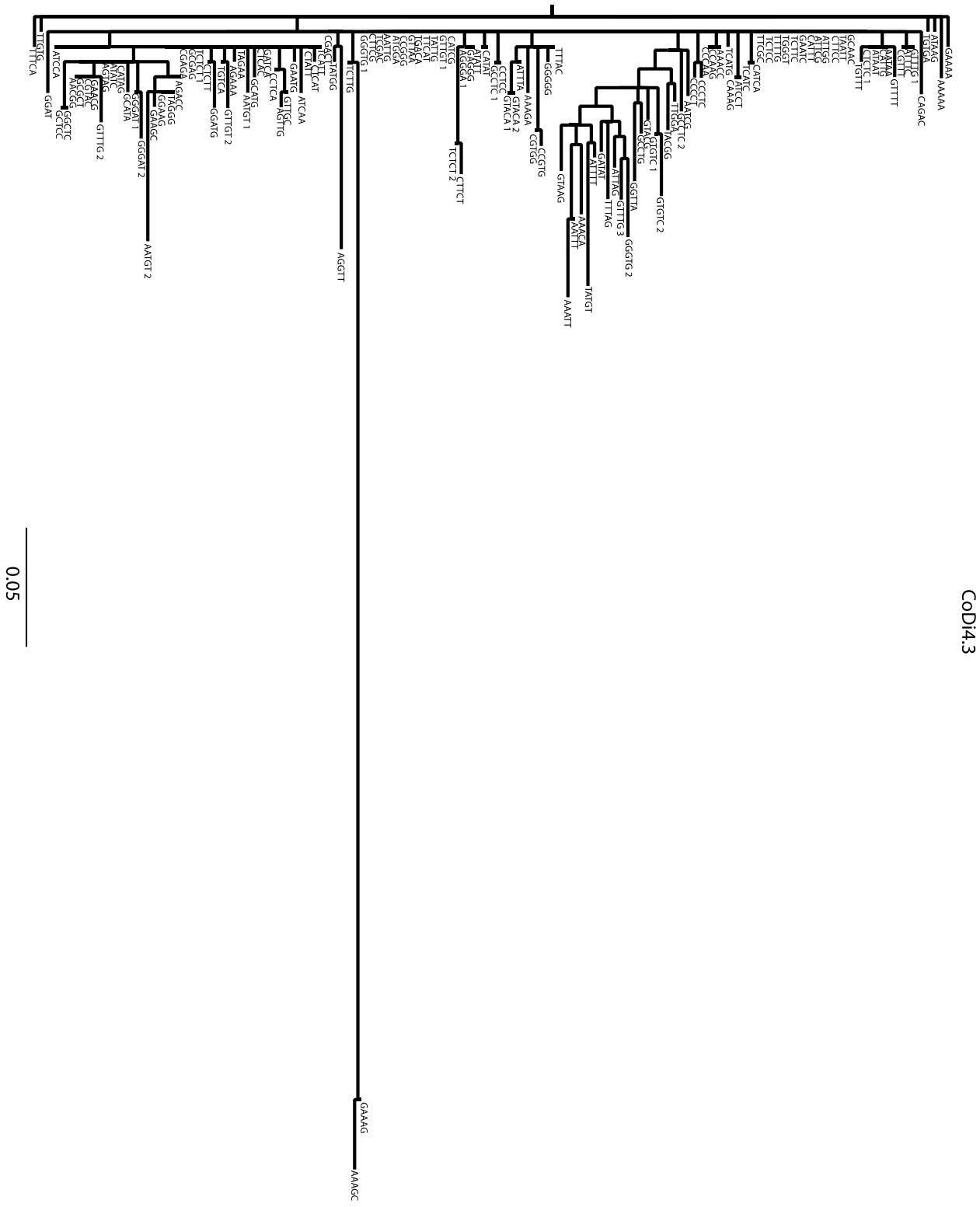
Appendix 2



Phylogenetic Tree of *P. tricornutum* TE CoDi4.1

Many of the copies have identical sequences, meaning that this is an active element. There is a set of 5 copies that have identical sequences and two further copies which also have identical copies. This family has a TSD of 4 nucleotides. There is only one solo copy and it is not the oldest copy in the element. The copy that appears the oldest has a branch length of 0.0531. There are 20 copies in this family.

Appendix 3

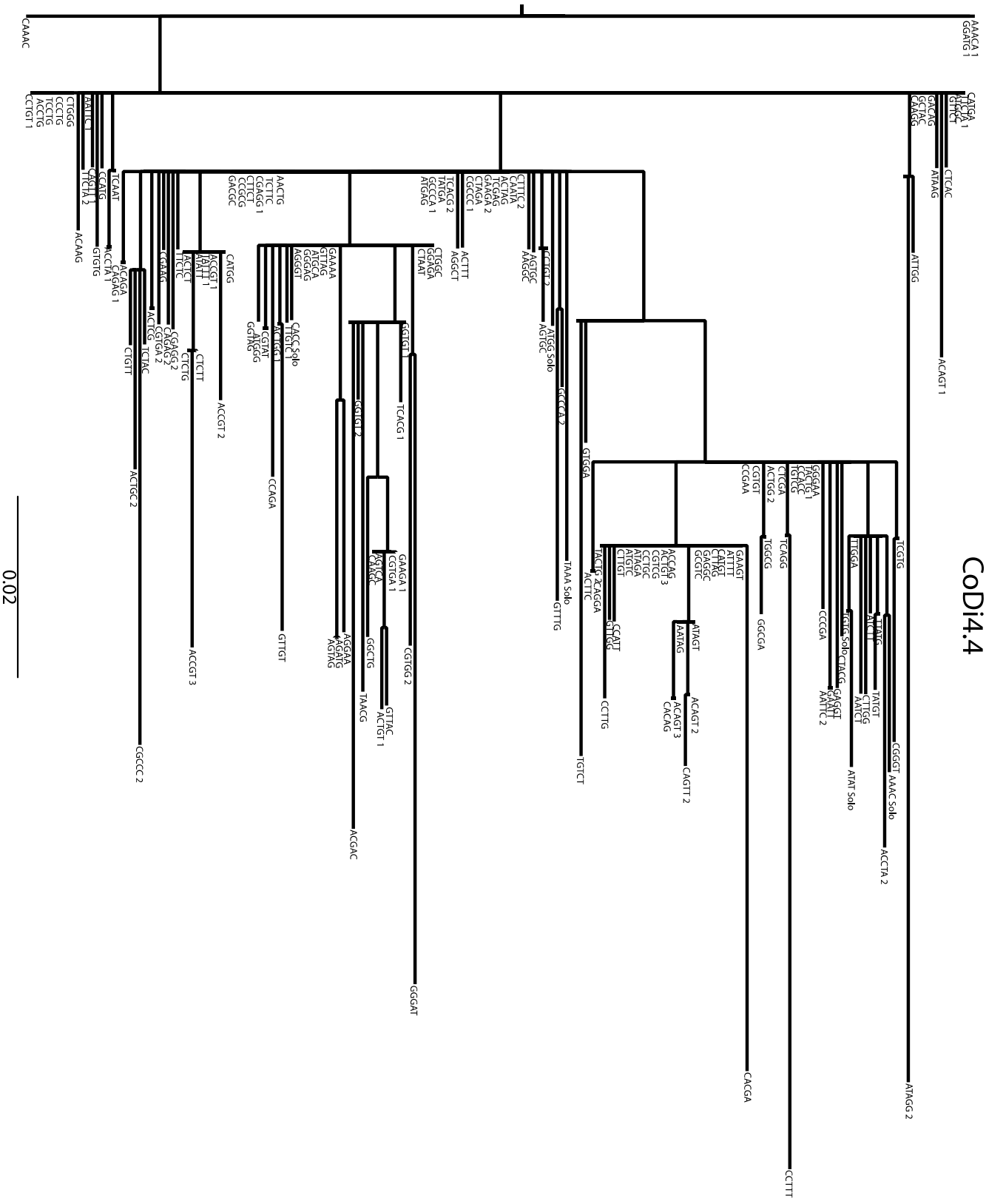


Phylogenetic Tree of *P. tricornutum* TE CoDi4.3

Many of the copies have identical sequences, meaning that this is an active element. There are three sets of 3 copies, two sets of 4 copies, a group of 15 copies, eight sets of 2 copies and

a set of 13 copies with identical sequences. There are no solo copies. There are two copies on branch lengths much longer than the other copies, with lengths of 0.44 for GAAAG and 0.4696 for AAAGC. There are 138 copies in this family.

Appendix 4

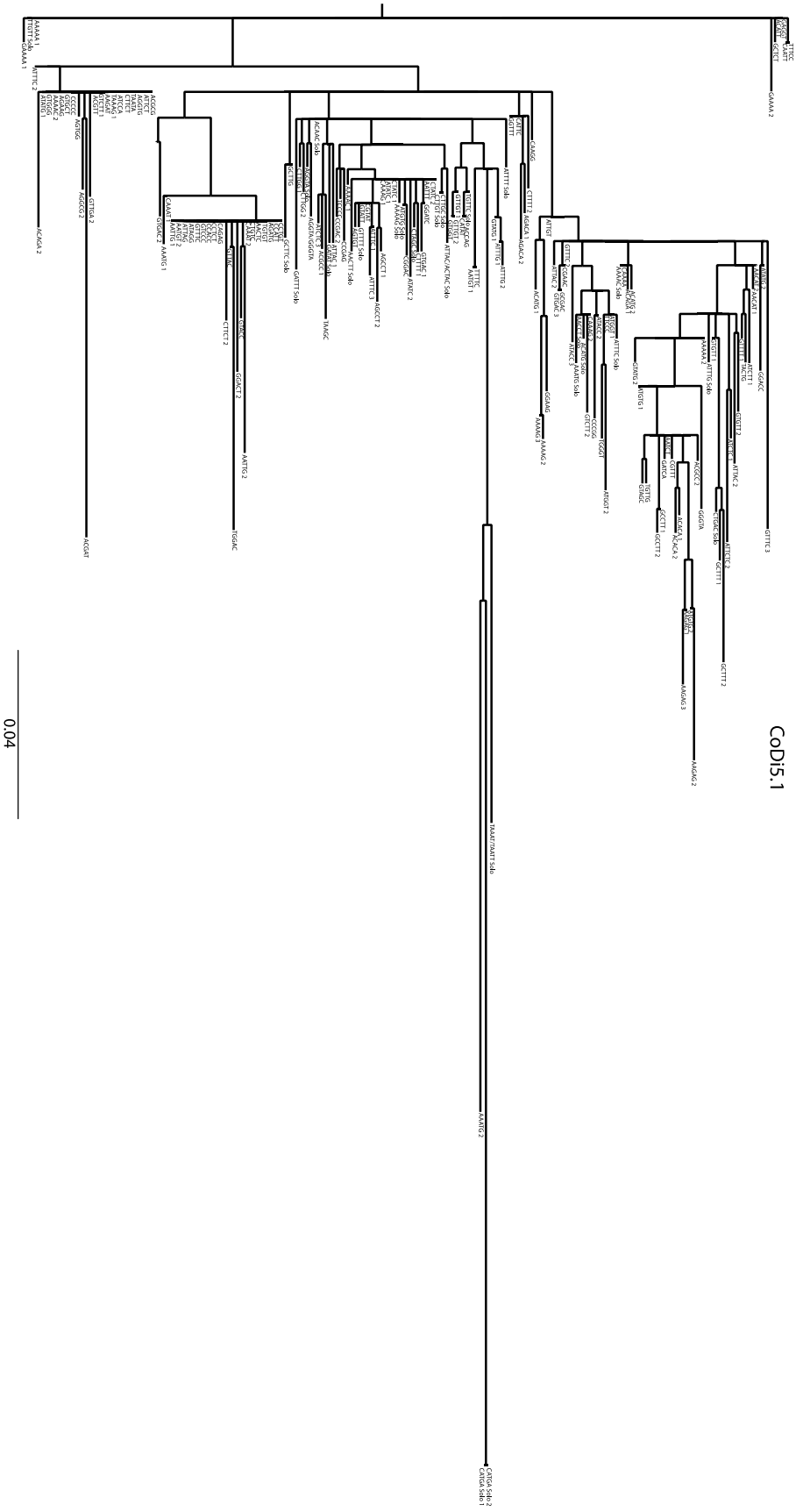


Phylogenetic Tree of *P. tricornutum* TE CoDi4.4

Many of the copies have identical sequences, meaning that this is an active element. There are eleven sets of 2 copies, four sets of 4 copies, two sets of 6 copies, two sets of 3 copies,

two sets of 10 copies, a set of 11 copies, a set of 13 copies and a set of 7 copies with identical sequences. There are 8 solo copies. Many of the copies are on long branch lengths so this may be an old element. The longest branch length is 0.1725. There are 202 copies in this family.

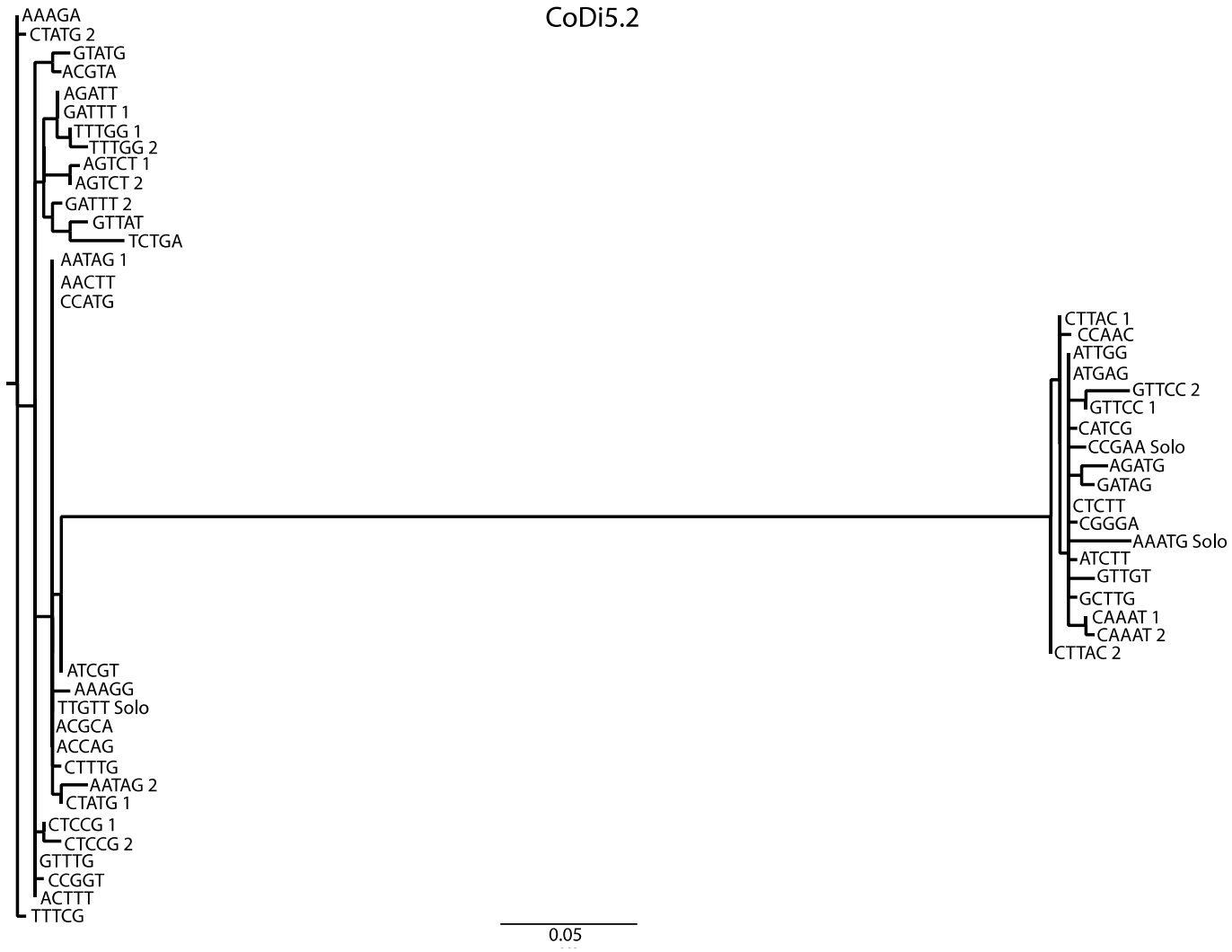
Appendix 5



Phylogenetic Tree of *P. tricornutum* TE CoDi5.1

Many of the copies have identical sequences, meaning that this is an active element. There are 28 solo copies. There are many copies on both short branch lengths and long branch lengths and the longest branch length is 0.205. There are 234 copies in this family.

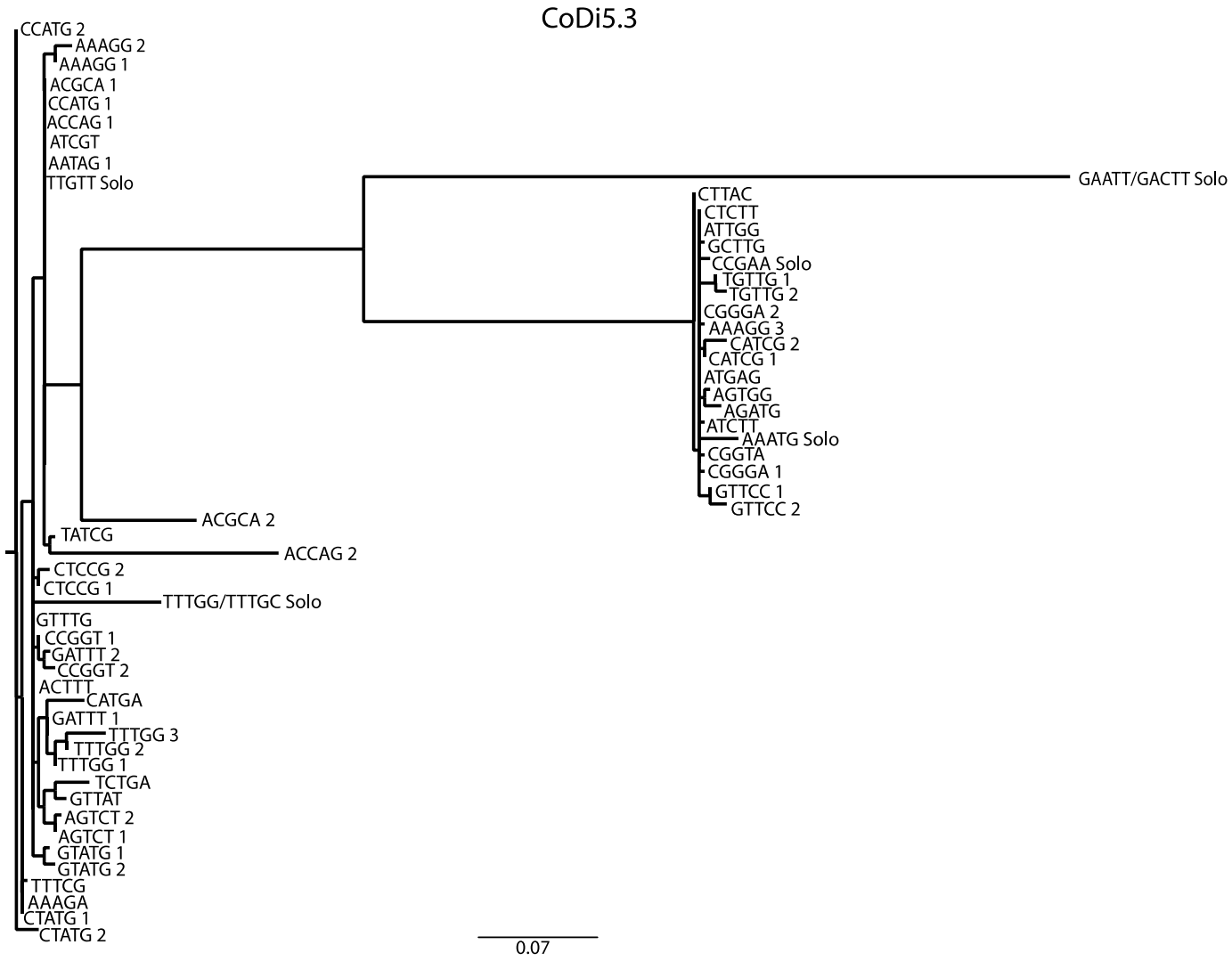
Appendix 6



Phylogenetic Tree of *P. tricornutum* TE CoDi5.2

Many of the copies have identical sequences, meaning that this is an active element. There are two sets of 2 copies with identical sequences, and two groups of 3 copies with identical sequences. There are 3 solo copies. Some of the copies are on short branch lengths and the remaining copies are all on a much longer branch length and are closely related to one another. The longest branch length is 0.4099. There are 49 copies in this family.

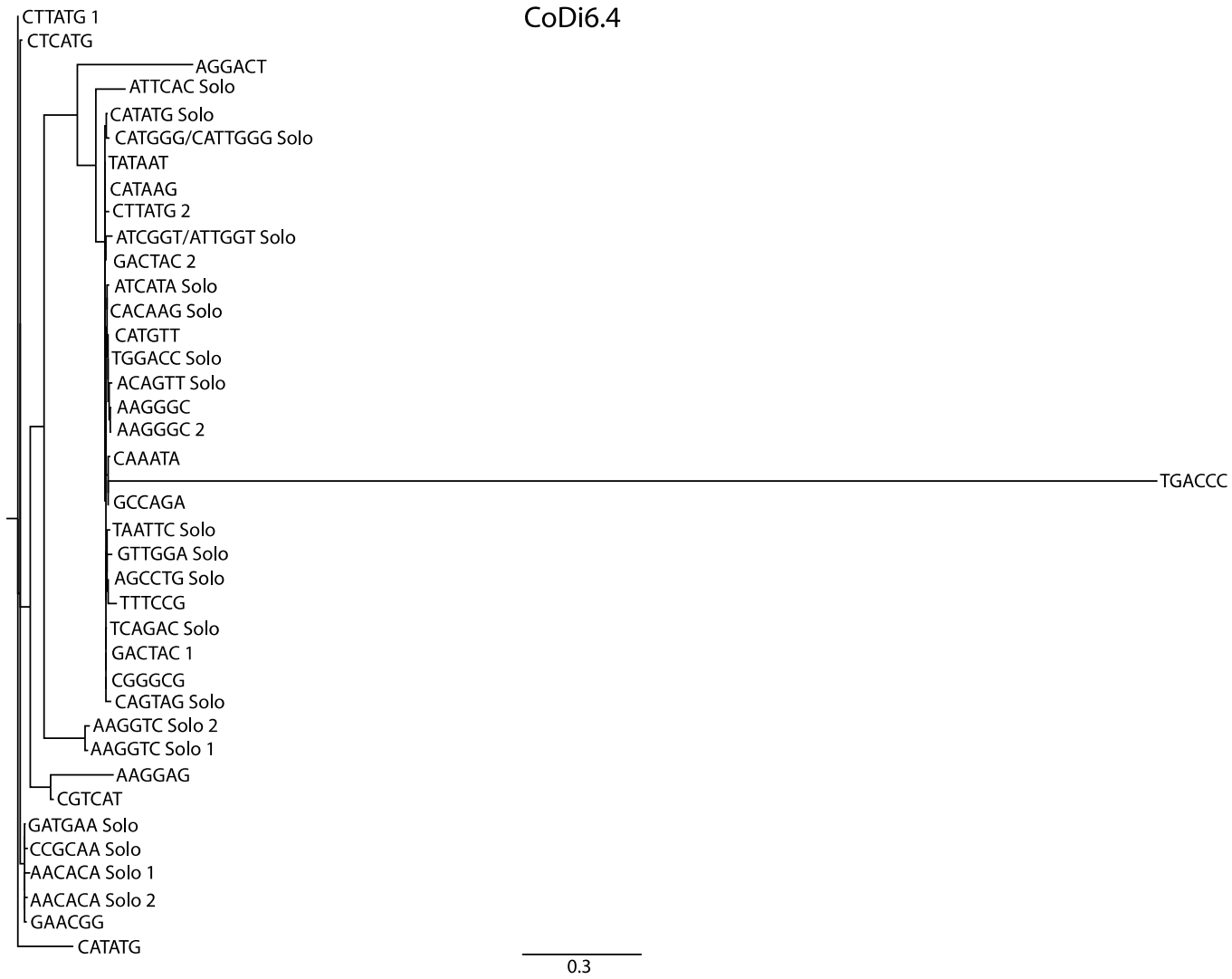
Appendix 7



Phylogenetic Tree of *P. tricornutum* TE CoDi5.3

Many of the copies have identical sequences, meaning that this is an active element. There are a set of 6 copies and a set of 2 copies with identical sequences. There are 5 solo elements. Most of the copies are on short branches and there is a group of copies all on much longer branches. There is also one solo copy on the longest branch, with a branch length of 0.6191. There are 62 copies in this family.

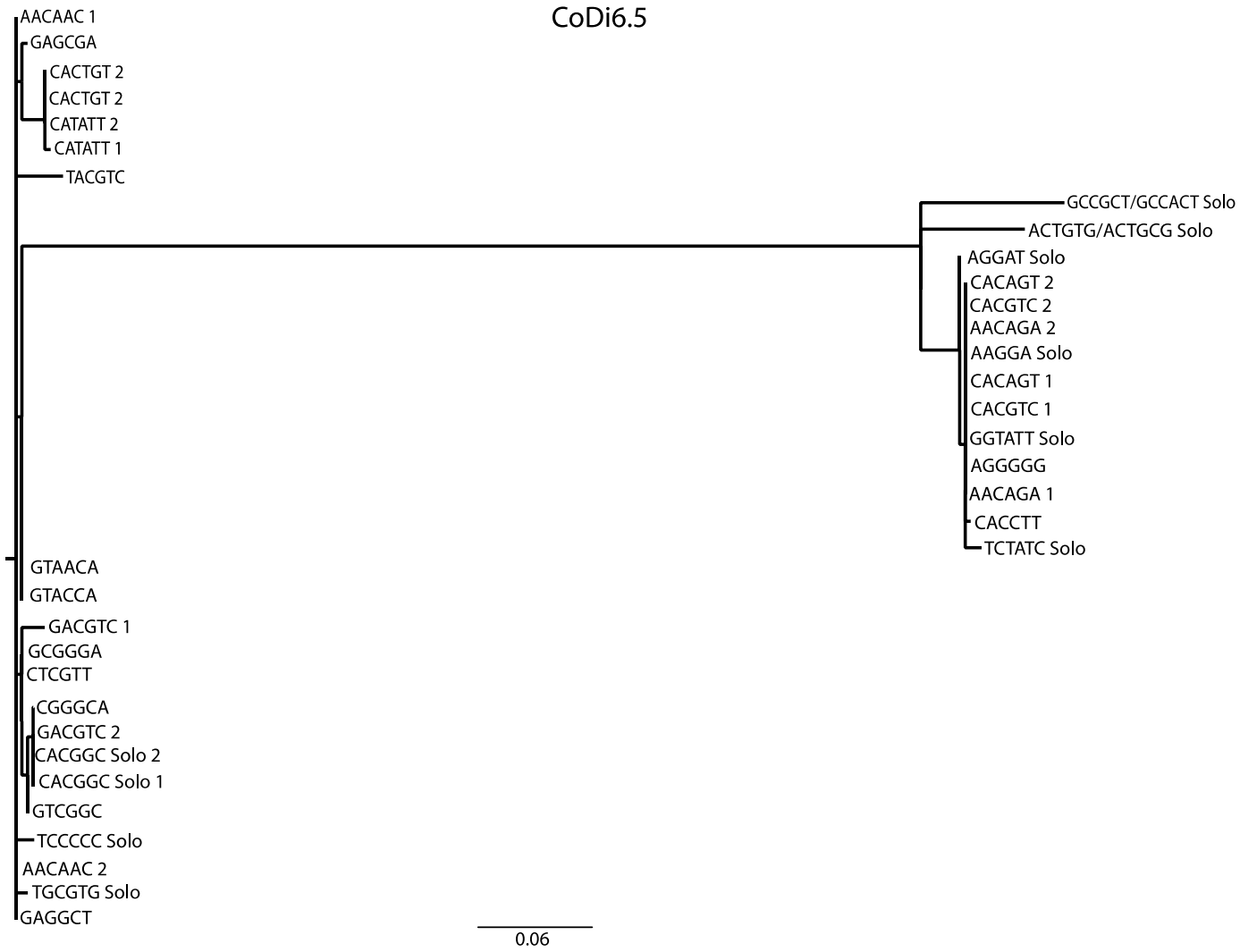
Appendix 8



Phylogenetic Tree of *P. tricornutum* TE CoDi6.4

Many of the copies have identical sequences, meaning that this is an active element. There are three sets of 2 copies with identical sequences. There are 19 solo copies. This family has a TSD of 6 nucleotides and the majority of the copies are closely related. There is one copy on a very long branch length of 2.8692. There are 38 copies within this family.

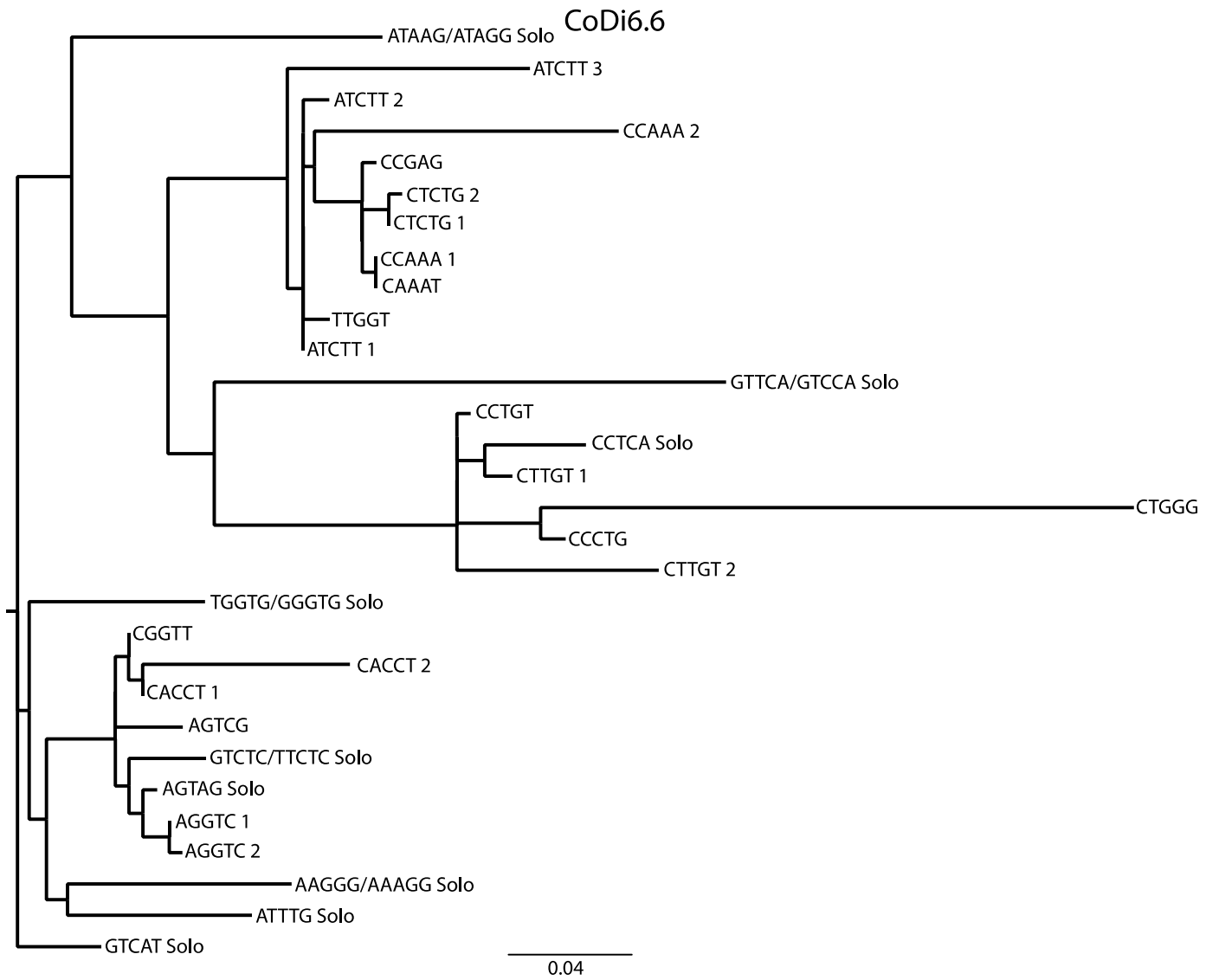
Appendix 9



Phylogenetic Tree of *P. tricornutum* TE CoDi6.5

Many of the copies have identical sequences, meaning that this is an active element. There are a set of 2 copies, a set of 3 copies, a set of 4 copies and a set of 9 copies with identical sequences. There are 10 solo copies and this family has a TSD of 6 nucleotides. Some of the copies are on short branch lengths and there is a clade of copies on long branches. The longest branch length is 0.5506. There are 36 copies within this family.

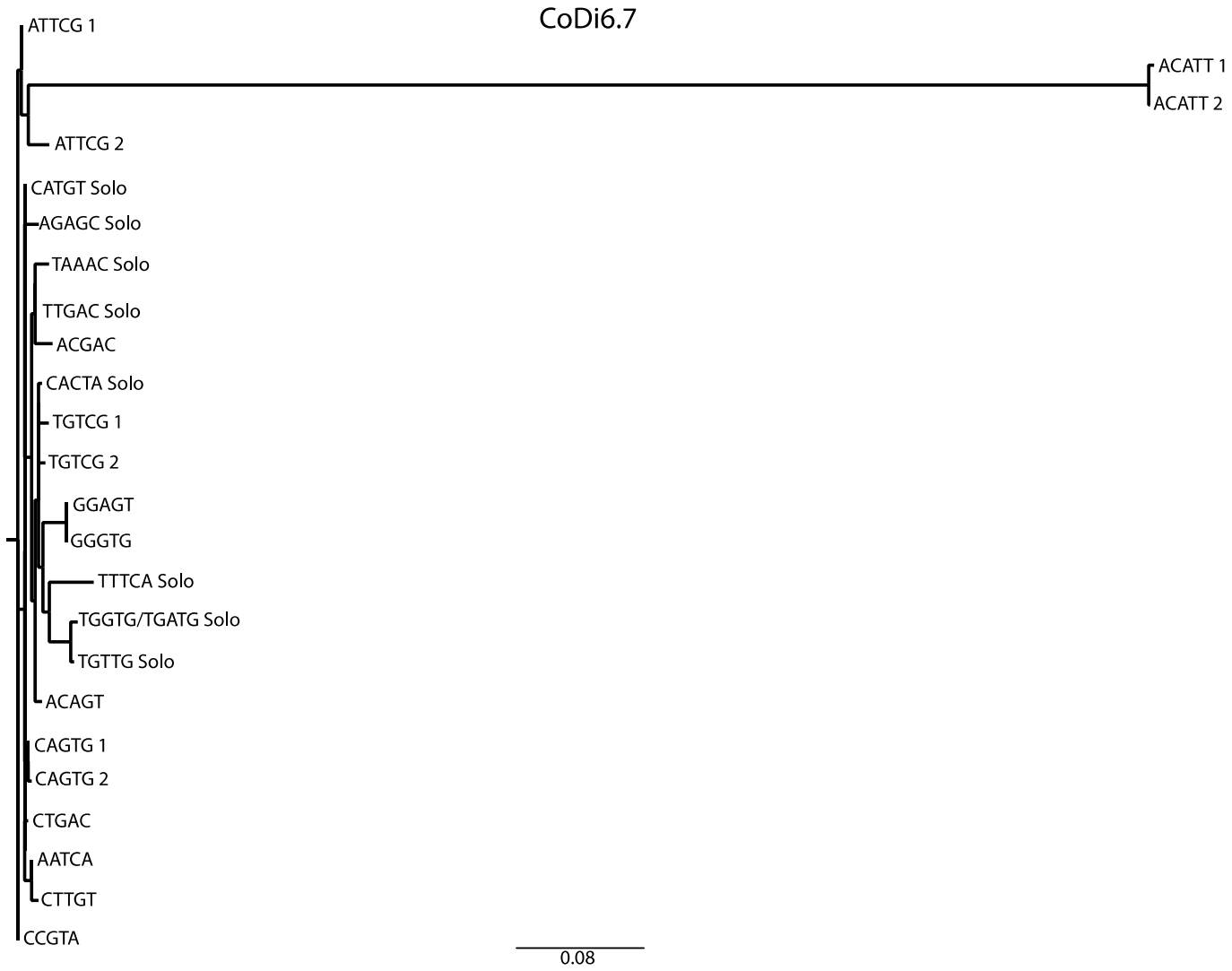
Appendix 10



Phylogenetic Tree of *P. tricornutum* TE CoDi6.6

The copies with TSDs of CCAAA 1 and CAAAT have identical sequences meaning that this is an active copy. There are 9 solo copies. The longest branch length is 0.362. There are 36 copies within this family.

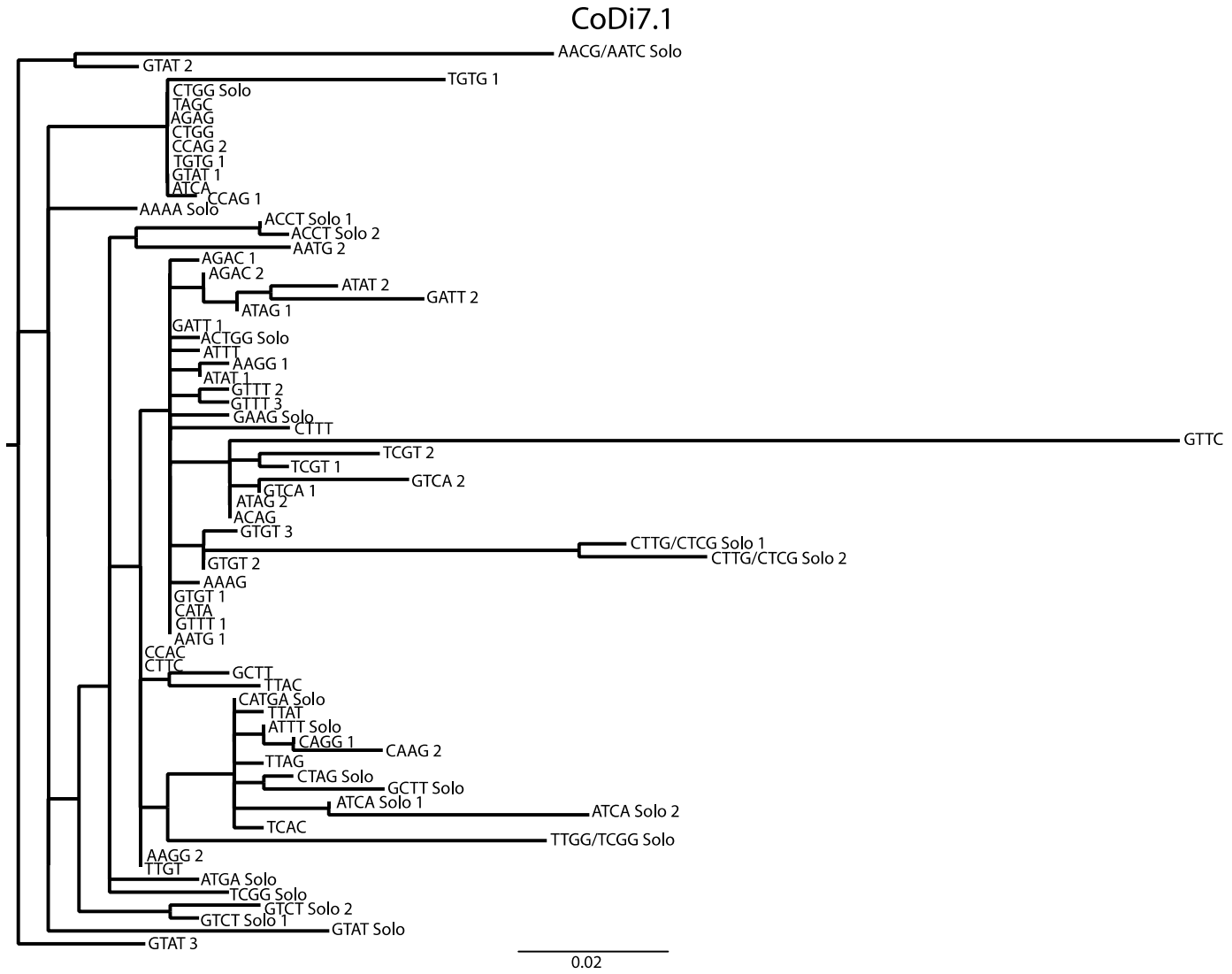
Appendix 11



Phylogenetic Tree of *P. tricornutum* TE CoDi6.7

This family has had recent transposing activity as many of the copies are very similar to one another. There are no sequences that are identical so we cannot be certain that the element is still active. There are 8 solo copies. Most of the copies are on short branch lengths except two closely related copies which are on long branches. The longest branch length is 0.7156. There are 25 copies within this family.

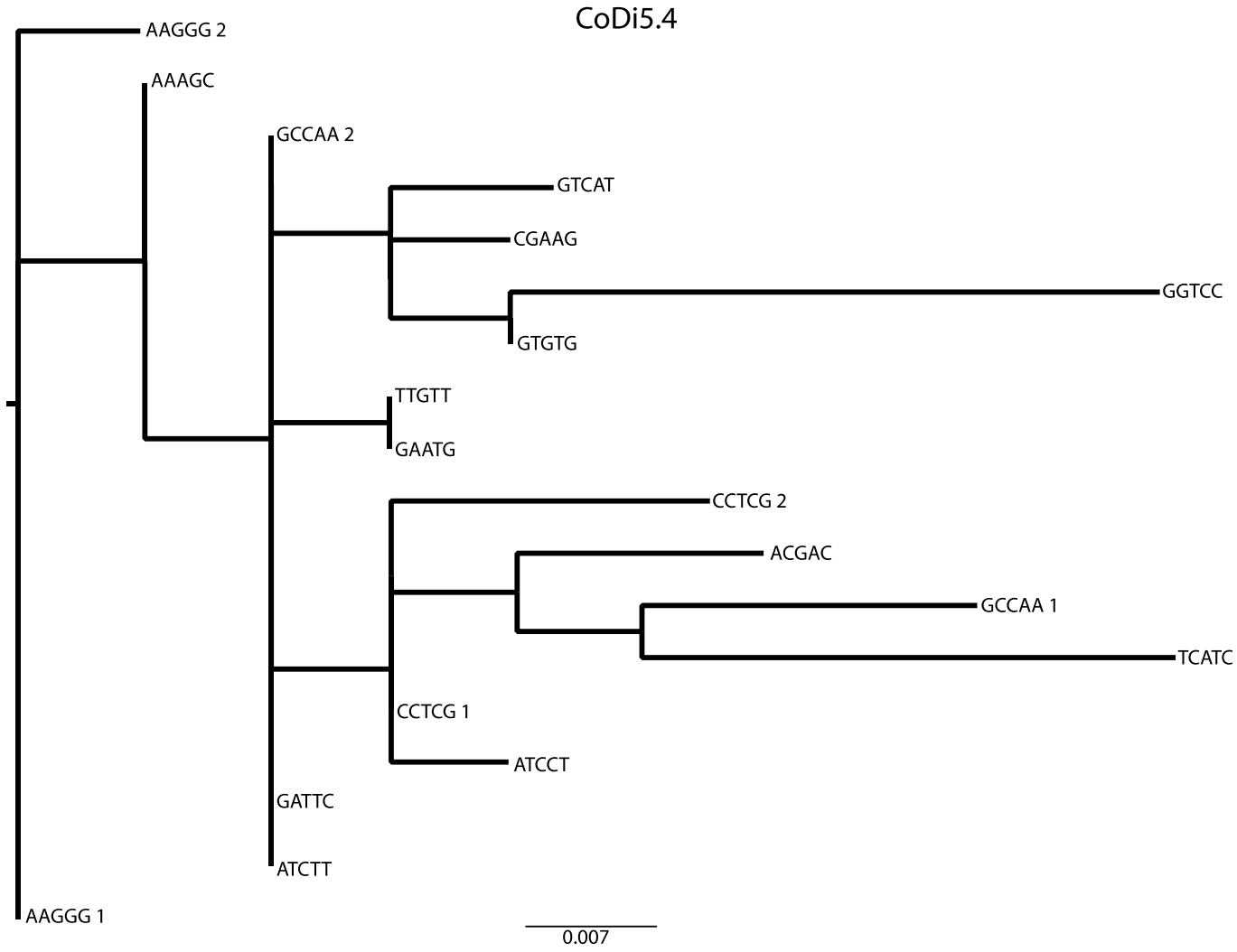
Appendix 12



Phylogenetic Tree of *P. tricornutum* TE CoDi7.1

Many of the copies have identical sequences, meaning that this is an active element. There are three sets of 2 copies, a set of 4 copies and a set of 8 copies with identical sequences. There are 21 solo copies. This family has a TSD of 4 nucleotides long. The copies are on a variety of branch lengths and the longest branch length is 0.1552. There are 71 copies within this family.

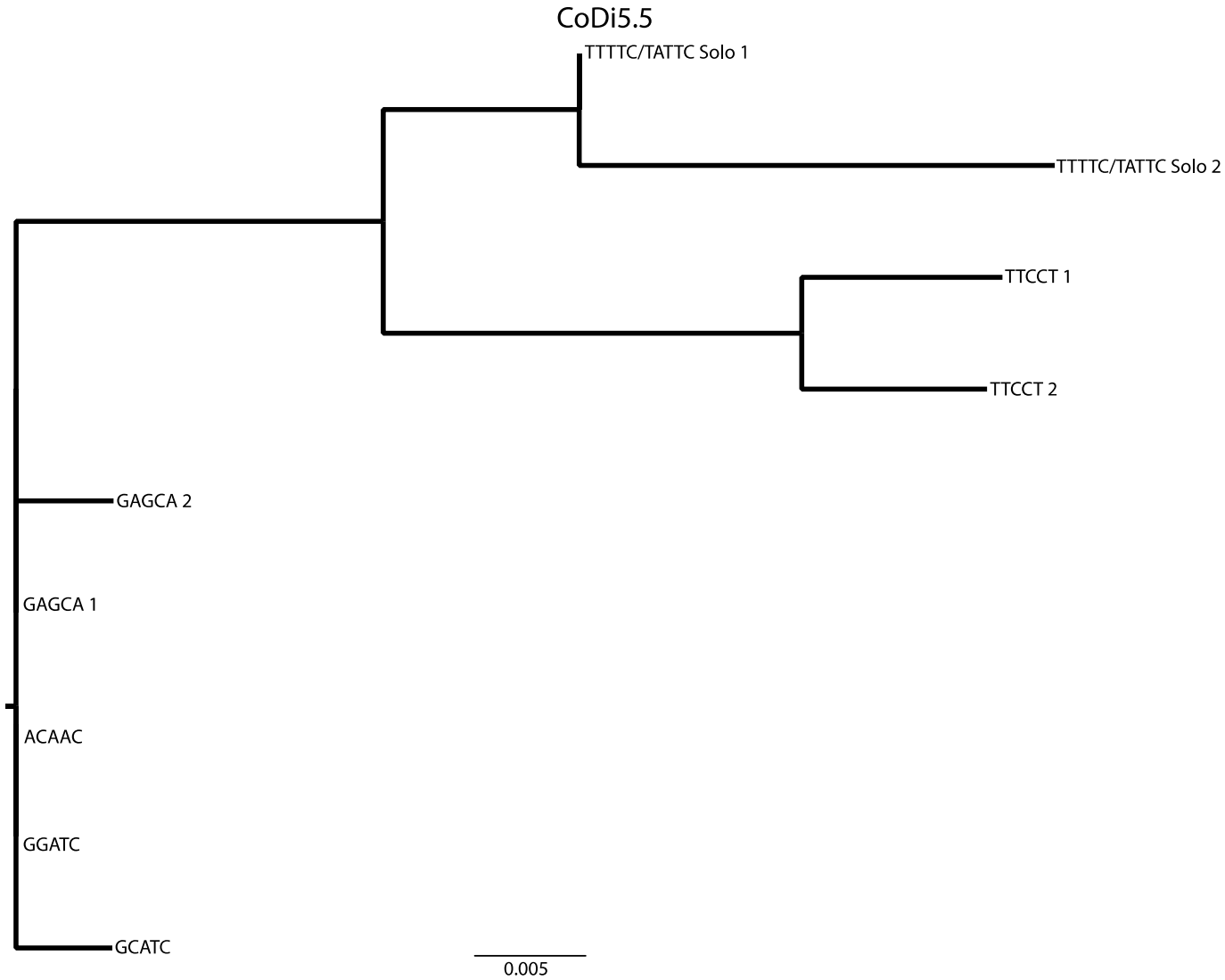
Appendix 13



Phylogenetic tree of *T. pseudonana* TE CoDi5.4

Some of the copies have identical sequences, meaning that this is an active element. There are two sets of 2 copies that have identical sequences. There are no solo copies. This family has a TSD of 4 nucleotides and the longest branch length is 0.0621. There are 18 copies within this family.

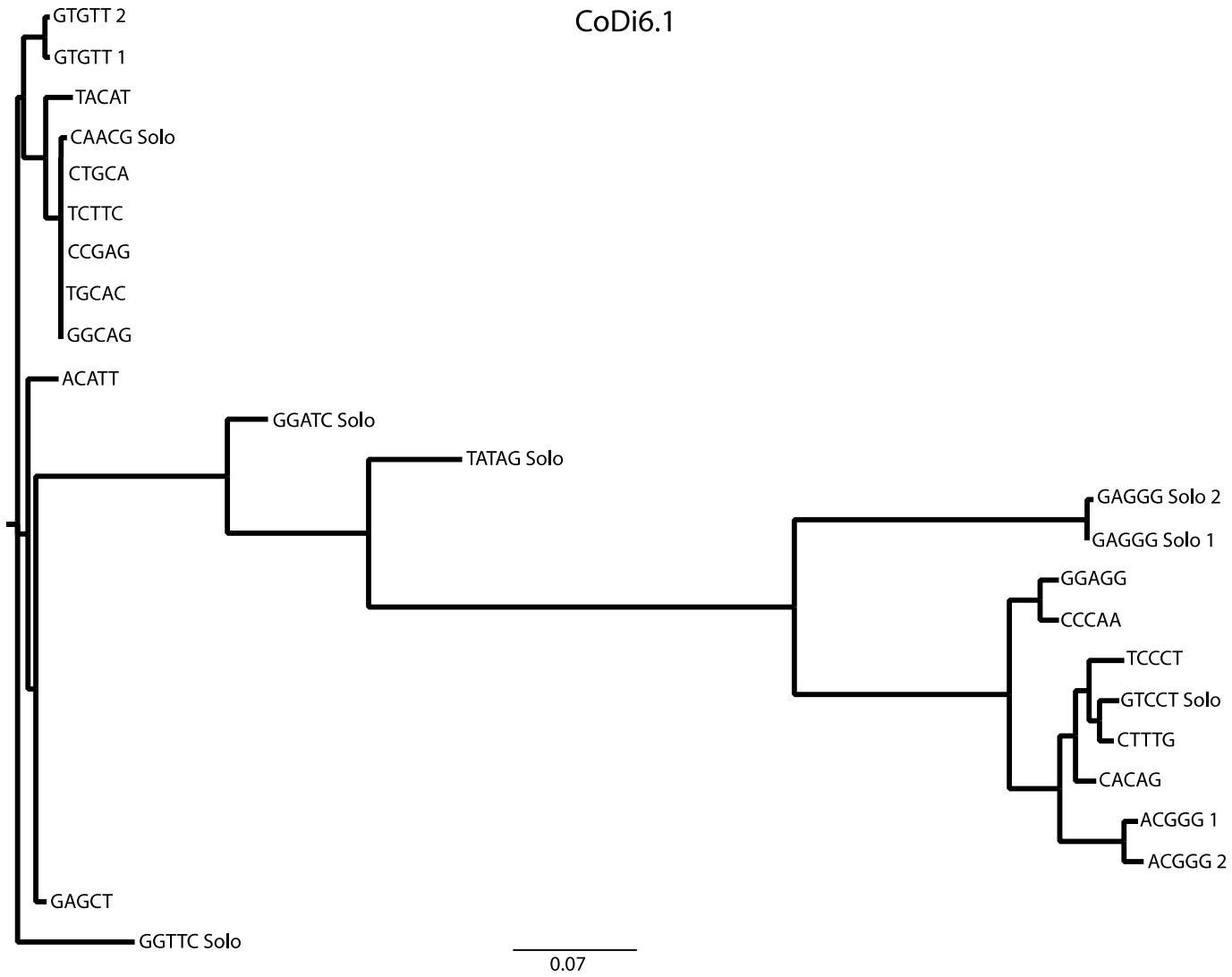
Appendix 14



Phylogenetic tree of *T. pseudonana* TE CoDi5.5

Some of the copies have identical sequences, meaning that this is an active element. The copies with the TSDs of GAGCA 1, ACAAC and GGATC have identical sequences. There are two solo copies within this family and they share the same TSD. This family has a TSD of 5 nucleotides and the longest branch length is 0.0467. There are 9 copies within this family.

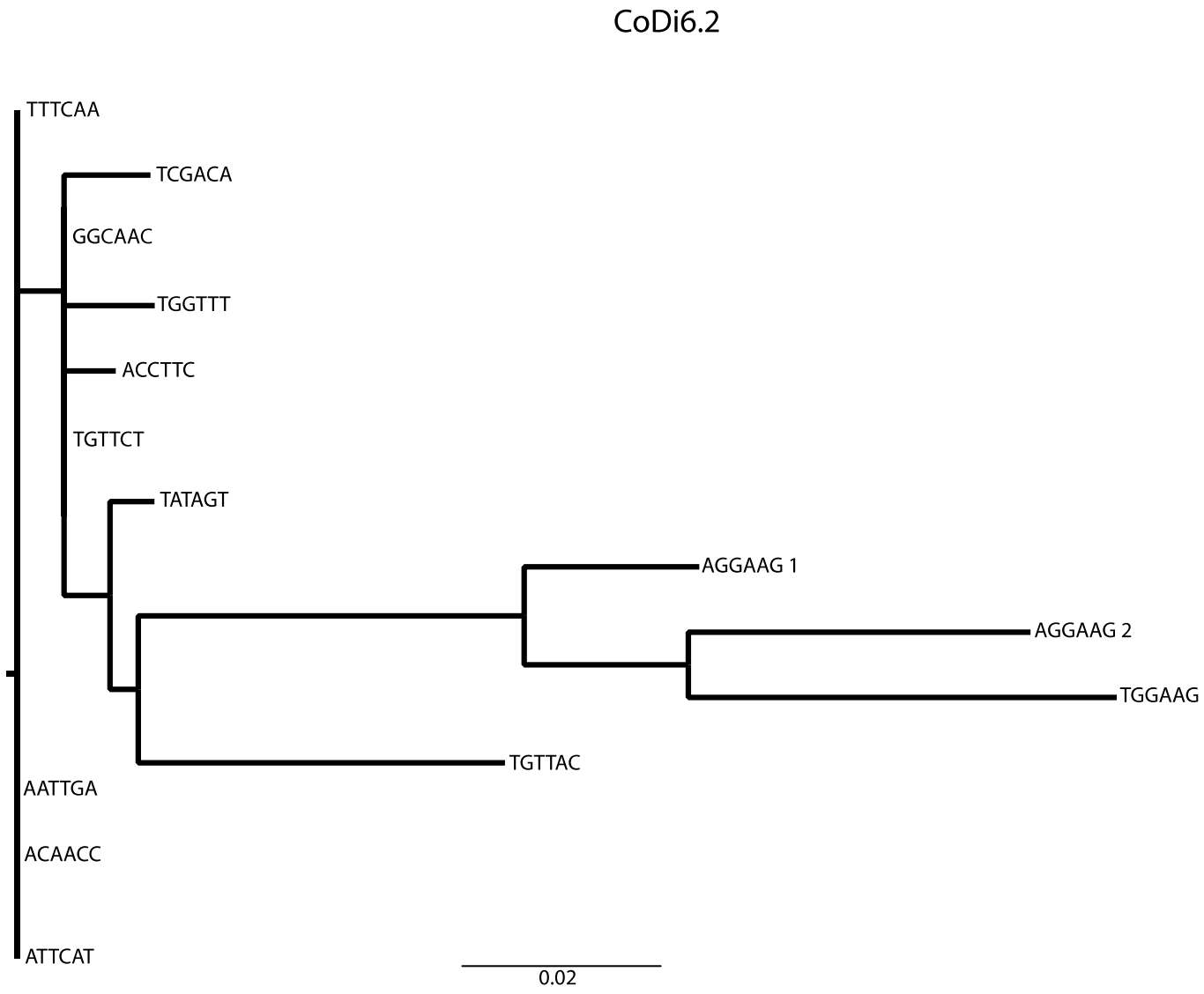
Appendix 15



Phylogenetic tree of *T. pseudonana* TE CoDi6.1

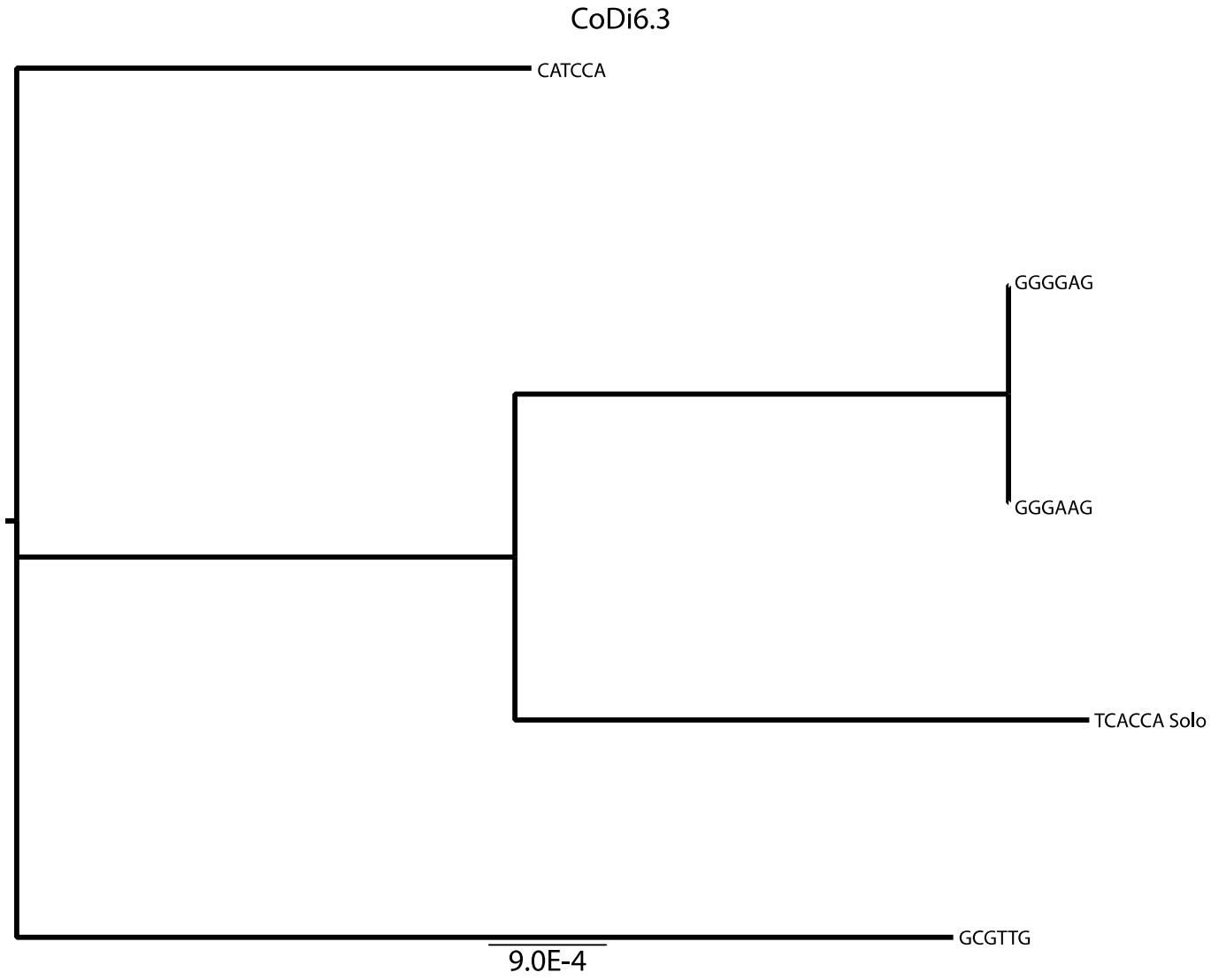
Some of the copies have identical sequences, meaning that this is an active element. There is a set of 5 copies with identical sequences. There are 5 solo copies. This family has a TSD of 5 nucleotides and the longest branch length is 0.637. There are 24 copies within this family.

Appendix 16



Phylogenetic tree of *T. pseudonana* TE CoDi6.2
TSDs of AATTGA, ACAACC and ATTCAT all have identical sequences. There are no solo copies. This family has a TSD of 6 nucleotides and the longest branch length is 0.1106. There are 15 copies within this family.

Appendix 17



Phylogenetic tree of *T. pseudonana* TE CoDi6.3
Some of the copies have identical sequences, meaning that this is an active element. The copies with the TSDs GGGGAG and GGGAAG have identical sequences. There is 1 solo copy. This family has a TSD of 6 nucleotides and the longest branch length is 0.0082 substitutions per site. There are 5 copies in this family.