The early evolution of tRNA modification in Eukaryotes

September 2020

Holly Dawson BSc (Hons)

Supervisors:

Dr Martin Carr and Dr Christopher Cooper

The department of Biological and Geographical Sciences

Thesis submitted to the

University of Huddersfield

In partial fulfilment of the requirements for the degree of Master of Sciences by Research

The work in the following research project is my own unless otherwise indicated and that the work of others (i.e. usually, published sources) has been fully acknowledged.

TABLE OF CONTENTS:

TABLE OF TABLES:

COPYRIGHT STATEMENT:

i) The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright The author of this thesis (including any appendices and/ or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii) Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii) The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

ACKNOWLEDGEMENTS:

I would like to thank my supervisor, Dr Martin Carr, for his limitless support and patience with me during the course of this research project. His expert knowledge and skills have been invaluable to me and I will be forever grateful for all I have gained from him.

I would also like to thank Dr Jade Southworth for her unwavering support and guidance throughout my research project. Her assistance in proof reading and advice on work has been incredibly important in making this work a reality.

ABSTRACT:

AIMS:

Deamination of tRNA molecules has already been shown in holozoan species with the evolution deemed to have occurred within the ancestor to the choanoflagellates, metazoans and filastereans. The aim of this project is to determine an estimate of the origin of evolution for this process, by providing evidence for this outside of the holozoans to the opisthokonts and further still, across the eukaryotic supergroup if possible. The determination of a time frame for this allows theories to be drawn regarding the possible expansion of this.

METHODS:

This research project used bioinformatic techniques modified from Southworth et al., (*2018*) in order to analyse the genomes of species: *Thecamonas trahens* (Apusozoa), *Fonticula alba* (Opisthokonta), *Leishmania tarentolae* (Excavata), *Chondrus crispus* (Archaeplastida), *Chlamydomonas reinhardtii* (Archaeplastida) and *Bigelowiella natans* (Rhizaria). The optimal codons for each species were determined using programs CodonW and SMALT in order to see whether these perfectly matched the most abundant tRNA species, or whether deamination is a possible explanation for why there is a lack of perfect matches. Other techniques such as tRNAscan-SE-2.0 analyses on whole genome sequences for some species under investigation, were utilised in order to see what tRNA genes are present, by analysing transcripts of the species.

Results:

BLAST analysis of modified and unmodified tRNA genes identified using the tRNAscan output of their genome files showed that both copies were present within the species *F. alba* and *T. trahens*. SMALT analyses of these transcriptome files identified multiple copies of both variations within additional eukaryotic species which fall on the opposite side of the eukaryotic root to the original species.

Conclusions:

Initial results of *T. trahens* and *F. alba* provided direct evidence for the evolution of deamination of tRNA being ancestral to the Opisthokonta group, and allowed this research to undertake analysis of multiple additional species outside of the opisthokonts, where direct evidence for deamination of tRNAs was identified within rhizarians (SAR), archaeplastids and excavates.

The aims this research project set out with have been achieved as the dataset generated by this work provides evidence for deamination being an ancestral eukaryotic trait. One explanation could be that this has undergone an expansion; from the deamination of the previously identified arginine, isoleucine, serine and threonine amino acid tRNAs in bacterial species, through their symbiosis with Archaea, forming the eukaryotes.

LIST OF TERMS AND ABBREVIATIONS:

BI – Bayesian Inference

BLAST – Basic Local Alignment Search Tool

EggNOG – A web page based program for functional annotation of sequences

Et al. – and others

Fop – Frequency of optimal codons

GC3s – Guanosine and cytosine composition of the third position of the codon

KOG – EuKaryotic Orthologous Groups

MAFFT – Multiple Alignment using Fast Fourier Transform

ML – Maximum Likelihood

Nc – Number of Effective Codons

PP – Posterior Probability

RAxML – Randomised Axelerated Maximum Likelihood

SRA – Sequence Read Archive

TAPSILVR – Threonine, Alanine, Proline, Serine, Isoleucine, Leucine, Valine & Arginine Amino Acids

tRNA – Transfer Ribonucleic Acid

INTRODUCTION:

THE EUKARYOTIC TREE OF LIFE –

The eukaryotes are a collection of species determined by the presence of a nucleus within their cells, the most common thought-of, of these being; plants, animals and fungi, however the majority of the diversity under this title are single-celled species. As a general rule the species contained within the domain; whether unicellular or multicellular share a large number of features and usually a common ancestry to explain these common features.

To group together those with similar molecular phylogenetic traits and resolve the eukaryotic tree of life somewhat the eukaryotes are broken down into "super-groups", proposed supergroups include; Stramenopiles, Alveolates and Rhizarians (SAR), Opisthokonts, Archaeaplastids, Amoebazoans and Amorphea (Adl et al., 2019).

At the basis of my research is a current representation of the eukaryotic tree of life taken from Derelle et al., (*2015*), (Figure 2) this tree has been utilised as a reference for current understand of the tree of life and hence the relationships of organisms within this (Figure 1). This is in order to direct my research to species which could be instrumental to the evolution of the mechanism. In this representation of the eukaryotic tree the root of the tree is shown to fall between the evolutionary groups Opimoda and Diphoda. The Opimoda contains the Opisthokonta and Amoebazoa groups, alongside a few enigmatic groups such as the placement of *Thecamonas trahens*, an Apusozoan, on its own long branch. The Opisthokonta group is an evolutionary level up from the Holozoa, the holozoans are an evolutionary group inclusive of the metazoans and also their closest single-celled relatives, including the choanoflagellates, however not inclusive of the Fungi group. The choanoflagellates are a lineage of unicellular eukaryotes within the Opisthonkonta; recovered as the closest living relatives of the evolutionary group, Metazoa. Choanoflagellate discovery came about due to their morphological similarity to some poriferan cells and the relationship as sister group to the Metazoan group was later confirmed (Carr et al., 2008). The Holozoa containing sister groups Metazoa and Choanoflagellatea means that any information gleaned in this area is vital to the early evolution of Metazoa; as this hints at some ancestral traits which may not be present in extant Metazoan species. Also included under the Opisthokonta is the sister group to Holozoa, the Holomycota: which contains Fungi, nucleariids and the species *Fonticula alba*. On the Diphoda side of the root there are groups such as the SAR major group (stramenopiles (Heterokonta), Alveolata and Rhizaria) and archaeplastids. Rhizarians are often characterised by their feeding groove, this is a split in the membrane which is used to ingest small particles that the organism's flagella beat into it, they usually have highly derived mitochondria which can be mistaken for a lack of mitochondria through their lack of classical shape. The species *Fonticula alba,* which has been utilised in this research is not included in the Derelle et al., (*2015*) tree however it's positioning inside the Opisthokonta group was established by Brown et al., (*2009*). Early branching species of the eukaryotic supergroup, such as *T. trahens,* were utilised. However initial results of this research project showed a more diverse range of species from both sides of the Opimoda/Diphoda root split were required, in order to further establish the evolution of this theory of tRNA deamination evolution. The excavate group was determined as a significantly important group for the purpose of this study. This was due to their positioning being established by multiple papers as on the opposite side of the root of the tree to the opisthokont group. The Excavata contains most importantly some infective human eukaryotic parasites such as *Giardia* & *Leishmania;* causing a diarrhoeal disease and Leishmaniasis respectively, their reclassification from Prokaryote to Eukaryote came due to their flagellar structures (Clark, 1868). This was shown in He et al., (*2014*), where the excavate group is placed on one side of the root and the remaining eukaryotes placed opposite on the other side. This placement was then reviewed by Derelle et al., (*2015*), where the placement of excavates was deemed incorrect and produced was an alternative phylogenetic tree. The tree displayed a root with the opisthokont group on one side with the Amoebozoa and other minor groupings; and on the other side the excavates plus the SAR major group. This placing of SAR with excavates, Archaeplastida and Rhizaria has been recovered will full support on more than one occasion. Despite the differing placements of the excavate group the fact that both groups concur that regardless of their positioning in relation to the root, that the excavates always fall on the opposite side to the opisthokonts. This means that if the process of deamination is present in the excavates, it is likely to have been present in the last common ancestor of both groups. Southworth et al., (*2018*) showed the process of deamination of tRNAs is present within the holozoan group; also showing evidence for the process in the filasterean species *Capsaspora owczarzaki,* a filasterean species of its monotypic genus, and choanoflagellate species *Monosiga brevicolis* and *Salpingoeca rosetta*. This supports that the process was present in the last common ancestor of these three protistan species.

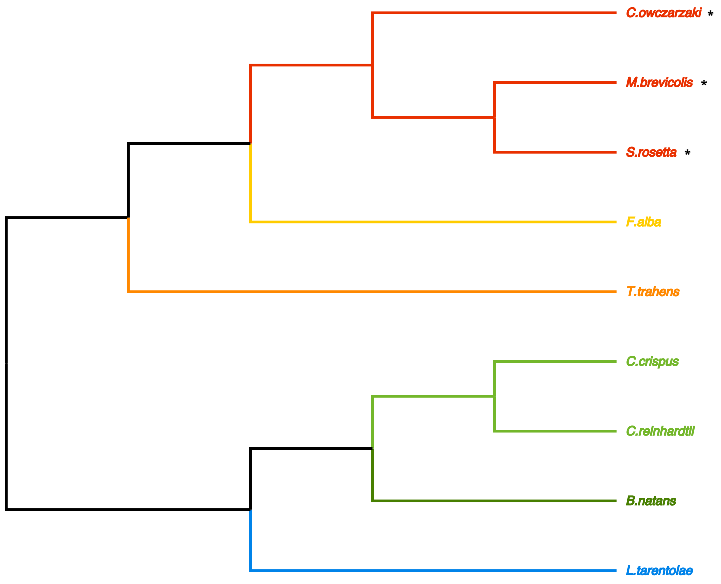


Figure 1: Representative cladogram displaying the relationships between the species under investigation, and species previously investigated within Southworth et al., *(2018)*, based upon phylogenies taken from; Derelle et al., *(2015)*, Brown et al., *(2018)*, Keeling & Burki, *(2019)*. Different supergroups are represented with differing colours: holozoans – red, opisthokont – yellow, apusozoan – orange, archaeplastids – light green, rhizarian – dark green, excavates – blue. Species where deamination of tRNAs has previously been identified are denoted by a black asterisk.

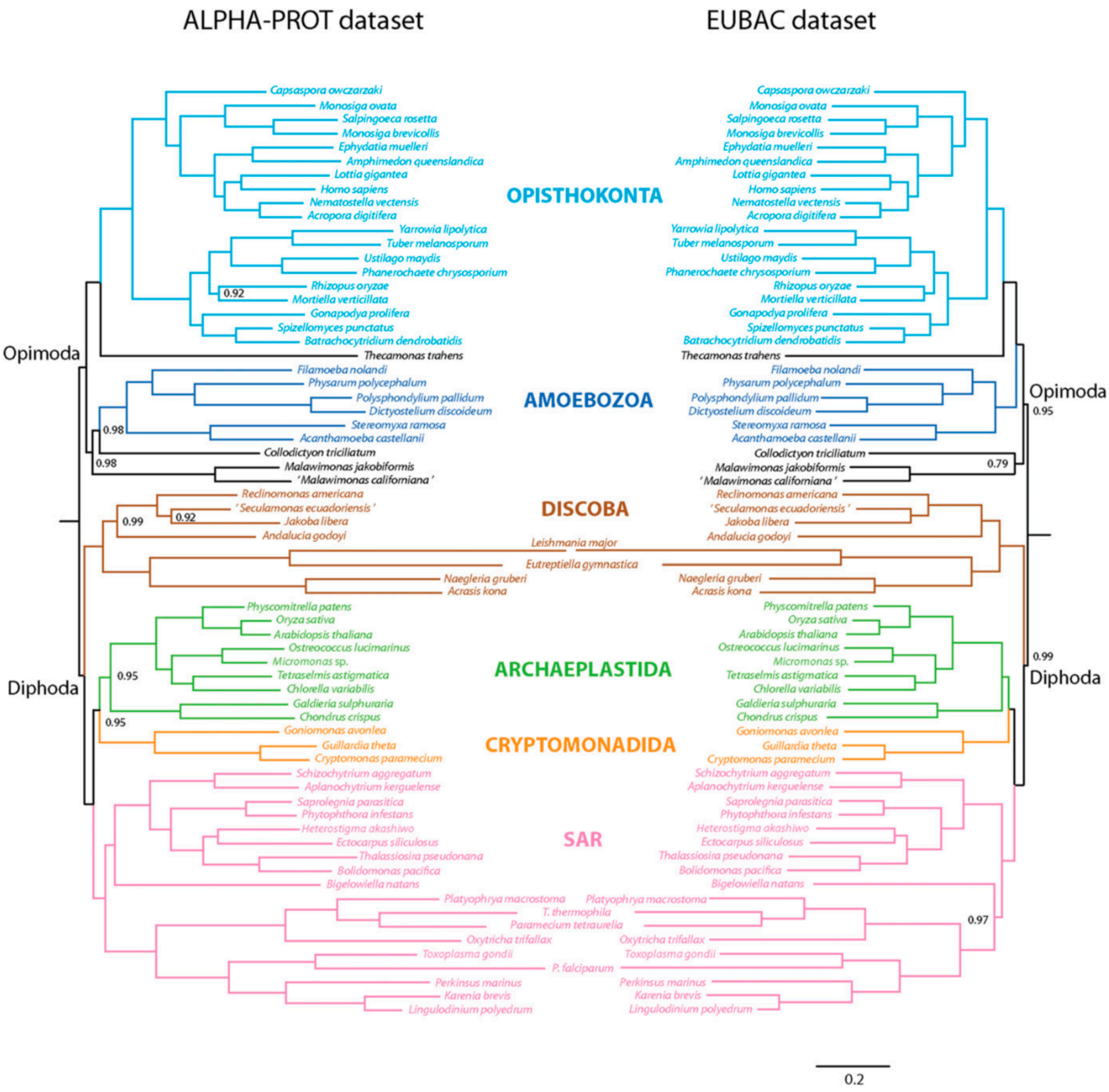


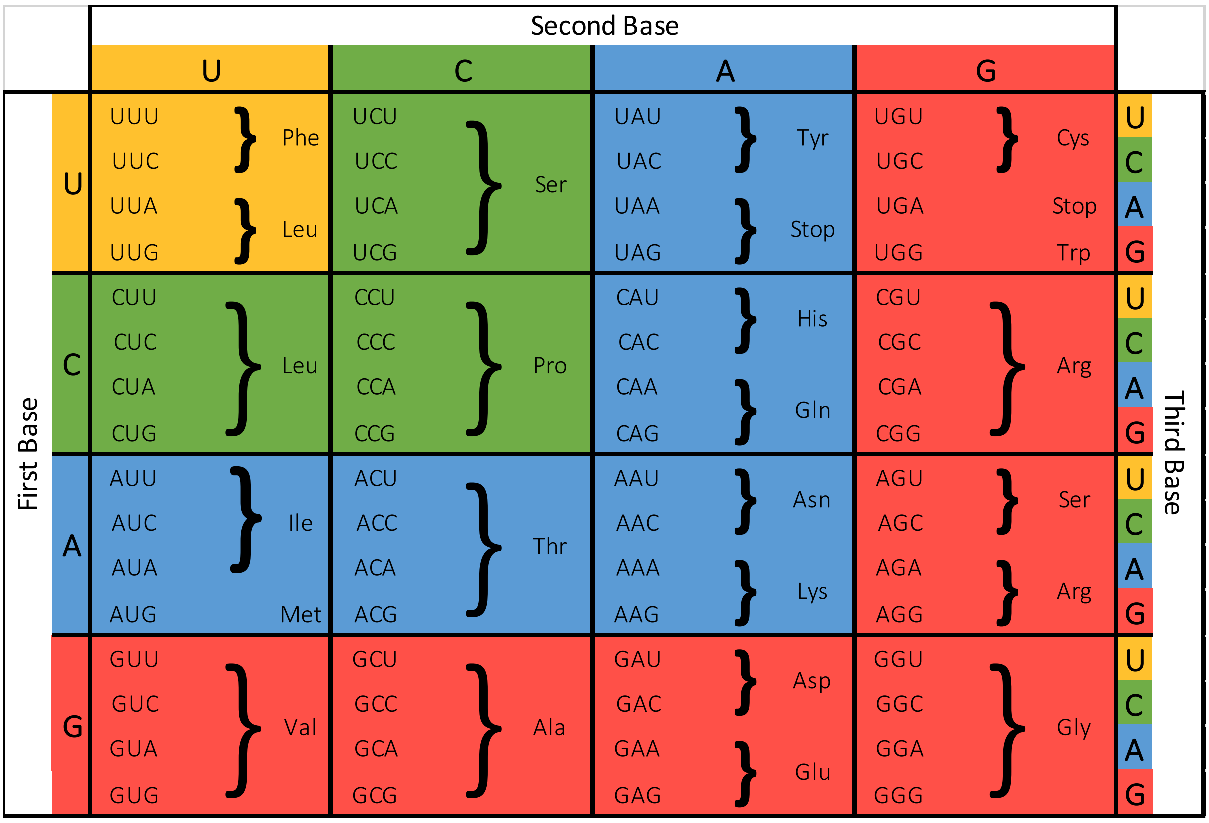
Figure 2: Diagram taken from Derelle et al., (*2015*) displaying Bayesian consensus trees created from both their ALPHA-PROT dataset and EUBAC dataset, left and right respectively, in order to display the Opimoda/Diphoda split at the base of the eukaryotic root. This phylogenetic tree was utilised as a reference during this research project in order to determine where species for further analysis should be picked from to ensure the results of this research show deamination of tRNA evolved within the eukaryotic supergroup.

Reconstructing Ancestral Traits –

Once a tree is constructed, it can be used for further analysis such as reconstructing ancestral traits. The common traits and conservation of functionally important genes seen within organisms of the same groups; such as those which encode the proteins which make up ribosomes (small sub-unit ribosomal proteins SSU) are hence vital to the survival of an organism. Investigating these allows strong conclusions to be drawn about phylogenetic placements of the species under investigation. This also allows for the ancestral traits to be investigated, by looking at conservation within members of these supergroups, when compared with other supergroups. One such trait which can be reconstructed is codon usage, codon usage is the phenomenon of how some codons are utilised more frequently than others that encode for the same amino acid. This can occur due to the degeneracy of the genetic code and the fact that most amino acids can be encoded by more than one amino acid due to the large number of codon combinations and the existence of only 20 amino acids. This degeneracy can be two- up to six-fold, the former occurring when only two of the four nucleotide bases in a specific position encode the same amino acid. The latter occurs where the third base could be any of the possible nucleotide bases and also two of the four codons are tolerated at the first position, this means a total of six codons are specific to this amino acid. Given that to build proteins using some of the higher degeneracy amino acids there can be the incorporation of up to six different codons this allows for some increase in efficiency of translation. However, these multiple codons are observed to not undergo equal use within the genome, and as such certain trends within codon usage are seen. These traits are seen to be reconstructed in Southworth et al., (2018) within the eukaryotes using the eukaryotic tree of life to do so.

DEGENERACY IN THE GENETIC CODE –

Triplets of nucleotide bases, known as codons, are the composite basis of the genetic code. This series of letters, in fact, encodes all the information required by an organism to synthesise functioning proteins (McInerney, 1997). Due to the degeneracy of the genetic code, the 20 amino acids used in all of life on the planet are encoded by a total of 64 possible codon combinations (McInerney, 1997). The degeneracy is displayed by the fact that it is possible for one amino acid to have up to six codons dictate its presence, however each codon can only represent one amino acid (Figure 3). Of these 64 codons, three encode stop codons and two encode methionine and tryptophan respectively, leaving 61 for the 20 amino acids in eukaryotic organisms. The 59 remaining codons all possess an alternate analogue encoding the same amino acid (McInerney, 1997). tRNA molecules are those molecules which provide a physical link from a triplet of DNA bases to an amino acid molecule, hence facilitating translation of the triplets into a functional protein (McInerney, 1997 Rafels-Ybern et al., 2019). Degeneracy in the genetic code is the ability of a singular codon to code for only one amino acid, however one amino acid can be encoded by multiple codons dependant on the extent of its degeneracy. Two fold degenerate amino acids can tolerate one out of three possible point mutations at their third position, these amino acids have their second and third positions containing either both purine or both pyrimidine bases. As a result some of these changes at the synonymous sites of two fold degenerate amino acid codons can end with a change in amino acid completely. Within four fold degeneracy, the first two nucleotides of a synonymous amino acid will encode for the same amino acid no matter the last nucleotide. For four fold degeneracy, changes at these synonymous first two positions can result in an amino change however changes within the third position still result in the same amino acid.



	Second Base								
First Base	U		C		A		G		Third Base
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 3: Diagram showing the base composition of the all possible codon combinations and their three letter amino acid code, this highlights the degeneracy of the genetic code as it can be seen that for most amino acids can be encoded by multiple triplets.

Codon Usage Bias & Mechanistic causes –

Codon usage is the phenomenon of how often these synonymous codons are used for each amino acid. It can be used to determine whether this phenomenon is driven by selection or mutation pressure, or a mix of both within the species studied. Codon usage is studied via numerous calculated figures.

Codon usage bias may operate via one of three different mechanisms: mutation pressure, genetic drift and natural selection (Sharp et al., 1988 dos Reis & Wernisch, 2009 Lerat et al., 2002). These mechanisms can work independently; however, they may work in conjunction with each other to affect the levels of codon bias in the host species. Research by Lerat et al., (*2002*) on the species: *Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Homo sapiens,* determined that codon usage was "not simply the outcome of mutational bias", but that selectional constraints would play a role in the trend seen in codon usage (Lerat et al., 2002). Earlier research from Sharp et al., (1988) also supports the viewpoint by commenting that "the pattern of codon usage for each gene must be a balance point between both the mutational pressures, and also the selection for translational accuracy and efficiency on the gene" (Sharp et al., 1988).

CUB could be due to their selection advantage or mutational pressures, as both have been observed to have an effect on this previously (Ikemura, 1985 Sharp et al., 2002 Vicario et al., 2007). However, these processes may not always result in a benefit for the organism, and in the case of mutation could decrease the translation accuracy and/ or efficiency as opposed to increasing it. An example of codon usage bias can be seen in Figure 4 which shows two *Nc* plots, where the effective number of codons of a gene are plotted against its GC content of the third position of its codon (Wright, 1990). An organisms' 'optimal codons' are so termed due to their complementation of the most highly expressed tRNA genes within the species; which has the effect of increasing the translational rate of an organism (Ikemura, 1981). This effect is even more pronounced in highly expressed genes, where selection pressures are linked with the functional importance of the genes, leading to a large CUB shift towards these GC ending codons.

Mutational pressures could also be a valid suggestion for the influential cause of these high frequency codons where natural selection is determined to not be the driving force (Kliman & Hey, 1994). Mutation pressures are seen when the force of random mutations occurring with high frequency are the causal factor of an observed bias. Comparing the GC- content at the synonymous third position within coding and non-coding DNA can display the relationship between the two across different genes. Were mutation pressure a driving force it would be expected that there would be a positive relationship between the two; suggesting local mutation rate is influencing both synonymous GC3s and non-coding GC content. However an observation of no causal relationship doesn't rule out the contribution of mutation pressure within codon usage bias but does indicate that local mutation is not responsible for the variation in GC3s (Kliman & Hey, 1994; Southworth et al., 2018).

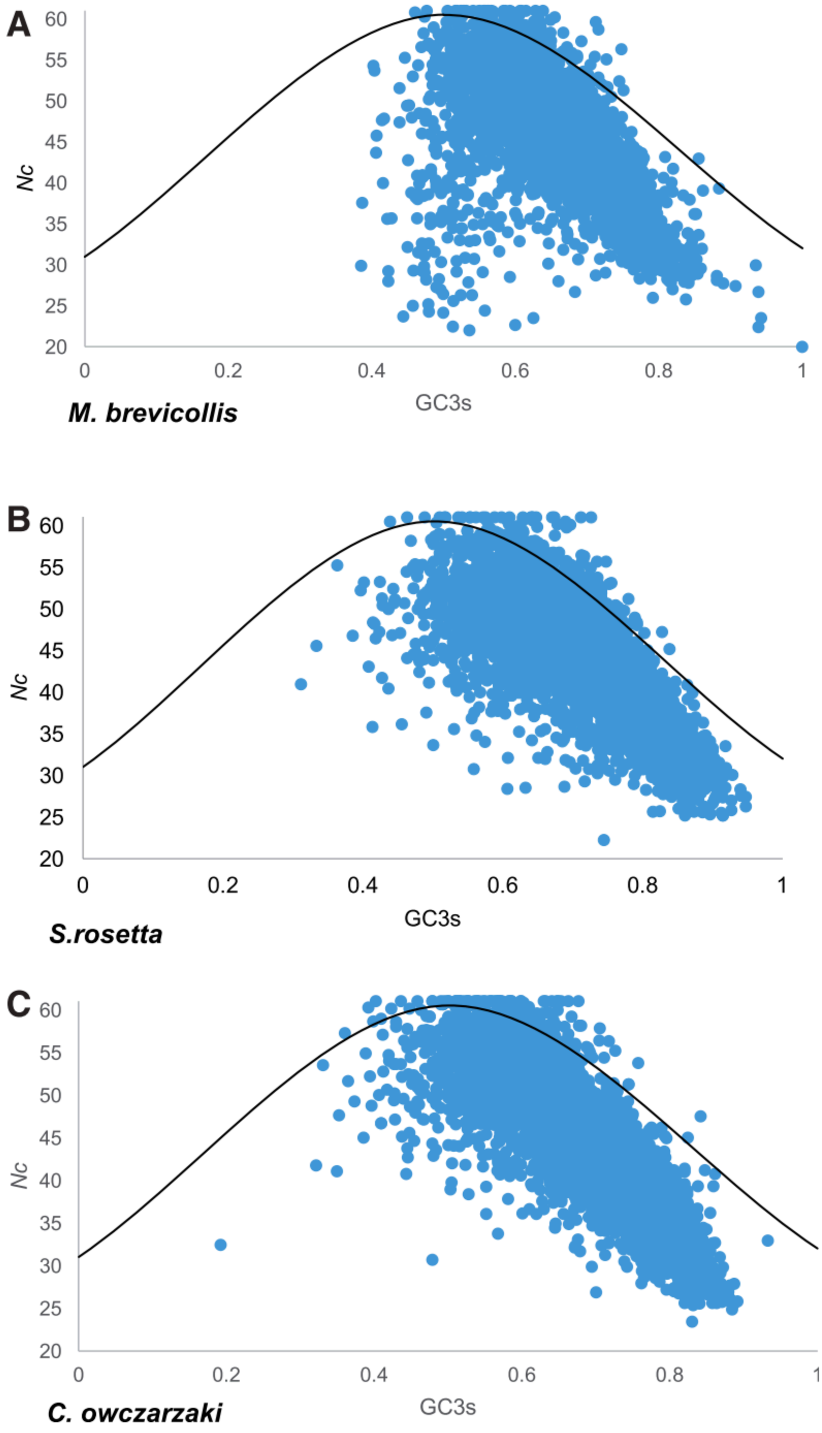


Figure 4: *Nc* plots for species: *M. brevicollis, S. rosetta* and *C. owczarzaki* with values of *Nc* on the Y-axis and GC3s on the X-axis with a curved line overlaying the plots which shows the expected position of genes which are evolving under a neutral mutation model, taken from Southworth et al., (*2018*).

Deamination of adenosine –

Deamination is the process of removal of an amino group (NH_3). In the case of deamination of adenosine this is carried out by two enzymes acting together as a heterodimer. Deamination of adenosine has been observed within the taxonomic group Metazoa, for example in *Homo sapiens*, during the translation of proteins (Rafels-Ybern et al., 2019). During this process the tRNAs involved in linking amino acids to the codon information in protein translation can have the 34th, 37th and 57th bases in their sequence deaminated (Figure 7) (Maraia & Arimbasseri, 2017 Rafels-Ybern et al., 2019). An adenosine base in the wobble position, the first nucleotide of the anti-codon at A34, can be deaminated with the help of a catalysing enzyme such as ADAT2/3 in eukaryotes or TadA in prokaryotic organisms (Su & Randau, 2011 Rafels-Ybern et al., 2018). This deamination causes structural changes to the base, converting it to the base inosine, leads to an alteration in the binding of the base to an alternative complementary partner. Deamination of an adenosine DNA base at position 34 of a tRNA, the first base in the anti-codon, results in a change of nucleotide base to that of inosine. Inosine as a base is structurally very similar to that of nucleotide base guanosine, and is hence termed an analogue of guanosine (Torres et al., 2014). Due to these structural similarities, inosine most often base pairs with nucleotide base cytosine, and hence alters the base pairing of a tRNA anticodon with an mRNA codon during translation (Figure 5) (Torres et al., 2014). This alteration is due to inosine's ability to also wobble pair adenosine and uracil on top of cytosine, giving three possible codon options instead of one (Rafels-Ybern, Torres, Grau-Bove, Ruiz-Trillo, & Ribas de Pouplana, 2018). This is seen in Figure 5, which demonstrates the binding of inosine, and how this enables the one deaminated tRNA molecules to bind three different bases and hence increase rate of translation.

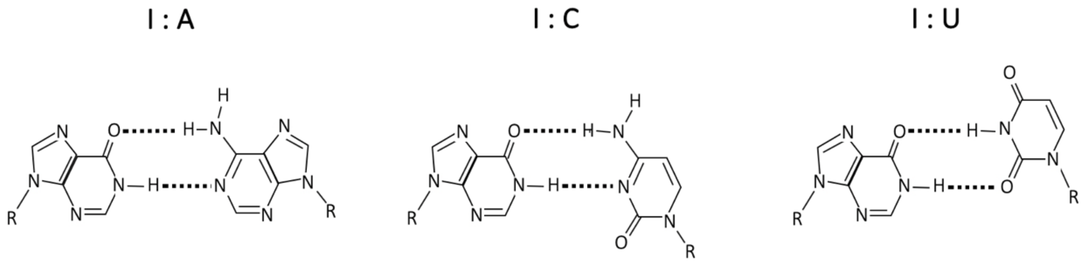


Figure 5: Diagram adapted from Torres et al., *2014* showing the structure of bonding between the deaminated base inosine and bases: cytosine as well as adenosine and uracil, with hydrogen bonds which bind the bases together represented by dotted lines.

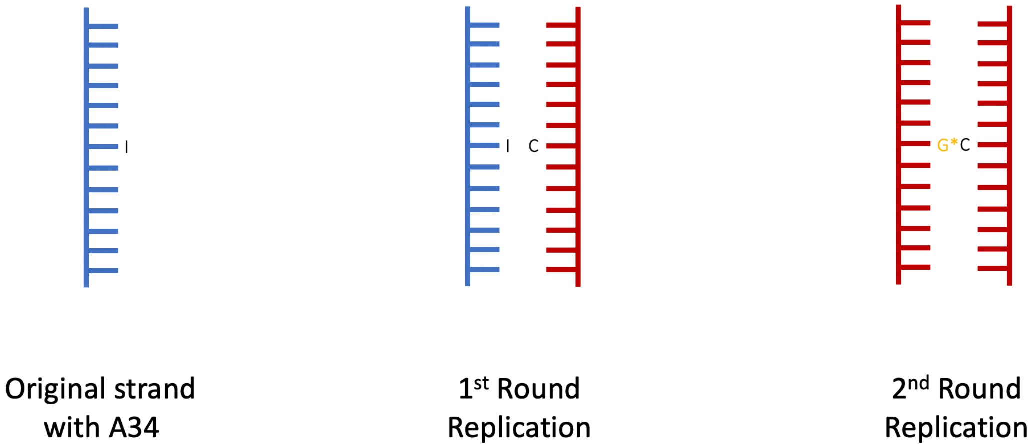


Figure 6: Diagram showing the process of DNA replication which results in the original strand containing inosine at position 34, is converted to a contain a guanosine nucleotide at the wobble position. This is to demonstrate how deamination of adenosine can be identified through BLAST searches to locate tRNA genes with guanosine at position 34; adenosine, which is converted to inosine, is then paired with cytosine during transcription. A second complimentary strand is manufactured to include a guanosine at position 34; this is to base pair with the cytosine which reveals evidence for deamination.

When DNA is undergoing transcription and being replicated the DNA bases in the original strands are paired with their complimentary nucleotide bases, as such the inosine base within the DNA sequence of the deaminated tRNA molecules is paired with a cytosine residue; it's preferential partner, during the first round of replication. Nucleotide bases for inosine do not exist within the cells and so the cytosine residue in the first new strand is paired with its preferential partner, guanosine, during the second round of replication. This leads to the existence of a tRNA molecule with guanosine at the wobble position from the replication of an A34 tRNA molecule (Figure 6). The presence of these identified tRNA genes with a guanosine at their wobble site within transcriptomes of a species can provide evidence for the process of deamination within the species.

The modification of A34 to inosine increases the ability of tRNAs to bind codons three-fold, hence the use of this tRNA modification reduces the number and also variety of tRNA genes required by the organism within its genome. This is because only one modified adenosine tRNA gene can complete the same role as three individual tRNA genes; adenosine, guanosine and uracil, which could provide a selective advantage in species which have more compact genomes.

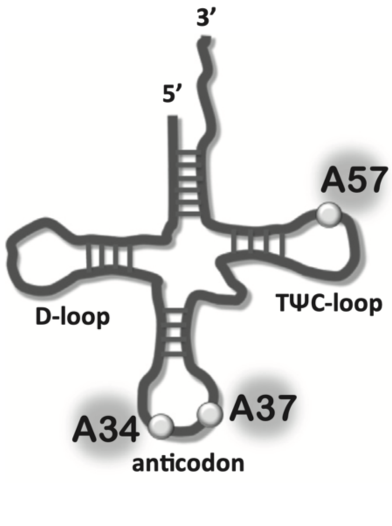


Figure 7: Diagram taken from Torres et al., (*2014*) displaying the structure of a tRNA molecule with the possible sites for adenosine deamination: A34, A37 and A57 highlighted. The adenosine at position 34 is the first nucleotide of the anti-codon which can be deaminated to inosine in order to bind a wider range of mRNA codons.

The adenosine to inosine modification has also been observed to produce a preference of the codon usage of an organism, whereby there is an increase in the bias towards GC ending codons (Maraia & Arimbasseri, 2017). This shift in bias is shown in that highly biased genes are seen to use GC ending codons over AT endings for the same amino acids. Bioinformatic techniques are used to determine the effective number of codons and the GC3s or GC content at the codon third position (Southworth et al., 2018). The preference of GC ending codons gives an increase in translation rates, due to the binding of these optimal codons to the modified inosine base, if these inosine tRNA molecules are the most abundant within the genome. This could be the driving force behind this codon bias, through being selectively favoured by the organism during evolution. The organism's increased translational efficiency, meaning increased growth rate through efficient protein synthesis, and hence reproductive success of the organism. The increased efficiency is due to the fact that the deaminated adenosine tRNA molecules complement the optimal codons for that organism; this facilitates rapid translation as the most abundant tRNA molecules are able to enter the ribosome swiftly with more being available within the cell. The less abundant tRNA molecules would take longer to enter the ribosome due to their rarity within the host cells. The optimal codons for a species are highly enriched within regions of the organism's genes which encode functional domains, as opposed to non-functioning regions (Akashi, 1994); this is consistent with selection on the translational accuracy of the organism's host machinery. The translation of functional domains is expected to be accurate, as the organism should be maximising its translational accuracy; as such selection should be stronger within these functional domains. *Precup & Parker (1987)* had previously shown that using the optimal codons within functional domains for species *Escherichia coli*, resulted in a ten fold increase in accuracy, through the correct amino acids being incorporated into the protein. Later in 1994, support was provided when it was shown that optimal codons were observed at a higher frequency within functional domains for the species *Drosophila melanogaster* (Akashi, 1994).

The process of the deamination of tRNA molecules is yet undetermined as to its evolutionary mechanism; but one possible suggestion is that it may possibly have evolved within the group Holozoa. This could potentially be highlighted by the discovery of this process on specific TAPSILVR amino acids in Metazoa, whose closest known relatives are the groups of the choanoflagellates and Filasterea (Rafels-Ybern et al., 2019 Southworth et al., 2018).

MEASURES OF CUB –

Measures of the number of effective codons (*Nc*) are used to determine how many of the possible codon combinations for the degenerate amino acids are actually utilised in practice. *Nc* is measured with a value of between 20 and 61; where 20 displays the most bias as this depicts a situation where each degenerate amino acid is biased towards using only one codon. A value of 61 shows the least bias as in this scenario each degenerate amino acid uses each codon at equal frequency, showing no bias towards a single optimal codon (Wright, 1990).
The calculations of GC3s are produced by looking at the guanosine and cytosine composition of the third synonymous base in triplets of DNA. In codon usage analysis GC3s is a standard measure of the proportion of codons with DNA bases guanosine or cytosine at their synonymous third position, as such this term has been used throughout. Cytosine binds inosine preferentially and as such its prevalence in the third position is a component of GC3s measure. As such this is the base interest for the investigations into GC3s measurements and allows the standard measures and terminologies must be upheld. The importance of their prevalence in this positioning is that these trends can show a direction of bias within the third base composition; allowing comparisons between species and also between base composition at coding and non-coding DNA regions within a genome. This displays a clear benefit to the process of tRNA deamination, as the efficiency of translation is increased, due to this less specific binding of codons allowing the host organism an increased growth rate.

There is a link between the efficiency of the protein translation and the deamination process where, in this case often C- ending optimal codons bind the major tRNA genes, which often happen to contain adenosine at their wobble position. The expression level of the tRNA genes which contain an adenosine at their wobble position is therefore linked to the codon usage bias, as those deaminated major tRNAs preferentially bind the C- ending optimal codons; which is a component of GC3s. Whilst these data are not available, in bacteria, it has been shown that tRNA expression level positively correlates with gene copy number (Ikemura, 1985). Therefore, it is possible to use the copy number as a proxy for tRNA molecule abundance, if deamination is occurring then it could be expected that the deaminated molecules would be the most abundant. Previous research by Southworth et al., (*2018*) determined that for three holozoan species, that the major tRNA genes all bound to these optimal GC- ending codons. As such a link is expected between the optimal codons, which bind the major tRNA anticodon, and expression level. If deamination occurs within the species then the prediction is that for the TAPSILVR amino acids they will have C- ending optimal codons, where the tRNA genes with an adenosine at the wobble position are the major tRNA genes.

RSCU is a measure of the deviation difference from even usage, the expected number of times a codon should appear when considering equal synonymous codon usage, and the usage measured. RSCU can be determined by its relation to the numerical value one; a value of below one indicates lower frequency of an individual codon and a value of above one indicates a codon was utilised at a higher frequency than would be expected in random codon usage.

Fop values, frequency of optimal codons, is a determination of the proportion of optimal codons within a gene when compared with the total number of codons within that gene. *Fop* values are calculated by dividing the number of optimal codons within a gene by the total number of codons within that gene (Ikemura, 1981). There are two methods of calculating optimal codons; using expression data and correspondence analyses. Within this research expression data was utilised over correspondence analyses as the latter these can give false positive results (Peden, 1999); if natural selection was not the driving force of this evolution then CodonW would still interpret this as a positive result, however this would be false.

Previous published codon usage Research –

Codon usage bias can be affected by both mutational and selection pressures. The theory for this was originally suggested by Clarke et al., (*1970*), no evidence was provided supporting this at the time due to a lack of data, as this paper is dated before the widespread availability of sequence data; however, there was no available data pointing to an alternative explanation. Clarke stated that there was 'no strong argument in favour of the "neutral" hypothesis' in the selection of base composition at the synonymous third position. Grantham et al., (*1980*) provided data to support this in the form of the usage data from 119 sequences, indicating that the neutral hypothesis was indeed not the driving force behind codon usage. This was followed by research from Ikemura in 1981 defining optimal codons as those which complement the most abundant tRNA molecules; which happened to correlate to the copy number of the tRNA genes within the species investigated (Ikemura, 1981). Later this was altered to state that optimal codons are those which complement the most highly expressed tRNA genes (Ikemura, 1981). A large component of the existing work on codon usage is based on multicellular eukaryotes (Chen et al., 2012 Duret, 2002 Smith & Eyre-Walker, 2001 Whittle & Extavour 2016), however most life on the planet is not multicellular. To improve this understanding of eukaryotic codon usage we need to look at bigger picture by studying unicellular species. *Homo sapiens* are recently evolved within eukaryotic life, however due to our interests in gaining knowledge about our own species, there has been a lack of research into the singular celled organisms from which we have evolved.

Southworth et al., (*2018*) determined that mutation pressure did not seem to be the driving force for codon bias in three species: *Monosiga brevicolis* (Choanoflagellatea), *Salpingoeca* *rosetta* (Choanoflagellatea) and *Capsaspora owczarzarki* (Filasterea). The paper also showed that the TAPSILVR: Threonine, Alanine, Proline, Serine, Isoleucine, Leucine, Valine and Arginine amino acid tRNA molecules require deamination in order to utilise the more abundant cytosine optimal codons. The identification of this process in both choanoflagellates and also *Capsaspora owczarzaki* for the TAPSILVR amino acids indicates that this deamination process was present in the last common ancestor of all three species. Eukaryotic species may have evolved from a symbiosis between archaeal and bacterial species (Williams et al., 2013), bacteria had previously been shown in existing research to deaminate only arginine tRNA molecules (Gerber & Keller 1999 Rafels-Ybern et al., 2018 Torres et al., 2014). However more recent research into bacterial deamination has identified tRNA genes for other amino acids; such as isoleucine with an adenosine base at its wobble position in species *Mycoplasma bovis*, a tenericute bacterium (Rafels-Ybern et al., 2019). Also identified was adenosine first codon tRNAs for isoleucine, serine and threonine within bacterial firmicute *Oenococcus oeni* (Rafels-Ybern et al., 2018). Investigations into archaeal deamination have determined there is no evidence for deamination of any tRNA molecules for any of the amino acids (Rafels-Ybern et al., 2018). This is implicative of the modification of tRNAs that we see today having been gained through the symbiosis of Archaea and Bacteria. From here it may have expanded to include all higher degenerate amino acids bar glycine, due to its instability with adenosine at the wobble position leading to its lack of deamination.

This is one explanation for the evolution of deamination of tRNA molecules within eukaryotes and is supported by the discovery of the deamination of arginine in bacteria, Further support is provided by the identification of other amino acid tRNA genes with adenosine at their wobble position, however requires some further investigation to establish this (Rafels-Ybern et al., 2018). The investigation of many additional species from all walks of life would aid this research by helping to rule out convergent evolution as a method of this process evolving. Understanding how widespread this process is within plants and animals can help to support theories for multicellular and unicellular transitions, to determine whether the emergence of multicellularity may have resulted in the loss of this process within certain lineages.

Codon Usage in Holozoan Protists –

Whilst codon usage has been extensively studied within Metazoa (Chen et al., 2012 Duret, 2002 Smith & Eyre-Walker, 2001 Whittle & Extavour 2016), there has been minimal work on this subject in the field of unicellular eukaryotes; which is a great detriment to current knowledge (Southworth et al., 2018). Southworth et al., (*2018*) showed conservation of codon usage bias for three protistan species, however, did not explore this further than the holozoan group. The paper determined there was a strong GC bias in all three species analysed; highlighted by higher GC3s shown in the highly biased gene categories, and lower GC3s in the low biased genes. An alternative explanation for codon usage bias is that the process is driven by mutational pressure; however, Southworth et al., (2018) investigated this and determined that GC3s values for the three species investigated show that this is unlikely to be the causal factor for this bias. Evidence against mutation bias included a lack of positive correlation between values of GC3s for host genes, when GC content within flanking DNA and non-coding DNA were analysed comparatively to the coding regions, indicating local mutation pressure does not influence the variation in GC3s (Southworth et al., 2018).

In order to gain a better understanding of earlier evolution of eukaryotes, and hence early evolution in general it is important to pursue this research within a broader range of groups. This is to determine where the evolution of the modification of translation and codon bias first occurred, by tracing their presence in many eukaryotic species. It is theorised that the eukaryotic supergroup evolved from a symbiotic relationship between the Archaea and bacterial lineages(Lake, 2015; Williams et al., 2013); this is observed by the presence of mitochondria in organisms whose DNA is of bacterial origin, and also the resemblance of eukaryotic ribosomes to those of Archaea (Lake, 2015). Investigations into Archaeal genomes has failed to identify any evidence of tRNA modification (Marck & Grosjean, 2002), however bacteria are known to modify tRNA genes for amino acids; arginine, isoleucine, serine and threonine in recent research (Rafels-Ybern et al., 2018 Rafels-Ybern et al., 2019). Provided by this is a possible explanation that the evolution of tRNA modification has been passed through the bacterial lineage at the base of the eukaryotic supergroup; and the process has expanded from there to include other tRNA molecules. This is supported by that fact that eukaryotes have multiple adenosine tRNA deaminating enzymes, for the deamination of all eight TAPSILVR amino acid tRNAs. However, bacteria are only shown to have one of these enzymes for deamination of its adenosine tRNA molecules (Rafels-Ybern et al., 2018).

Past research undertaken into the process of tRNA deamination has been carried out in very select groupings only; mainly focusing on multicellular organisms due to their relatively close genetic distance to the *Homo sapien* lineage, this is in the search for human treatments and medicines. However, in order to truly understand the evolution of the process research is required into early branching eukaryotic species. Research undertaken into the process of deamination in Archaea has shown that they do not deaminate any tRNA molecules, and past research into bacterial species had shown that bacteria only deaminate tRNA molecules for the amino acid arginine until recently. Rafels-Ybern., (*2018*) identified further tRNA genes such as; isoleucine and threonine, with adenosine at their wobble positions. Therefore, the mechanism for deamination of the TAPSILVR amino acids appears to be specific to eukaryotic organisms. My research has been carried out in order to attempt to determine exactly where and how this mechanism appears to have evolved. The presence of the mechanism in bacterial species, though limited to a smaller number of tRNA molecules, could suggest the mechanism was gained from bacteria through horizontal transfer in symbiotic species. Once the mechanism was in place within eukaryotic species it may have evolved to include other amino acid tRNA molecules through selectional pressures. Due to the lack of research into how this mechanism made the leap into eukaryotic lineages, my research has attempted to bridge this gap in current knowledge. This research began originally by looking at close relatives to the holozoan group, due to research by Southworth et al. 2018 into the mechanism within Holozoa; and as such species *Fonticula alba* and *Thecamonas trahens* were selected as close relatives for investigation.

Project Aims:

The aims of this research project are to identify the optimal codons of the species chosen within my study; *F. alba, T. trahens, C. crispus, C. reinhardtii, B. natans* and *L. tarentolae*, this will allow me to determine whether or not the optimal codons for each species match the major tRNA genes. This also includes identifying the range of tRNA genes within these species and seek out any evidence of tRNA modification at the wobble position. The overall aim of this project is to attempt to identify the origin of eukaryotic tRNA deamination.

METHODS:

Table 1: Table displaying the species names alongside their taxon identifiers, denoted by an asterisk if information taken from NCBI, accession numbers/locations and the webpage they were accessed from (Sayers et al., 2009).

SPECIES NAME	TAXON ID	ACCESSION NUMBER	WEBPAGE ACCESSED
F. alba	691883	GCA_000388065.2	Protists.ensemble.org
T. trahens	461836	GCA_000142905.1	Protists.ensemble.org
C. crispus	2769	GCA_000350225.2	Plants.ensemble.org
C. reinhardtii	3055	GCA_000002595.3	Plants.ensemble.org
B. natans	753081	GCA_000320545.1	Protists.ensemble.org
L. tarentolae	5689*	40	Tritrypdb.org

Nc &GC3s –

Values of *Nc* and GC3s for species: *T. trahens, F. alba, L. tarentolae, C. crispus, C. reinhardtii* and *B. natans* were calculated using the program CodonW. The input files were created by taking the gene files downloaded from the species respective webpages as shown in table 1 and running the program with parameters left as standard to obtain an output containing: the gene identifier, its *Nc* value and GC3s value.

RSCU –

Optimal codons were provided by CodonW correspondence analysis on relative synonymous codon usage, COA on RSCU, run with the default settings. The top and bottom 5% expressed gene sequences identified via expression data as stated in table 1, extracted from the cds file via Seqtk (https://github.com/lh3/seqtk), were analysed in CodonW with the genes concatenated in order to allow comparison of the codon usage between the high and low expressed optimal codons to determine the optimal codons as these are utilised more in the high expressed genes rather than the low expressed genes. Chi-squared values of these codon numbers were then calculated to determine whether there was a significant difference between the optimal codon's usage within high and low bias gene categories.

TRNA GENES –

tRNA genes were identified by taking the whole genomes and non coding RNA (ncRNA), downloaded from: www.protists.ensembl.org, www.plants.ensemble.org and www.tritrypdb.org, which were known to contain tRNA genes; and inputting them into tRNAscan-SE 2.0, an online and desktop program available at (http://lowelab.ucsc.edu/tRNAscan-SE/). These genomes were inputted into the tRNA scan system and the parameters were left as standard. The whole genomes analysed were for species: *Thecamonas trahens* (Apusuzoa)*, Fonticula alba* (Opisthokonta)*, Leishmania tarentolae* (Excavata)*, Chondrus crispus* (Archaeplastida)*, Chlamydomonas reinhardtii* (Archaeplastida) and *Bigelowiella natans* (Rhizaria). The output files contained the tRNA gene sequences of all tRNA's present within the genomes, which where extracted into Excel for further analysis between all genes (Lowe & Chan, 2016).

TRNA GENES CONTAINING GUANOSINE AT THE WOBBLE POSITION –

tRNA genes with guanosine at the wobble site were then searched for using NCBI blast's Nucleotide Blast, available at (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_L

OC=blasthome). The tRNA sequence from tRNA scan was used as a query and set to search the sequence read archive (SRA) by quoting the organism ID's in order to identify any genes present within the organisms' genome which match the query sequence (Altschul et al., 1990). These could then be used to positively confirm tRNA genes which contained adenosine bases. The tRNA sequences were then modified manually in TextWrangler to include a guanosine residue at the wobble position as opposed to the occurring adenosine and run again. The presence of a guanosine at the wobble sites proves the process of deamination is occurring as a deaminated adenosine (inosine base) will base pair preferentially with cytosine; when DNA replication occurs and cDNA is made the complimentary strand made during the first round would contain a cytosine base (Figure 8) , however when the second round occurs a guanosine would be incorporated to base pair with the cytosine as inosine cannot be incorporated. This was carried out as tRNA scan could have missed some tRNA genes, BLAST was used as it checks the DNA sequence only to determine the presence of tRNA genes with guanosine at the wobble position which may be present within the transcriptome but remain un-annotated. tRNA-scan is expected to be accurate however in the interests of carrying out thorough research and due to some of the interest species being unusual the results were checked with BLAST.

KOG CATEGORIES –

The functional KOG category of each gene was determined by use of EMBL's EggNOG, available at (http://eggnogdb.embl.de/#/app/home) (Huerta-Cepas et al., 2016). Carried out on genes for the species: *T. trahens*, *F. alba*, and *L. tarentolae,* to allow the spread of functional categories for each genome to be seen. The transcriptome file was taken from various websites as seen in table 1 and converted into protein sequence before being uploaded into the server, with the parameters were left as standard. The Eukaryotes database selected due to these sequences being from eukaryotic species. The numeric outputs of these analyses were inputted into excel and from there used to calculate proportional percentages of each category to determine the most enriched functional category for each species.

EXPRESSION DATA –

Expression data was calculated using the Sanger institute's SMALT tool, Sequence Mapping and Alignment Tool, (https://www.sanger.ac.uk/science/tools/smalt-0). Pairwise alignment of the CDS gene sequences with multiple transcriptome sequences, obtained from NCBI's Sequence Read Archive (SRA) for individual species, was carried out in order to determine the expression for individual genes. This was based on the relative number of reads of each gene within the transcriptomes being mapped onto the single copy of the gene within the consensus sequences. Once the read data was produced by SMALT then Tablet, a mapping viewer tool from The James Hutton Institute, (https://ics.hutton.ac.uk/tablet/) was employed in order to visualise the output files. This allowed visualisation of the read numbers mapped onto an index sequence, and also of the read values in a chart form. These values were exported into Microsoft's Excel for manipulation to allow comparison between methods of calculating expression data. Comparisons were only drawn within species and were not employed outside of species. SMALT was also utilised in order to map modified and unmodified versions of tRNA genes identified by tRNAscan-SE 2.0 as present within the genome onto a consensus file consisting of multiple transcriptome files merged into one, in order to determine whether there is presence of tRNA genes with an adenosine base at their wobble site in the first position of the anticodon. This was tested alongside a manually modified tRNA gene with the adenosine removed and replaced with a guanosine base with program TextWrangler. This is to identify any tRNA genes which match the C ending optimal codons, often seen in species, without the deamination of adenosine to inosine due to the conversion of inosine into guanosine during reverse transcription (Southworth et al., 2018).

RESULTS:

FONTICULA ALBA AND *THECAMONAS TRAHENS* –

Initial investigations into the evolution of tRNA modification began by looking at the genomes of *Fonticula alba* and *Thecamonas trahens*; these species were selected due to their status as close relatives of the species group Choanoflagellata, and the species *Capsaspora owczarzaki*. The results published by Southworth et al., 2018, showed evidence for deamination of the TAPSILVR amino acid tRNA molecules in choanoflagellate species' *Monosiga brevicolis, Salpingoeca rosetta* and also *Capsaspora owczarzaki*. These results showed the estimate for the evolution of tRNA deamination of TAPSILVR amino acid genes was the last common ancestor of the three species investigated.

Using data produced for both categories *Nc* and GC3s from program CodonW, utilising the CDS files for both species *T. trahens* and *F. alba;* plots of the number of effective codons (*Nc*) versus GC3s were created for both species. These plots were created in order to display the major trend in codon usage for each species. *Thecamonas trahens' Nc* plot shows a shift to the right of normal distribution. A shift such as this displays a bias towards GC ending codons and only a few outlying genes, falling outside of the major grouping of genes, are observed. (Figure 8). The *Nc* plot created for species *F. alba* also displays the majority of the genes following the same trend, with a strong positive correlation between the number of effective codons and GC3s. However, this species has a few more outliers falling into a less conserved grouping, opposing the trend, and showing a correlation between the *Nc* and GC3s (Figure 8). The outlying genes within *F. alba* were not investigated further, due to a time constraint on the project, however future work should include an investigation of this in order to expand the understanding of this grouping.

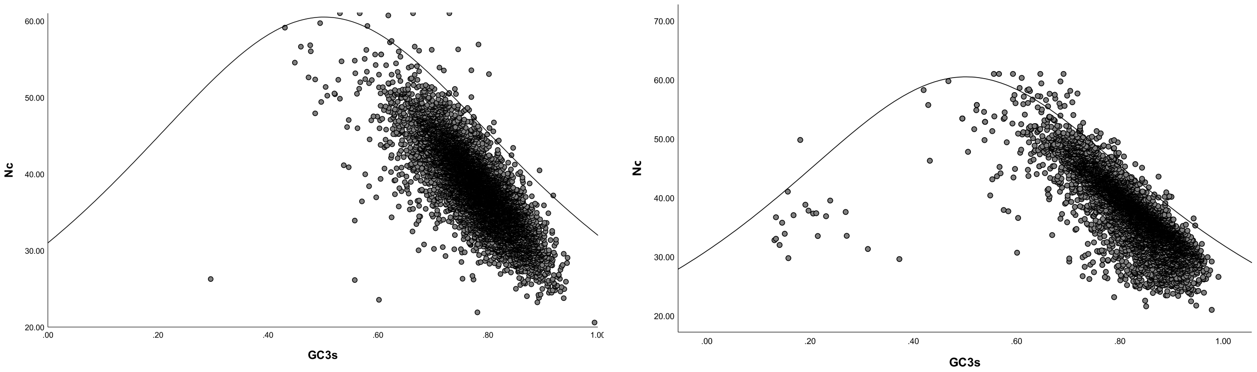


Figure 8: *Nc* plots produced by SPSS for both species *T. trahens* (left) and *F. alba* (right) showing the axes of both *Nc* on the Y-axis versus GC3s on the X-axis. Each scatter point represents one gene within the genome of the species with the GC3s and *Nc* of each gene determining its position in the graph relative to the curve; the curved line overlaying the plots shows the expected position of genes which are evolving under a neutral mutation model (Southworth et al., 2018). The observed trend seen of the points largely on the right hand side of the plot shows these species have a bias towards GC- ending optimal codons.

Correspondence analysis was undertaken for both species, which analyses major trends in the variation of the data presented; these were generated using correspondence analyses from the program CodonW and also the expression data from the program SMALT. For the high and low expressed genes, their optimal codons were determined using a chi-square test to determine if codon frequencies are statistically significantly different in highly and weakly expressed genes (Tables 4-5). These chi-squared values being below the critical value of P=0.01 (Peden, 1999), dictates that the hosts genome is selecting for efficiency during translation.

The analyses of the optimal codons via both expression level and also correspondence analyses produced identical results for the identification of optimal codons; this determines that for the species analysed the correspondence analyses on CodonW are enough to identify the optimal codons. The optimal codon analysis produced for each species during these analyses, are also consistent with the deamination of TAPSILVR amino acid tRNAs.

Table 2: Table displaying the optimal codons for species *F. alba* as determined by correspondence analyses on the genome. Displayed is the amino acid identity, the codons used for this amino acid, optimal codons are represented by an asterisk next to the codon. The frequency of the codon is displayed with the RSCU number first against the codon base, followed by the raw number of codons next to it in brackets.

Phe	UUU	0.07 (125) 0.88 (1166)	Ser	UCU	0.49 (498) 0.84 (1205)
	UUC*	1.93 (3539) 1.12 (1494)		UCC*	3.38 (3430) 1.56 (2242)
Leu	UUA	0.00 (0) 0.47 (661)		UCA	0.00 (1) 0.65 (934)
	UUG	0.01 (17) 1.07 (1515)		UCG	0.94 (953) 1.44 (2073)
	CUU	0.33 (398) 0.86 (1219)	Pro	CCU	0.14 (145) 0.70 (1574)
	CUC*	2.60 (3179) 1.52 (2160)		CCC*	2.46 (2587) 1.24 (2793)
	CUA	0.00 (2) 0.30 (430)		CCA	0.00 (4) 0.71 (1604)
	CUG*	3.05 (3729) 1.79 (2541)		CCG	1.40 (1475) 1.35 (3050)
Ile	AUU	0.42 (667) 1.05 (1021)	Thr	ACU	0.33 (441) 0.76 (853)
	AUC*	2.58 (4143) 1.42 (1380)		ACC*	3.54 (4748) 1.48 (1661)
	AUA	0.00 (1) 0.52 (507)		ACA	0.00 (6) 0.91 (1024)
Met	AUG	1.00 (2212) 1.00 (1853)		ACG	0.12 (165) 0.86 (964)
Val	GUU	0.44 (681) 0.73 (909)	Ala	GCU	0.44 (998) 0.71 (1835)
	GUC*	2.79 (4373) 1.26 (1568)		GCC*	3.32 (7603) 1.71 (4404)
	GUA	0.00 (1) 0.35 (434)		GCA	0.01 (12) 0.64 (1659)

	GUG	0.77 (1206) 1.67 (2086)		GCG	0.24 (546) 0.94 (2430)
Tyr	UAU	0.19 (240) 1.10 (787)	Cys	UGU	0.09 (56) 0.79 (726)
	UAC*	1.81 (2264) 0.90 (638)		UGC*	1.91 (1185) 1.21 (1119)
TER	UAA	2.46 (256) 0.62 (62)	TER	UGA	0.31 (32) 1.58 (159)
	UAG	0.23 (24) 0.80 (81)	Trp	UGG	1.00 (786) 1.00 (1242)
His	CAU	0.26 (266) 0.87 (1259)	Arg	CGU*	0.90 (783) 0.63 (841)
	CAC*	1.74 (1762) 1.13 (1624)		CGC*	4.69 (4080) 1.55 (2081)
Gln	CAA	0.01 (18) 0.64 (1111)		CGA	0.01 (6) 0.77 (1039)
	CAG*	1.99 (2956) 1.36 (2374)		CGG	0.40 (345) 1.71 (2300)
Asn	AAU	0.14 (267) 1.11 (1028)	Ser	AGU	0.04 (41) 0.51 (731)
	AAC*	1.86 (3581) 0.89 (824)		AGC*	1.15 (1168) 1.01 (1461)
Lys	AAA	0.00 (8) 0.86 (1011)	Arg	AGA	0.00 (0) 0.53 (706)
	AAG*	2.00 (4945) 1.14 (1340)		AGG	0.00 (4) 0.81 (1089)
Asp	GAU	0.44 (1139) 0.96 (1792)	Gly	GGU*	1.23 (1943) 0.62 (1276)
	GAC*	1.56 (4053) 1.04 (1938)		GGC*	2.56 (4059) 1.57 (3237)
Glu	GAA	0.01 (28) 0.62 (1168)		GGA	0.01 (18) 0.62 (1270)
	GAG*	1.99 (5468) 1.38 (2577)		GGG	0.20 (312) 1.19 (2456)

Table 3: Table displaying the optimal codons for species *T. trahens* as determined by correspondence analyses on the genome, formatted in the same style as table 2.

Phe	UUU	0.97 (4467) 1.34 (4273)	Ser	UCU	0.43 (1369) 0.55 (2012)
	UUC*	1.03 (4790) 0.66 (2096)		UCC*	1.64 (5164) 0.37 (1373)
Leu	UUA	0.01 (20) 0.04 (161)		UCA	0.34 (1072) 0.51 (1882)
	UUG	0.08 (315) 0.62 (2546)		UCG*	3.27 (10327) 2.43 (8890)
			Pro	CCU	0.34 (1144) 0.54 (1734)
	CUU	0.68 (2679) 0.87 (3593)		CCC*	1.32 (4461) 0.44 (1418)

	CUC*	4.40 (17412) 1.53 (6281)		CCA	0.25 (832) 0.68 (2171)
	CUA	0.05 (179) 0.24 (970)		CCG	2.10 (7080) 2.33 (7444)
	CUG	0.79 (3124) 2.71 (11134)	**Thr**	ACU*	0.47 (1620) 0.42 (1352)
Ile	AUU	0.51 (1946) 1.13 (2335)		ACC*	2.76 (9541) 0.86 (2761)
	AUC*	2.48 (9394) 1.69 (3476)		ACA	0.28 (959) 0.65 (2100)
	AUA	0.01 (37) 0.18 (371)		ACG	0.49 (1693) 2.06 (6612)
Met	AUG	1.00 (5566) 1.00 (6166)	**Ala**	GCU*	0.63 (5083) 0.51 (4869)
Val	GUU	0.44 (2032) 0.48 (2958)		GCC*	2.63 (21367) 0.82 (7789)
	GUC*	2.97 (13691) 0.91 (5584)		GCA	0.30 (2457) 0.69 (6583)
	GUA	0.03 (134) 0.22 (1366)		GCG	0.44 (3574) 1.99 (18962)
	GUG	0.56 (2569) 2.39 (14715)	**Cys**	UGU	0.20 (282) 0.57 (1120)
Tyr	UAU	0.13 (422) 0.56 (1311)		UGC*	1.80 (2578) 1.43 (2783)
	UAC*	1.87 (6241) 1.44 (3357)	**TER**	UGA	0.77 (137) 1.78 (311)
TER	UAA	1.43 (253) 0.53 (92)	**Trp**	UGG	1.00 (2677) 1.00 (3842)
	UAG	0.80 (141) 0.70 (122)	**Arg**	CGU*	0.77 (1816) 0.56 (1712)
His	CAU	0.30 (828) 0.75 (1832)		CGC*	4.56 (10714) 1.18 (3637)
	CAC*	1.70 (4649) 1.25 (3034)		CGA	0.09 (213) 0.75 (2290)
Gln	CAA	0.14 (495) 0.38 (1264)		CGG	0.50 (1178) 2.80 (8617)
	CAG*	1.86 (6433) 1.62 (5343)	**Ser**	AGU	0.04 (129) 0.51 (1863)
Asn	AAU	0.12 (439) 0.45 (1068)		AGC	0.28 (884) 1.63 (5960)
	AAC*	1.88 (7081) 1.55 (3634)	**Arg**	AGA	0.03 (71) 0.23 (708)
Lys	AAA	0.05 (304) 0.21 (649)		AGG	0.05 (114) 0.48 (1476)
	AAG*	1.95 (10911) 1.79 (5614)	**Gly**	GGU*	0.59 (2455) 0.51 (3055)
Asp	GAU	0.29 (2197) 0.69 (4582)		GGC*	3.13 (12925) 1.81 (10933)
	GAC*	1.71 (12963) 1.31 (8644)		GGA	0.13 (526) 0.64 (3870)

Glu	GAA	0.17 (1126) 0.27 (2030)		GGG	0.15 (608) 1.05 (6336)
	GAG*	1.83 (12024) 1.73 (13096)			

Table 4: Table displaying the optimal codons for species *T. trahens*, displayed is the raw number of optimal codons utilised for the amino acid within the high and low expressed genes, alongside the values for all the other codons utilised for that amino acid. Displayed alongside these are the chi-squared values for each amino acid, in order to determine whether their use within the high expressed genes is significantly higher than their use with the low expressed genes. The significant value for this analysis is P=0.01 therefore any analyses where the value is below this threshold determined there is a statistically significant difference between the two data categories.

	High Bias Genes		Low Bias Genes		Chi-Sqaured
Phe					
UUC	3235	2829	3410	5046	*P*<0.0001
Leu					
CUC	7620	6526	15458	19644	*P*<0.0001
Ile					
AUC	5519	1849	7438	3006	*P*<0.0001
Val					
GUC	6762	6518	12628	14343	*P*<0.0001
Ser					
UCC	2153	11677	3578	23751	*P*<0.0001
UCG	7961	5869	12391	14938	*P*<0.0001
Pro					
CCC	2052	6848	3423	14958	*P*<0.0001
CCG	5200	3700	9877	8504	*P*<0.0001
Thr					
ACC	5146	5145	7940	10435	*P*<0.0001
ACG	3233	7058	5287	13088	*P*<0.0001
Ala					
GCC	8970	10415	22680	34446	*P*<0.0001
Tyr					
UAC	3905	401	4660	1240	*P*<0.0001
His					
CAC	2674	790	5017	2321	*P*<0.0001
Gln					
CAG	5075	471	7408	1618	*P*<0.0001
Asn					
AAC	5360	467	5669	1242	*P*<0.0001
Lys					
AAG	8363	268	7134	825	*P*<0.0001
Asp					

GAC	8113	2121	12958	5947	*P*<0.0001
Glu					
GAG	8849	756	15759	2794	*P*<0.0001
Cys					
UGC	2356	416	2972	878	*P*<0.0001
Arg					
CGC	4696	5434	9789	13111	*P*<0.0001
Gly					
GGU	2144	10875	3449	19162	*P*<0.01
GGC	8589	4430	13483	9128	*P*<0.0001

Table 5: Table displaying the optimal codons for species *F. alba*, formatted in the same style as table 4.

	High Bias Genes		Low Bias Genes		Chi-Sqaured
Phe					
UUC	6403	921	2495	626	<0.0001
Leu					
CUC	5575	14054	2618	7316	0.0002
CUG	11324	8305	5391	4543	<0.0001
Ile					
AUU	1866	6436	609	2467	0.0023
Val					
GUU	1346	12140	432	5548	<0.0001
GUC	6509	6977	2334	3646	<0.0001
Ser					
UCU	1407	15767	401	8214	<0.0001
UCC	6667	10507	2477	6138	<0.0001
Pro					
CCU	1050	12936	709	9824	0.0211
CCC	5323	8663	3686	6667	<0.0001
CCG	7022	6964	5141	4692	0.0017
Thr					
ACU	1088	10260	415	5041	<0.0001
ACC	7937	3411	3232	2224	<0.0001
Ala					
GCU	2747	22745	1112	12733	<0.0001
GCC	17170	8322	8715	5130	<0.0001
Tyr					
UAC	3513	1103	1414	572	<0.0001
His					
n/a					
Gln					
CAG	6890	598	3283	740	<0.0001
Asn					
AAC	4731	1513	1249	808	<0.0001
Lys					

AAG	6873	306	1800	250	<0.0001
Asp					
GAU	3816	7335	1566	3692	<0.0001
Glu					
GAG	9968	679	3774	687	<0.0001
Cys					
n/a					
Arg					
CGU	1790	10760	619	7652	<0.0001
CGC	6550	6000	3188	5083	<0.0001
Gly					
GGU	3520	13950	817	9760	<0.0001

KOG (euKaryotic Orthologous Groups) analyses of the species': *T. trahens*, *F. alba* and *L. tarentolae* were carried out in order to analyse the distribution of different functional categories of each gene. There are four general categories of genes: Cellular processes and signalling, Information storage and processing, Metabolism, and poorly categorised genes. These are then sub-categorised into those defined by letters A through W, Y and Z and are described below in table 6. The results of this KOG analysis are also shown in table 5 and display the percentage of total genes for each category per species. As seen in table 6 one of the species analysed has the highest proportion of its genes within the Cellular Processing and Signalling category with a value of 15.87% for species *F. alba;* this value is assigned to the specific category O, whose roles include posttranslational modification of proteins to ensure correct functioning or ease of movement, protein turnover to ensure efficient translation and chaperones. The species *L. tarentolae,* who's highest proportion of genes are within the poorly categorised genes, with a percentage of 22.69 assigned to the function unknown category S meaning the genes have yet to be annotated as functional genes. *T. trahens* also follows the same trend with the highest percentage of its genes falling under the poorly categorised genes, specifically category S with a value of 16.35%. The lowest distribution of genes for species *F.*

alba is seen also in the Cellular processing and Signalling category; however, is under the subcategory N: Cell motility, at 0.08% of the entire genome encoding genes used in cell motility. *Thecamonas trahens* has a joint lowest gene distribution between subcategories W: Extracellular structure and Y: Nuclear structure which ensures the structures of the cell are functional; with a total of 0.09% each of the whole genes being relevant to these processes. Species *L. tarentolae* also has a joint lowest gene distribution through categories W & Y, with a total of 0.05% of the genes utilised in the Extracellular and Nuclear structure.

Table 6: Results of euKaryotic Orthologous Group analyses (KOG) carried out on species *T. trahens, F. alba* and *L. tarentolae*. Displayed utilising excel as a category percentage of total genes present within the genome for KOG categories A through W, Y and Z. Shown is the groups, their one letter codes and a description of each category used to segregate each gene within the species genome into for the eukaryotic species (Huerta-Cepas et al., 2016).

KOG Category	Percentage of Total Genes/%		
	F. alba	T. trahens	L. tarentolae
1) Cellular Processes and Signalling			
M: Cell wall/membrane/envelope biogenesis	0.7	2.53	1.84
N: Cell Mobility	0.08	0.22	0.09
O: Posttranslational modification, protein turnover, chaperones	15.87	11.57	11.18
T: Signal transduction mechanisms	10.1	14.3	10.59
U: Intracellular trafficking, secretion, and vesicular transport	7.62	6.1	5.67
V: Defence mechanisms	0.47	1.03	0.64
W: Extracellular structures	0.17	0.09	0.05
Y: Nuclear structure	0.14	0.09	0.05
Z: Cytoskeleton	2.62	4.04	3.29
2) Information Storage and Processing			
A: RNA processing and modification	6.96	3.52	4.73
B: Chromatin structure and dynamics	1.75	1.97	0.99
J: Translation, ribosomal structure and biogenesis	9.54	5.11	7.76
K: Transcription	5.44	3.79	2.45
L: Replication, recombination and repair	4.41	3.12	3.69
3) Metabolism			
C: Energy production and conversion	4.63	3.52	4.54
D: Cell cycle control, cell division, chromosome partitioning	2.82	2.49	2.21
E: Amino acid transport and metabolism	3.21	3.59	3.46
F: Nucleotide transport and metabolism	1.62	1.41	1.74
G: Carbohydrate transport and metabolism	3.18	3.36	3.04
H: Coenzyme transport and metabolism	1.7	0.94	1.34
I: Lipid transport and metabolism	4.69	5.04	4.09
P: Inorganic ion transport and metabolism	2.06	3.36	2.14
Q: Secondary metabolites biosynthesis, transport and	1.98	2.44	1.74
4) Poorly Categorized			
R: General function prediction only	-	-	-
S: Function unknown	8.26	16.35	22.69

BLAST analysis of tRNA genes identified from tRNA scan analysis of the genome of *F. alba,* found that the majority of adenosine wobble position tRNA genes identified were present within the transcriptome. There was an exception of one gene tRNA gene with anticodon AAT for Isoleucine, and one with anticodon AGT for Threonine (Table 7). These tRNA transcripts were manually modified in TextWrangler to include a guanosine base at the wobble position of the anticodon; and these transcripts were also run on a BLAST search. Located by this in *F.alba* was the presence of three Threonine tRNA genes, with the manually modified G in their first position. This was observed in three separate threonine tRNA genes, with a manually modified anticodon of GGT, from the original copies within the genome of anticodon AGT.

Table 7: Table showing the BLAST analyses of species *F. alba,* of unmodified and modified tRNA genes, with their original adenosine residues and modified guanosine bases. These were inputted into NCBI's blast to identify any copies within transcripts of the genome, any modified guanosine wobble site tRNA genes found are highlighted in green.

Amino	**tRNA**			
Acid	**Molecule**	**NCBI Code**	**Sequence**	**Transcriptome**
F. alba				
Ala	tRNA$^{Ala}_{AGC}$	SRA:SRR554359.2990242.2	GGGGGCGTAGCTCAAATGGTAGAGCGGCCGCTTAGCATGCGGCAGGcAGGGGGATCGATACCCTC	**Present**
	tRNA$^{Ala}_{GGC}$		GGGGGCGTAGCTCAAATGGTAGAGCGGCCGCTT**G**GCATGCGGCAGGcAGGGGGATCGATACCCTC	Absent
	tRNA$^{Arg}_{AGC}$	SRA:SRR554361.2180640.2	GGCCGCGTAGCTCAATTGGAtAGAGCACCTGACTACGGATCAGGAGGtTGTGTGTTCGAGCCGCAT	**Present**
	tRNA$^{Arg}_{GGC}$		GGCCGCGTAGCTCAATTGGAtAGAGCACCTGACT**G**CGGATCAGGAGGtTGTGTGTTCGAGCCGCAT	Absent
	tRNA$^{Arg}_{ACG}$	SRA:SRR554361.582077.2	GGCCGTGTAGCTCAATTGGAtAGAGCACCTGACTACGGATCAGGAGGtTGTGTGTTCGAGCCGCAT	**Present**

Arg	tRNA$^{Arg}_{GCG}$		GGCCGTGTAGCTCAATTGGAtAGAGCACCTGACT**G**CGGATCAGGAGGtTGTGTGTTCGAGCCGCAT	Absent
Ile	tRNA$^{Ile}_{AAT}$		GGCCCCATAGCTCAGTTGGTtAGTGGGTCGTGCTAATAACTTGGCCGtCGTCAGTTCGGTGCTGGCTG	Absent
	tRNA$^{Ile}_{GAT}$		GGCCCCATAGCTCAGTTGGTtAGTGGGTCGTGC**G**ATAACTTGGCCGtCGTCAGTTCGGTGCTGGCTG	Absent
	tRNA$^{Ile}_{AAT}$	SRA:SRR554359.2880346.1	GGCCCCATAGCTCAGTTGGTtAGAGCGTCGTGCTAATAACGCGAAGGtCGTCAGTTCGATCCTGGCT	**Present - Differs from**
	tRNA$^{Ile}_{GAT}$		GGCCCCATAGCTCAGTTGGTtAGAGCGTCGTGCT**G**ATAACGCGAAGGtCGTCAGTTCGATCCTGGCT	Absent
	tRNA$^{Ile}_{AAT}$	SRA:SRR554359.2880346.1	GGCCCCATAGCTCAGTTGGTtAGAGCGTCGTGCTAATAACGCGAAGGtCGTCAGTTCGATCCTGGCT	**Present**
	tRNA$^{Ile}_{GAT}$		GGCCCCATAGCTCAGTTGGTtAGAGCGTCGTGCT**G**ATAACGCGAAGGtCGTCAGTTCGATCCTGGCT	Absent
Leu	tRNA$^{Leu}_{AAG}$	SRA:SRR554361.9704391.2	GGTTGGATGGCCGAGTGGTtAAGGCGCCAGTTTAAGGCACTGGTGGGAAACCGCGTGGGTTCGAG	**Present**
	tRNA$^{Leu}_{GAG}$		GGTTGGATGGCCGAGTGGTtAAGGCGCCAGTTT**G**AGGCACTGGTGGGAAACCGCGTGGGTTCGAG	Absent
Pro	tRNA$^{Pro}_{AGG}$	SRA:SRR554361.5742821.1	GGGAAATTAGTCTAGTGGTATGATTCTCGCTTAGGGTGCGAGAGGtCCCGGGTTCGATTCCCGGATT	**Present**
	tRNA$^{Pro}_{GGG}$		GGGAAATTAGTCTAGTGGTATGATTCTCGCTT**G**GGGTGCGAGAGGtCCCGGGTTCGATTCCCGGAT	Absent
	tRNA$^{Pro}_{AGG}$	SRA:SRR554359.9811034.1	GGGAGATTAGTCTAGTGGTATGATTCTCGCTTAGGGTGCGAGAGGtCCCGGGTTCGATTCCCGGAT	**Present**
	tRNA$^{Pro}_{GGG}$		GGGAGATTAGTCTAGTGGTATGATTCTCGCTT**G**GGGTGCGAGAGGtCCCGGGTTCGATTCCCGGAT	Absent
Ser				
	tRNA$^{Ser}_{AGA}$	SRA:SRR554361.9267802.1	ATCACCGTGTCCGAGTGGTtAAGGAGTCCGATTAGAAATCGGATGGGCTCTGCCCGCGTAGGTTCAA	**Present**
	tRNA$^{Ser}_{GGA}$		ATCACCGTGTCCGAGTGGTtAAGGAGTCCGATT**G**GAAATCGGATGGGCTCTGCCCGCGTAGGTTCA	Absent

	tRNA$^{Thr}_{AGT}$	SRA:SRR554361.2483469.1	TAGCTCAGTGGTAGAGCGCCAGTCTAGTAAACTGGAGGtCGGGTGTTCGAtccnnnnnnnnnncccatT	**Present**
	tRNA$^{Thr}_{GGT}$	SRA:SRR554361.536567.1	TAGCTCAGTGGTAGAGCGCCAGTCT**G**GTAAACTGGAGGtCGGGTGTTCGAtccnnnnnnnnnncccatT	**Present**
	tRNA$^{Thr}_{AGT}$		GCTTGTTTAGCTCAGTGGTAGAGCGCCAGTCTAGTAAACTGGAGGtCGGGTGTTCGATCCACCCAA	Absent
	tRNA$^{Thr}_{GGT}$	SRA:SRR554361.536567.1	GCTTGTTTAGCTCAGTGGTAGAGCGCCAGTCT**G**GTAAACTGGAGGtCGGGTGTTCGATCCACCCAA	**Present**
	tRNA$^{Thr}_{AGT}$	SRA:SRR554361.2483469.1	GCTCGTTTAGCTCAGTGGTAGAGCGCCAGTCTAGTAAACTGGAGGtCGGGTGTTCGATCCACCCAA	**Present**
Thr	tRNA$^{Thr}_{GGT}$	SRA:SRR554361.536567.1	GCTCGTTTAGCTCAGTGGTAGAGCGCCAGTCT**G**GTAAACTGGAGGtCGGGTGTTCGATCCACCCAA	**Present**
	tRNA$^{Val}_{AAC}$	SRA:SRR554361.9583106.1	GTCCGAATGATGTAGATGGTtATCATATCTGTCTAACACACAGAATGtCCCAGGTTCGAGTCCTGGTT	**Present**
Val	tRNA$^{Val}_{GAC}$		GTCCGAATGATGTAGATGGTtATCATATCTGTCT**G**ACACACAGAATGtCCCAGGTTCGAGTCCTGGTT	Absent

Screening of Diphoda species for potential tRNA deamination –

Once the direct evidence for deamination of the TAPSILVR tRNA genes was established through initial investigations, and through the analysis of the species *T. trahens*, this process was shown to have evolved outside of the opisthokont grouping. Further research was then undertaken into groups falling on the other side of the eukaryotic root to the opisthokonts, in order to further establish the origin of this process. In all well supported phylogenetic representations of the origins of eukaryotic life, the excavate group falls on the opposite side of the root to the opisthokonts (Brown et al., 2018; Derelle et al., 2015). Selected for further investigation of the deamination were species *Leishmania tarentolae*, as a representative of the Excavata grouping. Also selected were archaeplastid species; *Chondrus crispus* and *Chlamydomonas reinhardtii*; another evolutionary group containing algal autotrophs. These autotrophs are placed alongside the excavates on the opposite side of the root to the opisthokont group. The final species analysed was *Bigelowiella natans*, a rhizarian species which is nestled within the SAR (stramenopiles, alveolates and Rhizaria) supergroup, again within a well established placement opposite the Opisthokont group. These species were chosen from multiple groups on the opposite side of the eukaryotic root; in order to limit the possibility of the deamination of tRNA being as a result of convergent evolution. The process could easily have evolved on two occasions within the opisthokonts and excavates; however, it is extremely unlikely to have independently evolved in opisthokonts, archaeplastids, rhizarians and also excavates.

NC PLOTS –

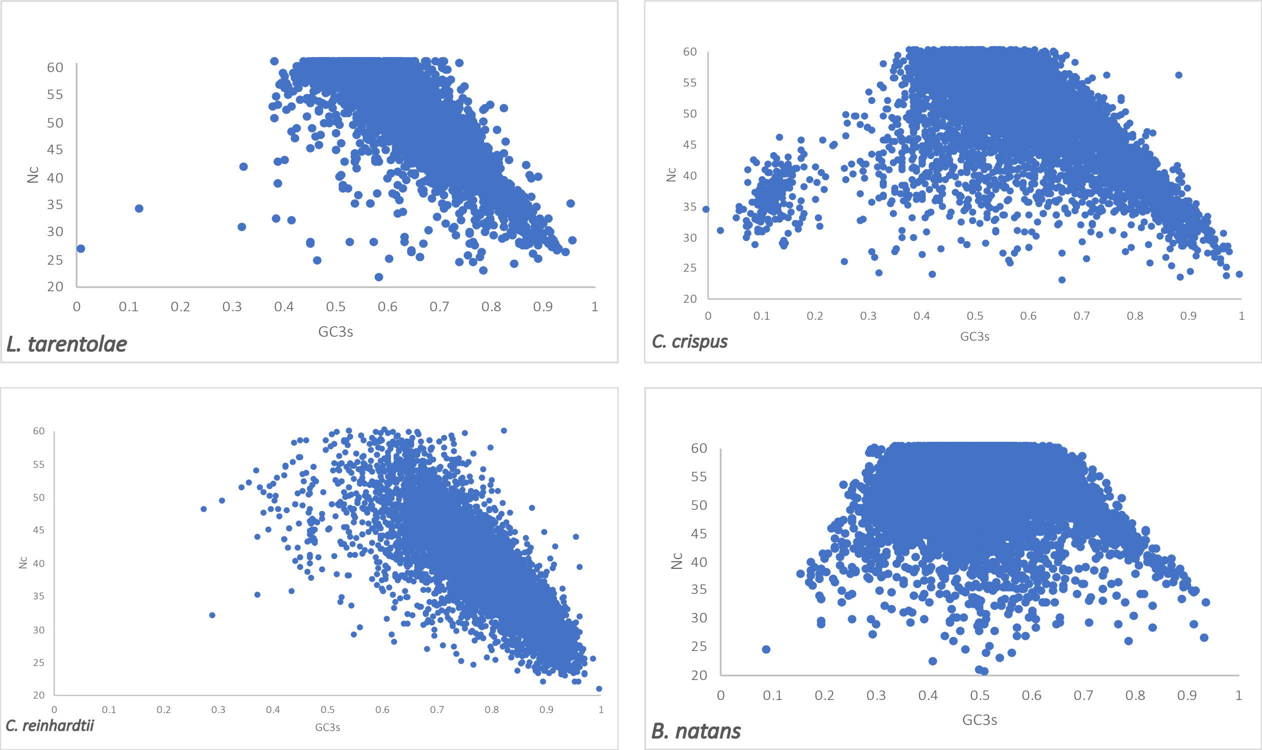


Figure 9: *Nc* plots of the data points for the number of effective codons (*Nc*) and GC content of the third codon position (GC3s) for each gene in the genome of the species under investigation of tRNA deamination. This includes excavate species *Leishmania tarentolae*, archaeplastid species *Chondrus crispus* and *Chlamydomonas reinhardtii* and final species under investigation, a representative of the SAR grouping, the rhizarian *Bigelowiella natans*.

Analyses of the number of effective codons were undertaken for all four species above and produced *Nc* plots, as with species *F. alba* and *T. trahens*, to allow comparison (Figure 9). *Leishmania tarentolae* also follows the trend of a bias towards GC ending codons; with only a few outlying genes falling outside of the strong positive correlation between *Nc* and GC3s, this left the plot visually similar to that of species *T. trahens*. Of the two species of archaeplastids *C. crispus* has the least conserved distribution, when regarding the increased bias towards GC ending codons observed in other species which are shown to deaminate their tRNA molecules.

To further the image provided by this research, these genes would have been investigated further to determine whether the large group of outlying genes are organelle genes; derived from mitochondria or chloroplasts. *C. crispus' Nc*/GC3s points follow more of a normal distribution as opposed to only a strong positive correlation between number of effective codons and GC composition of the third position. *C. reinhardtii* had an expected distribution of points; displayed is a GC- bias with GC3s increasing as the bias increases. This was observed with only a few genes straying from the trend, as is seen with other species analysed which are proven to deaminate tRNAs. The representative from the SAR group, *Bigelowiella natans,* appears to follow a normal distribution of points alongside *C. crispus;* with the highly biased genes showing either high GC content or high AT synonymous nucleotide content. Whilst there are some outlying genes, there are much larger numbers of these spread in a more normally distributed manner, conveying no bias towards codons ending either GC or AT; merely an even spread.

SMALT ANALYSIS –

Results were produced from the SMALT analysis of unmodified tRNA genes containing adenosine bases which were identified by tRNA scan; and manually modified tRNA genes containing guanosine bases at their wobble position, as opposed to the identified adenosine wobble position tRNA genes. The results of this analysis on species *L. tarentolae* identified all of the modified and unmodified tRNA gene transcripts, containing guanosine and adenosine respectively, for amino acids: alanine, arginine isoleucine, leucine, proline, serine threonine and valine were present (Appendices 3). Analysis of *C. crispus'* tRNA and altered tRNA genes mapped onto the merged transcriptomes of *C. crispus* (Table 8), located the majority of tRNA genes within the transcriptome; with the exception of seven tRNA genes. These included: one unmodified alanine tRNA genes with adenosine at their wobble site and one modified alanine tRNA gene with guanosine at its wobble site. Also not found within the transcriptomes were: one unmodified tRNA gene and two modified tRNA genes for arginine, and finally unmodified tRNA genes with adenosine at the wobble position for amino acids serine and threonine. Of the tRNA genes which were located in the *C. crispus* transcriptome there were notably: a modified threonine tRNA with a G at its wobble site present within the transcriptome, an unmodified valine tRNA with a transcript identified towards the bottom end of table 8. Also present were unmodified tRNA genes for amino acids: alanine, arginine and proline and modified tRNAs for serine and threonine. Identical analyses were carried out on *C. reinhardtii* (Appendices table 1), in which all tRNA genes bar one, for a modified valine tRNA gene containing guanosine at its wobble site, were present. The most notable results for this analysis were seven threonine tRNAs: two unmodified and five modified tRNAs with adenosine at their wobble position; and also seven tRNA genes which consisted of six arginine tRNA genes, three with unmodified and three modified with guanosine at their wobble site. The final gene was a modified valine tRNA gene. The species *B. natans* (Appendices table 2), had the most surprising results, with all bar four tRNA genes present within the transcriptome: these four included two tRNA genes for isoleucine and two for proline, one modified and one unmodified sequence in each case.

Table 8: Summary table showing the presence of unmodified and modified tRNA gene sequences for the TAPSILVR amino acids as determined by SMALT analysis. The species name is displayed with a Y representing instances where a copy was found within transcripts for that species and an N representing instances where no copies were found.

			Species			
			L. tarentolae	C. crispus	C. reinhardtii	B. natans
Amino Acid Code	T	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	A	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	P	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	S	Mod	Y	Y	Y	Y
		UnMod	Y	N	Y	Y
	I	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	L	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	V	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y
	R	Mod	Y	Y	Y	Y
		UnMod	Y	Y	Y	Y

tRNA SCREEN ANALYSIS –

The presence of tRNA genes with adenosine at their wobble position within the two fold degenerate amino acids: asparagine, aspartic acid, cysteine, histidine, phenylalanine, serine (for its ACU anticodon tRNA) and tyrosine was investigated. The results for the species: *T. trahens, F. alba, C. crispus, C. reinhardtii* investigated during this research project and also additional species: *Dictyostelium discoideum (Amoebozoa), Entamoeba histolytica (Amoebazoa), Entamoeba invadens (Amoebazoa), Capsaspora owczarzaki (Opisthokonta), Monosiga brevicollis (Opisthokonta), Salpingeoca rosetta (Opisthokonta)* and *Galderia sulphuraria (Archaeplastida)* are displayed in table 9. The data for these additional species was provided by Dr Martin Carr during personal communication; and is represented by a number of tRNA genes found for each amino acid in each of the species analysed. As seen in the table, not one tRNA gene with an adenosine at its wobble position was located for any of the two fold degenerate amino acids analysed, in either this research project or the species analysed by Dr Carr.

Table 9: Table showing the number of two fold degenerate amino acids with adenosine at their wobble position for species: *F. alba, T. trahens, C. crispus, C. reinhardtii*, which were investigated during the course of this research project. Also included were species: *Dictyostelium discoideum, Entamoeba histolytica, Entamoeba invadens, Capsaspora owczarzaki, Monosiga brevicollis, Salpingoeca rosetta* and *Galdieria sulphuraria* from Dr Martin Carr (personal communication); showing indirect evidence for the deamination by displaying the lack of adenosine at the wobble position of two fold degenerate amino acids.

	Species	Asn	Asp	Cys	His	Phe	Ser (ACU)	Tyr
Opimoda	**Amoebozoa**							
	Dictyostelium discoideum	0	0	0	0	0	0	0
	Entamoeba histolytica	0	0	0	0	0	0	0
	Entamoeba invadens	0	0	0	0	0	0	0
	Apusozoa							
	Thecamonas trahens	0	0	0	0	0	0	0
	Opisthokonta							
	Capsaspora owczarzaki	0	0	0	0	0	0	0
	Fonticula alba	0	0	0	0	0	0	0
	Monosiga brevicollis	0	0	0	0	0	0	0
	Salpingoeca rosetta	0	0	0	0	0	0	0

DIPHODA	**Archaeplastida**							
	Chlamydomonas reinhardtii	0	0	0	0	0	0	0
	Chondrus crispus	0	0	0	0	0	0	0
	Galdieria sulphuraria	0	0	0	0	0	0	0
		Asn	**Asp**	**Cys**	**His**	**Phe**	**Ser (ACU)**	**Tyr**

Table 10: Table displaying the tRNA genes identified by the program tRNAscanSE-2.0 within species *T. trahens*. Also displayed is the tRNA anticodon, the number of these genes present within the transcriptome, and whether or not this tRNA gene is a match for the optimal codon. The final column describes where the tRNA gene and optimal codon are a match only once the tRNA is deaminated.

AA	tRNA	No. of Genes	Match	Description
Ala	AGC	30	Y	The deaminated tRNA binds
	CGC	6	N	
	TGC	4	N	
Arg	ACG	35	Y	The deaminated tRNA binds
	CCG	3	N	
	CCT	1	N	
	TCT	3	N	
Asn	GTT	15	Y	
Asp	GTC	30	Y	
Cys	GCA	11	Y	
Gln	CTG	17	Y	
	TTG	4	N	
Glu	CTC	31	Y	
	TTC	5	N	
Gly	CCC	4	N	
	GCC	32	Y	
	TCC	1	N	
His	GTG	17	Y	
Ile	AAT	28	Y	The deaminated tRNA binds
	GAT	1	Y	

	TAT	1	N	
iMet	CAT	8		
Leu	AAG	13	N	
	CAA	1	N	
	CAG	11	Y	
	TAA	4	N	
	TAG	3	N	
Lys	CTT	38	Y	
	TTT	6	N	
Met	CAT	8	Y	
Phe	GAA	14	Y	
Pro	AGG	15	Y	The deaminated tRNA binds
	CGG	10	N	
	TGG	4	N	
Ser	AGA	21	Y	The deaminated tRNA binds
	CGA	6	N	
	GCT	8	Y	
	TGA	2	N	
Thr	AGT	23	Y	
	CGT	1	N	
	TGT	3	N	
Trp	CCA	11	Y	
Tyr	ATA	1	Y	The deaminated tRNA binds
	GTA	14	Y	
Val	AAC	25	Y	The deaminated tRNA binds
	CAC	6	N	
	TAC	2	N	

STRENGTH OF SELECTION ON CODON USAGE IN *F. ALBA* AND *T. TRAHENS* –

Odds ratios were calculated for all species investigated during the course of this research project. These odds ratios were weighted according to the proportion of high bias optimal codons, to all other codons, for each of the two fold degenerate amino acid species: alanine, arginine, cysteine, histidine, glutamine, glutamic acid, lysine, phenylalanine and tyrosine. This was as a result of the Southworth et al., (*2018*) paper which examined 3 holozoan species and showed that the strength of selection (S) varied between three species investigates; with *C. owczarzaki* showing a much higher S value than the choanoflagellate species. This research therefore examines the selection between its two opimoda species under investigation: *F. alba* and *T. trahens*. These S values then had their natural logs taken, allowing comparison of all, and can be seen in Table 11. Seen in the results of these analyses for *F. alba*, these weighted natural logs ranged from values of 1.036 to 6.438 for amino acids histidine and lysine respectively; and all values produce an average of 3.1. Values for species *T. trahens* ranged from 0.916 within glutamic acid and up to 1.694 for aspartic acid with an average of 1.1.

Table 11: Table displaying the odds ratio for each of the nine degenerate amino acids within both species *F. alba* and *T. trahens*. Also displayed is the weighted odds ratio taking into account the number of high bias category codons in total; and finally the natural logarithm was taken for each of the weighted averages, allowing comparison between amino acids.

	F. alba			*T. trahens*		
	Odds Ratio	Weighted Avg.	Ln(Odds Ratio)	Odds Ratio	Weighted Avg.	Ln(Odds Ratio)
Phe	22.10	30.05	3.40286266	2.19	2.33	0.845868268
Tyr	11.64	7.91	2.06812778	5.78	4.45	1.492904096
His	5.14	2.82	1.03673688	3.39	2.14	0.760805829
Gln	76.85	61.48	4.11871192	3.07	2.46	0.900161350
Asn	16.73	17.4	2.85647021	4.74	4.12	1.415853163
Lys	466.36	624.93	6.43763964	4.15	5.35	1.677096561
Asp	3.29	4.61	1.52822786	3.13	5.44	1.693779061
Glu	88.51	131.88	4.88189242	1.66	2.50	0.916290732
Cys	13.73	4.67	1.54115907	3.68	1.21	0.190620360
	Weighted Avg.		3.09686983			1.099264380
			3.1			1.1

DISCUSSION:

CODON USAGE IN EUKARYOTES –

The results of this research project provided support for the theory that codon usage is very well conserved across the eukaryotic supergroup. The results of this investigation have identified direct evidence of the deamination of tRNA molecules, within species which fall across the root of the eukaryotic tree of life from the opisthokont group; in which deamination of tRNAs was already established by Southworth et al., (*2018*). The species in these analyses were hence chosen as they fall across the root from the opisthokont group, and presence of deamination in species *F. alba* and *T. trahens* pushes the estimate of evolution to the origin of the Opimoda. The Opimoda is close to the base of the eukaryotic origin and further establishes the evolutionary extent of this process. The additional species utilised: *L. tarentolae, C. crispus, C. reinhardtii* and *B. natans*, are all outside of the opisthokont group. Evidence of the presence of the deamination of tRNA molecules in any of these species, indicates that the process was also present in the last common ancestor of these species. The process being present within the last common ancestor to all of these species means it can therefore be pushed further back, from ancestral to just opisthokonts, to be classed as ancestral to the eukaryotic lineage (Figure 10). Were this not the case, an unexpected level of evolution would have to have taken place in order to provide this process within both branches of the eukaryotic lineages. That would be required due to the fact that a minimum of 4 individual cases of evolution of the deamination process would have had to occur, in order to have been identified within this research project. This is supported by the fact that until recently bacterial species had only been shown to deaminate one amino acids tRNA molecule, arginine; however not any of the other TAPSILVR amino acids. Rafels-Ybern., et al (*2018*) disputed this by showing some bacterial species also contained tRNA genes for other amino acids such as: isoleucine, serine and threonine. One possible mode of evolutionary explanation for the eukaryotic supergroup, is that they have evolved from a symbiosis between bacterial and archaeal species. Support for the evolution of tRNA deamination occurring within the eukaryotic line is taken from the fact that bacteria only deaminate select tRNAs. Archaea do not deaminate any of the tRNA molecules, indicating that at the base of the eukaryotes a large expansion must have taken place (Marck & Grosjean, 2002). This expansion would be required in order to provide such a large range of tRNA deamination, of all eight TAPSILVR amino acid tRNAs, from a symbiosis where only four amino acids tRNAs are deaminated. The expansion of deamination of tRNAs can also be assumed to be an important aspect of the evolution of the ancestral eukaryotic translational machinery which has been retained across the tree. Were this not the case and alternatively bacteria had lost this complexity, this would have had to occur in large numbers of bacterial lineages independently in order to account for this; which provides a much less parsimonious explanation for these events.

The investigations into strength of selection could potentially give an insight into the effective population size of the species under investigation, the results shown in table 11 display that species *F. alba* has a higher strength of selection than *T. trahens*. This could potentially indicate that *F. alba* has a higher population size and hence this affects the strength of selection; however, there are other factors which may affect this to be considered. Recombination may be an affecting factor; this is because the more recombination that occurs the more efficient natural selection is. There is a potential that *T. trahens* undergoes less sexual reproduction, and as a result less recombination occurs, which leads to a lower S value; and so these values do not necessarily mean that *T. trahens* has a lower effective population size.

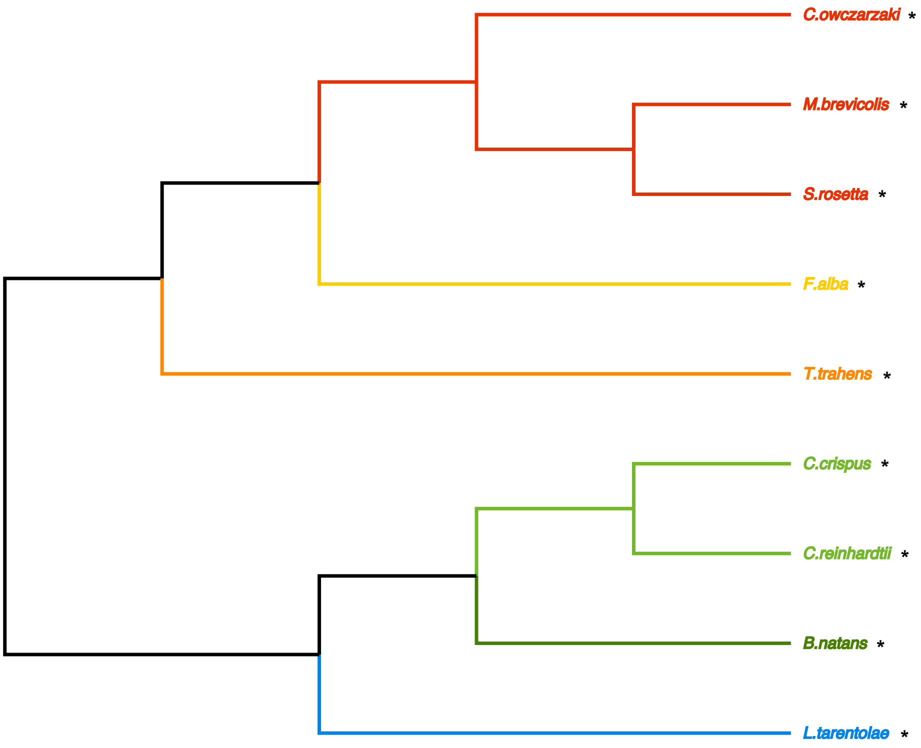


Figure 10: Representative cladogram displaying the relationships between the species under investigation, and species previously investigated within Southworth et al., *(2018)*, based upon phylogenies taken from; Derelle et al., *(2015)*, Brown et al., *(2018)*, Keeling & Burki, *(2019)*. Different supergroups are represented with differing colours: holozoans – red, opisthokont – yellow, apusozoan – orange, archaeplastids – light green, rhizarian – dark green, excavates – blue. Species where deamination of tRNAs has now been identified are denoted by a black asterisk.

The use of deamination of tRNAs within translation of proteins appears to be of benefit to the host organism, this could have positive effects such as allowing the host to have a smaller, more compact genome. This could have been through the reduction of multiple genes for the multiple tRNAs required for translation, and variation of tRNA types within their genomes;

allowing unused genetic material to be removed over time through selection. This reduction is as a result of a smaller repertoire of tRNA genes within the host genome, as there is a lack of tRNA genes encoding tRNA molecules with guanosine at the wobble position. When inosine is present in tRNA molecules its ability to wobble bind with not only cytosine, but through its structural similarity to guanosine bases adenosine and uracil also; allowing inosine to do the job of guanosine tRNAs, hence rendering them redundant within the genome.

The absence of deamination within the tRNA genes of all of the two-fold degenerate amino acids, serves as an indication that deamination of tRNA from two fold degenerate amino acids is strongly deleterious. From a deamination point of view the conversion of adenosine to inosine leads to an increase in the binding ability of codons, if two fold degenerate amino acid tRNAs were able to have their tRNA genes deaminated then these tRNAs would be capable of binding codons from two different amino acids. The binding of tRNAs to three amino acid codons (A-, C- and U- ending) would hence break down the non-overlapping nature of the genetic code; and hence problems within three or more fold degenerate amino acids. This could provide an explanation as to why, if deamination of two fold amino acids is not observed. Deamination of two fold degenerate amino acids would cause the binding of tRNAs to three amino acid codons (A-, C- and U- ending); hence breaking down the non-overlapping nature of the genetic code. This research is supported by a lack of any two fold degenerate amino acid tRNAs identified with adenosine at their wobble position. Further, indirect evidence is provided by the amino acid serine, serine is a six fold amino acid which presents itself in two “blocks”. One of these blocks acts as a four fold degenerate amino acid and the other acts as a two fold degenerate amino acid, however both are from very different regions of the genetic code.

Within these blocks the four fold degenerate serine codons are seen to deaminate their A34 tRNA genes, however the two fold degenerate serine codons don't undergo deamination. This further supports the proposed deleterious nature of deamination within two fold degenerate amino acid tRNAs and also that this deamination of tRNAs is non-specific. Table 6 displays supporting evidence for this, where the analyses failed to identify any tRNA genes with adenosine at their first anticodon position, within the TAPSILVR amino acids; this was for species on both sides of the opimoda/diphoda split of the eukaryotic lineage. The absence of deamination of A34 tRNA genes, compared to their abundance in three to six fold degenerate amino acid tRNAs, suggests they are deleterious. then there is less reason to avoid the adenosine wobble position tRNAs within these two fold genes. A plausible deleterious effect is the non-specific deamination of adenosine, resulting in a lack of specificity in the tRNA molecules. These deaminated tRNAs could code for more than one of the amino acids, and hence breaking down the genetic code resulting in the incorporation of the incorrect amino acid.

BLAST analyses of tRNA genes for species *F. alba* determined that for this species most of the tRNA genes for TAPSILVR amino acids were present within transcripts of the genome, with an adenosine residue at their wobble site. Only three of the manually modified guanosine containing tRNA genes were located within transcripts of the species. This successful identification of the majority of adenosine tRNA genes within the species *F. alba,* and the lack of most G containing tRNA genes using NCBI Blast, determines that deamination of these adenosine wobble position tRNAs must be taking place within the host. The deamination of A34 tRNAs allows binding of the inosine base to the cytosine ending optimal codons. Were this not the case translation of proteins would be hugely hindered by the fact that the adenosine wobble site tRNAs are unable to bind the GC ending optimal codons. The presence of three of the manually modified A34 tRNA genes with guanosine substituted at their first anticodon position were detected which confirms that the deamination of adenosine occurs in *F. alba.* The genome can therefore be more streamlined, as less of these guanosine first position tRNAs are present, giving a more compact, efficient genome. The fact that the complementary C-ending codons make up the majority of the optimal codons, provides evidence for the deamination of tRNA genes being beneficial to the host species.

The optimal codons produced for each species during these analyses are consistent with the deamination of TAPSILVR amino acid tRNAs; and hence this allows this research to fulfil its aim of identifying an estimate of evolution of tRNA deamination to ancestral to the opisthokonts. It is then pushed back further, to the base of eukaryotic evolution, somewhere between 1.4 to 1.8 billion years ago billion years ago (Kumar, Stecher, Suleski, & Hedges, 2017; Nei, Xu, & Glazko, 2001; Parfrey, Lahr, Knoll, & Katz, 2011) This means that ancestral codon usage has been retained by many unicellular eukaryotes for approximately 1.5 billion years. If the species were to change their codon usage away from this trend towards C-ending codons, then these codons would no longer be complementary to the major tRNA genes within the organism, and as such selection would resist these changes. If the tRNA genes were to change so that the most abundant tRNA genes were no longer those with adenosine at their wobble position, then these would not be able to bind to the high frequency codons. These potential changes to the translational process may not evolve unless selection pressures weaken and therefore and mutation pressure and genetic drift become more influential in codon usage.

SMALT analysis of the modified and unmodified tRNA genes for species *C. crispus, C. reinhardtii* and *B. natans,* revealed direct evidence for the deamination of tRNA molecules with adenosine at their wobble position. This direct evidence is in the form of the presence of both the adenosine and also manually modified tRNA genes containing guanosine, within the transcriptomes of all three of the species tested. Due to their positioning within the eukaryotic supergroup on the opposite sides of the root to the opisthokonts, this direct evidence of deamination for these species further supports the theory that the process is ancestral to the eukaryotes; this data provides new insights into the translational processes of the ancestral eukaryote. The evidence is consistent with the tRNAs of all eight TAPSILVR amino acids being deaminated. Therefore, the process appears to have evolved in the eukaryotic stem-group – between eukaryogenesis and the radiation of the extant eukaryotes.

The optimal codons as identified by both correspondence analyses and also produced by expression data of the high and low expressed genes determine that there is a requirement for the deamination of tRNA within the species analysed. This is due to the optimal codons for the species all being C- ending; in order to bind these C- ending codons the tRNA genes must either contain an inosine base at their wobble site or a guanosine base. The absence of the guanosine wobble position tRNA genes was established during this research, and hence to bind the C- ending optimal codons the most abundant tRNA genes with adenosine at their wobble position must be deaminated.

tRNAscanSE-2.0 analyses of the tRNA genes present within these species provides indirect evidence for the deamination of tRNA molecules within the species; as the TAPSILVR amino acids always have the highest gene copy numbers of all of the tRNA genes within the genome, these are the genes with adenosine at their wobble position. This highlights the fact that deamination is occurring outside of the holozoan group and also the Opisthokonta.

FUTURE WORK:

In the future, to extend this research project further, bioinformatic analyses of multiple other species of the eukaryotic lineage from all representative groups is required; in order to enhance the current picture of the evolution of tRNA modification. A larger screen could also allow any instances of loss of deamination to be identified or alternatively discover whether the mechanism has been retained within all eukaryotic groups. Also to be analysed would be the genomes of some bacterial species from the prokaryotic side of the evolutionary tree; in order to investigate the extent of deamination of adenosine tRNA molecules. An ideal group to investigate would be the alphaproteobacterial, this is because the classical mitochondrion within todays eukaryotes appears to be derived from the mitochondria observed within these species (Andersson et al., 1998; Gray, 2012; Schmitz-Esser et al., 2010). Screening a broad range of these alphaproteobacterial would allow the extent of bacterial deamination to be further investigated. The original position of the evolution of deamination of these select amino acids could be determined, which could allow the exact origin of the deamination process to be pin pointed. It would also be of scientific value to investigate any loss of the deamination process within lineages of eukaryotes. The tRNA genes for all two fold degenerate amino acids would have been investigated, in order to confirm that adenosine is never present at the wobble site of these amino acid tRNAs. The role of the deamination enzymes involved in tRNA deamination in eukaryotes would also be an area to pursue in further research. The determination of the extent of activity of these genes through knock out experiments could be useful in investigating the effects of a lack of deamination on growth rate; and whether this leads to a decrease through impaired protein translation. The frequency of c ending codons for the TAPSILVR amino acids should be investigated in order to determine whether these are significantly enriched within high bias genes when compared with low bias genes and also between domains and non-domains of genes. The final area requiring development would be to investigate whether deamination is present within species which do not have the C- ending optimal codons which are observed within the species involved in this research project. This would allow investigation into whether or not the presence of a bias towards AT- ending codons affects the deamination of tRNA genes at all. Deep sequencing of tRNA molecules could also provide further insight by determining what proportion of the total number of tRNA molecules with A34 undergo deamination, as this could affect the resulting most abundant tRNA molecules.

A potential weakness of this research was that the time limit imposed didn't allow for an investigation into the effects of mutation pressure on the high frequency codons of species; and how these can ensure the increase in translational efficiency through the binding of deaminated tRNAs. This area requires further work in order to understand if mutation pressure does play a role within the bias of GC3s values for the species under investigation.

BIBLIOGRAPHY:

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., … Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Akashi, H. (1994). Synonymous Codon Usage. *Genetics Society of America*, *136*, 927–935.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., … Kurland, C. G. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, *396*(6707), 133–140. https://doi.org/10.1038/24094

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. https://doi.org/10.1093/gbe/evy014

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*, *105*(43), 16641–16646. https://doi.org/10.1073/pnas.0801667105

Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., … Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(7), E693–E699. https://doi.org/10.1073/pnas.1420657112

Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology*, *4*(9), 1476–1481. https://doi.org/10.1101/cshperspect.a011403

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., … Bork, P. (2016). EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293. https://doi.org/10.1093/nar/gkv1248

Ikemura, T. (1981). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of Molecular Biology*, *151*(3), 389–409. https://doi.org/10.1016/0022-2836(81)90003-6

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*, *2*, 13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335

Kliman, R. M., & Hey, J. (1994). The effects of mutation and natural selection on codon bias in the genes of drosophila. *Genetics*, *137*(4), 1049–1056.

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Lake, J. A. (2015). Eukaryotic origins. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 1–5. https://doi.org/10.1098/rstb.2014.0321

Lowe, T. M., & Chan, P. P. (2016). tRNAscan-SE On-line: Search and Contextual Analysis of Transfer RNA Genes. Retrieved from http://lowelab.ucsc.edu/tRNAscan-SE/

Maraia, R. J., & Arimbasseri, A. G. (2017). Factors that shape eukaryotic tRNAomes: Processing, modification and anticodon-codon use. *Biomolecules*, *7*(4)*,* 26*.* https://doi.org/10.3390/biom7010026

Marck, C., & Grosjean, H. (2002). tRNomics: Analysis of tRNA genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. *Rna*, *8*(10), 1189–1232. https://doi.org/10.1017/S1355838202022021

McInerney, J. O. (1997). Codon usage patterns in Trichomonas vaginalis. *European Journal of Protistology*, *33*(3), 266–273. https://doi.org/10.1016/S0932-4739(97)80004-1

Nei, M., Xu, P., & Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(5), 2497–2502. https://doi.org/10.1073/pnas.051611498

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13624–13629. https://doi.org/10.1073/pnas.1110633108

Peden, J. F. (1999). Analysis of Codon Usage, PhD Thesis, University of Nottingham.

Rafels-Ybern, À., Torres, A. G., Camacho, N., Herencia-Ropero, A., Roura Frigolé, H., Wulff, T. F., … Ribas de Pouplana, L. (2019). The Expansion of Inosine at the Wobble Position of tRNAs, and Its Role in the Evolution of Proteomes. *Molecular Biology and Evolution*, *36*(4), 650–662. https://doi.org/10.1093/molbev/msy245

Rafels-Ybern, À., Torres, A. G., Grau-Bove, X., Ruiz-Trillo, I., & Ribas de Pouplana, L. (2018). Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. *RNA Biology*, *15*(4–5), 500–507.

https://doi.org/10.1080/15476286.2017.1358348

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin., V., … Ye, J. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *37*(SUPPL. 1), 5–15. https://doi.org/10.1093/nar/gkn741

Schmitz-Esser, S., Tischler, P., Arnold, R., Montanaro, J., Wagner, M., & Rattei, T. (2010). The genome of the amoeba symbiont ``Candidatus Amoebophilus asiaticus'' reveals common mechanisms for host cell interaction among amoeba-associated bacteria. *J Bacteriol*, *192*(4) 1045-1047. https://doi.org/10.1128/JB.01379-09

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. https://doi.org/10.1093/molbev/msy157

Su, A. A. H., & Randau, L. (2011). REVIEW A to I and C to U Editing within Transfer RNAs, *76*(8), 932–937.

Torres, A. G., Piñeyro, D., Filonava, L., Stracker, T. H., Batlle, E., & Ribas De Pouplana, L. (2014). A-to-I editing on tRNAs: Biochemical, biological and evolutionary implications. *FEBS Letters*, *588(*23*),* 4279-4286*.* Federation of European Biochemical Societies. https://doi.org/10.1016/j.febslet.2014.09.025

Williams, T. A., Foster, P. G., Cox, C. J., & Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*, *504*(7479), 231–236. https://doi.org/10.1038/nature12779

Wright, F. (1990). The “effective number of codons” used in a gene. *Gene*, *87*(1), 23–29.

APPENDICES:

Table 1: Table showing the output of SMALT analysis of unmodified and modified tRNA gene sequences for species *C. reinhardtii*, with adenosine and their wobble site and manually altered to include guanosine at their first anticodon position respectively. Displayed are the read numbers also for each tRNA gene; this is a measure of the number of times that gene sequence was identified within the transcriptomes, and hence mapped onto the consensus file.

CONTIG	LENGTH	PRESENCE
AlaMod_1	96	Y
AlaUnMod_1	96	Y
AlaMod_2	96	Y
AlaUnMod_2	96	Y
AlaMod_3	97	Y
AlaUnMod_3	97	Y
AlaMod_4	96	Y
AlaUnMod_4	96	Y
AlaMod_5	96	Y
AlaUnMod_5	96	Y
AlaMod_6	96	Y
AlaUnMod_6	96	Y
AlaMod_7	95	Y
AlaUnMod_7	95	Y
AlaMod_8	95	Y
AlaUnMod_8	95	Y
ArgMod_1	73	Y
ArgUnMod_1	73	Y
ArgMod_2	73	Y
ArgUnMod_2	73	Y
ArgMod_3	120	Y
ArgUnMod_3	120	Y
ArgMod_4	87	Y
ArgUnMod_4	87	Y
ArgMod_5	87	Y
ArgUnMod_5	87	Y
ArgMod_6	85	Y
ArgUnMod_6	85	Y
ArgMod_7	120	Y
ArgUnMod_7	120	Y
IleMod_1	74	Y
IleUnMod_1	74	Y
IleMod_2	74	Y
IleUnMod_2	74	Y

IleMod_3	74	Y
IleUnMod_3	74	Y
LeuMod_1	80	Y
LeuUnMod_1	80	Y
ProMod_1	84	Y
ProUnMod_1	84	Y
ProMod_2	84	Y
ProUnMod_2	84	Y
ProMod_3	83	Y
ProUnMod_3	83	Y
ProMod_4	82	Y
ProUnMod_4	82	Y
ProMod_5	94	Y
ProUnMod_5	94	Y
ProMod_6	84	Y
ProUnMod_6	84	Y
ProMod_7	84	Y
ProUnMod_7	84	Y
ProMod_8	84	Y
ProUnMod_8	84	Y
SerMod_1	81	Y
SerUnMod_1	81	Y
SerMod_2	81	Y
SerUnMod_2	81	Y
ThrMod_1	98	Y
ThrUnMod_1	98	Y
ThrMod_2	74	Y
ThrUnMod_2	74	Y
ThrMod_3	103	Y
ThrUnMod_3	103	Y
ThrMod_4	103	Y
ThrUnMod_4	103	Y
ThrMod_5	103	Y
ThrUnMod_5	103	Y
ValMod_1	101	Y
ValUnMod_1	101	Y
ValMod_2	101	N
ValUnMod_2	101	Y
ValMod_3	101	Y
ValUnMod_3	101	Y
ValMod_4	101	Y
ValUnMod_4	101	Y
ValMod_5	97	Y
ValUnMod_5	97	Y

Table 2: Table showing the output of SMALT analysis of unmodified and modified tRNA gene sequences for species *B. natans*, presented in the style of Appendices table 1.

Contig	Length	Presence
AlaMod_1	73	Y
AlaUnMod_1	73	Y
ArgMod_1	73	Y
ArgUnMod_1	73	Y
IleMod_1	74	N
IleUnMod_1	74	N
IleMod_2	73	Y
IleUnMod_2	73	Y
LeuMod_1	85	Y
LeuUnMod_1	85	Y
LeuMod_2	78	Y
LeuUnMod_2	78	Y
LeuMod_3	83	Y
LeuUnMod_3	83	Y
ProMod_1	71	Y
ProUnMod_1	71	Y
ProMod_2	72	N
ProUnMod_2	72	N
ProMod_3	72	Y
ProUnMod_3	72	Y
SerMod_1	82	Y
SerUnMod_1	82	Y
SerMod_2	84	Y
SerUnMod_2	84	Y
ThrMod_1	73	Y
ThrUnMod_1	73	Y
ValMod_1	74	Y
ValUnMod_1	74	Y
ValMod_2	74	Y
ValUnMod_2	74	Y

Table 3: Table showing the output of SMALT analysis of unmodified and modified tRNA gene sequences for species *L. tarentolae*, presented in the style of Appendices table 1.

Contig	Length	Presence
AlaMod_1	73	Y
AlaUnMod_1	73	Y
ArgMod_1	72	Y
ArgUnMod_1	72	Y

IleMod_1	74	Y
IleUnMod_1	74	Y
LeuMod_1	82	Y
LeuUnMod_1	82	Y
ProMod_1	72	Y
ProUnMod_1	72	Y
SerMod_1	81	Y
SerUnMod_1	81	Y
ThrMod_1	72	Y
ThrUnMod_1	72	Y
ValMod_1	74	Y
ValUnMod_1	74	Y

Table 4: Table showing the output of SMALT analysis of unmodified and modified tRNA gene sequences for species *C. crispus*, presented in the style of Appendices table 1.

CONTIG	LENGTH	PRESENCE
AlaMod_1	73	Y
AlaUnMod_1	73	Y
AlaMod_2	87	Y
AlaUnMod_2	87	N
AlaMod_3	86	N
AlaUnMod_3	86	Y
ArgMod_1	74	Y
ArgUnMod_1	74	Y
ArgMod_2	106	Y
ArgUnMod_2	106	N
ArgMod_3	106	N
ArgUnMod_3	106	Y
ArgMod_4	86	N
ArgUnMod_4	86	Y
IleMod_1	74	Y
IleUnMod_1	74	Y
LeuMod_1	80	Y
LeuUnMod_1	80	Y
ProMod_1	72	Y
ProUnMod_1	72	Y
SerMod_1	82	Y
SerUnMod_1	82	N
ThrMod_1	73	Y
ThrUnMod_1	73	Y
ThrMod_2	73	Y
ThrUnMod_2	73	N

ValMod_1	74	Y
ValUnMod_1	74	Y

Table 5: Table displaying the optimal codons for species *C. crispus*, displayed is the raw number of optimal codons utilised for the amino acid within the high and low expressed genes, alongside the values for all the other codons utilised for that amino acid. Displayed alongside these are the chi-squared values for each amino acid, in order to determine whether their use within the high expressed genes is significantly higher than their use with the low expressed genes. The significant value for this analysis is P=0.01 therefore any analyses where the value is below this threshold determined there is a statistically significant difference between the two data categories.

	High Bias Genes		Low Bias Genes		Chi-Squared
Phe					
UUC	2733	1471	2291	2087	P<0.0001
Leu					
CUC	3080	5830	2137	8850	P<0.0001
Ile					
AUC	2906	2235	2204	3093	P<0.0001
Val					
GUC	3406	4272	2050	7128	P<0.0001
Ser					
UCC	2105	6038	1651	8491	P<0.0001
Pro					
CCC	1843	4455	1407	4544	P<0.0001
CCG	2217	4081	1694	4257	P<0.0001
Thr					
ACC	2302	3850	1689	5771	P<0.0001
Ala					
GCC	4082	5751	2265	7701	P<0.0001
Tyr					
UAC	1928	855	1962	1253	P<0.0001
His					
CAC	1387	1044	1873	1634	P<0.01
Gln					
CAG	2214	1254	2032	2155	P<0.0001
Asn					
AAC	2644	1118	2233	1960	P<0.0001
Lys					
AAG	4970	1306	3228	2458	P<0.0001
Asp					
GAC	3546	2429	3540	3284	P<0.0001
Glu					
GAG	4245	2348	3774	3295	P<0.0001

Cys					
UGC	1326	489	1502	1184	P<0.0001
Arg					
CGU	1304	5381	1555	8291	P<0.0001
CGC	2783	3902	1946	7900	P<0.0001
Gly					
GGU	1946	5907	1844	6719	P<0.0001
GGC	3114	4739	2078	6485	P<0.0001

Table 6: Table displaying the optimal codons for species *L. tarentolae,* formatted in the style of table 5.

	HIGH BIAS GENES		LOW BIAS GENES		CHI-SQUARED
Phe					
UUC	2421	1034	3546	2759	P<0.0001
Leu					
CUG	3569	4291	5851	11686	P<0.0001
Ile					
AUC	2567	1251	3214	2956	P<0.0001
Val					
GUG	4029	3053	6515	7162	P<0.0001
Ser					
UCC	1538	5623	3295	15423	P<0.0001
UCG	1623	5538	3587	15131	P<0.0001
AGC	1865	5296	4418	14300	P<0.0001
Pro					
CCC	1261	3595	2841	9737	P<0.0001
CCG	2033	2823	4141	8437	P<0.0001
Thr					
ACG	2321	3217	3919	8586	P<0.0001
Ala					
GCC	2914	6291	5530	14682	P<0.0001
GCG	3353	5852	6051	14161	P<0.0001
Tyr					
UAC	2277	427	3546	1199	P<0.0001
His					
CAC	2014	540	3867	1893	P<0.0001
Gln					
CAG	2910	498	5414	1996	P<0.0001
Asn					
AAC	2625	604	3817	1684	P<0.0001
Lys					
AAG	4773	534	4220	1648	P<0.0001
Asp					
GAC	3017	1339	4994	3274	P<0.0001

Glu					
GAG	4454	885	6794	2893	P<0.0001
Cys					
UGC	1700	493	2974	1398	P<0.0001
Arg					
CGC	3646	3095	4637	8571	P<0.0001
Gly					
GGC	3547	3053	4947	6683	P<0.0001

Table 7: Table displaying the optimal codons for species *C. reinhardtii,* formatted in the style of table 5.

	High Bias Genes		Low Bias Genes		Chi-Squared
Phe					
UUC	6505	1233	10781	3443	P < 0.0001
Leu					
CUG	16341	6527	34031	20546	P < 0.0001
Ile					
AUC	6327	2594	9225	4121	P < 0.005
Val					
GUC	4828	14273	9387	30638	P < 0.0001
Ser					
UCC	4230	14308	9701	40310	P < 0.0001
UCG	4462	14076	9024	40987	P < 0.0001
Pro					
CCC	7968	9383	16886	34625	P < 0.0001
Thr					
ACC	7656	6505	14719	19942	P < 0.0001
Ala					
GCC	15740	22196	35940	66177	P < 0.0001
Tyr					
UAC	5440	684	9838	1836	P < 0.0001
His					
CAC	4515	671	11339	2850	P < 0.0001
Gln					
CAG	9633	1142	25079	3880	P < 0.0001
Asn					
AAC	6732	597	11582	2245	P < 0.0001
Lys					
AAG	14137	632	11439	1598	P < 0.0001
Asp					
GAC	10738	2263	20338	6104	P < 0.0001
Glu					
GAG	14726	861	21572	3049	P < 0.0001
Cys					

UGC	3568	485	8679	2092	P < 0.0001
Arg					
CGU	1739	14895	3972	25226	P < 0.0001
CGC	9459	7175	17550	9317	P < 0.0001
Gly					
GGC	18518	8703	38827	25079	P < 0.0001