University of HUDDERSFIELD
Inspiring tomorrow's professionals

An analysis of the selection, usage and bias of optimal codons in the excavate protozoan parasite *Trichomonas vaginalis*

Mathew Lowe

Supervisor 1: Dr Martin Carr

Supervisor 2: Dr Jarek Bryk

MSc by Research

I confirm that, unless stated otherwise, I undertook all work presented in this report. Any work carried out by other parties has been fully acknowledged.

Acknowledgments

I would like to take this opportunity to express my deepest gratitude to all who helped me during this work. Most importantly I would like to thank Dr Martin Carr whose patience, wisdom and eagerness to help with any issue I encountered (particularly when I needed to use software that was only compatible with Mac or Linux based operating systems such as SeqTK) made the process of completing this work smooth and relatively stress free. Your knowledge and expertise were invaluable and your ability to help me find patterns in my data that I myself could not has helped me immeasurably. You pushed me to reach outside my comfort zone and to try new technologies and when in doubt seek help from others in the post-graduate office. Secondly, I would like to offer my sincere appreciation to Dr Jarek Bryk, who at certain points during this work, when I was unable to find a solution to a particular problem, was able to devise a means by which to solve this issue. Your help and input have been greatly appreciated and I would have progressed far slower without it.

I would also like to offer my sincere thanks to all the other post-graduate researchers in the bioinformatics suite at the University of Huddersfield, who have helped me immensely throughout this process, especially when dealing with systems or programs I was not that familiar with.

Abstract

Introduction

Trichomonas vaginalis (*T. vaginalis*) is a single-celled eukaryotic parasite and excavate protist. The genome of *T. vaginalis* was fully sequenced in 2007 and this has presented an opportunity to study various features of its genome such as introns/exons, transposable elements etc. However, while the structure of its genome has been studied by a number of research groups, the codon usage and genomic bias of this organism has not had a significant amount of research dedicated to it. This study attempted to look in more detail at the codons of the organism with regards to selection, bias and usage as well as the structure and evolutionary factors that may be affecting these preferences. As well as this, the organism's genome was studied using current methods and resources to determine if the existing understanding of *T. vaginalis's* genome structure are accurate.

Methodology

The original genome dataset for *T. vaginalis* was analysed in order to verify its gene content as well as identify which genes were truly coding, after which the optimal codons were determined and their bias analysed using codonW in the coding genes and transposable elements. An analysis of mutation pressure was also performed by determining the ratio of coding GC3s to non-coding GC's in genes with high and weak expression levels, as well as a phylogenetic analysis on the candidate genes for adenosine deaminase acting on tRNA (ADAT) to determine if these genes are present and actively transcribed.

Results and Conclusions

Many of the genes may potentially be false positives due to their lack of functional domains and failure to fall into a recognised KOG category. As well as this, the overall GC3 distribution was low, likely due to the AT bias of the organism's genome. The candidate optimal codons were also identified and matched most of the corresponding tRNAs in both the coding and full gene datasets. The few exceptions may potentially be the result of the deamination of adenosine in their corresponding tRNAs. This is evidence that natural selection is having an influence on the selection of OCs in *T. vaginalis* and is consistent with mutation/drift overriding the selection of many of the codons in the dataset, skewing them towards AT, and contributing to the AT bias of the *T. vaginalis* genome. The optimal codon frequency (Fop) in the coding genes was higher in their domain regions, suggesting that selection operates at the level of translational accuracy. In the phylogenetic analysis, ADAT1 gene phylogeny and tree topology matched the expected species phylogeny, while the ADAT2 and 3 phylogenies only did so after adding an outgroup of prokaryote tRNA TadA genes. This suggests that these genes are the true ADAT genes and implies that the deamination of adenosine may indeed be occurring in the tRNAs and influencing the selection of OCs. The analysis of mutation pressure also indicated that mutation pressure may have an influence on codon usage as there was a statistically significant difference between coding GC3s and non-coding GC content. As well as this, the Fop analysis of the domain and non-domain regions, also suggests that selection operates at the level of translational accuracy. With regards to the transposable elements, while overall, they showed no relationship between GC3s and Fop, when separated into their respective families, the LTR retrotransposons alone displayed a positive relationship. This indicates that the retrotransposons do select for optimal codons, but the other transposons do not.

Table of Contents

1.0 Introduction

1.1 DNA, RNA and amino acid selection: a brief history of the genetic code

Deoxyribose nucleic acid (DNA) is a molecule that is comprised of two molecular chains of nucleic acids, or nucleotides which are coiled around one another to form a helix. These nucleotides are the organic molecules which code for the amino acids that form the proteins that make up the cells of an organism and perform the various functions essential to that organism's survival. DNA, as a molecule, was first isolated by Friedrich Miescher, a Swiss physician in 1869, and was dubbed nuclein, due to it being found within the nucleus of cells (Dahm, 2008). Later in 1878, Albrecht Kossel discovered that nuclein (DNA) contained a non-protein component, which was then isolated, along with its 5 primary nucleobases and dubbed nucleic acid (Jones, 1953). However, the full structural components of DNA were not fully characterised until 1929, when Phoebus Levene, who also discovered and isolated the sugar, base and phosphate nucleotides in ribonucleic acid (RNA) twenty years prior (Levene and Jacobs, 1909), discovered that DNA also consisted of a ribose sugar, however unlike RNA this sugar was deoxygenated and thus dubbed deoxyribose. Levene later proposed the tetranucleotide hypothesis, which suggested that DNA might consist of a series of 4 units, (which were dubbed nucleotides), that were strung together through their phosphate groups to form a series of bases that formed a short chain that repeated in a fixed order (Cohen and Portugal, 1974). Even during these early days of nucleic acid research there were some suggestions of the existence of a molecule that was the basis for all the proteins and structures that form an organism, most notably in 1927, when Nikolai Koltsov proposed the existence of a large, hereditary molecule that consisted of 2 strands that mirror one another, containing all the information for the hereditary traits seen between family groups, and which replicate in a semi-conservative manner, with each strand using the other as a template (Koltzoff, 1928).

While understanding the structure of DNA provided the first evidence that this was potentially the molecule Koltsov had proposed, the first evidence that DNA contains genetic information came as a result of an experiment performed by Frederick Griffith in 1928, where he demonstrated that traits that were associated with the smooth form of *Pneumococcus* could be 'passed on' or transferred to the rough form of the organism when killed smooth bacteria were mixed with living rough bacteria (Griffith, 1928; Lorenz and Wackernagel, 1994). The results of this experiment suggested that traits from one organism were being incorporated by another of the same species via a molecular transfer of information, suggesting that there is a molecular mechanism behind the traits seen in organisms and Griffith suggested that DNA was the molecule that facilitated this transfer of information (Griffith, 1928). For much of the early twentieth century it was thought that RNA was only present in plants and DNA, only present in animals, with DNA being predominantly involved in buffering the pH of the cell nucleus. Later, in 1937, the first diffraction patterns were produced of DNA which demonstrated that it had a regular structure similar to the structure proposed by Koltsov in 1927, and in 1943, the Avery, MacLeod and McCarty experiment demonstrated that DNA was the molecule that facilitated the bacterial transformation seen by Griffiths in 1928 and thus held the genetic information of an organism, rather than proteins as had been initially assumed (Avery et al, 1944). In 1952, this was confirmed by Hershey and Chase, who demonstrated that DNA was the genetic material for the enterobacteria phage T2 and then later in 1953, James Watson and

Francis Crick identified the full structure of DNA and determined that the molecule formed a double helix consisting of two deoxyribose molecules that were antiparallel, wrapped around one another (Watson and Crick, 1953).

Later research by Crick and his co-workers demonstrated that the genetic code is comprised of a series of non-overlapping trinucleotide bases called codons, and in 1961 Nirenberg and Matthaei were able to characterise and demonstrate the function of a codon with regards to amino acid synthesis by translating an RNA sequence consisting of only triplets of the nucleotide uracil in vitro and discovering that the polypeptide that was synthesised consisted of only phenylalanine. They thus concluded that the nucleotide base 'UUU' codes for the amino acid phenylalanine exclusively (Nirenberg and Matthaei, 1961). Most of the further codons and their corresponding amino acids were characterised by Gardner et al, throughout the 1960's following a series of similar experiments, with the final genetic code being fully characterised by Har Gobind Khorana (Sakmar, 2012). At the time, only 2 forms of RNA were fully characterised ribosomal RNA (rRNA) and messenger RNA (mRNA), but later in 1965, Holley et al, were able to determine the structure and demonstrate the activity of the third form of RNA, transfer RNA (tRNA), which acted as an adapter molecule between the mRNA and the amino acids that form proteins during the process of translation, and helped to fully characterise the process of protein synthesis from DNA being transcribed into mRNA and that mRNA binding to a tRNA molecule that mirrors the mRNA molecule and acts as an adapter for the amino acids that form the protein.

1.2 Codon usage bias: selection or mutation?

Despite there being only 20 amino acids (with a few rare organisms possessing a 21st selenocysteine amino acid) there exist 64 different potential codons that make up the genetic code. This disparity between potential codons and the quantity of amino acid that make up proteins gives the genetic code a great deal of redundancy, allowing a number of amino acids to be encoded by more than one potential codon, suggesting a large degree of degeneracy within the genome. However, while initially this degeneracy would appear to suggest a certain lack of specificity with regards to which codons an organism uses for certain amino acids, researchers have found that in many organisms a pattern of codon usage exists, with many organisms displaying a 'bias' towards some codons for an amino acid over others, with these codons being dubbed 'optimal codons' (Ikemura, 1985). However, while most organisms do demonstrate a bias towards some codons over others, it is not universal to all organisms, as which codons an organism prefers varies from one species to another. As well as this there are some organisms, particularly those which have a slow rate of growth (Dong et al, 1996), where codon bias would appear to be absent, for example *Helicobacter pylori*, a prokaryote which displays little if any bias with regards to its codon usage (Atherton, Sharp and Lafay, 2000).

How and why these biases arise, while well established in some species, in most others is unknown, as there are a multitude of different potential factors and mechanisms that have been demonstrated to have an influence on bias, and which has the greater influence varies from one species to another. However, the current consensus is that, at least in the case of translational optimisation, there is a general balance between mutational bias and natural selection (Crow and Kimura, 1970; Crow and Pogson, 2013). While it has been suggested that codon usage bias may have influenced tRNA evolution and vice versa, with codons that have a high frequency throughout the genome driving up the expression of their

corresponding tRNAs, potentially as a result of selection (Saint-Léger et al, 2016), the experimental evidence to support this is limited. One interpretation as to why codon usage bias exists is that the optimal codons help improve the rate as well as accuracy of translation. As a result, it is expected that translational selection is stronger in genes that are highly expressed and as a result, should also be stronger in organisms that have a faster rate of growth compared to those with a slower rate of growth, such as *Helicobacter pylori* (Atherton, Sharp and Lafay, 2000; Sharp, Emery and Zeng, 2010). In those organisms, it would appear that codon preference is instead influenced by mutational bias (Atherton, Sharp and Lafay, 2000). Over the years a number of other potential factors outside of gene expression level have been proposed that appear to influence codon usage bias, such as guanine – cytosine (GC) content/skew due to mutation pressure, amino acid conservation, RNA stability etc (McInerny, 1998).

While the mechanisms behind codon usage bias and why it evolved remain controversial, there are two prevailing theories that offer a potential explanation. The first is the selectionist theory, which postulates that codon bias contributes to the efficiency and/or accuracy of protein expression and therefore undergoes positive selection. Which may explain why more frequent codons tend to be recognised by more abundant tRNA molecules, as well as the correlation between the availability of optimal codons, tRNA expression levels and gene copy number (GCN) (Sharp et al, 1993). However, while the rate at which amino acids are incorporated in more frequent codons is much higher than that of rarer codons, codon frequency has not been shown to directly influence the rate of translation and therefore, there is likely no direct advantage to having a bias towards more abundant codons (Hershberg and Petrov, 2009). However, there are some potential indirect advantages, such as an increase in the cellular concentration of free ribosomes as well as an increase in translational elongation which, in turn, may provide an increase in the initiation rate for mRNAs (Sharp et al, 1993; Hershberg and Petrov, 2009).

The other prevailing theory is the mutational bias theory, which posits that codon bias exists as a result of the non-random pattern of mutation within the genomes of certain organisms, as some codons have been shown to undergo more changes than others which results in them displaying lower equilibrium frequencies i.e. they become rarer (Sharp et al, 1993; Hershberg and Petrov, 2009). Mutation bias has been shown to vary from one organism to another and there is growing evidence that the level of GC content throughout the organism's genome is the primary parameter that explains the differences between the various organisms with regards to mutation bias (Hershberg and Petrov, 2009). This model tends to argue against selective forces acting on the coding regions of DNA, instead proposing that the codons that are more abundant and thus used more frequently, are the result of these codons displaying a low mutation rate compared to the rarer codons, resulting in them displaying higher equilibrium frequencies (Hershberg and Petrov, 2009). However, this theory still cannot explain why optimal codons are recognised by tRNAs that tend to be more abundant (Dong et al 1996), as this should not be influenced by mutation rate or the equilibrium frequencies of the optimal codons. As well as this, the model does not account for the tendency of selection coefficients, that have been associated with optimal codons, to be quite small. This potentially results in mutation pressure and genetic drift overwhelming selection with regard to codon usage in organisms with relatively small effective population sizes (dos Reiss & Wernisch, 2008). While both models explain some of the observations seen by researchers over the years with regards to codon usage bias, they

don't individually provide a complete explanation for why codon usage bias exists. As such, another model for codon usage bias was adopted more recently that balanced the existing evidence for the two theories, "the mutation-selection-drift balance model", which suggests that selection favours the optimal codons over those codons that are rarer, however, the rarer codons persist due to a combination of genetic drift and mutation pressure. It also suggests that selection is generally weak, but scales with the expression level as well as the functional constraints of the organisms coding sequences (Hershberg and Petrov, 2009).

1.3 *Trichomonas vaginalis*: unlocking an unusual genome

T. vaginalis is a single-celled eukaryote and the primary causative agent of the sexually transmitted infection (STI) trichomoniasis in humans, a disease which affects both males and females and is one of the most prevalent sexually transmitted infections that is curable in the world (Rowley et al, 2019; Harp and Chowdhury, 2011). While the molecular aspects of the disease's pathogenicity have been explored in great detail for a number of years, there has been little progress with regards to understanding the genome of the organism itself until quite recently. This has been largely attributed to a number of recent developments in genomic analysis techniques and technology which have in turn helped enhance the understanding of *T. vaginalis* and its genome, most notably the advent of advanced gene sequencing technologies and their ever-increasing cost effectiveness and level of accuracy. These endeavours as well as the advent of databases such as the TrichDB genome database (Aurrecoechea et al, 2008) have contributed a great deal to the understanding of *T. vaginalis's* genome structure and expression. The most notable development however, was the complete sequencing of the *T. vaginalis* genome by Carlton et al (2007), who determined that, while it had a fairly large genome for a single-celled eukaryote (approximately 160Mb), the gene annotations suggested that there are over 60,000 potential genes (~3X more genes than the human genome) (Carlton et al, 2013; Harp and Chowdhury, 2011). However, this large genome size was later attributed to gene duplication events throughout *T. vaginalis*'s genome, causing it to largely consist of pseudogenes and non-coding RNAs which also contributed to its repetitive nature (Woehle et al, 2014).

Over the years since the sequencing of the *T. vaginalis* genome an effort has been made to study and characterise the numerous genes and coding regions within the genome of the organism (Woehle et al, 2014; Smith and Johhnson, 2011; Conrad et al, 2012). However, there have been a number of difficulties researchers have faced when trying to study its genome, most notably its highly repetitive nature (Carlton et al, 2007; Carlton et al, 2013), which makes accurate sequencing difficult resulting in some regions with several gaps throughout. There are a number of potential factors that could be influencing this repetition in the organism's genome, for example the gene duplication events that comprise the majority of the genome as outlined by Woehle et al (2014). One other potential factor is the large number of transposable elements that are present throughout the *T. vaginalis* genome, which may also be one of the potential factors influencing the genomes bias towards adenine (A) and thymine (T) as opposed to guanine (G) and cytosine (C) nucleotides (Conrad et al, 2013).

1.4 Codon usage statistics and gene expression

Determining which genes primarily form part of *T. vaginalis*'s exome requires the identification of genomic regions that possess characteristics that have been associated with

actively transcribed genes in other eukaryotes, which can be done using a gene annotation server such as the EggNOG gene annotation server (Huerta-Cepas et al, 2019), a new resource which allows the rapid and accurate identification of features and regions that are homologous with those found in genes that are actively transcribed in other eukaryotes/organisms. These key features are then used by the server to group the sequences into specific categories known as COG (clustered orthologous group) or, in the case of eukaryotes, KOG (eukaryotic orthologous group) categories (Tatusov et al, 2000; Tatusov et al, 2003; Novoa et al, 2019). One other characteristic that could potentially indicate that these genes may be actively transcribed is a high level of G or C nucleotides at synonymous third positions or GC3s, as these have also been identified as a potential candidate feature of protein coding regions in the genome of other organisms (Elhaik and Tatarinova, 2012); however, which nucleotides are preferred at synonymous sites can also be influenced by other factors, such as mutation pressure and the organisms effective population size (Epstein et al, 2000; Hershberg and Petrov, 2009). These GC3s can be analysed and determined in a number of ways but are most often done using CodonW (Peden, 1999), a program that allows the analysis of the content of genes to determine the presence of a number of key features including GC3s. However, a high GC3 level on its own does not indicate a gene is protein coding as there a number of organisms that have relatively low GC3s (Grosjean et al, 1996); which may be true for *T. vaginalis* as well due to its genome's bias towards AT (Meade et al, 1997). Many of the new resources that have become available for genome analysis would provide a good method of determining if the original genome analysis by Carlton et al (2007) was accurate, whether the gene content of *T. vaginalis*'s genome is truly as large as the original analysis would suggest and whether these genes all possess the features that are associated with coding regions in other eukaryotes and group together as expected.

While all organisms tend to use all potential codons to some degree throughout their genome during transcription, many organisms have been shown to demonstrate a preference for one or two particular codons over others (Ikemura, 1981). There are a number of methods to determine which codons an organism prefers to use, however, not all organisms and not all genes within organisms have a bias towards certain codons. It must therefore be determined whether a bias exists and the most commonly used method of determining this is to measure the organism's effective gene codon number (Nc). This is a measure of whether the codons in an organism's genome are used equally or if certain codons are preferred over others. The values for Nc tend to range from a minimum of 20, indicating that there is a strong bias towards specific codons, to a maximum value of 61, indicating all codons are used equally and there is little or no preference (Wright, 1990).

There are two primary methods of identifying the candidate optimal codons for a particular genome: in the absence of expression data, a correspondence analysis can be performed where each codon is placed on one of 3 axes, with the first axis being used to explain the majority of the overall variation between genes (and thus representing a high degree of bias), those genes which have an overall higher frequency at the more biased end of the axis (position one) are considered to be optimal (Peden, 1999; Takeia, 2016). This method is not ideal as the optimal codons generated must also correlate with the gene's expression level to be considered genuine. The other more accurate and reliable method is to analyse the codon usage statistics in genes based on expression level, however, this requires existing expression data for that organism's genome. Once these candidate optimal codons have

been determined, their frequency throughout the genome or within particular genes (Fop) can be determined (Ikemura, 1981; Drummond and Wilke, 2008). The frequency throughout the genome is another measure of bias, with a range from 1.0 (indicating that the candidate codons are used exclusively) to 0.0 (indicating that these codons are not used at all) (Ikemura, 1981).

1.5 Potential factors influencing *T. vaginalis* codon usage: accuracy and efficiency

Not all organisms demonstrate selection at the level of codon usage, as there are several that do not have optimal codons or use codons fairly evenly, particularly those that have fairly slow growth rates such as *H. pylori* (Atherton, Sharp and Lafay, 2000). Those that do however, tend to select these optimal codons based on both efficiency and accuracy such as in the case of *E. coli* and *S. cerevisiae* (Gingold and Pilpel, 2011). While all genes in these organisms demonstrate both forms of selection, this is not universal across the whole genome, as in some genes selection for one can be stronger than the other (Gingold and Pilpel, 2011) and in the case of *T. vaginalis*, this has not been fully determined. As well as this, the heavy AT bias of its genome (Conrad et al, 2013) may potentially have an influence on accuracy and efficiency with regards to the selection of optimal codons. However, most organisms tend to prioritise accuracy in order to reduce their genomes evolutionary rate, often possessing a number of mechanisms to resist or control this as mistranslation can have a significant constraint on the viability of the resulting proteins, and thus the organism's survival in that environment (Drummond and Wilke, 2008).

However, while accuracy during translation is important as it ensures the correct protein is produced and that said protein is able to function correctly, with the genome possessing a number of mechanisms to ensure this (Steiner and Ibba, 2019; Gingold and Pipel, 2011; Drummond and Wilke, 2008), the genome also requires efficiency during translation, as it ensures that the correct protein can be translated quickly and efficiently when needed. Codon usage in some organisms such, as *E. coli*, have been shown to influence this as when the availability of the tRNA for a codon that is biased in particular genes with a high expression level (and is therefore more abundant) is reduced in favour of its less optimal counterparts, it results in a reduction in translational efficiency across the genome, but particularly within genes that are highly expressed, which in turn results in a reduction in the cellular fitness of the organism (Frumkin et al, 2018). This suggests that genes with a high expression level and codon usage bias also rely on translational efficiency and, more significantly, that the codon usage bias of these genes has an influence on the translational efficiency across the whole proteome (Frumkin et al, 2018).

While maintaining a balance of efficiency and accuracy is important, they are not prioritised equally in all genes or genomes. Some genes and some organisms prioritise accuracy while others appear to prioritise efficiency or a balance between the two (Gingold and Pilpel, 2011). In the case of *T. vaginalis*, this has never been fully determined, however, research into a number of trypanosomatids by Horn (2008) demonstrated a strong relationship between selection of optimal codons at the level of translation and the expression of key proteins. This may be true for *T. vaginalis* as well, however, a more extensive analysis is necessary to fully determine whether this is the case.

1.6 Codon usage bias in the transposable elements

Transposable elements are DNA sequences which have the capability of altering their position within a particular genome, a phenomenon known as transposition (McClintock, 1950), and in doing so, generate a number of effects such as reversing (and sometime generating) particular mutations in the region to which they transpose, altering the genetic identity of a cell as well as duplicating genetic material resulting in an increase in the size of the organism's genome (Bourque et al, 2018). The percentage of an organism's genome that consists of Transposable elements can vary from one species to another with some, like maize having approximately 90% of their genome consist of transposable elements (McClintock, 1950), while the human genome only consists of 66-69% transposable elements (de Koning et al, 2011). While transposable elements are generally considered 'selfish' i.e. they are capable of enhancing their own transmission even at the expense of other genes or genetic elements in the genome, even if this has a negative impact on the fitness of the organism or wastes genomic resources with no positive gain (Werren, Nur and Wu, 1988), there are many transposable elements that play an important role in both genome function as well as evolution.

Many organisms' genomes consist predominantly of transposable elements, and *T. vaginalis* is no exception as approximately 60% of its genome is comprised of transposable elements, which may account for the repetitiveness of its genomes (Conrad et al, 2013). There has been a large effort in recent years to analyse transposable elements in single-celled organisms, particularly single-celled eukaryotes such as the study by Jiang and Govers (2006) which found that the long tandem repeat (LTR) retrotransposon transposable element families in the stremenopile genus *Phytophthora* displayed a bias towards codons that were GC-ending and mirrored the genes present in the host sequence, with similar results found by Southworth et al (2019) in choanoflagellate *S. rosetta*, which also showed evidence for selection at the level of codon usage in the LTR retrotransposons. As well as this, both Southworth et al (2019) and Jiang and Govers (2006) revealed that the transposable elements which had the highest copy-number (indicating greater success during transposition) also had the highest degree of codon usage bias, similar to how the genes with the highest expression levels in an organism's genome also demonstrate the strongest codon usage bias (Hershberg and Petrov, 2009). While transposable elements are widespread throughout the *T. vaginalis* genome and have been associated with its strong AT bias and repetition (Conrad et al, 2013), whether there is evidence for selection in the codons of the transposable elements has yet to be determined. This therefore presents an opportunity to study another aspect of *T. vaginalis*'s genome and, in particular, to determine whether these transposable elements also demonstrate codon bias similar to what has been observed in other organisms, and as transposable elements are believed to be the primary contributing factor influencing the AT bias of the *T. vaginalis* genome, it is also possible that they may play a potential role in influence the codon usage bias of *T. vaginalis*.

1.7 Determining the codon usage and bias of *Trichomonas vaginalis*

The first attempt at studying the codon usage of *T. vaginalis* was by McInerny (1997), who proposed that there was evidence for bias with regards to the selection of optimal codons in *T. vaginalis* and identified several potential candidate optimal codons. However, the dataset used in the study was limited as the genome of *T. vaginalis* had not yet been fully

sequenced. The full genome of *T. vaginalis* was sequenced by Carlton et al (2007), yet despite this and despite trichomoniasis being one of the most prevalent and widespread sexually transmitted infections in the world (WHO, 2008), its codon usage and bias have not been fully characterised since McInerny's (1997) original analysis. As well as this, despite single-celled species making up the vast majority of the genomic diversity seen within the eukaryotic phyla, there has been little if any attempt at determining their codon usage and bias (as most research is heavily focussed on either bacterial cells, multi-cellular organisms or fungi). Given the large number of advances in DNA and genome analysis in the decades since then, now is an opportune time to perform a more extensive codon analysis of the organism than that performed by McInerny (1997), this time using the, now fully sequenced, genome of *T. vaginalis* to determine if the codons proposed by McInerny were truly optimal. To that end, this study attempted determine any trends in the newly sequenced genome by generating an Nc plot and investigating them. Following this, an attempt was made to fully characterise the optimal codons of the organism and compare them to those optimal codons proposed by McInerny (1997).

Another potential factor that may influence codon usage is the deamination of adenosine by adenosine deaminase. However, it is unclear whether this mechanism (which predominantly favours C-ending codons) is still required in a genome with such a strong AT bias. In order to determine if the genes that encode one group of enzymes responsible for this are still present, a phylogenetic analysis was performed on the candidate genes for adenosine deaminase acting on tRNA (ADAT) alongside the ADAT genes of a number of other organisms to determine if they were genuine. As well as this, a series of other analyses was performed on the codons to determine to determine whether there was evidence for selection for optimal codons at the level of translational accuracy and efficiency as well as whether mutation pressure has an influence on codon usage. As a number of other organisms have demonstrated evidence for codon usage bias in a number of their transposable elements, and the AT bias in *T. vaginalis* has been attributed to the high number of transposable elements in its genome (Conrad et al, 2013), an analysis was also performed on the transposable elements in *T. vaginalis* to determine if they also demonstrate this bias. Through these analyses, a greater and more comprehensive understanding can be made of the nature of codon usage within *T. vaginalis*, as well as add insight into the codon usage of eukaryotes, in particular eukaryotic parasites, and the potential factors that may influence codon usage in their genome.

2.0 Materials and methods

2.1 Genome filtering

Initially the full *T. vaginalis* coding sequence (CDS) was downloaded from the TrichDB website v46 (Aurrecoechea et al, 2008) and, using the codonW genome analysis program v1.4.2 (Peden, 1999), GC3 and effective codon number (Nc) values were calculated. Results were then used to separate outliers with GC3 scores above 0.6 from the initial CDS dataset and work out the standard deviation (SD) as well as averages for both GC3 values and Nc, which were then used to work out the distribution. A series of scatter plots of GC3 vs Nc as well as Nc vs GC3 were then generated and a distribution curve for the data superimposed over them so that the results could be visualised. A further curve visualising the distribution under no selection was also superimposed to aid in the interpretation of the data (Wright, 1990). The graphs were then used to identify genes with GC3 scores above 0.6 which then had their FASTA sequences extracted and run through the EggNOG gene annotation server v5.0.0 (Huerta-Cepas et al, 2019), in order to determine if they possess any functional domains which could suggest that the gene is functional and therefore potentially protein coding. The genes were then BLASTed (Basic Local Alignment Search Tool – a tool developed by the National Centre for Biotechnology Information (NCBI) in 1991 that automates the alignment, comparison and annotation of gene sequences to determine if they share any similarities with recognised genes sequences from other organisms within their genome database and are therefore potentially genuine) to determine if they are indeed non-coding before removing them from the dataset. The tRNA genes for *T. vaginalis* were then downloaded from the TrichDB and uploaded to EggNOG in order to identify which were the most abundant; the results were then compared to McInerny's (1997) predicted OCs.

The full CDS was uploaded to EggNOG after separating sequences into smaller groups (as HMMR gene mapper will only accept ≤5000 sequences at a time) and gene annotation was attempted in order to identify coding sequences which correspond to a recognised KOG category and separate them from the non-coding sequences, without a recognised KOG category. After an initial failed attempt, due to the presence of a number of sequences containing strings of 'NNN' (indicating unknown nucleotide), sequences with unknown nucleotides were removed and mapping attempted again. Gene sequences were then separated based on their KOG category and genes with either a KOG category of 'S' (indicating unknown function) or no KOG category at all were BLASTed using the NCBI gene database to determine if they correspond to any recognised proteins. The remaining genes with a recognised KOG category (KOG^+) were separated from the main dataset and sequences without a recognised KOG category (KOG^-) were uploaded to NCBI in groups of 10,000 and BLASTed using BLASTn, in order to determine if they possessed any functional domains or had any orthologues among the other taxa. Those that lacked either of these were determined to be potentially non-coding and excluded.

2.2 Determining optimal codons, frequency and selection

KOG^+ genes were then inserted into codonW in order to calculate their GC3 and Nc values. Results were then inserted into Excel where their SD, average and distribution were calculated and compared to the full dataset. Sequence read archive (SRA) data from NCBI along with the original coding genes were used to generate sequence read data which was then inserted into SMALT in order to determine the number of reads per gene CDS and the results were then inserted into and visualised using TABLET v1.19.09.03 (Milne et al, 2013).

The resulting data was then inserted into Excel where read number was divided by sequence length in order to determine the normalised expression level for each gene. The genes were then ordered by expression level and the top 5% with the highest expression level and the bottom 5% with the lowest expression level were isolated and extracted using SeqTK v1.3. Genes were then run through codonW v1.4.2 (Peden, 1999) in order to determine if any sequences end in a partial codon, and if they did they were trimmed back to avoid frameshifting the dataset. The top and bottom 5% genes were then run through codonW v1.4.2 (Peden, 1999) again in order to extract the BLK files (image/graphics type files produced as a by-product of all codonW analyses, that contain the number of copies of each codon present in that dataset as well as the RSCU values for each codon) which are then inserted into Excel in order to determine which codons were predicted to be optimal (determined by analysing which codon had a higher RSCU and frequency in the highly expressed compared to weakly expressed genes at 1% significance). A Chi^2 was then performed on all predicted OCs in order to determine the level of significance and the results were compared to the major tRNAs (as the tRNA anticodons have a specific codon to which they bind during translation and if the predicted optimal codons match those in the tRNAs this is evidence for them being genuine) as well as McInerny's (1997) predicted optimal codons. The results were then compared to the predicted optimal codons from the full dataset.

A correspondence analysis (CoA) was performed on the coding as well as full dataset using codonW v1.4.2 (Peden, 1999), and the results from the 'CoA' output file (which places each codon into one of three potential categories: 1 (rare), 2 (normal) or 3 (optimal)) that returned a value of 3 from this analysis were compared to the predicted optimal codons in both the coding and full datasets to determine if they match. In an effort to determine which sequences were enriched for OCs, the gene names for all the KOG+ genes were extracted and, using seqTK v1.3, had their names linked to their corresponding gene sequence before running each category sequence file through codonW v1.4.2 (Peden, 1999) where their Nc, GC3 and Fop were determined. In order to investigate if there is evidence for selection for translational accuracy in the optimal codons, the top 100 genes from the KOG+ dataset with the highest expression level, had their sequences extracted and checked on the NCBI website to determine if they returned any annotations indicating whether they possessed any functional domains. Those that did had their sequences translated to protein using ExPASy (Swiss Institute of Bioinformatics) and the domain region extracted from the rest of the sequence. Non-domain regions were then concatenated and both the domain and non-domain sequences were uploaded into codonW v1.4.2 (Peden, 1999) where a Fop analysis was performed to determine if Fop was higher in the domain regions (which would potentially be evidence for selection based on translational accuracy in these genes).

2.3 Phylogenetic analysis of adenosine deaminases acting on tRNA (ADAT)

An attempt was made to identify if *T. vaginalis* possessed any ADAT genes by searching NCBI, and extracting protein sequences and BLASTing them using BLASTp on the NCBI genome database under five other distantly related organisms: *Homo sapiens* (Metazoa), *Drosophila melanogaster* (Metazoa), *Arabidopsis thaliana* (Streptophyta), *Emiliania huxleyi* (Haptophyta) and *Ectocarpus siliculossus* (stramenopiles). The relationship and percentage similarity to each organism was analysed in order to determine if the three provisional *T. vaginalis* ADAT genes were genuine. A series of multiple sequence alignments were generated for these ADAT genes using Multiple Alignment with Fast Fourier Transform

(MAFFT) (Katoh, 2002; Madeira et al, 2019), with ADAT1 being aligned separately from ADAT2 and 3, which were aligned together in order to see if they formed monophyletic groups. These alignments were then run through the XSEDE supercomputer on the Cyberinfrastructure for Phylogenetic Research (CIPRES) gene alignment server v3.3 (Miller, Pfeiffer & Schwartz, 2011) for a maximum of 12 hours using the L-INS-I alignment program (as there are <200 sequences and this program is more accurate with smaller datasets). This generated an alignment file (FASTA format), which was reformatted using Notepad++ v6.5.1 into the Nexus and Phylip alignment formats. The Nexus alignment was run through the MrBayes Bayesian phylogenetic inference program v3.2.7a (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003; Ronquist et al, 2012) for 96 hours and 5,000,000 generations, using a mixed amino acid rate matrix, a burnin value of 1,250 and a sampling frequency of 1000. The results were then visualised using Figtree v1.4.4 (Rambout, 2009). Bootstrap values were determined by performing a maximum likelihood analysis on the phylogeny using the Phylip file, the Randomised Axelerated Maximum Likelihood (RAxML) program v8.2.12 (Stamatakis, 2014) and the XSEDE supercomputer on the CIPRES gene annotation server for 96 hours, using the DAYHOFF protein substitution matrix and 1,000 bootstrap iterations. The bootstrap values as well as posterior probabilities allow the nodes to be evaluated.

In order to improve the alignments, a number of additional species were added to the phylogeny and the identity of the ADAT genes were redetermined, this time by identifying them using the *Homo sapiens* ADAT gene sequences and BLASTing these sequences against *T. vaginalis* on the NCBI genome database. The resulting ADAT gene sequences were then used to identify corresponding sequences in *Fonticula alba* (a nuclearioid amoeba, that was an opisthokont as well as part of the Holomycota along with *S. cerevisiae*) as well as 20 other organisms from a range of different taxa (fig 2.1) which were then added to the original dataset. However, further organisms were required to refine the results, and thus the following organisms were added to the dataset: *Leptomonas seymouri* (Discoba)*, Leptomonas pyrrhocoris* (Discoba)*, Tritrichomonas foetus* (Metamonada)*, Paramecium tetraurelia* (SAR) and *Angomonas deanei* (Discoba). As well as this, in order to improve the tree topology of the ADAT1 phylogeny and to determine whether the ADAT1 genes for *T. vaginalis* are monophyletic as expected, the ADAR gene from *H. sapiens* were added (as these are paralogues and could act as an outgroup) and used as a root. In order to improve the tree topology in the ADAT2 and 3 phylogenies as well as determine whether they too are monophyletic, the two datasets were combined and re-rooted using the TAD genes from the following bacteria: *Mycobacterium tuberculosis, Escherichia coli, Shigella spp, Citrobacter rodentium* and *Rhodococcus quingshengii*.

<table>
<tr><th colspan="6">Additional Taxa</th></tr>
<tr><th>Opisthokonta</th><th>Apusuzoa</th><th>Amoebozoa</th><th>Archaeplastida</th><th>SAR</th><th>Excavata</th></tr>
<tr><td>Salpingoeca rosetta</td><td rowspan="5">Thecamonas trahens</td><td rowspan="3">Acanthamoeba castellanii</td><td>Chondrus crispus</td><td>Paramecium tetraurelia</td><td>Leishmania major</td></tr>
<tr><td>Capsaspora owczarzaki</td><td rowspan="2">Glycine max</td><td>Tetrahymena thermophila</td><td>Trypanosoma cruzi</td></tr>
<tr><td>Sphaeroforma arctica</td><td>Albugo laibachii</td><td>Giardia lamblia</td></tr>
<tr><td>Saccharomyces cerevisiae</td><td rowspan="2">Dictyostelium discoideum</td><td rowspan="2">Chlamydomonas reinhardtii</td><td>Phaeodactylum tricornutum</td><td rowspan="2">Naegleria gruberi</td></tr>
<tr><td>Fonticula alba</td><td>Thalassiosira pseudonana</td></tr>
</table>

Fig 2.1: Additional organisms added to phylogeny in order to improve the accuracy of results and better display the evolutionary relationships between the different ADAT genes.

2.4 The influence of mutation pressure on codon usage bias

An attempt was also made to determine whether there is evidence for mutation pressure influencing the frequency of optimal codons in *T. vaginalis*. This was done by extracting the top and bottom 100 genes, based on expression level, in the *T. vaginalis* coding gene dataset and extracting the flanking DNA sequence 200bp upstream and downstream of the CDS, ensuring that part of another gene was not also extracted. This was determined by examining the image for each of the genomic region for each gene on NCBI to see if there are any gene coding regions for neighbouring genes within each 200bp region either side of the gene being analysed; if any of the DNA in the regions flanking the CDS of each gene corresponded with another gene, that gene was excluded. Once the flanking DNA had been extracted, they were concatenated and the genes were then separated into their corresponding dataset (top or bottom 100). These datasets were then converted to FASTA format and inserted into codonW v1.4.2 (Peden, 1999) where their GC content was analysed.

The resulting data then had a t-test performed on it with an α value of 0.5 in order to determine the statistical significance of the results. As well as this, a box and whisker plot of the two datasets was also generated in order to observe the distribution as well as visualise the differences between the two. Due to this analysis being fairly crude, and localised variation in GC content potentially having an influence on the results, the GC3 expression levels of the coding region of the DNA were plotted against the GC content of the non-coding regions either side of the DNA in the top 100, bottom 100 and combined coding and non-coding datasets. After this a scatter graph was produced for each so that the relationship between the coding and non-coding regions could be determined. A positive relationship would suggest that the variation in GC3s could, to some degree, be attributed to local mutation pressure. In order to further characterise this, the same analysis was performed on the 100 genes with mid-expression and the GC content results were then compared to the top and bottom 100 gene dataset to determine if there is any significant variation in the GC content of their non-coding regions. A one-way ANOVA analysis was performed on the combined top, mid and bottom 100 gene datasets comparing non-coding GC content in order to determine if there was a statistically significant difference between the GC content of the 3 datasets.

2.5 Codon usage in the transposable elements of *T. vaginalis*

NCBI was searched for *transposase* sequences by searching for *T. vaginalis* transposase and the resulting sequences were downloaded as both proteins and nucleotides. The SMALT DNA mapping and alignment tool v0.7.4 (Genome Research Ltd, 2010) was then used to map the number of reads for each sequence, which was then divided by the sequence length in order to normalise the expression level of each *transposase* sequence. The nucleotide sequence for each *transposase* was then inserted into the protein repeat masker server (repeatmasker.org) in order to identify which transposon protein group each sequence falls into, and then grouped into their corresponding transposable element family/superfamily. The sequences of each group were then loaded into codonW v1.4.2 (Peden, 1999) and, using the original *T. vaginalis* CoA file, had their optimal codon frequency (Fop) determined. A scatter plot of Fop against expression level for each group was then generated in order to determine if the frequency of optimal codons increased with expression level, which would suggest that selection of optimal codons does occur in the transposable elements and would be evidence for selection at the level of translational efficiency.

In order to identify if any of the transposase sequences contained any conserved domains the sequences for each protein group were BLASTed using NCBI. Those that did had their domain and non-domain regions isolated from the sequence and concatenated before both being inserted into codonW v1.4.2 (Peden, 1999) where their Fop was analysed to determine if the optimal codons identified earlier appeared more frequently in the domain regions or the non-domain regions of the sequence. Following this, the results were compared for each group and analysed statistically to determine their significance.

3.0 Results

3.1 GC3s, Nc and tRNA genes

The full genome of *T. vaginalis* from the TrichDB v46 (Aurrecoechea et al, 2008) had a total of 95,100 annotated genes and the results from the Nc distribution (fig 3.1) are in line with past research suggesting that the full annotated gene dataset has a strong bias towards AT as the distribution curve indicates that most genes have a GC3 score below 0.5. The overall average for GC3s in the dataset was 0.29 with a standard deviation of 0.09, and the overall average for Nc was 43.70 with a standard deviation of 5.82. This remained true even when the outliers with a GC3 score above 0.6 were removed (fig 3.2) (determined to be outliers as when uploaded to EggNOG they produced few if any results, which was likely due to not possessing any functional domains).

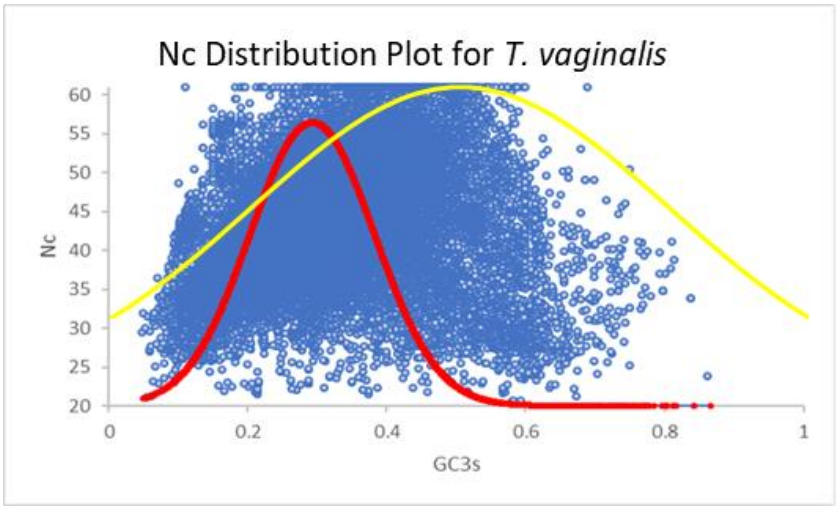

Figure 3.1: distribution plot for *T. vaginalis* demonstrating Nc distribution in relation to GC3. overall distribution for *T. vaginalis* = . Nc distribution under no selection according to Wright (1990) =

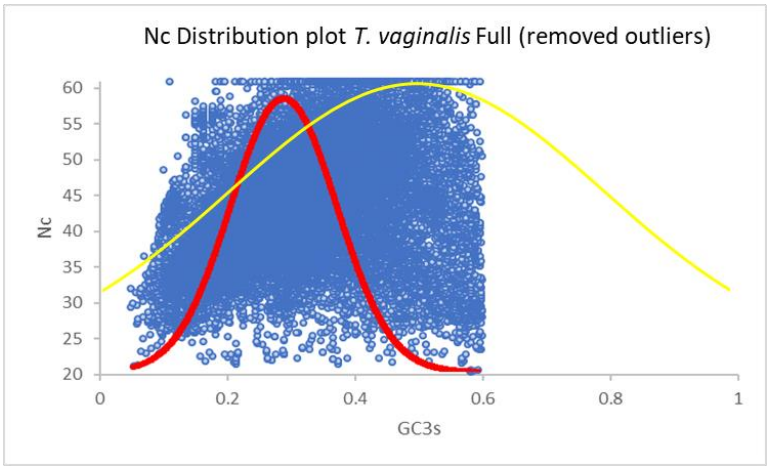

**Figure 3.2: Nc distribution plot for *T. vaginalis* after removal of outliers with GC3 scores above 0.6. Nc distribution under no selection according to Wright (1990) = **

The results of the tRNA amino acid analysis all matched the predicted optimal codons suggested by McInerny, 1997, with the amino acids that possess multiple potential optimal codons and their codon preferences outlined in table 3.1. All other amino acids possessed only a single potential optimal codon, with the exception of glutamine (Q) and lysine (K), both of whom had only two potential optimal codons.

tRNA Analysis Results - Predicted Optimal Codons (OC) in Multiple Codon Amino Acids (AA)					
AA	Number of tRNA Genes Identified	Number of Potential Codons	Most Abundant Codon (Anticodon)	Total in tRNA Genes	% Frequency
Ala	28	3	AGC (GCU)	19	67.86
Arg	30	5	ACG (CGU)	16	51.61
Glu	38	2	TTC (GAA)	24	63.16
Gly	28	3	GCC (GGC)	18	62.07
Ile	30	2	GAT (AUC)	20	66.6
Leu	37	5	AAG (CUU)	19	51.3
Phe	17	2	GAA (UUC)	15	88.24
Pro	16	3	CCA (TGG)	11	68.76
Ser	30	6	AGA (UCU)	12	40.0
Thr	25	3	TGT (ACA)	18	72.0
Val	25	3	GAC (GUC)	21	84.0

Table 3.1: Predicted optimal codons in multiple codon amino acids. Provides total number of tRNA genes identified for each amino acid as well as which codon/anticodon for each amino acid has the largest number of tRNA genes associated with it and its percentage frequency in relation to the other potential codons for that amino acid.

3.2 Genome filtering and identifying the true coding genes of *T. vaginalis*

The results from EggNOG Mapper2, indicated that only approximately 45,000 of the genes in the database could be categorised into any functional classes, suggesting that more than half the genes in the TrichDB may not be genuine genes for *T. vaginalis*. After removing those genes that had a KOG category of 'S' (indicating unknown function), only 13,422 genes remained. The genes with no KOG category returned no results after being BLASTed on NCBI; as for those with a category of 'S', all returned with the same result "partial mRNA, hypothetical protein", with some having extremely low query coverage and percentage identity and almost all being incomplete on both ends, while the 13,422 genes, that were suspected to be the coding genes for *T. vaginalis*, all corresponded to a recognised sequence or protein.

The Nc and GC3 analysis of the KOG^{+} genes required the removal of 14 sequences as codonW v1.4.2 (Peden, 1999) was unable to produce any Nc values for them which later, after BLASTing, was determined to be due to these sequences being incomplete and 5 at the end being transposable elements (potentially resulting in them lacking degeneracy groups, thereby preventing Nc values from being determined). However, a new Nc plot was ultimately generated and then compared to the full dataset as outlined in fig 3.3. The results of this analysis demonstrated a clear reduction in the overall distribution of the data as well as a fairly significant shift towards GC compared to the full dataset. As well as this, unlike the full dataset were the overall average for GC3 was 0.29, the KOG^{+} dataset's average was

0.40. This suggests that while an AT bias still exists within the KOG$^+$ genes, it is considerably weaker compared to the full original gene dataset.

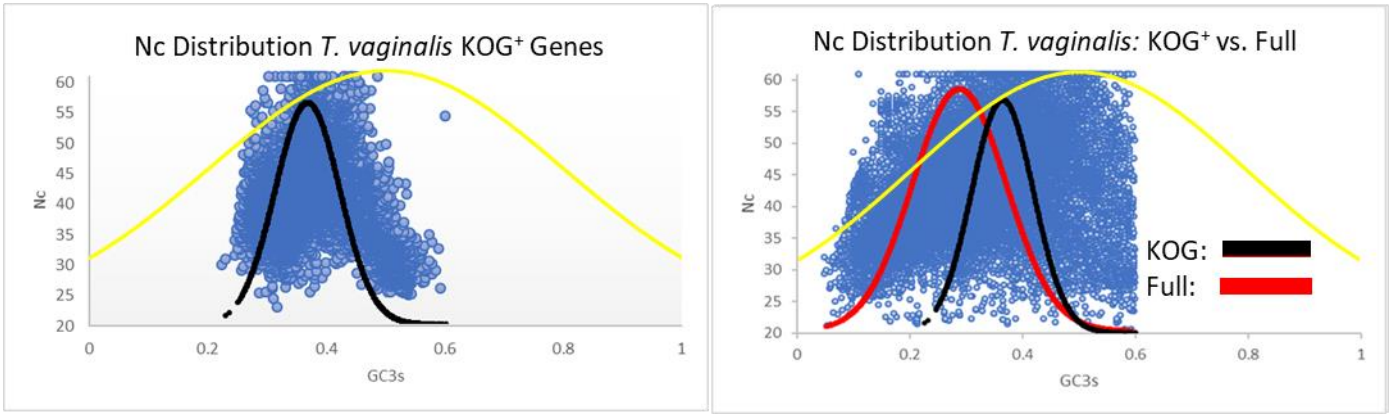

Figure 3.3: GC3 distribution plot for *T. vaginalis* KOG$^+$ genes (left) and combined with original full dataset (right). Nc distribution under no selection according to Wright (1990) =

3.3 Identifying potential optimal codons

The optimal codon analysis of the major tRNAs (table 3.3) and the analysis of the top and bottom 5% of the KOG$^+$ genes, based on expression level (table 3.2), was able to identify the potential candidate OCs. When compared to the major tRNAs, the majority of the OCs in the top 5% matched those anticodons with the highest frequency, with some exceptions, such as aspartic acid, (which only had had one anticodon (GUC) in the major tRNA dataset, but the analysis of the top and bottom 5% indicated that GAU was the primary optimal codon rather than GAC) and valine (where GUC was the anticodon with the highest frequency in the tRNAs, but the analysis of the top and bottom 5% suggested that GUU was also an optimal codon, despite there being no AAC anticodon in the major tRNAs). Aspartic acid in particular also had the weakest p-value when the Chi^2 was performed, being only 0.0178 as opposed to all the other amino acids whose p-value was <0.0001, indicating that it likely does not have any true optimal codons.

While most amino acids had only a single potential optimal codon (based on their frequency in the top 5% coding dataset), some such as alanine, glycine, leucine, arginine, serine and valine possessed multiple potential optimal codons. After performing a further Chi^2 analysis on these additional codons, they were determined to also potentially be optimal. However, while the majority of amino acids primary OCs matched the predicted OCs from the major tRNAs there were some exceptions, namely leucine and serine, whose predicted OC in the major tRNAs was instead determined to likely be a secondary OC when analysing the top and bottom 5% gene dataset. As well as this, one of alanine's OC GCC as well as arginine's OC AGA, had no anticodon in the major tRNAs despite having a high frequency in the top 5% of the KOG$^+$ genes (as well as the bottom 5% in arginine's case).

KOG⁺ Gene Analysis: Top and Bottom 5%									
AA	Codons	Frequency		Chi²	AA	Codons	Frequency		Chi²
		Top	Bottom				Top	Bottom	
A	GCU	2.29	1.02	2537.013 p = <0.0001	M	AUG	1	1	
	GCC	1.02	0.39						
	GCA	0.65	1.67	958.119 p = <0.0001					
	GCG	0.04	0.91						
C	UGC	1.8	1.15	889.805 p = <0.0001	N	AAC	1.52	0.63	4363.240 p = <0.0001
	UGU	0.2	0.85			AAU	0.48	1.37	
D	GAU	1.27	1.23	39.105 p = 0.0178	P	CCA	3.58	1.72	4588.922 p = <0.0001
	GAC	0.73	0.65			CCU	0.27	0.87	
						CCC	0.06	0.42	
						CCG	0.09	0.99	
E	GAA	1.25	1.23	0.815 p = 0.3666	Q	CAG	1.35	0.25	4844.631 p = <0.0001
	GAG	0.75	0.77			CAA	0.65	1.75	
F	UUC	1.75	0.73	5170.796 p = <0.0001	R	CGU	2.27	0.99	1018.898 p = <0.0001
						CGC	2.19	0.98	
	UUU	0.25	1.27			CGA	0.05	0.3	933.024 p = <0.0001
						CGG	0.02	0.06	
						AGA	1.32	2.84	1496.215 p = <0.0001
						AGG	0.15	0.83	
G	GGC	1.88	0.81	1337.933 p = <0.0001	S	UCC	2.14	0.83	1614.125 p = <0.0001
						AGU	0.27	0.87	
	GGU	1.55	1.59	87.490 p = <0.0001		AGC	0.38	0.44	31.013 p = <0.0001
						UCU	1.57	1.59	
	GGA	0.57	1.27			UCA	1.55	1.55	321.775 p = <0.0001
	GGG	0.01	0.33			UCG	0.09	0.71	
H	CAC	1.47	0.08	1172.230 p = <0.0001	T	ACA	3.29	2.51	1047.852 p = <0.0001
	CAU	0.53	1.2			ACU	0.39	0.81	
						ACC	0.26	0.14	
						ACG	0.05	0.54	
I	AUC	2.1	0.95	4299.117 p = <0.0001	V	GUU	1.99	1.83	37.679 p = <0.0001
						GUC	1.8	0.75	
	AUU	0.83	1.36			GUA	0.19	0.94	2056.905 p = <0.0001
	AUA	0.08	0.68			GUG	0.03	0.48	
K	AAG	1.66	0.55	10788.820 p = <0.0001	W	UGG	1	1	
	AAA	0.34	1.45						
L	CUC	2.72	0.91	3225.919 p = <0.0001	Y	UAC	1.48	0.42	4547.030 p = <0.0001
	UUA	0.52	1.11						
	UUG	0.33	1.3						
	CUU	2.33	1.54	607.515 p = <0.0001		UAU	0.52	1.58	
	CUA	0.06	0.82						
	CUG	0.04	0.31						

Table 3.2: Summary of results of optimal codon frequency analysis and Chi² of top and bottom 5% of *T. vaginalis* KOG⁺ genes based on expression level. OC 1 [ ], OC 2 [ ], OC 3 [ ]

Anticodon Frequency in the Major tRNA's					
AA	Anticodons	Number of tRNA Genes	AA	Anticodons	Number of tRNA Genes
A	AGC	19	M	AUG	22
	CGC	2			
	UGC	7			
C	GCA	16	N	GUU	17
D	GUC	26	P	AGG	3
				CGG	2
				UGG	11
E	CUC	14	Q	CUG	9
	UUC	24		UUG	9
F	AAA	2	R	ACG	16
				CCG	2
				CCU	4
	GAA	15		UCG	1
				UCU	8
G	CCC	3	S	ACU	2
				AGA	12
	GCC	18		CGA	3
				GCU	5
	UCC	8		GGA	1
				UGA	7
H	GUG	10	T	AGU	5
				CGU	2
				UGU	18
I	GAU	20	V	CAC	1
				GAC	21
	UAU	10		UAC	3
K	CUU	20	W	CCA	5
	UUU	10			
L	AAG	19	Y	GUA	13
	CAA	5			
	CAG	2			
	UAA	9			
	UAG	2			

Table 3.3 Results of identified tRNA genes. Indicates number of tRNA genes per codon per amino acid (AA).

3.4 KOG category enrichment for OCs

While there is a distinct lack of research into the extent to which protein coding genes and genes that fall into one of the KOG categories in single-celled eukaryotes are enriched for OCs, some previous research, such as that of Southworth et al (2018), has demonstrated that this is definitely the case in some single-celled eukaryotes such as *Monosiga brevicollis.* Their research has demonstrated that the KOG category genes of *M. brevicollis* are definitely enriched for OCs, with KOG categories C and J being particularly highly enriched. Similar results were obtained in the KOG category gene analysis of *T. vaginalis* with the analysis of the KOG^{+} genes Fop, GC3 and Nc revealing that while, *T. vaginalis* has a low degree of bias towards optimal codons in most KOG categories, even when considering its preferred codons from the Fop results outlined in table 3.4. and with the Fop also being quite low in all KOG categories, there were three notable exceptions, categories B (genes involved in chromatin structure and dynamics), C (genes involved in energy production and

conservation) and J (genes associated with translation, ribosomal structure and biogenesis), all three of which had a high Fop (all above 0.5) compared to the other KOG categories, as well as the lowest overall Nc values (indicating a high degree of bias). KOG category J, in particular, held the highest level of bias as well as OC enrichment, with the highest Fop and GC3 values, as well as the lowest value for Nc of any other KOG category (table 3.4). The analysis of the KOG gene domains in the top 100 genes with the highest expression level revealed that the majority (~55%) of genes demonstrated a higher Fop in the domain region compared to the non-domain region (with a difference >0.05).

KOG Category	Fop	Nc	GC3
A	0.357	44.427	0.275
B	0.560	38.647	0.354
C	0.535	38.799	0.372
D	0.390	44.474	0.291
E	0.462	41.612	0.334
F	0.476	42.131	0.337
G	0.336	43.716	0.265
H	0.430	43.535	0.316
I	0.431	42.268	0.312
J	0.595	**37.341**	**0.408**
K	0.381	43.530	0.280
L	0.355	43.959	0.306
M	0.393	42.859	0.302
N	0.333	41.573	0.248
O	0.378	42.338	0.293
P	0.415	43.833	0.322
Q	0.382	42.811	0.286
T	0.361	44.602	0.295
U	0.415	43.114	0.324
V	0.366	42.638	0.292
W	0.437	43.780	0.293
Y	0.435	46.495	0.356
Z	0.413	41.404	0.321

Table 3.4 KOG category codon usage statistics results measuring optimal codon frequency (Fop), effective codon number (Nc) and GC nucleotides at synonymous 3rd positions (GC3s).

t-Test: Paired Two Sample for Means		
	Dom	***NoDom***
Mean	0.776	0.685
Variance	0.014	0.029
Observations	99	99
Pearson Correlation	0.602	
Hypothesized Mean Difference	0	
df	98	
t Stat	6.650	
P(T<=t) one-tail	8.392E-10	
t Critical one-tail	1.661	
P(T<=t) two-tail	**1.678E-09**	
t Critical two-tail	1.984	

Average SD (Domains) = **0.013**

Average SD (Non-Domains) = **0.029**

Table 3.5 t-tests for significance and standard deviation in domain vs non-domain regions in top 100 highly expressed KOG category genes.

3.5 Phylogenetic analysis of ADAT candidate genes

The results of the phylogenetic analysis suggest that the ADAT 2 and 3 genes are potentially genuine as the tree topology for these genes matched the expected species phylogeny for both ADAT 2 and 3 when rooted with the prokaryote TAD genes, particularly in the topology for the excavate species phylogeny (Fig 3.4). A similar result was also observed in the ADAT 1 genes when rooted with the *Homo sapiens* ADAR genes, which served as an appropriate outgroup as the *Homo sapiens* ADAR genes are paralogues whilst the ADAT genes tend to be homologues and were shown to be monophyletic as expected. However, whilst the topology of the gene phylogeny for ADAT 1 did not match the expected species phylogeny as well as ADAT 2 and 3, it still displayed an overall topology within the various phyla that matched what was expected, with some of the phylogenies also having relatively strong branch support, most notably among the Opisthokonts (fig 3.4). However, there were some exceptions such as *D. discoideum* which grouped together with *H. sapiens* and *D. melanogaster* with strong branch support (0.96), however this is likely due to it being an amoebozoan, making it more closely related to animals than the other phyla (fig 3.4). While *T. vaginalis* ended up grouping together with *P. tricornutum* (a stramenopile and diatom) in the ADAT1 phylogeny, this relationship lacked any significant branch support as it only produced a posterior probability of 0.54. It did still group together with most of the other excavates, as for the ADAT2 and 3 phylogenies, and once the phylogeny was rerooted with the TAD genes from the prokaryotes, *T. vaginalis* separated into the correct phylogenies along with the other excavates, in particular *T. foetus*, a genus of *Tritrichomonas* that is closely related to *T. vaginalis*.

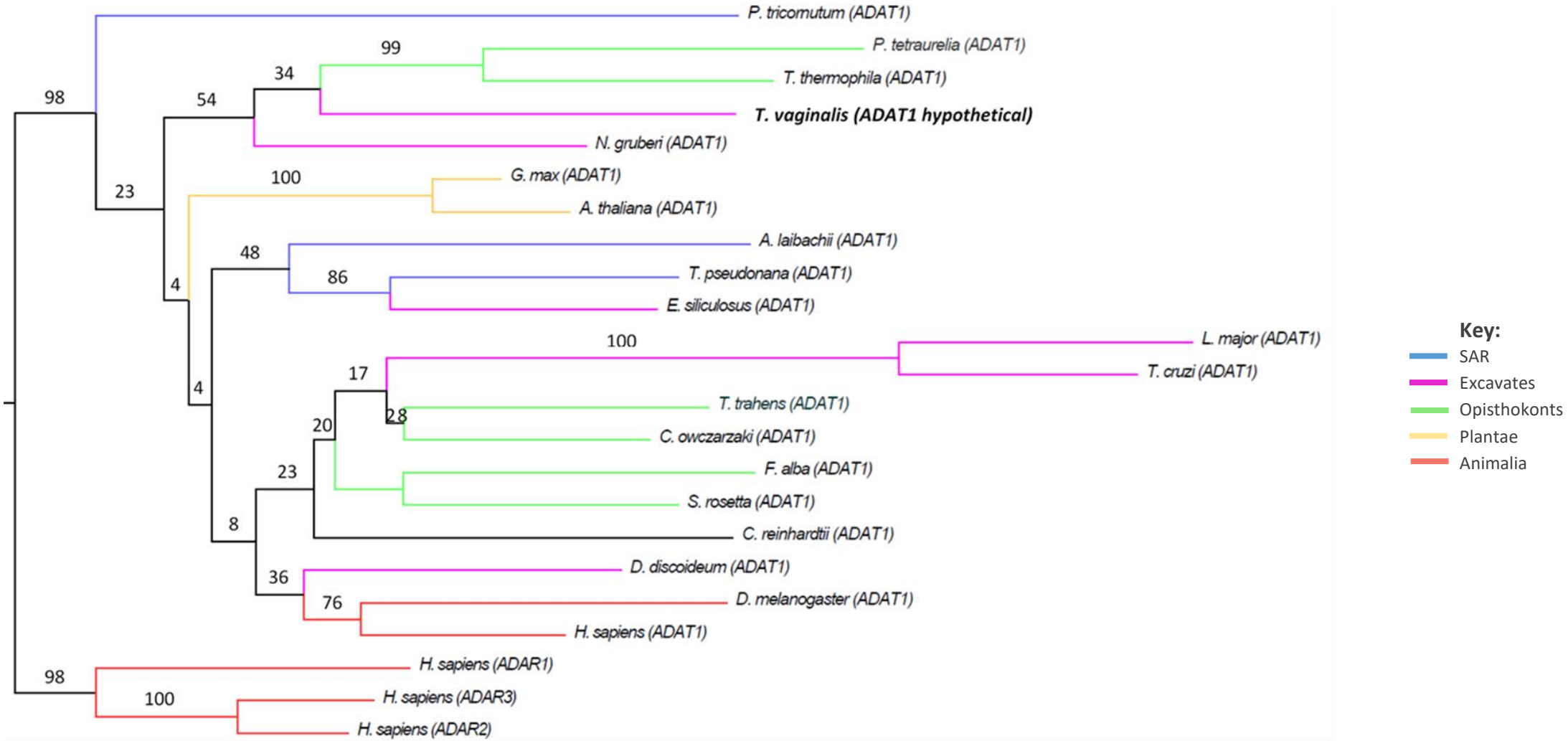

Figure 3.4.1 ADAT1 phylogeny using RaxML (maximum likelihood) and visualised using Figtree v1.4.4. Rooted with *H. sapiens* ADAR (adenosine deaminase acting on RNA) genes. *T. vaginalis* highlighted in bold and bootstrap support values outlined on branches.

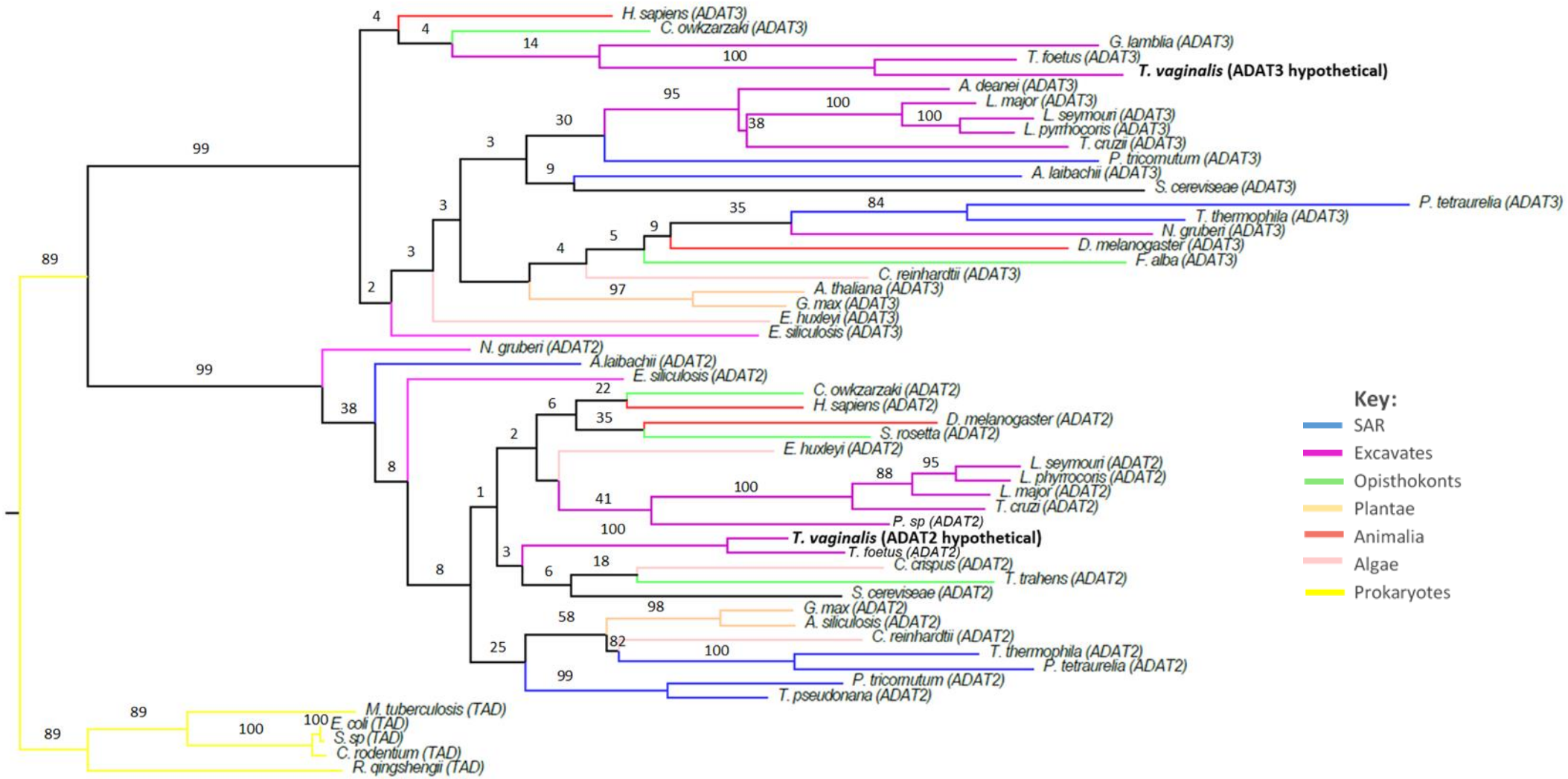

Figure 3.4.2 ADAT2 +3 phylogeny using RaxML (maximum likelihood) and visualised using Figtree v1.4.4. Rooted with the TAD (tyrosine adenosine deaminase) genes from several prokaryotes. *T. vaginalis* highlighted in bold and bootstrap support values outlined on branches.

3.6 Mutation pressure's effect on codon usage bias in *T. vaginalis*

Mutation pressure is a process where a series of recurrent mutations which occur throughout the genome of an organism produce a measurable change in the base composition frequency in that organism's genome. The effect of mutation pressure is generally quite weak, with most organisms displaying genetic drift as a more significant evolutionary force (Sueoka, 1988; Kliman and Hey, 1994). However, in many organisms it has been observed that if the mutation rate is high enough it can start to influence the pattern of codon usage in that organism's genome, in particular decreasing or increasing the direction of bias towards particular codons (Kilman and hey, 1994). In *T. vaginalis*'s genome the AT bias of the overall genome would suggest that there is a stronger directional mutation rate than in other organisms and the results of the codon bias analysis suggest this as well, as, despite the AT bias of the *T. vaginalis* genome, the codons that were determined to be optimal are predominantly GC ending. If this is indeed the case, then mutation pressure may have an influence on the codon usage across the different regions of the *T. vaginalis* genome as has been observed in other organisms such as *Brassica campestris* (Paul et al, 2018). If there is a statistically significant difference between the GC content in the regions surrounding the genes that are highly expressed and the regions surrounding the genes that are weakly expressed then this would suggest that mutation pressure has a degree of influence over the variation in GC3s observed in *T. vaginalis*.

The results of the two-sample paired t-test of mean variance that was performed on the top and bottom 100 genes produced a p-value of 0.668 (table 3.5). This suggests that there is no statistically significant difference between the highly and weakly expressed genes with regards to local mutation pressure, suggesting that mutation pressure is not significantly influencing the differences in codon usage between these two categories. The boxplot for the two datasets (fig 3.5) also displayed no significant difference in their average variation. While the top 100 genes had a narrower distribution, the median average GC content for both varied by only approximately 0.02. When the mid expression genes were added and a new boxplot produced (fig 3.6), the datasets appeared to have an average GC content of between 0.26 and 0.28. However, the one-way ANOVA p-value (0.0020) indicated that there was now a statistically significant difference between the results and this is likely due to the inclusion of the mid-expression genes as they had a noticeably lower GC content average in the flanking DNA compared to the top and bottom 100.

When the GC content of the non-coding flanking DNA was compared to the GC3s of the coding DNA, the top 100 highly expressed and bottom 100 weakly expressed genes produced a positive relationship, while the mid expression genes produced a negative relationship (fig 3.8). However, the t-test indicated that there was no statistically significant difference between non-coding GC content and coding GC3s in the mid expression genes (p=0.325), and while the bottom 100 genes p-value of 0.036 suggests that its results are statistically significant, it is quite weak (fig 3.7). Only the top 100 gene dataset's results were deemed to be highly statistically significant ($p = 2.877 \times 10^{-31}$).When the datasets were then combined they produced a strong positive relationship (fig 3.8), and this may be due to the GC3 averages in the top 100 gene dataset, as, in fig 3.7, the mid and bottom expression genes only varied, in terms of GC3 averages, by ~0.01, whereas the GC3 averages in the top 100 genes were significantly higher, differing from the mid and bottom 100 by ~ 0.2 (fig 3.7). As well as this, the t-test for significance indicated that the results were statistically

significant with a p-value of 8.625×10^{-12} (fig 3.9), potentially suggesting that the effect of local mutation pressure varies depending on a gene's expression level, as the AT bias observed within the genome is more significant the lower the expression level, potentially suggesting that mutation pressure is influencing the direction of bias towards AT in the genes that have a weaker expression level.

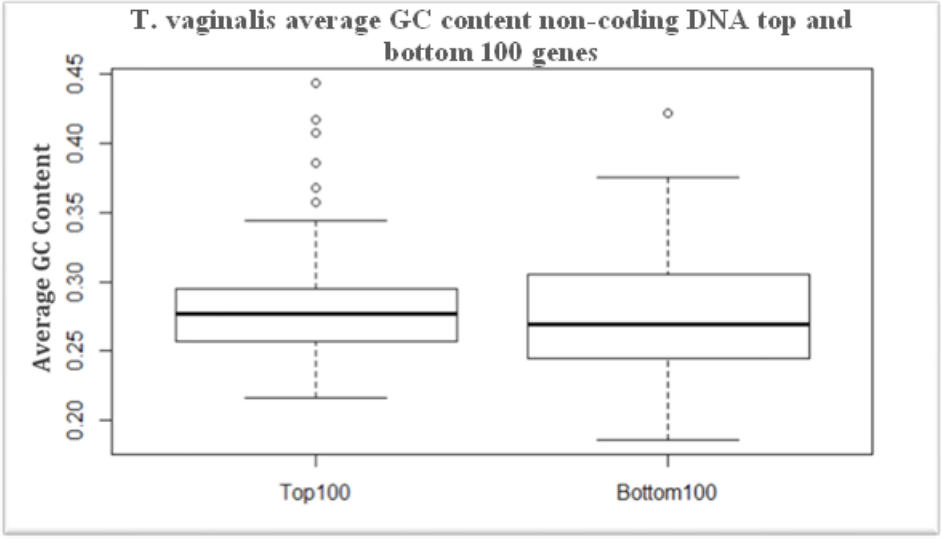

Figure 3.5 Results of analysis of non-coding flanking DNA GC content averages in top and bottom 100 genes based on expression level

t-Test: Paired Two Sample for Means		
Top, Bottom (Flanking GC)	***Top100***	***Bottom100***
Mean	0.280	0.278
Variance	0.00160	0.0020
Observations	100	100
Pearson Correlation	-0.0338	
Hypothesized Mean Difference	0	
df	99	
t Stat	0.430	
P(T<=t) one-tail	0.334	
t Critical one-tail	1.660	
P(T<=t) two-tail	0.668	
t Critical two-tail	1.984	

Table 3.5 Results of t-test for statistical significance: flanking DNA GC content averages in top and bottom 100 genes based on expression level

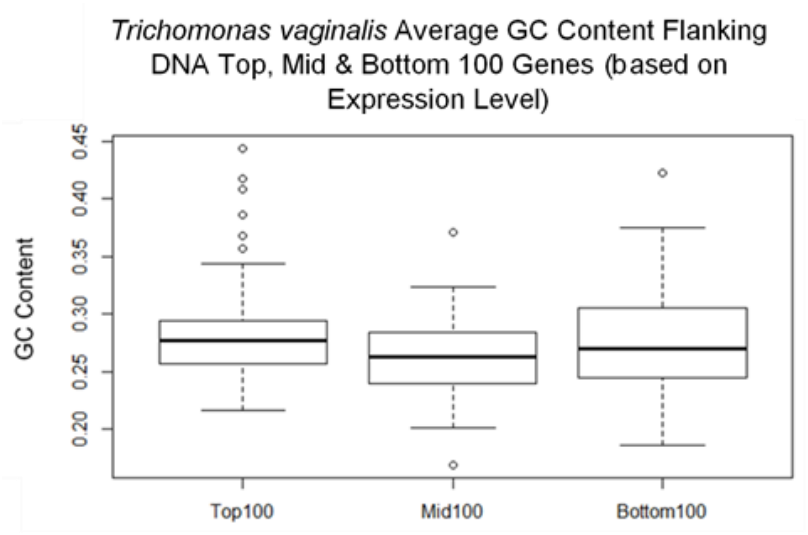

Figure 3.6 Results of analysis of Flanking DNA GC content averages in top, mid and bottom 100 genes based on expression level

SUMMARY				
Groups	***Count***	***Sum***	***Average***	***Variance***
Top 100	100	28.0150	0.2802	0.0016
Mid 100	100	26.1750	0.2618	0.0011
Bottom 100	100	27.7520	0.2775	0.0020

ANOVA						
Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	0.0198	2	0.0099	6.3242	0.0020	3.0262
Within Groups	0.4651	297	0.0016			
Total	0.4849	299				

Table 3.6 One-way ANOVA results and summary for Top, Mid and Bottom 100 genes flanking DNA GC content averages and statistical significance. α = 0.05.

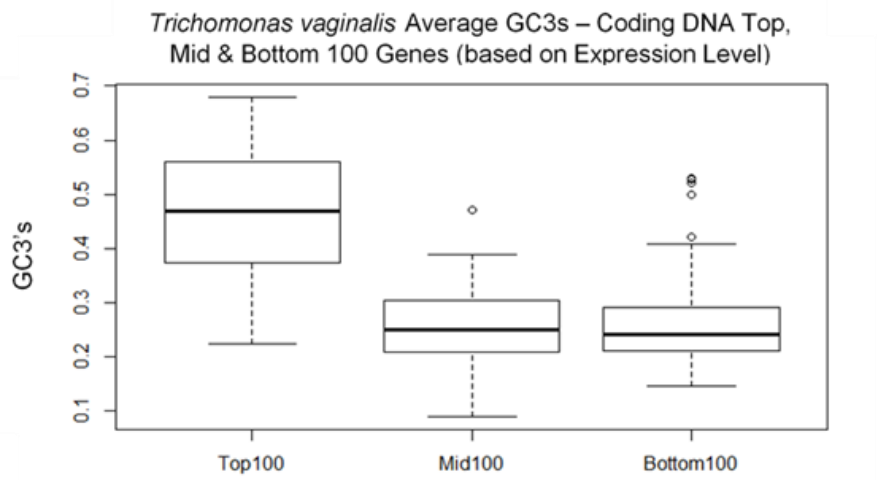

Figure 3.7 Results of analysis top, mid and bottom 100 genes coding regions (based on expression level)

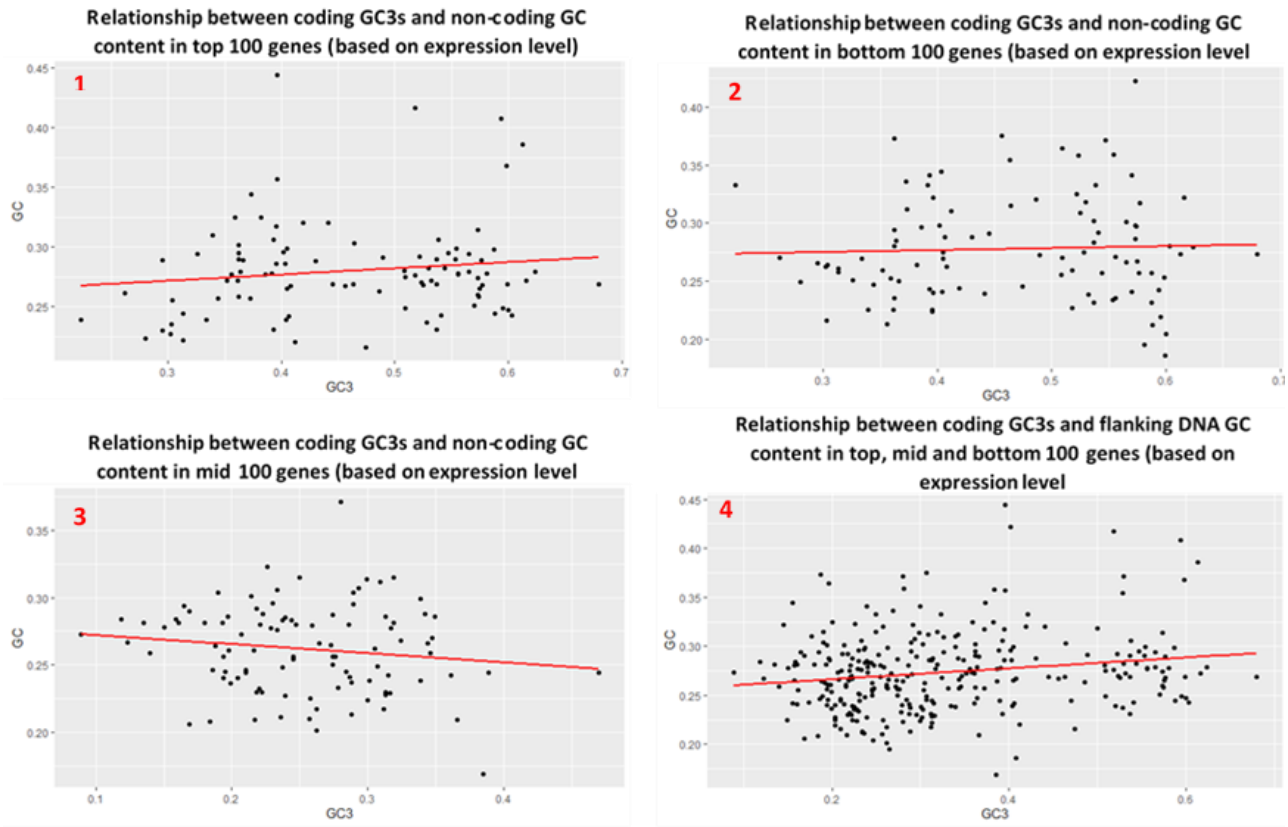

Figure 3.8 comparison of the relationship between the coding DNA GC3s and flanking DNA non-coding GC content in the top 100 highly expressed [1], Bottom 100 weakly expressed [2], mid 100 [3] and combined expression genes [4].

t-Test: Paired Two Sample for Means	**Top100**	
	GC	*GC3*
Mean	0.28	0.46
Variance	0.0016	0.0112
Observations	100	100
Pearson Correlation	0.138	
Hypothesized Mean Difference	0.5	
df	99	
t Stat	17.076	
P(T<=t) one-tail	1.438E-31	
t Critical one-tail	1.6604	
P(T<=t) two-tail	2.877E-31	
t Critical two-tail	1.984	

t-Test: Paired Two Sample for Means	**Mid 100**	
	GC	*GC3*
Mean	0.262	0.254
Variance	0.001	0.004
Observations	100	100
Pearson Correlation	-0.138	
Hypothesized Mean Difference	0	
df	99	
t Stat	0.988	
P(T<=t) one-tail	0.163	
t Critical one-tail	1.660	
P(T<=t) two-tail	0.325	
t Critical two-tail	1.984	

t-Test: Paired Two Sample for Means	**Bottom 100**	
	GC	*GC3*
Mean	0.278	0.260
Variance	0.002	0.007
Observations	100	100
Pearson Correlation	0.286	
Hypothesized Mean Difference	0	
df	99	
t Stat	2.128	
P(T<=t) one-tail	0.018	
t Critical one-tail	1.660	
P(T<=t) two-tail	0.036	
t Critical two-tail	1.984	

t-Test: Paired Two Sample for Means		
Top, Mid, Bottom	*GC*	*GC3*
Mean	0.273	0.326
Variance	0.002	0.017
Observations	300	300
Pearson Correlation	0.180	
Hypothesized Mean Difference	0	
df	299	
t Stat	-7.109	
P(T<=t) one-tail	4.31266E-12	
t Critical one-tail	1.650	
P(T<=t) two-tail	8.625E-12	
t Critical two-tail	1.968	

Table 3.7 t-test results (α = 0.5) for Top 100, Bottom 100, Mid 100 and combined coding GC3 vs flanking DNA GC content. P-value is highlighted for reference.

3.7 The selection of OCs in the transposable elements of *T. vaginalis*

The analysis of the transposable element protein sequences resulted in the sequences being placed into either the long tandem repeat (LTR) retrotransposon family *Gypsy*, *Copia* and *IS3EU* or the following DNA transposable element families: *Kolobok-T2, Merlin/Maverick, MULE-MuDR, PIF-Harbing, TcMar-Marin* or *TcMar-Fot1*. The analysis of each transposable element family indicated that, while 3 had a positive relationship between Fop and expression level (namely the LTR retrotransposons as well as the *Maverick* and *Merlin* DNA transposons), most either had a negative relationship between Fop and expression level (*Kolobok-T2*, *Pif-Harbing* and *TcMar-Marin*) or no relationship, namely *TcMar-fot1* (Fig 3.10). However, the *Maverick* and *Merlin* transposon datasets only consisted of four transposons total, which makes drawing conclusions from these datasets difficult, thus they were excluded from any further analysis. When the entire dataset is considered, however, there is little evidence of any relationship between the expression level of DNA transposable elements and the frequency of optimal codons, however, the LTR retrotransposons did display a positive relationship indicating that in these transposons, as expression level increases, so too does optimal codon frequency.

When exploring the relationship between the domain regions and non-domain regions of the transposons, only eleven transposons produced a Fop value above 0.5 in the domain regions compared to only two that produced this in the non-domain regions. However, when comparing the overall average Fop in the domain and non-domain regions for the transposable elements (Fig 3.9), the domain region produced an average Fop of 0.338 while the non-domain regions produced a Fop of 0.326. While this difference is relatively small, the t-tests for significance produced a p-value of 2.3 × 10E-4 (table 3.8) suggesting that the difference between the domain and non-domain regions is statistically significant. These results provide evidence that there may be selection for optimal codons in the transposable elements in *T. vaginalis* and that this selection may be biased towards translational accuracy. Further, the bar chart outlining the differences between the average Fop in the domain and non-domain regions of each transposable element type (fig 3.11) indicated that on average, all transposable element families had a higher Fop in their domain regions, with the exception of *DNA/Pif-Harbing* and *DNA/TcMar-Marin,* which had a higher Fop in their non-domain regions. However, when analysed as a whole, irrespective of transposon type, 53.82% of the transposable element sequences had a higher Fop in the domains compared to 45.15% which had a higher Fop in the non-domains. While this would appear to be a significant difference, many of the sequences only had a difference of ~0.009 or less. When the transposable element families that only have a small number of sequences such as *DNA/TcMar-Tc1, DNA/Maverick* and *DNA/Merlin* were removed, only two transposable element types have a higher Fop in the non-domain regions, compared to four in the domain regions, and the t-tests performed on these transposable elements indicated that all results were statistically significant. (table 3.9).

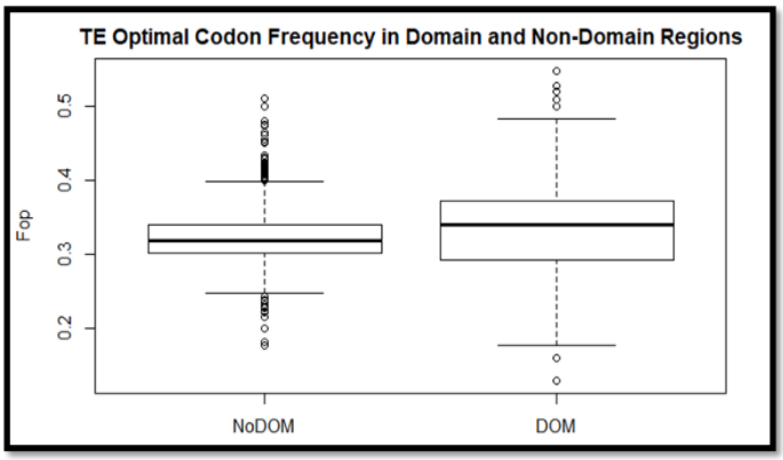

Figure 3.9 boxplot of Fop in domain and non-domain regions of the transposable elements

t-Test: Paired Two Sample for Means		
	Variable 1	*Variable 2*
Mean	0.334	0.325
Variance	0.004	0.002
Observations	680	680
Pearson Correlation	0.366	
Hypothesized Mean Difference	0	
df	679	
t Stat	3.708	
P(T<=t) one-tail	0.000	
t Critical one-tail	0	
P(T<=t) two-tail	0.00023	
t Critical two-tail	0.675	

Table 3.8 t-tests of statistical significance of transposable element Fop in domain vs non-domain regions

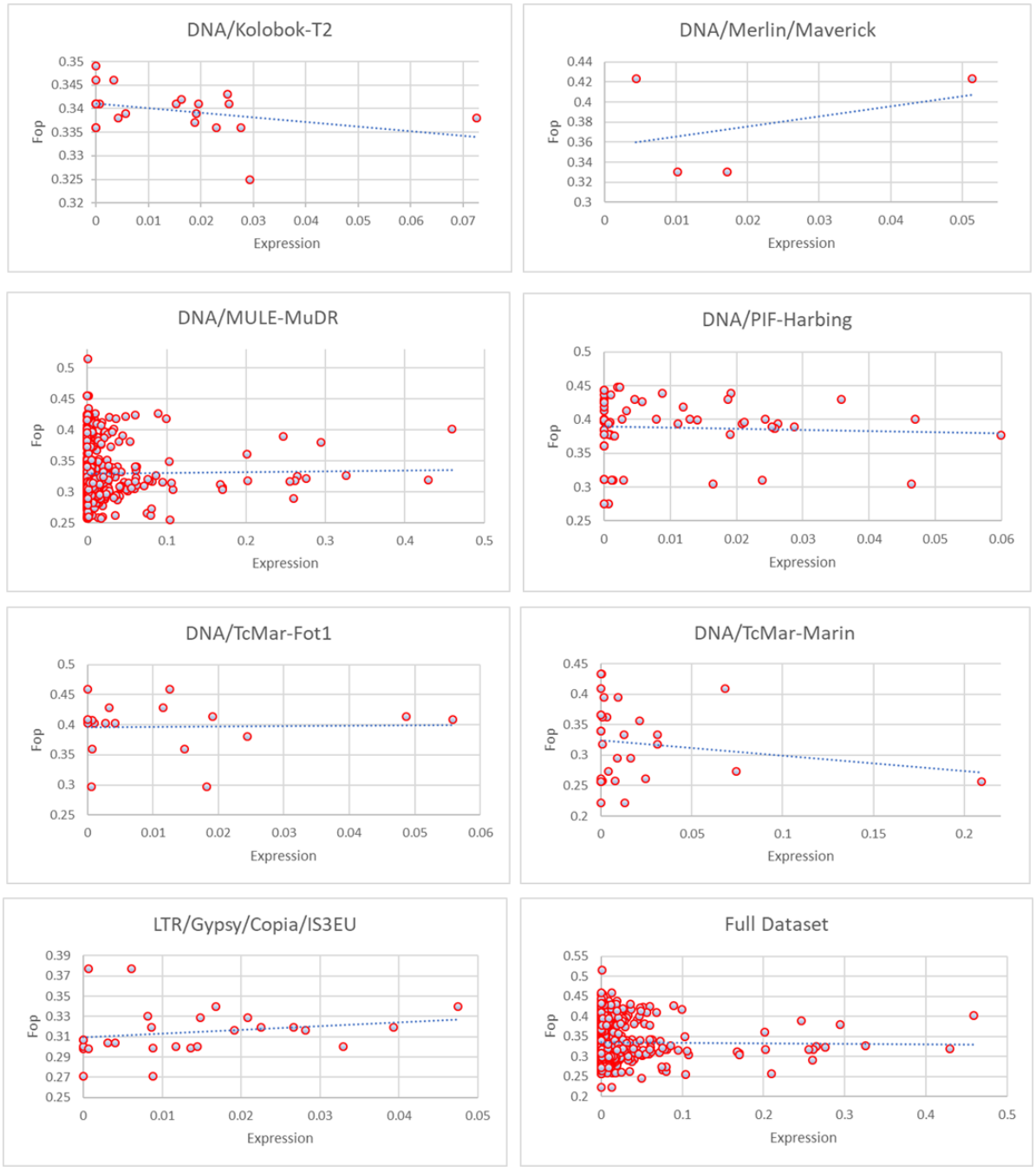

Figure 3.10 Scatter plots of relationship between Fop and expression level in the transposable elements by transposon family/type and the full transposon dataset.

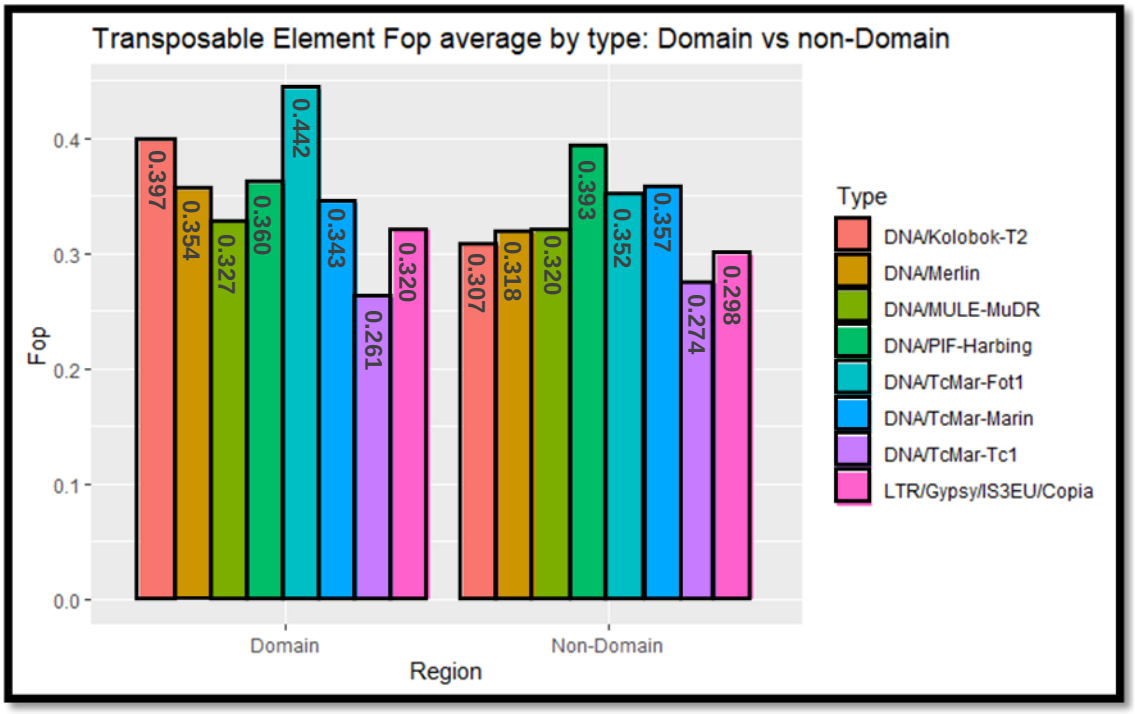

Figure 3.11 bar chart visualising domain vs non-domain Fop by transposable elements by family/type. Average Fop given with each transposable element type.

t-Test: Paired Two Sample for Means — DNA/PIF-Harbing

	Variable 1	*Variable 2*
Mean	0.359937	0.39251
Variance	0.002115	0.00280
Observations	63	63
Pearson Correlation	0.474340	
Hypothesized Mean Difference	0	
df	62	
t Stat	-5.065004	
P(T<=t) one-tail	0.000002	
t Critical one-tail	1.669804	
P(T<=t) two-tail	3.93E-06	
t Critical two-tail	1.998972	

t-Test: Paired Two Sample for Means — DNA/TcMar-Marin

	Variable 1	*Variable 2*
Mean	0.34263	0.35675
Variance	0.00442	0.00838
Observations	8	8
Pearson Correlation	0.60448	
Hypothesized Mean Difference	0	
df	7	
t Stat	-0.54161	
P(T<=t) one-tail	0.30245	
t Critical one-tail	1.89458	
P(T<=t) two-tail	0.60489	
t Critical two-tail	2.36462	

t-Test: Paired Two Sample for Means — DNA/MULE-MuDR

	Variable 1	*Variable 2*
Mean	0.32869	0.32073
Variance	0.00288	0.00098
Observations	529	529
Pearson Correlation	0.27201	
Hypothesized Mean Difference	0	
df	528	
t Stat	3.37592	
P(T<=t) one-tail	0.00039	
t Critical one-tail	1.64774	
P(T<=t) two-tail	0.00079	
t Critical two-tail	1.96447	

t-Test: Paired Two Sample for Means — LTR/Gypsy/IS3EU/Copia

	Variable 1	*Variable 2*
Mean	0.31982	0.29786
Variance	0.00080	0.00117
Observations	28	28
Pearson Correlation	0.02653	
Hypothesized Mean Difference	0	
df	27	
t Stat	2.64966	
P(T<=t) one-tail	0.00665	
t Critical one-tail	1.70329	
P(T<=t) two-tail	0.01330	
t Critical two-tail	2.05183	

t-Test: Paired Two Sample for Means — DNA/Kolobok-T2 (only 4!)

	Variable 1	*Variable 2*
Mean	0.3965	0.307
Variance	0.00519	0.00067
Observations	4	4
Pearson Correlation	-0.25915	
Hypothesized Mean Difference	0	
df	3	
t Stat	2.16597	
P(T<=t) one-tail	0.05945	
t Critical one-tail	2.35336	
P(T<=t) two-tail	0.11891	
t Critical two-tail	3.18245	

t-Test: Paired Two Sample for Means — DNA/TcMar-Fot1

	Variable 1	*Variable 2*
Mean	0.44175	0.35175
Variance	0.00978	0.00337
Observations	20	20
Pearson Correlation	-0.11707	
Hypothesized Mean Difference	0	
df	19	
t Stat	3.34263	
P(T<=t) one-tail	0.00171	
t Critical one-tail	1.72913	
P(T<=t) two-tail	0.00342	
t Critical two-tail	2.09302	

Table 3.9 t-tests of average domain and non-domain Fop results for each transposon type with higher Fop in domain regions (■), and higher Fop in non-domain regions (■). Variable 1 = domain average, variable 2 = non-domain average.

4.0 Discussion

4.1 Reassessing the genome of *T. vaginalis*

All initial research into the *T. vaginalis's* genome, after it had been fully sequenced, has presented it as unusual, primarily due to its heavy AT bias (McInerny, 1997) which was also observed in this research's analysis of the full original dataset (fig 3.1), but also due to its large potential gene content, appearing to possess well over 95,000 potential genes (Carlton et al, 2013; Harp and Chowdhury, 2011). However, the results from the EggNOG gene mapper v5.0.0 (Huerta-Cepas et al, 2019) suggest that the *T. vaginalis* gene content may not be as large as was proposed initially by Harp and Chowdhury (2011). Due to EggNOG v5.0.0 (Huerta -Cepas et al, 2019) being unable to assign more than half the proposed genes in the TrichDB a recognised KOG category, they may be, as Whoele et al (2014) suggested, be pseudogenes and non-coding RNAs which have been incorrectly annotated. However, while it is more likely that the gene annotations on the TrichDB were erroneous, it is also possible that the genes that failed to return any annotations are genuine but possess a novel domain structure. As well as this, it is also possible that the non-expressed genes are only expressed in cell types that were not present/used in the cultures that were used in the transcriptome analysis. The erroneous annotations proposal is, for now, more plausible and this may have happened due to the repetitive nature of the *T. vaginalis*'s genome, as this would likely make determining the correct coding genes quite difficult. However, despite the vast majority of the genes potentially being pseudogenes/intergenic DNA, (which initially would explain the large AT bias of the original dataset), the bias didn't disappear, as even once its genome had been reduced to only the 13,422 coding genes, bringing it in line with other excavates like *T. brucei* whose genome consists of ~10,700 genes (Berriman et al, 2005), there was still a bias towards AT, albeit significantly smaller compared to the original sequenced dataset by Harp and Chowdhury (2011).

The synonymous AT bias observed in *T. vaginalis*, while potentially reduced by this research, is still present, however the large shift in GC3 average distribution between the original and suspected coding datasets (Fig 3.3.), is good evidence that these genes are potentially protein coding, as most other eukaryotes tend to display a bias towards G or C at their synonymous third position in their protein coding genes. Similar results have been observed in some trypanosomes, another group of excavates related to *T. vaginalis*, which do not display a strong bias in either direction but what bias does exist, tends to be towards AT whereas the excavate groups *Leishmania* and *Crithidia* tend to display a fairly strong bias towards GC (Subramanian and Sarkar, 2015). As for the Nc averages for *T. vaginalis*, these were similar to what was observed in some choanoflagellates which tended to have an Nc average of ~44-46 when under relatively strong selection and only having a small skew towards GC (Southworth et al, 2018). Altogether, this presents the *T. vaginalis* genome as more similar in size of presumed gene content to many other excavate genomes and although there is a bias towards AT at synonymous sites this bias is not as significant as initially proposed (Harp and Chowdhury, 2011).

4.2 *T. vaginalis* optimal codons:

4.2.1 Discrepancies between major tRNA genes and optimal codons

In most organisms, the major tRNA genes can be used as a good way of predicting the potential OCs of that genome (Ikemura, 1985), however, in *T. vaginalis*, while the CoA was

able to identify the primary OCs (table 3.2), they did not all match the predicted OCs from the major tRNAs (table 3.3). The most notable of these was aspartic acid and valine, both of whom had OCs whose anticodon did not even appear in the major tRNA dataset, as well as leucine and serine, who had additional OCs which simply differed from what was expected/predicted. These discrepancies between the predicted OCs in the tRNAs and the codons which display the highest frequency in the top and bottom 5% of the coding dataset does potentially suggest that forces other than purely selection may have a noticeable influence on codon usage and bias in *T. vaginalis*'s genome, as selection for optimal codons is clearly not purely down to the abundance of the major tRNA genes for that codon (Atherton, Sharp and Lafay, 2000; Sharp et al, 2008). While selection is likely still the primary driving force behind the determination of optimal codons, other forces such as G:U wobble binding may also have an influence (albeit small) on the selection of optimal codons which are less efficient, as these codons are still more advantageous than non-optimal codons. As well as this, while original studies into the relationship between the abundance of major tRNA genes and the selection of optimal codons was able to identify a correlation, these studies were predominantly performed in bacteria, which have far fewer tRNA genes than eukaryotes (Atherton, Sharp and Lafay, 2000; Sharp et al, 2008). However, while a eukaryotic genome may possess a large number of tRNA genes, all genes are not necessarily expressed at equal levels, and as well as this, some may be pseudogenes. Therefore, while there may initially be a correlation between the number of genes and expression level, this may not always hold true. Despite this, the majority of the amino acids matched the major tRNA dataset and this is evidence for translational accuracy at the level of codon usage.

While these discrepancies differ from what was expected, there are a few potential explanations. Initially it was determined that there is more than one optimal codon for some of the amino acids, and a further analysis of the dataset confirmed this in the amino acids: alanine, leucine, arginine, valine and serine. However, In the case of valine, the lack of a major tRNA gene for its optimal codons corresponding anticodon could be explained by the anticodon for GUC also potentially bind to GUU through 'wobble binding', where 'G' in the anticodon also binds to 'U' in the synonymous position in the codon via non-Watson & Crick base pairing (Murphy and Ramakrishnan, 2004). The same could also apply with regards to GAU in aspartic acid (which also lacked a corresponding major tRNA gene), and while this is not as efficient as the binding of GUC or GAC, these alternative codons may be favoured due to the AT mutation pressure of the *T. vaginalis* genome (Conrad et al, 2013).

In the case of the other amino acids whose OC did not match the tRNA dataset, namely leucine, arginine, serine and alanine, there was likely potentially more than one OC. However, when analysing these secondary OCs their high degree of bias in the top and bottom 5% dataset did not correspond with a high abundance in the corresponding major tRNA genes dataset. One example is that of arginine, as the CGC secondary OC (which was favoured at an almost equal rate to the CGU primary OC) lacked any major tRNA genes, much like valine and aspartic acid. It is unlikely that the bias towards CGC in arginine is due to 'wobble binding' like in the case of valine and aspartic acid, as CGC lacks any AT nucleotides to support this, and it is unlikely that selection would favour this type of binding in so many amino acids.

One explanation for this phenomenon which would also give *T. vaginalis* another way of overcoming the limitations imposed on translation by the AT bias of its genome (Conrad et

al, 2013) whilst still maintaining its accuracy is through the deamination of adenosine in the tRNAs to inosine at the wobble position, which has a larger binding profile than adenosine as it can bind to cytosine, adenosine and uracil (Grosjean et al, 1996; Rafels-Ybern et al, 2019). This phenomenon was first identified by Holley et al (1965) in yeast tRNA where inosine replaced adenosine at the wobble position of the IGC anticodon in alanine; following this adenosine deamination has been observed in a number of other organisms, with varying frequency as some organisms, such as *M. capricolum* only possess a single inosine while others have multiple inosines throughout their major tRNA genes (Andachi et al, 1987; Grosjean et al, 1996). This would potentially explain some of the optimal codon results observed in the TAPSILVR amino acids, as many possess optimal codons that suggest that their tRNA's are undergoing deamination.

4.2.2 Codon usage in other excavates and KOG category enrichment

When comparing the codon usage of *T. vaginalis* to that of other excavates there is some noticeable variation in codon usage bias between them. Overall the codons that appeared more frequently in *T. vaginalis* top and bottom 5% genes based on expression level (table 3.2) tended to match those preferred codons among *Leishmania* and *Crithidia*, in particular CGC (Arg), AUC (Ile), GCC (Ala), UUC (Phe) and GUC (Val) (Subramanian and Sarkar, 2015). While the codon usage of the trypanosomes primarily matched *T. vaginalis* in which codons were less preferred such as CUG/UUA (Leu), GUA (Val), ACC (Thr) and UCG (Ser), with some codons that were preferred by the trypanosomes, such as GUG (Val), being rare in *T. vaginalis*, which may potentially be due to inosine not binding to guanosine, causing codons with G at the wobble position in valine to be rare. This is somewhat unexpected as the trypanosomes tend to display a slight bias towards A and T with regards to their codon usage, much like *T. vaginalis,* while the *Leishmania* and *Crithidia* genomes tend to display a bias towards G or C (Subramanian and Sarkar, 2015).

While the codons for *T. vaginalis* did display particular preferences in most amino acids, suggesting a certain degree of bias, this was not the case when analysing the KOG category domain regions, as the majority of the KOG categories demonstrated a relatively low overall bias. However, the 3 KOG categories that were the exception, B, C and J, whilst still having a low overall GC3s average had much higher averages for Fop and much lower averages for Nc (table 3.4), suggesting that many of the genes within these KOG categories are significantly enriched for optimal codons, particularly category J. Category J predominantly consists of genes that code for structural components of the ribosome as well as a number of key genes involved in protein translation, all of which are known to be highly enriched in other organisms such as *M. brevicollis* which also demonstrated a higher OC enrichment level in KOG categories C and J (Southworth et al, 2018). Category C on the other hand, consists predominantly of genes that encode a number of enzymes, particularly hydrogenases and dehydrogenases, while category B predominantly consists of genes that encode proteins involved in heterodimerization as well as encoding a number of histones, in particular histone H2A. Histones tend to be highly expressed and also tend to evolve under strong purifying selection on their amino acid sequence (Duan et al, 2019), and due to this, histones are candidates to be evolving under both translational efficiency and accuracy with regards to their codon usage, which is consistent with the findings in this research. These are key proteins/enzymes and structural components that make up the genetic machinery of an organism and may explain why these particular KOG category genes are so highly

enriched for optimal codons. Many of these results are similar to those observed by Southworth et al (2018) in *M. brevicollis,* which was also significantly enriched for optimal codons in the KOG categories C and J. Had time permitted, further research into the expression levels of the genes that make up these three KOG categories would have been undertaken to determine if they truly are more highly expressed in *T. vaginalis* than the genes in other KOG categories (as their high GC3 expression would suggest). This would also provide evidence for *T. vaginalis* genome adhering mostly to the mutation-selection-drift balance model for codon usage proposed by Hershberg and Petrov (1996).

4.2.3 Adenosine deaminase acting on tRNA (ADAT)

The AT bias of *T. vaginalis*'s genome may potentially place a certain degree of pressure on its genome, particularly with regards to the selection of optimal codons. The deamination of adenosine in the tRNAs to inosine (Grosjean et al, 1996; Gerber and Keller, 1999; Murphy and Ramakrishnan, 2004) would explain some of the results seen in the optimal codons and may well be a solution to overcoming the restrictions the AT bias of its genome potentially places on translation. In order to determine if the deamination of adenosine was indeed occurring in the tRNAs an analysis was necessary to determine if the specific adenosine deaminase enzyme that is responsible for this phenomenon was present in *T. vaginalis*. Some single-celled eukaryotes do tend to display this phenomenon and possess the genes for the adenosine deaminase enzyme (Rafels-Ybern et al, 2019) although this has only been explored in a relatively small number of single-celled eukaryotes. The variant of adenosine deaminase which specifically deaminate adenosine in the tRNAs is encoded by the ADAT 1, 2 and 3 genes (Grosjean et al, 1996). While all three of these genes were potentially identified via BLASTing on NCBI, as they had not been characterised as ADAT genes, it was necessary to perform a phylogenetic analysis to determine if they are indeed potential candidates for adenosine deaminase acting on tRNA (Fig 3.4).

The phylogenies suggest that the ADAT genes are present within *T. vaginalis*'s genome, and therefore suggest that the deamination of adenosine may be occurring in the tRNAs. If this is the case then it would explain the bias that *T. vaginalis* has towards some of the optimal codons in the codon usage dataset of the reduced genome (table 3.2). This would also mirror what has been observed in other excavates such as *T. brucei*, one of the more well-known single-celled eukaryotes and excavates where the conversion of adenosine to inosine has been observed (Spears et al, 2011). However, while the results of the phylogenetic analyses provide evidence that these genes may be present within the *T. vaginalis* genome, and the results from the major tRNA analysis indicate that these genes are potentially functional, a further lab analysis of the gene products will need to be performed to fully confirm this. However, as many of the codons in the top and bottom 5% dataset matched those in *Crithidia* and *Leishmania* and those that matched were also ADAT codons (Subramanian and Sarkar, 2015), suggesting the possibility of shared translational mechanisms between these species.

4.3 Mutation pressure's influence on codon usage and the AT bias of the *T. vaginalis* genome

One of the potential key factors that could influence the selection of optimal codons is the pressure placed on the genome by recurrent mutations, as has been observed in a number of other organisms (Prosenjit et al, 2018; Jian-hua et al, 2014) as well as some of the other

excavates, including the trypanosomes, whose mutation pressure, like with *T. vaginalis*, pushes their genome towards AT (Subramanian and Sarkar, 2015). In the case of *T. vaginalis*, it was hypothesised that mutation pressure, particularly from the AT bias of its genome, may have an influence on the selection of optimal codons, and the results of the ANOVA (after including the mid-expression genes) and the analysis of domain GC3 expression in relation to non-domain GC content seem to support this, albeit predominantly for the mid-expression genes

The results of these analyses suggest that local mutation pressure does not differ significantly between genes with a high expression level and those with a weak expression level. Due to both local base composition and synonymous base composition displaying a skew towards AT, this suggests that mutation pressure is influencing the direction of bias with regards to codon usage (albeit more so in synonymous base composition than local base composition). However, the negative relationship between non-coding GC content and coding GC3s in the genes with mid-level expression and its weak statistical significance suggests that, in the regions of the genome where these genes are located, mutation pressure may not have a significant effect on codon usage or is being counteracted by selection. Overall this suggests that mutation pressure does potentially influence synonymous codon usage in the highly and weakly expressed genes, but is likely being counteracted by selection in the genes with mid-level expression. As well as this, while local non-coding GC content does vary between the high, mid and low expressed genes, this difference is small, with an average difference of 0.1-0.2 (table 3.6), and only gains statistical significance when adding the mid expression genes.

4.4 Evidence for codon usage bias in the transposable elements of *T. vaginalis*

Conrad et al (2013) suggested that the AT bias in the *T. vaginalis* genome may be due to the abundance of transposable elements throughout, and it has been observed in some choanoflagellates such as *S. rosetta* that selection for optimal codons does occur, particularly in the LTR retrotransposons, in contrast to many other eukaryotes which didn't display this (Southworth et al, 2019). The results of the analysis performed on the transposable elements suggest that *T. vaginalis* transposable elements do potentially select for OCs, particularly in the LTR retrotransposons as they alone displayed a positive relationship between Fop and expression level (similar to *S. rosetta*), while most others displayed a negative relationship or no relationship (fig 3.10). As well as this, the results of the analysis of the domain and non-domain regions of the transposable elements suggests selection for translational accuracy.

The research into the transposable elements of the choanoflagellates by Southworth et al (2018), in particular those of *S. rosetta*, indicated that the transposable elements had favoured codons that are adapted to the organisms own translational machinery, however, in *T. vaginalis*, the overall Fop values are much lower compared to those observed in *S. rosetta,* with all transposable elements except *DNA/TcMar-Fot1* being below 0.4. As well as this, the transposable elements that displayed a positive relationship in fig 3.10 also tended to have more transposable elements with higher expression levels, indicating that in some as expression level increased so did optimal codon frequency and therefore bias, which was also observed by Southworth et al (2018) in *S. rosetta's* transposable elements. However, there were two which also displayed a negative relationship, namely the two DNA transposon families *TcMAr-Marin* and *Kolobok-T2,* both of which indicated that as

expression level increased OC frequency decreased. The only transposable element families to display a significant positive relationship where the LTR transposable elements and the DNA transposon families *Merlin/Maverick* (fig 3.10).

4.5 Conclusions and future research

In conclusion, *T. vaginalis*'s genome may be not nearly as gene rich as originally proposed (Carlton et al, 2013; Harp and Chowdhury, 2011) and most of the genes listed in the TrichDB (Aurrecoechea et al, 2017) may not be genuine genes and are instead potentially intergenic DNA and/or pseudogenes as suggested by Woehle et al (2014). Of the remaining genes, only 13,422 were determined to likely be actively transcribed indicating that *T. vaginalis* exome is not as abnormally gene rich as past research would suggest, even for a single-celled eukaryote and is instead more in line with other eukaryotic genomes such as *T. brucei* (Berriman et al, 2005). With regards to the optimal codons, most matched those proposed by McInerny in 1997, however, there were some OCs which either lacked any corresponding anticodon in the major tRNAs or whose predicted OC from the major tRNAs differed in the coding gene dataset. This is potentially due to a number of other mechanisms which are influencing codon usage bias such as adenosine deamination in the wobble position of the major tRNAs as well as wobble binding of some synonymous codons. This may suggest that while the major tRNA genes are a good method of identifying candidate OCs in bacteria, they are less reliable in eukaryotes.

As well as this, several of the amino acids also appeared to possess more than one potential optimal codon which, despite possessing a lower RSCU value, more closely matched the predicted OCs from the major tRNA dataset. Once all candidate OCs had been identified, the analysis of the effective codon number (Nc), which measures the overall bias of the genome, indicated that *T. vaginalis*'s genome has a relatively low degree of overall bias with only the genes in KOG categories B, C and J displaying any significant bias towards these potential optimal codons. Most bias observed in *T. vaginalis*'s genome was skewed more towards AT than GC, although unlike previous research this bias is relatively small. One potential explanation for some of the discrepancies between the major tRNAs and the coding gene datasets optimal codons, is that the adenosine in some of the major tRNAs is undergoing deamination to inosine, allowing a number of codons that would otherwise be non-optimal to be used more frequently. As well as this, it would appear that *T. vaginalis* may possess the ADAT genes that code for adenosine deaminase within its exome and the phylogenetic analysis performed on these potential genes suggests as much, particularly for ADAT 2 and 3.

With regards to the KOG categories, only three display a level of bias that was significantly higher than the others (albeit still relatively low overall), categories B, C and J, which is similar to the results observed by Southworth et al (2018) in the single-celled eukaryote *M. brevicollis*. As for the other KOG categories, their low GC3s as well as relatively low Fop (indicating they are not enriched for OCs) further suggests that there is not very much bias overall in the genome. This bias may be influenced by a number of things, such as mutation pressure which this analysis found may be influencing synonymous base composition as well as local base composition. As well as this, while the difference between non-coding GC content in the high and low expression genes lacked statistical significance, the relationship between synonymous GC3s and non-synonymous GC content in the high, mid and low expression genes was statistically significant. Overall the low overall GC3s in the *T. vaginalis*

genome suggests that mutation pressure is acting against the selective pressure on OCs, resulting in a depression of GC3s across the *T. vaginalis* genome, but particularly within those genes with mid-level expression. As for the transposable elements, much like what was observed by Southworth et al (2018) in *S. rosetta*, many of the transposon families (particularly the LTR retrotransposons) returned a higher Fop in their domain vs non-domain regions, indicating that many of *T. vaginalis*'s transposable elements possess optimal codons. In the case of mutation pressure, there is evidence that mutation pressure is affecting genes across the genome as the genome is predominantly AT biased despite the OCs being predominantly GC-ending. However, the lack of variation between non-coding GC's across the genome suggests that mutation pressure is not driving variation in the GC3s. With regards to selection of optimal codons, there is evidence for both accuracy and efficiency acting on translation.

There are a number of potential avenues researchers can explore to further the findings outlined in this research, most notably, with regards to adenosine deaminase acting on tRNA: while the phylogenetic analysis outlined in chapter 3.5 does suggest that the ADAT genes are present and correlate to those selected in this analysis, to confirm if they are indeed the ADAT genes, a lab analysis could be conducted to sequence the products of these genes in order to determine if they are indeed adenosine deaminase. A small RNA sequencing experiment can also be performed on the tRNAs, similar to that performed by Southworth et al (2019), in order to determine if the adenosine had been modified as, if adenosine had been modified to inosine, it would identify a number of tRNAs that are not present in the genome. As well as this, in order to determine whether the non-Watson and Crick base pairing hypothesis proposed for valine's OC is indeed correct, an analysis of the hydrogen bonding between the codons and anticodons in tRNA could be performed to determine whether the codon is binding to the anticodon in the tRNA correctly (indicating some other mechanism is being used) or whether it is binding to the valine primary OC. The analysis of mutation pressure in this study, whilst relatively rudimentary and using a fairly limited dataset due to time constraints, did demonstrate that mutation pressure is not a major driver of variation in GC3s. However, a more robust analysis with a larger dataset would likely give more definitive and clear results and could determine whether mutation pressure contributes to minor variation in the patterns of codon usage as well.

References

1. Akashi, H. (1994). Synonymous Codon Usage in *Drosophila melanogaster*: Natural Selection and Translational Accuracy. *Genetics, 136*(3), 927-935. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205897/
2. Andachi, Y., Yamao, F., Iwami, M., Akira, M., & Syozo, O. (1987). Occurrence of Unmodified Adenine and Uracil at the First Position of Anticodon in Threonine tRNAs in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences, 84*(21), 7398 - 7402. doi:10.1073/pnas.84.21.7398
3. Anders, F. (2008). Impact of Bias Discrepancy & AA Usage on Estimates of the Effective Number of Codons Used in a Gene, & a Test for Selection on Codon Usage. *Gene, 410*(1), 82-88. doi: 10.1016/j.gene.2007.12.001
4. Atherton, J. C., Sharp, P. M., & Lafay, B. (2000). Absence of Translationally Selected Synonymous Codon Usage Bias in *Helicobacter pylori*. *Microbiology, 146*(4), 851-860. doi: 10.1099/00221287-146-4-851
5. Aurrecoechea, C., Brestelli, J., Brunk, B. P., Carlton, J. M., Dommer, J., Fischer, S. ... Wang, H. (2008). GiardiaDB and TrichDB: Integrated Genomic Resources for the Eukaryotic Protist Pathogens *Giardia lamblia* and *Trichomonas vaginalis*. *Nucleic Acids Research, 37*(1), 526-530. doi: 10.1093/nar/gkn631

6. Avery, O. T., Macleod, C. M., & McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Deoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III. *The Journal of Experimental Medicine, 79*(2), 137-158. doi: 10.1084/jem.79.2.137
7. Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C. ... El-Sayed, N. M. (2005). The Genome of the African Trypanosome *Trypanosoma brucei*. *Science, 309*(5733), 416-422. doi: 10.1126/science.1112642
8. Bourke, G., Burns, K. H., Ghering, M., Gorbunova, V., Seluanov, A., Hammell, M. ... Feschotte, C. (2018). Open Access Ten Things You Should Know About Transposable Elements. *Genome Biology, 19*(1), 199-212. doi: 10.1186/s13059-018-1577-z
9. Campbell, P. N. (1989). What Mad Pursuit: A Personal View of Scientific Discovery. *Biochemical Education, 17*(3), 163. doi: 10.1016/0307-4412(89)90119-2
10. Carlton, J. M., Hirt, R. P., Silva, J. C., Delcher, A. L., Schatz, M., Zao, Q. ... Coombs, G. H. (2007). Draft Genome Sequence of the Sexually Transmitted Pathogen *Trichomonas vaginalis*. *Science, 315*(5809), 207-212. doi: 10.1126/science.1132894
11. Cohen, J. S., & Portugal, F. H. (1974). The Search for the Chemical Structure of DNA. *Conn Med., 38*(10), 551-557. Retrieved from https://collections.nlm.nih.gov/catalog/nlm:nlmuid-101584575X186-doc
12. Comeron, J. P., & Aguadé,, M. (1998). An Evaluation of Measures of Synonymous Codon Usage Bias. *Journal of Molecular Evolution, 47*(3), 268-274. doi: 10.1007/PL00006384
13. Conrad, M. D., Bradic, M., Warring, S. D., Gorman, A. W., & Carlton, J. M. (2012). Getting Trichy: Tools and Approaches to Interrogating Trichomonas vaginalis in a Post-Genome World. *Trends in Parasitology, 29*(1), 17-25. doi: 10.1016/j.pt.2012.10.004
14. Conrad, M. D., Gorman, A. W., Schillinger, J. A., Fiori, P. L., Arroyo, R., Malla, N. ... Carlton, J. M. (2013). Extensive Genetic Diversity, Unique Population Structure & Evidence of Genetic Exchange in the Sexually Transmitted Parasite *Trichomonas vaginalis*. *PLoS Neglected Tropical Diseases, 6*(3), e1573. doi: 10.1371/journal.pntd.0001573
15. Crick, F. H. C. (1966). Codon—Anticodon Pairing: The Wobble Hypothesis. *Journal of Molecular Biology, 19*(2), 548-555. doi: 10.1016/S0022-2836(66)80022-0
16. Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General Nature of the Genetic Code for Proteins. *Nature, 191*(4809), 1227-1232. doi: 10.1038/1921227a0
17. Cristalli, G., Costanzi, S., Lambertucci, C., Lupidi, G., Vittori, S., Volpini, R., & Camaioni, E. (2001). Adenosine Deaminase: Functional Implications and Different Classes of Inhibitors. *Medicinal Research Review, 21*(2), 105-128. doi: 10.1002/1098-1128(200103)21:2<105::AID-MED1002>3.0.CO;2-U

18. Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the Early Years of Nucleic Acid Research. *Human Genetics, 122*(6), 565-581. doi: 10.1007/s00439-007-0433-0
19. de Koning, A. P., Wanjun, G., Todd, A., Castoe, M. A., Batzer, M. A., & Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics, 7*(12), e1002384. doi: 10.1371/journal.pgen.1002384
20. Doherty, A., & McInerny, J. O. (2013). Translational Selection Frequently Overcomes Genetic Drift in Shaping Synonymous Codon Usage Patterns in Vertebrates. *Molecular Biology and Evolution, 30*(10), 2263-2267. doi: 10.1093/molbev/mst128
21. Donders, G. G. G., Depuydt, C. E., Bogers, J. P., & Vereecken, A. J. (2013). Association of *Trichomonas vaginalis* and Cytological Abnormalities of the Cervix in Low Risk Female. *PloS one, 8*(12), e86266. doi: 10.1371/journal.pone.0086266
22. Dong, H., Nilsson, L., & Kurland, C. J. (1996). Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* at Different Growth Rates. *Journal of Molecular Biology, 260*(5), 649-663. doi: 10.1006/jmbi.1996.0428.
23. dos Reiss, M., & Wernisch, L. (2008). Estimating Translational Selection in Eukaryotic Genomes. *Molecular Biology and Evolution, 26*(2), 451-461. doi: 10.1093/molbev/msn272
24. Drummond, D. A., & Wilke, C. O. (2008). Mistranslation – Induced Protein Misfolding as a Dominant Constraint on Coding Sequence Evolution. *Cell, 134*(2), 341-352. doi: 10.1016/j.cell.2008.05.042
25. Duan, J., Zhu, L., Dong, H., Zheng, X., Jiang, Z., Chen, J., & Tian, X. C. (2019). Analysis of mRNA Abundance for Histone Variants, Histone- and DNA-Modifiers in Bovine in vivo and in vitro Oocytes and Embryos. *Scientific reports*, 9(1), 1217 - 13. https://doi.org/ 10.1038/s41598-018-38083-4
26. Dunne, R.L, Dunne, L.A, Upcroft, P, O'donogue, P.J, Upcroft , J.A. (2003). Drug Resistance in the Sexually Transmitted Protozoan *Trichomonas vaginalis*. *Cell Research,* 13(4), 239-250. doi: 10.1038/sj.cr.7290169
27. Duret, L., & Mouchiroud, D. (1999). Expression Pattern &, Surprisingly, Gene Length Shape Codon Usage in *Caenorhabditis*, *Drosophila*, & *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America, 96*(8), 4482-4487. doi: 10.1073/pnas.96.8.4482
28. Elhaik, E., & Tatarinova, T. (2012). GC3 Biology in Eukaryotes and Prokaryotes. *arXiv e-prints*, , arXiv:1203.3929. https://doi.org/10.5772/33525
29. Epstein, R. J., Lin, K., & Tan, T. W. (2000). A Functional Significance for Codon Third Bases. *Gene, 245*(2), 291-298. doi: 10.1016/S0378-1119(00)00042-1
30. Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon Usage of Highly Expressed Genes Affects Proteome-Wide Translation Efficiency. *Proceedings of the National Academy of Sciences of the United States of America, 115*(21), E4940 - E4949. doi: 10.1073/pnas.1719375115
31. Gardner, I. R. S., Whaba, A. J., Basilio, C., & Miller, R. S. (1963). The Dependence of Cell-Free Protein Synthesis in *E. coli* Upon Naturally Occurring or Synthetic Polyribonucleotide's. *Proceedings of the National Academy of Science, 46*(6), 880-885. doi: 10.1073/pnas.49.6.880

32. Genome Research Ltd. (2010). SMALT [Aligns DNA Sequencing Reads with a Reference Genome]. (0.7.4). Retrieved from https://www.sanger.ac.uk/tool/smalt-0/
33. Gerber, A. P., & Keller, W. q. (1999). An Adenosine Deaminase That Generates Inosine at the Wobble Position of tRNAs. *Genomics, 286*(5442), 1146-1149. doi: 10.1126/science.286.5442.1146
34. Gingold, H., & Pilpel, Y. (2011). Determinants of Translation Efficiency and Accuracy. *Molecular Systems Biology, 7*(1), 481. doi: 10.1038/msb.2011.14
35. Govers, F., & Gijzen, M. (2006). Phytophthora Genomics: the Plant Destroyers Genome Decoded. *Molecular Plant-Microbe Interactions, 19*(12), 1295-1301. doi: 10.1094/MPMI-19-1295
36. Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A. (1980). Codon Catalog Usage & the Genome Hypothesis. *Nucleic Acids Research, 8*(1), 197. doi: 10.1093/nar/8.1.197-c
37. Griffith, F. (1928). The Significance of Pneumococcal Types. *Epidemiology & Infection, 27*(2), 113-159. doi: 10.1017/S0022172400031879
38. Grosjean, H., Auxilien, S., Constantinesco, F., Simon, C., Corda, Y., Becker, H. F. ... Fourrey, J. L. (1996). Enzymatic Conversion of Adenosine to Inosine and to N1-methylinosine in Transfer RNAs: A review. *Biochimie, 78*(6), 488-501. doi: 10.1016/0300-9084(96)84755-9
39. Harp, D. F., & Chowdhury, I. (2011). Trichomoniasis: Evaluation to Execution. *European Journal of Obstetrics & Gynecology and Reproductive Biology, 157*(1), 3-9. doi: 10.1016/j.ejogrb.2011.02.024
40. Herbert, A. (1996). RNA Editing, Introns and Evolution. *Trends in Genetics, 12*(1), 6-9. doi: 10.1016/S0022-2836(66)80022-0
41. Hershberg, R., & Petrov, D. A. (2008). Selection on Codon Bias. *Annual Reviews Genetics, 42*(1), 287-299. doi: 10.1146/annurev.genet.42.110807.091442.
42. Hershberg, R., & Petrov, D. A. (2009). General Rules for Optimal Codon Choice. *PLoS Genetics, 5*(7), e1000556. doi: 10.1371/journal.pgen.1000556
43. Hershey, A. D., & Chase, M. (1952). Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J Gen Physiol, 36*(1), 39-56. doi: 10.1085/jgp.36.1.39
44. Holley, R. W. (1965). Structure of an Alanine Transfer Ribonucleic Acid. *JAMA, 194*(8), 868-871. doi: 10.1001/jama.1965.03090210032009
45. Holley, R. W., Everett, G. A., Madison, J. T., & Zamir, A. (1965). Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid*. *Journal of Biological Chemistry, 240*(5), 2122-2127. 0.
46. Horn, D. (2008). Codon Usage Suggests That Translational Selection has a Major Impact on Protein Expression in Trypanosomatids. *BMC Genomics, 9*(1), 2. doi: 10.1186/1471-2164-9-2
47. Huelsenbeck, J. F., & Ronquist, F. (2001). MRBAYES: Bayesian Inference of Phylogenetic Trees. *Bioinformatics (Oxford, England), 17*(8), 754 - 755. doi: 10.1093/bioinformatics/17.8.754
48. Huelsenbeck, J.P., and F. Ronquist. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.

49. Ikemura, T. (1985). Codon Usage and tRNA Content in Unicellular and Multicellular Organisms. *Molecular Biology and Evolution, 2*(1), 13-34. doi: 10.1093/oxfordjournals.molbev.a040335
50. Janssen, B. D., Chen, Y. P., Molgora, B. M., Wang, S. E., Simoes-Barbosa, A., Johnson, P. J., & Wang, H. (2018). CRISPR/Cas9-Mediated Gene Modification and Gene Knock Out in the Human-Infective Parasite *Trichomonas vaginalis*. *Scientific Reports, 8*(1), 270-14. doi: 10.1038/s41598-017-18442-3
51. JF Peden (1999). CodonW [Codon Usage Analysis Software]. (1.4.2). Retrieved from codonw.sourceforge.net
52. Jones, M. E. (1953). Albrecht Kossel, a Biographical Sketch. *Yale J Biol Med, 26*(1), 80-97.
53. Katoh, K. (2002). MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research, 30*(14), 3059 - 3066. doi: 10.1093/nar/gkf436
54. 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. 0.
55. Koltzoff, N. (1934). The Structure of the Chromosomes in the Salivary Glands of *Drosophila*. *Science, 80*(2075), 312-313. doi: 10.1126/science.80.2075.312
56. Sakmar, T. P. (2012). Har Gobind Khorana (1922–2011): Pioneering Spirit. *PLoS Biol*, 10(2), e1001273. https://doi.org/10.1371/journal.pbio.1001273
57. Levene, P. A., & Jacobs, W. A. (1909). Further Studies on the Constitution of Inosinic Acid. *Proceedings of the Society for Experimental Biology and Medicine, 6*(3), 90. doi: 10.3181/00379727-6-42
58. Lorenz, M. G., & Wackernagel, W. (1994). Bacterial Gene Transfer By Natural Genetic Transformation in the Environment. *Microbiol Rev, 58*(3), 563-602. doi: 10.1128/MMBR.58.3.563-602.1994
59. Lynn, D. L., Singer, G. A. C., & Hickey, D. A. (2002). Synonymous Codon Usage is Subject to Selection in Thermophilic Bacteria. *Nucleic Acids Research, 30*(19), 4272-4277. doi: 10.1093/nar/gkf546
60. Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N. ... Lopez, R. (2019). MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research, 47*(W1), W636 - W641. doi: 10.1093/nar/gkz268
61. McClintock, B. (1950). The Origin and Behaviour of Mutable Loci in Maize. *Genome Biology, 36*(6), 344-355. doi: 10.1073/pnas.36.6.344
62. McInerny, J. O. (1997). Codon Usage Patterns in *Trichomonas vaginalis*. *European Journal of Protistology, 33*(3), 266-273. doi: 10.1016/S0932-4739(97)80004-1
63. McInerny, J. O. (1998). Replicational and Transcriptional Selection on Codon Usage in *Borrelia burgdorferi*. *Proceeding of the National Academy of Sciences U.S.A, 95*(18), 10698–10703. . doi: 10.1073/pnas.95.18.10698
64. Mclelland, R. S., Sangaré, L., Hassan, W. M., Lavreys, L., Mandaliya, K., Kiarie, J. ... Baeten, J. M. (2007). Infection with *Trichomonas vaginalis* Increases the Risk of HIV – 1 Acquisition. *The Journal of Infectious Diseases, 195*(5), 698-702. doi: 10.1086/511278

65. Meade, J. C., Shah, P. H., & Lushbaugh, W. B. (1997). *Trichomonas vaginalis*: Analysis of Codon Usage. *Experimental Parasitology, 87*(1), 73-74. doi: 10.1006/expr.1997.4185
66. Miller, M., Pfeiffer, W., & Schwartz, T. (2011). The CIPRES Science Gateway: a Community Resource for Phylogenetic Analyses. *Proceedings of the 2011 TeraGrid Conference, 11*, 1-8. doi: 10.1145/2016741.2016785
67. Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L. ... Marshall, D. (2013). Using Tablet for Visual Exploration of Second-Generation Sequencing Data. *Briefings in Bioinformatics, 14*(2), 193-202. Retrieved from https://ics.hutton.ac.uk/tablet/
68. Munagala, N. R., Wang, C. C., Gorman, A. W., & Carlton, J. M. (2003). Adenosine is the Primary Precursor of all Purine Nucleotides in *Trichomonas vaginalis*. *Molecular & Biochemical Parasitology, 127*(2), 143-149. doi: 10.1016/S0166-6851(02)00330-4
69. Murphy IV, F. V., & Ramakrishnan, V. (2004). Structure of Purine-Purine Wobble Base Pair in the Decoding Centre of the Ribosome. *Nature Structural & Molecular Biology, 11*(12), 1251-1254. doi: 10.1038/nsmb866
70. Nirenberg, M. W., & Matthaei, J. H. (1961). The Dependence of Cell-Free Protein Synthesis in E. coli Upon Naturally Occurring or Synthetic Polyribonucleotide's. *Proceeding of the National Academy of Sciences U.S.A, 47*(10), 1588-1602. doi: 10.1073/pnas.47.10.1588.
71. Novoa, E. M., Jungreis, I., Jaillon, O., & Kellis, M. (2019). Elucidation of Codon Usage Signatures Across the Domains of Life. *Molecular Biology and Evolution, 36*(10), 2328–2339. doi: 10.1093/molbev/msz124
72. Paul, P., Malakar, A. K., & Chakraborty, S. (2018). Compositional Bias Coupled with Selection and Mutation Pressure Drives Codon Usage in *Brassica campestris* genes. *Annual Reviews Genetics, 27*, 725-733. doi: 10.1007/s10068-017-0285-x
73. Paul, P., Malakar, A. K., & Chakraborty, S. (2018). Compositional Bias Coupled with Selection and Mutation Pressure Drives Codon usage in *Brassica campestris* Genes. *Food Science and Biotechnology, 27*(3), 725-733. doi: 10.1007/s10068-017-0285-x
74. Peden, J. F. (1999). *Analysis of Codon Usage* (PhD Thesis). Retrieved from University of Nottingham
75. Quax, T. E. F., Claassens, N. J., Söll, D., van der Oost, J., & Gould, S. B. (2015). Codon Bias as a Means to Fine Tune Gene Expression. *Molecular Cell, 59*(2), 149-161. doi: 10.1016/j.molcel.2015.05.035
76. Radonjić , I.V, Mitrović, S.M, Džamić , A.M, Arsić-Arsenijević , V.S, Kranjčić-Zec , I.F. (2003). Correlation Between Clinical Symptoms and Diagnosis of Trichomoniasis in Female. *Medicinski Pregled,* 56(5-6), 227-231. doi: doi.org/10.2298/MPNS0306227R
77. Rafels-Ybern, À., Torres, A. G., Camacho, N., Herencia-Ropero, A., Frigolé, H. R., Wulff, T. F. ... De Pouplana, L. R. (2019). The Expansion of Inosine at the Wobble Position of tRNAs, and it's Role in the Evolution of Proteomes. *Molecular Biology and Evolution, 36*(4), 650-662. doi: 10.1093/molbev/msy245
78. Rafels-Ybern, A., 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. doi: 10.1080/15476286.2017.1358348
79. Rambaut, A - http://tree. bio. ed. ac. uk/software/figtree/, 2009
80. Rodin, P., King, A. J., Nicol, C. S., & Barrow, J. (1960). Flagyl in the Treatment of Trichomoniasis. *The British Journal of Venereal Diseases, 36*(3), 147-151. doi: 10.1136/sti.36.3.147
81. Ronquist, F., & Huelsenbeck, J. F. (2003). MrBayes 3: Bayesian Phylogenetic Inference Under Mixed Models. *Bioinformatics (Oxford, England), 19*(12), 1572 - 1574. doi: 10.1093/bioinformatics/btg180
82. Ronquist, F., Teslenko, M., Paul van der Mark, Ayres, D. L., Darling, A., Höhna, S., . . . Naturvetenskapliga fakulteten. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61(3), 539-542. doi:10.1093/sysbio/sys029
83. Rowley, J., Vander Hoorn, S., Korenromp, E., Low, N., Unemo, M., Abu-Raddad, L. J. ... Taylor, M. M. (2019). Chlamydia, Gonorrhoea, Trichomoniasis and Syphilis: Global Prevalence and Incidence Estimates, 2016. *Bulletin of the World Health Organization, 97*(8), 548-562. doi: 10.2471/BLT.18.228486
84. Saint-Léger, A., Bello, C., Dans, P. D., Torres, A. G., Novoa, E. M., Camacho, N. ... de Pouplana, L. R. (2016). Saturation of Recognition Elements Blocks Evolution of New tRNA Identities. *Sci Adv, 2*(4), e1501860. doi: 10.1126/sciadv.1501860
85. Salim, H. M. W., Ring, K. L., & Cavalcanti, A. R. O. (2007). Patterns of Codon Usage in Two Ciliates That Reassign the Genetic Code: *Tetrahymalesa thermophila*, & *Paramecium tetraurelia*. *Protist, 159*(2), 283-298. doi: 10.1016/j.protis.2007.11.003
86. Schaub, M., & Keller, W. (2002). RNA Editing by Adenosine Deaminases Generates RNA and Protein Diversity. *Biochimie, 84*(8), 791-803. doi: 10.1016/S0300-9084(02)01446-3
87. seqtk, Toolkit for Processing Sequences in FASTA/Q Formats. Available from: https://github.com/lh3/seqtk.
88. Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that Influence the Evolution of Codon Bias. *Philosophical Transactions. Biological sciences, 365*(1544), 1203 - 1212. doi: 10.1098/rstb.2009.0305
89. Sharp, P. M., Stenico, M., Peden, J. F., & Lloyd, A. T. (1993). Codon Usage: Mutational Bias, Translational Selection, or Both? *Biochemical Society Transactions, 21*(4), 835-841. doi: 10.1042/bst0210835
90. Smith, A., & Johnson, P. (2011). Gene Expression in the Unicellular Eukaryote *Trichomonas vaginalis*. *Research in Microbiology, 162*(6), 646-654. doi: 10.1016/j.resmic.2011.04.007
91. Southworth, J., Grace, C. A., Marron, A. O., Fatima, N., & Carr, M. (2019). A genomic Survey of Transposable Elements in the Choanoflagellate *Salpingoeca rosetta* Reveals Selection on Codon Usage. *Mobile DNA, 10*(1), 1-19. doi: 10.1186/s13100-019-0189-9
92. Spears, J. L., Rubio, M. A. T., Gaston, K. W., Wywial, E., Strikoudis, A., Bujnicki, J. M. ... Alfonzo, J. D. (2011). A Single Zinc Ion Is Sufficient for an Active *Trypanosoma brucei* tRNA Editing Deaminase*. *Proceedings of the National Academy of Sciences, 286*(23), 20366-20374. doi: 10.1074/jbc.M111.243568

93. Steiner, R. E., & Ibba, M. (2019). Regulation of tRNA-Dependent Translational Quality Control. *IUBMB Life, 71*(8), 1150 - 1157. doi: 10.1002/iub.2080
94. Subramanian, A., & Sarkar, R. R. (2015). Comparison of Codon Usage Bias Across Leishmania and Trypanosomatids to Understand mRNA Secondary Structure, Relative Protein Abundance and Pathway Functions. *Genomics, 106*(4), 232-241. doi: 10.1016/j.ygeno.2015.05.009
95. Sueoka, N. (1988). Directional Mutation Pressure and Neutral Molecular Evolution. *Proceedings of the National Academy of Sciences, 85*(8), 2653-2657. doi: 10.1073/pnas.85.8.2653
96. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V. ... Natale, D. A. (2003). The COG Database: an Updated Version Includes Eukaryotes. *BMC Bioinformatics, 4*(1), 41. doi: 10.1186/1471-2105-4-41
97. Tatusov, R. L., Galperin, M. Y., DA, N. a. t. a. l. e., & Koonin, E. V. (2000). The KOG Database: a Tool for Genome-Scale Analysis of Protein Functions and Evolution. *Nucleic Acids Research, 28*(1), 33-36. doi: 10.1093/nar/28.1.33
98. Tekaia, F. (2003). Genome Data Exploration Using Correspondence Analysis. *Bioinformatics and Biology Insights, 10*(1177-9322), doi: 10.4137/BBI.S39614
99. Twu, O., de Miguel, N., Lustig, G., Stevens, G. C., Vashisht, A. A., Wohlschlegel, J. A., & Johnson, P. J. (2013). *Trichomonas vaginalis* Exosomes Deliver Cargo to Host Cells and Mediate Host: Parasite Interactions. *PLoS Pathogens, 9*(7), e1003482. doi: 10.1371/journal.ppat.1003482
100. Watson, J. D., & Crick, F. H. (1953). Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid. *Nature, 171*(4356), 737-738. doi: 10.1038/171737a0
101. Werren, J. H., Nur, U., & Wu, C. L. (1988). Selfish Genetic Elements. *Trends in Ecology and Evolution, 3*(11), 297-302. doi: 10.1016/0169-5347(88)90105-x
102. Woehle, C., Kusdian, G., Radine, C., Graur, D., Landan, G., & Gould, S. B. (2014). The Parasite *Trichomonas vaginalis* Expresses Thousands of Pseudogenes and Long Non-Coding RNAs Independently from Functional Neighbouring Genes. *BMC Genomics, 15*(1), 906. doi: 10.1186/1471-2164-15-906
103. World Health Organisation. (2008). *Global Incidence and Prevalence of Selected Curable Sexually Transmitted Infections: 2008*. Retrieved from https://doi-org.libaccess.hud.ac.uk/10.1016/S0968-8080(12)40660-7
104. Wright, F. (1990). The Effective Number of Codons Used in a Gene. *Gene, 87*(1), 23-29. doi: 10.1016/0378-1119(90)90491-9
105. Yang, Z., & Nielsen, R. (2008). Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. *Molecular Biology and Evolution, 25*(3), 1537-1719. doi: 10.1093/molbev/msm284