Under-representation of CTAG in the bacterial genome

A Thesis

Submitted in partial fulfilment of the requirements

For the award of the degree of

Masters of Science

In Biotechnology

by

SONIA

Roll no. 302101037

Under the supervision of

DR. VIKAS HANDA

Assistant Professor

Department of Biotechnology

Thapar Institute of Engineering and Technology

Patiala, Punjab- 147004

July 2023

Certificate

This is to certify that the thesis entitled, **Under-representation of CTAG in the bacterial genome** being submitted by Sonia (Reg. No. 302101037), in partial fulfilment of the requirements for the award of the degree of Master of Science in Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab is a bonafide work carried out under the guidance and conception of Dr. Vikas Handa and that no part of this thesis has been submitted for the award of any other degree.

25.08.23

DATE: 26:07:2023

Dr. Vikas Handa

Supervisor

Candidate Declaration

I hereby certify that the project work entitled, **Under-representation of CTAG *in the* bacterial genome** in partial fulfillment of the requirements for the award of the degree of Master of Science in Biotechnology and submitted is an authentic record of my work carried out during the period January 2023 to June 2023 under the guidance of Dr. Vikas Handa, Assistant professor, Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

Date: 26:07:2023

Sonia

This is to certify that the above statement made by the student is correct to the best of our knowledge and belief.

Date: 26:07:2023

Dr. Vikas Handa

Supervisor

ACKNOWLEDGEMENT

I am grateful to each one who has helped me throughout the entire project for its successful completion. First, thanks to Almighty for giving me strength and support so that the project could be completed peacefully.

With great reverence, I express my warmest gratitude to Dr. Vikas Handa, Assistant Professor, Department of Biotechnology, Thapar Institute of Engineering and Technology who agreed to take upon and guided this dissertation. I have no words to express my heartfelt thanks to him for his illuminating guidance, unfailing encouragement, supervision, and keen interest during this dissertation.

I would like to express my heartfelt respect to Dr. M S Reddy, Head of the Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala for his kind suggestions and foresightedness.

I have immense gratitude towards my parents for their affection and faith. Also, I would like to thank my siblings for giving me strength and loving support. I would like to acknowledge my lab friends Aapoorva and Tamnayee Basu who helped me in learning small things with great perfection. The whole credit goes to all the people who had their unshakeable faith in me which has always motivated me.

Sonia

Sonia

Table of Content

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
CG	Guanine Cytosine
IS	Insertion Sequences
Dam	DNA adenine methyltransferase
DCM	DNA cytosine methyltransferase
TA	Thymine Adenine
AT	Adenine Thymine
VSR	Very Short repair
VSP	Very Short Patch
O/E	Observed/Expected
C.V.	coefficient of variation
oriC	replication origin

Under-representation of CTAG in the bacterial genome

ABSTRACT

DNA sequences in the genomes have several examples of inhomogeneities which point towards their non-random nature. One such case is under-representation of CTAG tetranucleotide in bacterial genomes. This work aims to study the reasons behind the under-representation of CTAG in bacterial genomes. A randomly selected set of 23 bacterial genomes, largely from diverse prokaryotes, showed that CTAG is under-represented (Observed / Expected value <1.0). In the majority of the selected genomes, CTAG had the lowest O/E value. It was found that TAG, a submotif of CTAG has one of the lowest O/E values in most of the bacteria. The under-representation effect is further enhanced by the occurrence of a cytosine at the 5' end of TAG which makes it CTAG and explains its under-representation. CTAG distribution in the genome is also very uneven. While most of the genome has very low occurrence, there are certain short regions where CTAG is found in relatively much higher abundance. Such regions coincide with the sequence with highly uneven base composition. This study demonstrates base composition as one of the factors causing highly heterogeneous distribution of CTAGs in the bacterial genomes and indicates that CTAGs play some important biologic role in bacteria.

Key words: Under-representation, Inhomogeneity, CTAG, bacterial genome, termination codon.

CHAPTER 1

INTRODUCTION

The diversity of life, especially at cellular and molecular levels, is a remarkable feature among living organisms. The diversity in the living world is based on the genetic information of the living organisms. All living organisms harbor this information encoded in the sequence of DNA (RNA in the case of RNA viruses). The unit of genetic information is gene, a defined segment of DNA that encodes specific proteins. The genome of a cellular organism consists of hundreds to tens of thousands of genes, each encoding a unique protein. These proteins are responsible for most of the cellular functions.

The structure of DNA can be described as two long polynucleotide chains bound by hydrogen bonds between their nucleic acid moiety and wrapped around each other to form a double helix. The backbone of DNA is made up of repeating deoxyribose sugar-phosphate units. Attached to each deoxyribose molecule is one of four possible bases, adenine (A), cytosine (C), guanine (G), and thymine (T). In DNA each nucleotide carries any one of the four bases. Thus, in turn, the polynucleotide chain consists of a sequence of bases in DNA, which is capable of storing genetic information. The structure of DNA, with its sugar-phosphate backbone and variable bases, plays a crucial role in storing, replicating and encoding genetic information in living organisms. This is how a chain of DNA forming chromosomes consists of a long sequence of bases that stores an enormous amount of genetic information in cells.

Cellular organisms are classified into two different types, prokaryotes and eukaryotes. Prokaryotes, lacking a nucleus, possessed typically a single linear or circular chromosome. In case of eukaryotes, their genetic information is stored in multiple linear chromosomes which are present in membrane bound organelle called nucleus. Eukaryotic genomes are much larger in size when compared to prokaryotes. As a result, eukaryotic genomes have much higher number of gene and exhibit more complex organization.

The genomes of eukaryotic organisms, especially the higher eukaryotes, are much larger than prokaryotes but contain greater proportions of non-coding sequences. The non-coding sequences of eukaryotes consist of introns, dispersed repeat sequences such as transposons and tandem repeats. The proportion of repetitive DNA increases with increase in the complexity of the genome. The number of the gene carrying intron as well as the average size of the intron increases with increases in the genomic complexity of eukaryotic organism.

Genomic DNA has non-repetitive and repetitive sequences, with the ratio between the two varying across various organisms. In prokaryotes, most of the DNA consists of non-repetitive sequences. In higher eukaryotes large amounts of DNA falls in the category of repetitive DNA. Repetitive DNA is classified into two groups that are tandem repeats and dispersed repeats. In unicellular eukaryotes, most of the DNA is a non-repetitive sequence. Up to half of the DNA in animal cells contains moderately and highly repetitive components. In plants and amphibians, up to 80% of the genome containing moderately and highly repetitive components, thereby reducing the non-repetitive DNA in small components.

The complexity of the organism is related to the non-randomness of the genomic DNA sequences (Häring and Kypr, 1999). The existence of repeat sequences results in a higher relative abundance of the repeat sequence units in the genome. Similarly, because of specific functionality or vulnerability to mutations, certain DNA sequence motifs may become rarer in the course of evolution. Such factors can cause genomic DNA sequences to show compositional inhomogeneity, from local signals to differences at the isochore level (Karlin *et al.,* 1994). Genomic DNA exhibit various types of sequence inhomogeneities. A striking example of inhomogeneities is the under-representation of CG in the mammalian genome. Tandem repeats are another prominent example of inhomogeneities as their units are overrepresented. The sequence inhomogeneities may also arise due to presence of coding sequences, promoters and other regulatory elements of the genes. As a result, genomic inhomogeneity is widespread and manifests itself on numerous dimensions in prokaryotic as much as eukaryotic genes (Burge *et al.,* 1992).

In higher eukaryotic, the suppression of CG dinucleotides motif often attributed to methylation, deamination, and mutation processes (kelvin *et al.,* 1997). Most CpG sites in eukaryotic DNA, including chromosomal and organelle DNA, may be disadvantageous, and that the methylase helps expedite changes that are desirable in their own right. This would explain why eukaryotes have not developed an effective way to stop the methylase from causing mutations. The limitations imposed by high advantages to CG deficits (kelvin *et al.,* 1997Some authors have proposed that CG deficiencies in DNA sequences, could be attributed to certain structural limitations at the DNA level. This structural organization of DNA involves various forms of packing, such as nucleosomes, which differ between different types of living organisms like eukaryotes (like plants

and animals), viruses, prokaryotes (like bacteria), and organelles (like mitochondria) (Burge *et al.,* 1992). Since they are repeated often their components are over-representative in the genome. Not only are motifs which are over-represented but there are also some cases in which motifs are under-represented or suppressed in the genome. A striking example of under-representation is CG dinucleotide in higher eukaryotes.

This over- and under-representation of motifs is not only limited to the eukaryotic genomes but it is also found in bacteria. One of the most prominent cases is the suppression of CTAG tetranucleotide sequence. It is a tetranucleotide palindromic sequence that is found to be under-represented in majority of the bacterial genomes (Burge *et al.,* 1992). CTAG is remarkably under-represented relative to its reverse palindromic sequence GATC (Sharawy *et al.,*2021).

The under-representation and over-representation of certain DNA sequence motifs in a genome may be used as a cue to some factors or functional significance that leads to the bias. These factors include tendencies of DNA conformations, as well as the influence of context-dependent processes such as mutation, repair, and modification (Abe *et al.,* 2006). However, this work is an attempt to explore whether there could be any other important biologically significant reasons behind the under-representation of CTAG in bacterial genomes.

CHAPTER 2

REVIEW OF LITERATURE

Genomic sequences are non-random that may be observed by looking at its compositional inhomogeneities. A simple form of genomic sequence heterogeneity can observe by analyzing G+C contents of DNA in different parts of genome, viz. isochore analysis. Some other examples of sequence heterogeneity are suppression of CGs in mammals and GATC overrepresentation in OriC region of *E. coli*.

Prokaryotic and eukaryotic DNA show distinct differences in their composition. Prokaryotes generally have a higher proportion of coding DNA compared to eukaryotes. However, eukaryotes possess a more advanced DNA composition characterized by extensive non-protein coding regions. Moreover, eukaryotic DNA molecules are larger and utilize nucleosomes for compaction, resulting in pronounced small-scale (approximately 200 bp) and long-range correlation effects that are absent in bacteria (Bohlin & Skjerve, 2009).

Eukaryotic genomes exhibit greater compositional heterogeneity compared to prokaryotic genomes, as deviations from expected values are more substantial in eukaryotes. The composition of genomic sequences is highly variable and is strongly linked to the GC content in all multicellular eukaryotes, regardless of genome size. In the case of the human genome, it exhibits significant compositional heterogeneity within and between individual chromosomes. While all multicellular

eukaryotic genomes studied in the report show compositional heterogeneity, they also contain segments with uniform GC content, called isochores. Generally, GC-poor isochores tend to be longer compared to GC-rich ones. A characteristic feature of the mammalian genome is presence of GC-rich regions with relatively higher O/E value of CpGs, called CpG islands, typically found near the 5' end of housekeeping genes. Rest of the genome has strong under-representation of CpGs. (Eyre-Walker, 1992; Nekrutenko & Li, n.d.; Saitou, 2013)..

2.1 Over- and under-represented motifs in DNA sequences

It has been reported that analyze of a wide range of living organisms showed patterns in the occurrences of short sequences of DNA. These sequences can be di-, tri-, and tetranucleotides within and between genomic sequences. The primary focus was on identifying instances of significant over- and under-representation of these short oligonucleotides (Burge *et al.,* 1992).

The dinucleotide TA is consistently found in lower amounts across phages, bacteria, viruses, eukaryotes, and non-mammalian mitochondrial genomes. It is one of the least represented dinucleotides in most sequence sets, except for the mitochondrial genomes. On the other hand, the reversed dinucleotide AT does not show a consistent pattern in its relative abundance. Similarly, the dinucleotide CG is significantly under-represented in the vertebrate genomes as well as in the mitochondrial genomes (Burge *et al.,* 1992).

In vertebrates, the decrease in CG content is commonly attributed to a process involving methylation, deamination, and mutation. This process leads to the conversion of CG/CG to

TG/CA. However, it is not clear whether pure mutation pressure alone is the primary driving force behind this phenomenon, especially when active CG methylases are present and can increase the mutation rate. This explanation is unlikely to apply to the majority of bacteria that have been studied, and it also fails to account for the widespread CG suppression observed in mammalian mitochondria, which lack the typical methylase activity. Other factors, such as high dinucleotide stacking energy, supercoiling, or chromatin packing, may contribute to the CG deficits by imposing structural limitations. These CG deficits could potentially offer a selective advantage (Karlin *et al.,* 1997).

The dinucleotide AC/GT are not commonly found in both eukaryotic nuclear and mitochondrial genomes. However, the CA/TG pair is commonly found in eukaryotic sequences, but it is not as common in mitochondrial genomes. The GCA/TGC pair is a trinucleotide that is typically under-represented in phages, viral genomes, and eukaryotic sequences. However, it's worth noting that this under-representation does not apply universally. When it comes to termination codons, the CTA/TAG pair shows a tendency towards under-representation in nuclear eukaryotes and all bacteria. The other two stop codons, TAA and TGA, are not typically under-represented in most organisms (Burge *et al.,* 1992).

Tetranucleotide four base pair palindromes (CTAG) are typically found less frequently in many phages and bacterial sequences. Other two tetranucleotide palindromes (GTAC and TGCA) are more common in eukaryotic nuclear and mitochondrial sequences. Interestingly, these two palindromes are significantly less abundant in bacteriophage genomes (Burge *et al.*, 1992).

Another remarkable example is GATC abundance in bacterial genomes. The Dam methyl transferase recognizes GATC sequence and methylates the Adenine at 6^{th} position. This enzyme is a component of the Dam methylation system, which, during DNA replication, methylates the adenine residue inside the GATC sequences. Many biological processes, including the regulation of gene expression, DNA replication, repair, and recombination, depend on the methylation process (Cohen et al., 2016).

The distribution of the GATC methylation pattern on the bacterial chromosome is an intriguing feature. Peaks appear at around one-third of the way from the termination area on both replichores, and it follows a clear pattern. Dam-positive bacteria exhibit this pattern consistently, while Dam-negative bacteria do not, indicating a direct connection to the Dam methylation system (Sobetzko et al., 2016). Furthermore, the replication origin (oriC) and nearby genes are locations where GATC sites are very common. The fact that GATCs are overrepresented in the area around oriC shows that this structure is crucial for the proper functioning of this area and that the SeqA protein plays a role in protecting the origin.

GATC site are also abundantly found in the genes involved in DNA replication and repair when compared to the genes of other functional categories. Other than GATC, no other tetranucleotide sequence is found over-represented in these genes. This suggests that GATC may have an unknown regulatory function in the processes of DNA replication and repair.

Dam enzyme plays important roles in many physiological processes via the methylation of the GATC, a frequent and significant DNA methylation site in bacteria. Its importance in preserving

appropriate cellular processes is further underlined by the unique position of GATC sites on the bacterial chromosome, especially close to the replication origin and genes involved in replication and repair (Karlin, 2005; Sobetzko et al., 2016).

2.2 The CTAG Tetranucleotide Sequence

The CTAG tetranucleotide sequence is exceptionally rare in various genomic regions of *E. coli.* Research has consistently shown that CTAG occurs far less frequently than would be expected for a random tetranucleotide sequence. Its scarcity is particularly pronounced in protein-coding regions of the genome. However, CTAG is surprisingly abundant in genes that code for structural RNAs, especially within the small portion of the genome responsible for coding tRNAs (Blattner *et al.,* 1997). CTAG remains one of the least common tetranucleotides in the species, predominantly found in intergenic areas and rarely observed within genes (Balaceanu *et al.,* 2019; Blattner *et al.,* 1997). Overall, CTAG is considered the rarest tetranucleotide in most bacterial DNA, demonstrating its distinctive distribution pattern across genomic subsets (Balaceanu *et al.,* 2019; Stibitz & Garletts, 1992).

2.3 CTAG under-representation in bacterial genomes

The tetranucleotide CTAG is found to be under-represented in both bacterial and eukaryotic sequences (Kandel *et al.,* 1996). One possible explanation for this phenomenon is the presence of overlapping oppositely oriented TAG stop codons in CTAG, which might have a detrimental effect. However, the tetranucleotide TTAA, which also contains stop codons in both orientations, does not show the same under-representation (Burge *et al.,* 1992). The CTAG-containing elements

are known to be binding regions for core repressors in certain operons and genes, and they are frequently found near rRNA gene clusters. However, in the rest of the genome, their distribution appears to be relatively uniform (Sivaraman *et al.,* 2005).

The consensus binding site for the trpR repressor, ACTAGTTAACTAGT, contains two copies of CTAG. Evidence from the crystal structure of the trp repressor/operator complex suggests that these two CTAGs "kink" when bound by trpR, and this structural feature might be disadvantageous to DNA stability under certain conditions. Moreover, the under-representation of CTAG may be a result of the DCM methylase/short patch repair mechanism. The CCAGG sequence, which can evolve to CTAGG by deamination, is the target of the DCM methylase. The repair mechanism converts T-G mismatches back to C-G, but if it is not perfectly specified, it can incorrectly convert valid CTAG occurrences, increasing the rarity of CTAG (Burge *et al.,* 1992).

The under-representation of CTAG might be related to structural defects caused by the "kinking" or specific functional roles associated with this tetranucleotide (Karlin *et al.,* 1997). This process is also affected by the VSP DNA repair system, which converts T: G mismatches to C: G pairs and causes a decrease in the number of sequences with CTAG motifs. It is observed that CTAG motifs are most under-represented in protein-coding regions, which suggests interference with secondary structure or trpR binding might be a selective explanation, though mutational explanations are not ruled out (Karlin *et al.,* 1997; McVean & Hurst, 2000).

The CTAG-containing cleavage sequences could be used to accurately define genetic boundaries with significantly reduced requirements for time and resources compared to whole genome

sequence comparisons (Tang *et al.,* 2014). The Very Short Patch repair system tends to eliminate the CTAG sequence, which results in rare sites for endonucleases such as BlnI, XbaI, and NheI in the genomes of *Salmonella* (Tang *et al.*, 2017).

Furthermore, researchers have employed a bioinformatics tool called SOM (Self-Organizing Map) to classify genomic sequences into biological categories based solely on oligonucleotide frequency. oligonucleotide sequences that serve as genetic signals, particularly those involved in regulating gene expression. they observed that the occurrence levels of these important functional signals deviated significantly from what would be expected based on the random distribution of nucleotides in the genome. By considering specific oligonucleotides that were both under-represented and overrepresented in each genome, they identified several factors that contribute to this bias. These factors include the inherent tendencies of DNA to adopt certain conformations, as well as the influence of context-dependent processes such as mutation, repair, and modification (Abe *et al.,* 2006).

In prokaryotes, the low occurrence of palindromic tetranucleotides, such as CTAG, GATC, GTAC, and CATG, may partly reflect avoidance of restriction systems. The counts and distributions of these palindromic nucleotides with the same nucleotide content show notable patterns (Karlin *et al.,* 1998).In summary, the under-representation of CTAG in bacterial and eukaryotic genomes may result from a combination of factors, such as potential structural defects, specific functional roles, and interactions with DNA repair systems. Oligonucleotide frequencies play important roles in biological processes and can be utilized for species separation and genetic signal identification (Karlin *et al.,* 1998).

2.4 Factor Influencing CTAG Under-representation in bacterial genomes

2.4.1 DNA methylation

DNA methylation is a crucial process in the regulation of eukaryotic genomes, although its role in pathogenic prokaryotes remains poorly understood. Unlike in eukaryotes, where DNA methylation primarily involves 5-methylcytosine (5mC), bacteria exhibit three forms of DNA methylation: N6-methyladenine (6mA), N4-methylcytosine (4mC), and 5mC. Among these, 6mA is the most prevalent in bacterial genomes. The specific effects of 4mC, a unique epigenetic marker found in bacteria and archaea, are not yet well-defined, as most epigenetic research focuses on eukaryotes (Gaultney *et al.*, 2020).

Methylated cytosines have been found to induce tighter wrapping of DNA around the nucleosome histone core, affecting the structure of nucleosomes (Lee & Lee, 2012). Additionally, methylated cytosines serve as binding sites for Methyl-CpG binding proteins, which recruit repressive epigenetic factors (Nan *et al.*, 1998).

DNA methylation can be rapidly influenced by environmental factors such as diet, hormones, stress, drugs, or exposure to environmental chemicals. This suggests that the regulatory mechanisms governing DNA methylation can be activated in response to the environment, leading to modifications in gene expression (Hämälä *et al.*, 2022; Keil & Lein, 2016).

In bacteria, DNA methylation primarily functions in restriction-modification systems. However, the discovery of the Dam methyl transferase in *E. coli* revealed that methylation also plays

regulatory roles in the cell, including the control of DNA replication and transcription. These findings highlight the multifaceted nature of DNA methylation in bacterial organisms, indicating its involvement in both restriction-modification systems and cellular regulatory processes (Reisenauer *et al.*, 1999).

2.4.2 Restriction modification System

The presence and distribution of specific recognition sites for type II restriction-modification enzymes in a given species indicate a direct link between the avoidance of certain palindromic sequences and the presence of restriction-modification systems with corresponding specificity. It suggests that the bacterial DNA has encountered a wide range of restriction enzymes, possibly due to horizontal gene transfer facilitated by mobile genetic elements like plasmids and prophages (Gelfand & Koonin, 1997).

The DNA sequence containing CTAG is uncommon in enteric bacteria. Both XbaI (with the restriction site TCTAGA) and BlnI (CCTAGG) enzymes possess this rare sequence, and their cleavage sites occur less frequently than expected for enzymes with a 6-base pair specificity, unlike other enzymes of this type (Liu *et al.*, 1993). This scarcity is due to the significant under-representation of the CTAG tetranucleotide in bacterial DNA.

The CTAG tetranucleotide is not commonly found in bacterial DNA. It is frequently seen as a central part of the consensus sequence for insertion, which implies that SpeI and XbaI sites are likely to be preferred sites for insertion. This preference can influence the variability of digestion patterns produced by SpeI and XbaI enzymes when there are duplications or deletions of these

sites caused by the insertion or removal of genetic elements called IS481 and IS1002 (Stibitz, 1998). There are two enzymes, XbaI and BlnI, that can cut DNA at sites containing this rare sequence. However, these cutting sites are not as common as expected for enzymes that usually recognize six-base pair sequences, unlike other similar enzymes. Scientists believe that another gene of vsr is responsible for removing the CTAG sequence when it appears as a mismatch in the DNA. In other words, when there is an error in the DNA sequence, the vsr gene product repairs it by getting rid of the CTAG part. Researchers have found that the CTAG sequence, along with some other targets of the vsr enzyme, is not as abundant as it should be in the DNA. At the same time, the smaller pieces of DNA that result from the vsr repair process are more abundant than expected (Liu *et al.,* 1993).

2.4.3 Repair Mechanism

VSP (Very Short Patch or VSR (very short repair)) repair is a mechanism that corrects specific types of DNA mismatches, specifically T: G mismatches, by converting them into C: G pairs. These mismatches often occur in certain sequence contexts where cytosine methylation is present in DNA. The repair process helps reduce the harmful effects of 5-methylcytosine deamination, which can lead to the conversion of 5-methylcytosine (5mC) to thymine (T). By repairing these mismatches, VSP repair prevents mutagenic changes (Bhagwat & McClelland, 1992).

VSP repair is not only limited to fixing mismatches caused by 5-methylcytosine deamination, but it can also repair other types of mismatches. For example, when there is a T:G mismatches that doesn't occur due to this particular process. In cases where the original DNA base pair was T:A,

the VSP repair process changes the T to a C, resulting in a new base pair, C:G (Bhagwat & McClelland, 1992).

Researchers used a database of *E. coli* sequences to perform a Markov chain analysis to investigate the impact of VSP repair on DNA sequence composition. They examined whether repair at certain sites affected the abundance of specific tetranucleotides. The results of the study supported the idea that a specific type of DNA repair process, known as VSP repair, reduces the number of sequences in the genome that contains 'T' (like CTAG), while increasing the number of sequences that contains 'C' (like CCAG).This has implications for the scarcity of CTAG-containing restriction enzyme sites observed in the genomes of enteric bacteria (Bhagwat & McClelland, 1992).

Overall, this study highlights VSP repair as a significant factor in shaping the sequence composition of bacterial genomes. By selectively repairing T:G mismatches and influencing the abundance of specific tetranucleotides, VSP repair plays a role in maintaining genomic stability and impacting the distribution of DNA sequences in bacterial population (Bhagwat & McClelland, 1992)

2.4.4 Transposable elements

Transposable elements (TEs) are mobile DNA sequences that may shift their location within a genome, occasionally causing or reversing mutations and changing the genetic identity and genome size of the cell (Bourque *et al.*, 2018). TEs, which can introduce new genes or cause the loss of existing ones, may be responsible for the under-representation of CTAG in bacterial

genomes (Humayun *et al.,* 2017). TE activity is governed by complicated selection regimes within species, and actively transposing elements might accelerate genome evolution as well as promote genome expansion (Oggenfuss *et al.,* 2021). Many transposable element insertion processes include making staggered cuts at short sequences, resulting in "target site duplication" (TSD), which is commonly believed to be the transposition signature. The prevalence of transposable elements has long been considered as evidence of a highly developed and mutually beneficial system (Humayun *et al.,* 2017). Therefore, gene loss or the addition of novel genes caused by transposable elements may result in the under-representation of CTAG in bacterial genomes.

2.4.5 Context dependent mutation

Context dependent mutation are Replication error rates which includes the misincorporation of nucleotide and transient misalignments (Karlin *et al.,* 1994). CTAG under-representation in bacterial genomes is influenced by context-dependent mutation. Here are some factors that contribute to this phenomenon:

- Mutation rate: Mutation are the key source of genetic diversity, were considered to be determined by properties of relative generic sequence. Even though it is generally thought that site-specific mutation rates are independent of the local sequence context, a growing amount of evidence shows otherwise (Sung *et al.,* 2015) such as in eubacteria the avoidance of CTAG, which is caused by mutation of C → A (Karlin *et al.*, 1994).
- Selection efficiency: Context-dependent mutation has a stronger impact on genome architecture in species with small effective populations, which is consistent with diminished selection effectiveness in these organisms (Sung *et al.,* 2015).

2.5 Termination codon

Three codons, TGA, TAA, and TAG, are universally recognised by the genetic code as termination codons that signal to terminate the protein synthesis. Termination codon use can differ significantly between organisms and even within the same gene. It is likely that environmental adaptation and other factors contributed to this variation in stop codon patterns (Wong et al., 2008).

The three stop codons TAG, TGA, and TAA are included in the standard codon table for bacteria (UAG, UGA, and UAA on mRNA). However, their usage can be influenced by factors such as the GC content of the genome. TGA is the preferred stop codon for genes with high GC content, while TAA is preferred for genes with low GC content due to research demonstrating that the TAG codon usage is generally suboptimal in bacteria (Kirilov et al., 2013; Korkmaz et al., 2014; Wei et al., 2016).

It is important to note that the usage of stop codons in bacteria is not consistent and can vary across different species. Some studies suggest that both TAG and TGA are used less frequently in bacteria and archaea under GC pressure. The specific choice of stop codon depends on a complex interplay of evolutionary processes, including mutation bias, selection, and GC content pressure (Povolotskaya et al., 2012).

Codon usage varies extensively between species and even within genomes due to these evolutionary processes. When examining the evolution of termination codon usage, it's important

to consider "stop codon switch events," which involve changes between the three termination codons: TAA ↔ TGA, TAA ↔ TAG, and TGA ↔ TAG. However, it's unlikely that intermediate states using sense codons at the usual termination site are tolerated. Among the three termination codons, TAA stands out because it can withstand two types of errors: a single change that mutates TAA to TGA or TAG. In other words, TAA is less likely to be affected by certain mutation events compared to the other termination codons. On the other hand, TGA → TAA and TAG → TAA switches are also resistant to such mutations, but no single nucleotide change allows for direct switches between TGA and TAG (assuming double mutants are extremely rare). Because of these properties, TAA may be considered optimal for termination codon usage. Moreover, transitions between TGA and TAG must pass through TAA regardless of which one is preferred (Ho & Hurst, 2021).

2.6 CTAG cluster

In the genome of *Salmonella*, there are specific regions that contain a large number of sequences composed of the nucleotides CTAG. These regions are typically found near the starting point and ending point of DNA replication. These CTAG clusters are consistently present across different lineages of *Salmonella*, indicating their importance in the evolutionary history and genetic boundaries of the bacterium (Tang et al., 2014; Tang, Zhu, et al., 2017) .

CHAPTER 3

RESEARCH GAP

There are many reasons why CTAG sequences are under-represented in bacterial genomes, such as DNA methylation and modification systems, context-dependent mutations, transposable elements, and repair processes. This study aims to investigate potential biologically significant reasons behind the under-representation of CTAG in bacterial genomes, addressing a crucial research gap in the current understanding of bacterial genome composition.

Researchers like Burge and his colleagues have hypothesised that CTAG sequences that include overlapping, oppositely oriented TAG stop codons might have a detrimental effect. However, no specific evidence has been shown to indicate that TAG (a rare termination codon) is a key factor determining CTAG under-representation in the genome.

OBJECTIVES

- Analysis of CTAG distribution in bacterial genomes
- To determine the biological reasons behind the under-representation of CTAG

CHAPTER 4

METHODOLOGY

4.1 Genome sequence

Retrieve the FASTA sequence of and GenBank data of 23 different bacterial species, largely from different families and 5 eukaryotes with the partial sequence length (~5mbp) with the help of the NCBI web server. (https://www.ncbi.nlm.nih.gov/nuccore)

Table1: List of 23 diverse prokaryotes and 5 eukaryotes

S. NO	ORGANISMS	STRAINS	Family	ACCESSION NO.	SIZE (bp)
	PROKARYOTES				
1	***Agrobacterium tumefaciens***	12D1	*Rhizobiaceae.*	CP033031.1	3027766
2	***Anabaena cylindrica***	PCC 7122	*Nostocaceae*	NC_019771.1	6395836
3	***Bacillus subtilis***	168	*Bacillaceae*	CP010052.1	4215619
4	***Bordetella pertussis***	Tohama I	*Alcaligenaceae*	BX470248.1	4086189
5	***Corynebacterium diphtheriae***	NCTC 13129	*Corynebacteriaceae*	BX248353.1	2488635
6	***Deinococcus radiodurans***	ATCC 13939	*Deinococcaceae*	CP038663.1	2644543
7	***Erwinia carotovora***	SCRI1043	*Pectobacteriaceae*	BX950851.1	5064019
8	***Escherichia coli***	str. K-12	*Enterobacteriaceae*	NC_000913.3	4641652
9	***Haemophilus influenzae***	Hi375	*Pasteurellaceae.*	CP009610.1	1850897
10	***Klebsiella pneumoniae***	WRC19_AI1572C	*Enterobacteriaceae*	CP079634.1	4522788
11	***Lactobacillus acidophilus***	LA-G80-111	*Lactobacillaceae*	CP054559.1	1991976
12	***Leptospira interrogans***	FMAS_PN2	*Leptospiraceae*	CP092151.1	4302184
13	***Mycoplasmoides genitalium***	G37	*Metamycoplasmataceae*	NC_000908.2	580076
14	***Neisseria gonorrhoeae***	AT159	*Neisseriaceae*	CP097846.1	2232771
15	***Pseudomonas aeruginosa***	PAO1	*Pseudomonadaceae*	AE004091.2	6264404
16	***Salmonella enterica***	P-stx-12	*Enterobacteriaceae*	CP003278.1	4768352
17	***Serratia marcescens***	KS10	*Yersiniaceae*	CP027798.1	5199559
18	***Shigella flexneri 2a***	2457T	*Enterobacteriaceae*	AE014073.1	4599354
19	***Staphylococcus aureus***	JP080	*Staphylococcaceae*	AP017922.1	2729352
20	***Streptococcus agalactiae***	NGBS128	*Streptococcaceae*	CP012480.1	2074179
21	***Thermus aquaticus***	Y51MC23	*Thermaceae*	CP010822	2158963
22	***Xanthomonas oryzae***	NJ01	*Xanthomonadaceae*	CP092971.1	4966135
23	***Yersinia pestis***	CO92	*Yersiniaceae*	AL590842.1	4653728
	EUKARYOTES				
1	***Homo sapiens*** **(chromosome 21)**		*Hominidae*	NC_000021.9	20000001-24500000
2	***Saccharomyces cerevisiae*** **(chromosome IV)**	S288C	*Saccharomycetaceae*	NC_001136.10	1531933
3	***Arabidopsis thaliana*** **(chromosome 1)**		*Brassicaceae*	NC_003070.9	20000001-24500000
4	***Mus musculus*** **(chromosome 1)**		*Muridae*	NC_000067.7	8000001-12500000
5	***Sus scrofa*** **(chromosome 1)**	Duroc	*Suidae*	NC_010443.5	20000001-24500000

4.2 DETERMINATION OF TRI- AND TETRA-NUCLEOTIDE FREQUENCIES

A. Sequence Manipulation Suite (SMS):

DNA Stats is the web server that calculates the number of occurrences of each residue in the sequence and its probability and is used for the count of the mono nucleotide (G, A, T, C) in the genome. The result is used to find the probability of each selected residue for further analysis. (https://www.bioinformatics.org/sms2/dna_stats.html)

B. Chaos Game Representations of Frequencies (FCGR):

FCGR is the web server that determines the frequency of oligonucleotides in the DNA sequence and is was used to find the frequency of the tri and tetra-nucleotides in the studied genomes. The result is used to find the probability of each selected residue for further analysis.

(http://gscompare.ehu.eus/tools/CGR-FCGR/index.php?FCGR)

C. The **observed/expected** value is used to assess the frequency of a specific DNA sequence motif in a genome. It compares the observed number of occurrences of a particular motif to the number of occurrences that would be expected based on the nucleotide composition of the genome

$$\text{O/E Value} = \frac{\text{OBSERVED VALUE}}{\text{Product of Probabilities of bases at each position} \times \text{ length of the DNA sequence}}$$

D. In Excel, logical functions are used to perform various calculations and operations based on specific conditions. Some commonly used logical functions are IF; COUNT; COUNTIF; RIGHT; MID.

4.3 DISTRIBUTION OF CTAGs IN BACTERIAL GENOME

A. FUZZNUC TOOL

FUZZNUC is the embossed web server that Searches for patterns in nucleotide sequences. Also, represent the position and frequency of specific pattern. and used for the position and frequency of the CTAG sequence.

(https://embossgui.sourceforge.net/demo/fuzznuc.html)

B. The statistical tests applied in the analysis:

- **POISSON DISTRIBUTION:**

The Poisson distribution is a statistical probability distribution used to model the number of events occurring within a fixed interval of time or space. Calculate the probability CTAG of different arrivals in a given space using the Poisson distribution.

$$f(x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

Where,

e - the base of the logarithm

x - a Poisson random variable

λ - an average rate of value

4.4 PYTHON CODE FOR DETERMINING MOTIF FREQUENCIES

a) FOR EXTRACTING THE SUBSTRING OF THE GENES FROM THE WHOLE GENOME BY USING PYTHON CODE

```
import csv
import openpyxl
csv_file_path = "/csv_file.csv"
output_file_path = "/excel_file.xlsx"
def combine_all_rows (csv_file_path):
  combined_text = ""
  with open (csv_file_path, "r") as csv_file:
    csv_reader = csv.reader (csv_file)
    for row in csv_reader:
      for cell in row:
        combined_text += ''. join (char for char in cell if char.isalpha ())
  return combined_text
substrings_to_extract = [ ]
def extract_substrings (text, substrings):
  extracted_substrings = [ ]
  for start_index, end_index in substrings:
    substring = text[start_index:end_index]
    extracted_substrings.append (substring)
```

```
  return extracted_substrings
def save_to_excel (csv_file_path,excel_file_path,extracted_substrings):
  wb = openpyxl.load_workbook (excel_file_path)
  sheet = wb.active
  with open (csv_file_path, "r") as csv_file:
    csv_reader = csv.reader (csv_file)
    next (csv_reader)
    row_index = 2
    for extracted_substring in extracted_substrings:
      cell = sheet.cell (row=row_index, column=6)
      cell.value = extracted_substring
      row_index += 1
  wb.save (output_file_path)
  wb.close ()
combined_text = combine_all_rows (csv_file_path)
print (f"Combined alphabets in all cells: {combined_text}")
extracted_substrings = extract_substrings (combined_text, substrings_to_extract)
save_to_excel (csv_file_path, output_file_path, extracted_substrings)
print ("Extracted substrings saved to Excel file.")
```

b) USING THE PYTHON CODE FOR REVERSE COMPLEMENT THE COMPLEMENT SEQUENCE

```
import pandas as pd
from Bio.Seq import Seq
excel_file_path = '/Excel file.xlsx'
excel_file = pd.ExcelFile (excel_file_path)
dfs = {}
for sheet_name in excel_file.sheet_names:
  df = pd.read_excel (excel_file, sheet_name=sheet_name)
  for index, row in df.iterrows ():
    seq_type = row['SENSE/ COMPLEMENT']
    sequence = row['SEQUENCE']
    if seq_type == 'complement':
      rev_complement = str (Seq (str (sequence)).reverse_complement ())
    elif seq_type == 'sense':
      rev_complement = str (Seq (str (sequence)).complement ())
    else:
      rev_complement = sequence
    df.at[index, 'Reverse Complement'] = rev_complement
  dfs[sheet_name] = df
with pd.ExcelWriter ('/excel file.xlsx') as writer:
  for sheet_name, df in dfs.items ():
    df.to_excel (writer, sheet_name=sheet_name, index=False)
```

```
print ("Reverse complement sequences have been written to the Excel file.")
```

c) USING A PYTHON CODE TO COUNT THE OLIGONUCLEOTIDE WITHIN THE CODING REGION

```
import pandas as pd
from openpyxl import Workbook
data = pd.read_excel ('/excel file.xlsx')
wb = Workbook ()
sheet = wb.active
sheet['A1'] = 'DATA'
sheet['B1'] = 'TAA'
sheet['C1'] = 'ATT'
sheet['D1'] = 'TAG'
sheet['E1'] = 'CTA'
sheet['F1'] = 'TGA'
sheet['G1'] = 'TCA'
sheet['H1'] = 'CTAG'
sheet['I1'] = 'GTAG'
sheet['J1'] = 'TTAG'
sheet['K1'] = 'ATAG'
sheet['L1'] = 'CTAA'
sheet['M1'] = 'CTGA'
sheet['N1'] = 'TCAG'
```

```
sheet['O1'] = 'CTAT'
sheet['P1'] = 'CTAC'
for i, j in data.iterrows ():
  stringdata = j[0]
  sheet.cell (row=i+2, column=1).value = stringdata
  sheet.cell (row=i+2, column=2).value = stringdata.count ('TAA')
  sheet.cell (row=i+2, column=3).value = stringdata.count ('ATT')
  sheet.cell (row=i+2, column=4).value = stringdata.count ('TAG')
  sheet.cell (row=i+2, column=5).value = stringdata.count ('CTA')
  sheet.cell (row=i+2, column=6).value = stringdata.count ('TGA')
  sheet.cell (row=i+2, column=7).value = stringdata.count ('TCA')
  sheet.cell (row=i+2, column=8).value = stringdata.count ('CTAG')
  sheet.cell (row=i+2, column=9).value = stringdata.count ('GTAG')
  sheet.cell (row=i+2, column=10).value = stringdata.count ('TTAG')
  sheet.cell (row=i+2, column=11).value = stringdata.count ('ATAG')
  sheet.cell (row=i+2, column=12).value = stringdata.count ('CTAA')
  sheet.cell (row=i+2, column=13).value = stringdata.count ('CTGA')
  sheet.cell (row=i+2, column=14).value = stringdata.count ('TCAG')
  sheet.cell (row=i+2, column=15).value = stringdata.count ('CTAT')
  sheet.cell (row=i+2, column=16).value = stringdata.count ('CTAC')
wb.save ('/excel file _result.xlsx')
```

CHAPTER- 5

RESULT

Genome sequence heterogeneity may be discerned in the form of under- and over-representation of DNA sequence motifs. The basis of such under- and over-representation of certain motifs seems to be evolutionary in nature. One of the most conspicuous sequence heterogeneity that has been found in majority of the bacterial genomes is the under-representation of CTAG. There are several reasons that have been reported for the under-representation of the CTAG, such as, their involvement in very short repair (vsr) patches, restriction site and DNA methylation. In addition to those possible reasons the presence of TAG submotif, one of the three termination codons, in CTAG has been investigated as a possible reason for the under-representation of in this work.

5.1 Genome analysis for abundance of CTAG in prokaryotes

A randomly selected set of 23 bacterial genomes, largely from different families have been analyzed to study CTAG under-representation. The CTAG tetranucleotide motif was found to be under-represented (Observed / Expected value < 1.0) in all the 23 bacterial genomes (Table 2). The CTAG O/E value was found to be the least among all the possible 256 tetranucleotides in 15 out of 23 the bacterial genomes. However, in rest of the 8 bacterial genomes, there were few to several tetranucleotides which had O/E value lower than CTAG. In these bacteria, the CTAG O/E value was also much higher than the 15 bacteria. A few eukaryotic genomic sequences were also studied as a reference in which higher O/E values of CTAG as well as its higher rank among other tetranucleotides was observed (Fig. 1). Additionally, O/E value of CTAG were compared to the 23 other tetranucleotide permutation of identical base composition in all the bacterial genome

(Table 3). The O/E value of CTAG were very low in comparison to that of all other permutation (Fig. 2)

Table2: Frequency of bases and O/E value of CTAG in prokaryotic genomes

ORGANISMS PROKARYOTES LEAST ABUNDANCE	G	A	T	C	G+A+T+C	OBSERVED CTAG	EXPECTED CTAG	O/E CTAG	No. of tetranucleotide having O/E value less than CTAG
Agrobacterium tumefaciens	893292	618175	608785	907514	3027766	734	10991	0.067	0
Bacillus subtilis	915119	1188075	1193138	919287	4215619	3109	15918	0.195	0
Bordetella pertussis	1388992	659351	659647	1378199	4086189	1074	12204	0.088	0
Deinococcus radiodurans	885992	434938	435568	888045	2644543	900	8059	0.112	0
Erwinia carotovora	1290827	1244797	1238227	1290168	5064019	3938	19766	0.199	0
Escherichia coli	1177437	1142742	1141382	1180091	4641652	885	18122	0.049	0
Klebsiella pneumoniae	1285458	969403	970252	1297675	4522788	1002	16959	0.059	0
Neisseria gonorrhoeae	583734	534178	528783	586076	2232771	585	8682	0.067	0
Pseudomonas aeruginosa	2066633	1056134	1038950	2102687	6264404	1545	19396	0.080	0
Salmonella enterica	1244732	1141055	1144705	1237860	4768352	1024	18563	0.055	0
Serratia marcescens	1549024	1045635	1052774	1552026	5199559	1095	18827	0.058	0
Shigella flexneri	1168009	1131104	1126890	1173323	4599326	1258	17954	0.070	0
Staphylococcus aureus	451354	913288	916527	448183	2729352	4211	8328	0.506	0
Xanthomonas oryzae	1587647	896360	899363	1582765	4966135	1972	16540	0.119	0
Yersinia pestis	1114185	1219520	1217353	1102670	4653728	3396	18097	0.188	0
NOT LEAST ABUNDANCE									
Anabaena cylindrica	1241193	1961115	1953384	1240144	6395836	14091	22538	0.625	74
Corynebacterium diphtheriae	665651	578357	579393	665234	2488635	5093	9627	0.529	1
Haemophilus influenzae	353709	573742	570833	352613	1850897	2007	6442	0.312	1
Lactobacillus acidophilus	351462	650770	649863	339881	1991976	4036	6391	0.631	29
Leptospira interrogans	750968	1396611	1398192	756413	4302184	5765	13930	0.414	4
Mycoplasmoides genitalium	92306	200544	195711	91515	580076	1575	1699	0.927	125
Streptococcus agalactiae	372580	672332	660796	368471	2074179	5437	6835	0.795	65
Thermus aquaticus	732676	346223	342558	737506	2158963	2230	6368	0.350	29
EUKARYOTES									
Homo sapiens	791391	1464022	1460946	783641	4500000	11409	14556	0.784	73
Saccharomyces cerevisiae	291364	476761	474471	289337	1531933	3409	5304	0.643	6
Arabidopsis thaliana	793260	1463491	1449697	793552	4500000	10188	14656	0.695	29
Mus musculus	885326	1372148	1364646	877880	4500000	14072	15971	0.881	91
Sus scrofa	848472	1409994	1387789	853745	4500000	14034	15555	0.902	98

*A randomly selected sample genomic sequences (>5 mbp) of eukaryotes (*Homo sapiens, Saccharomyces cerevisiae, Arabidopsis thaliana, Mus musculus* and *Sus scrofa*) were analyzed.

Fig. 1: Comparing the O/E value of CTAG in the genome of a randomly selected organism

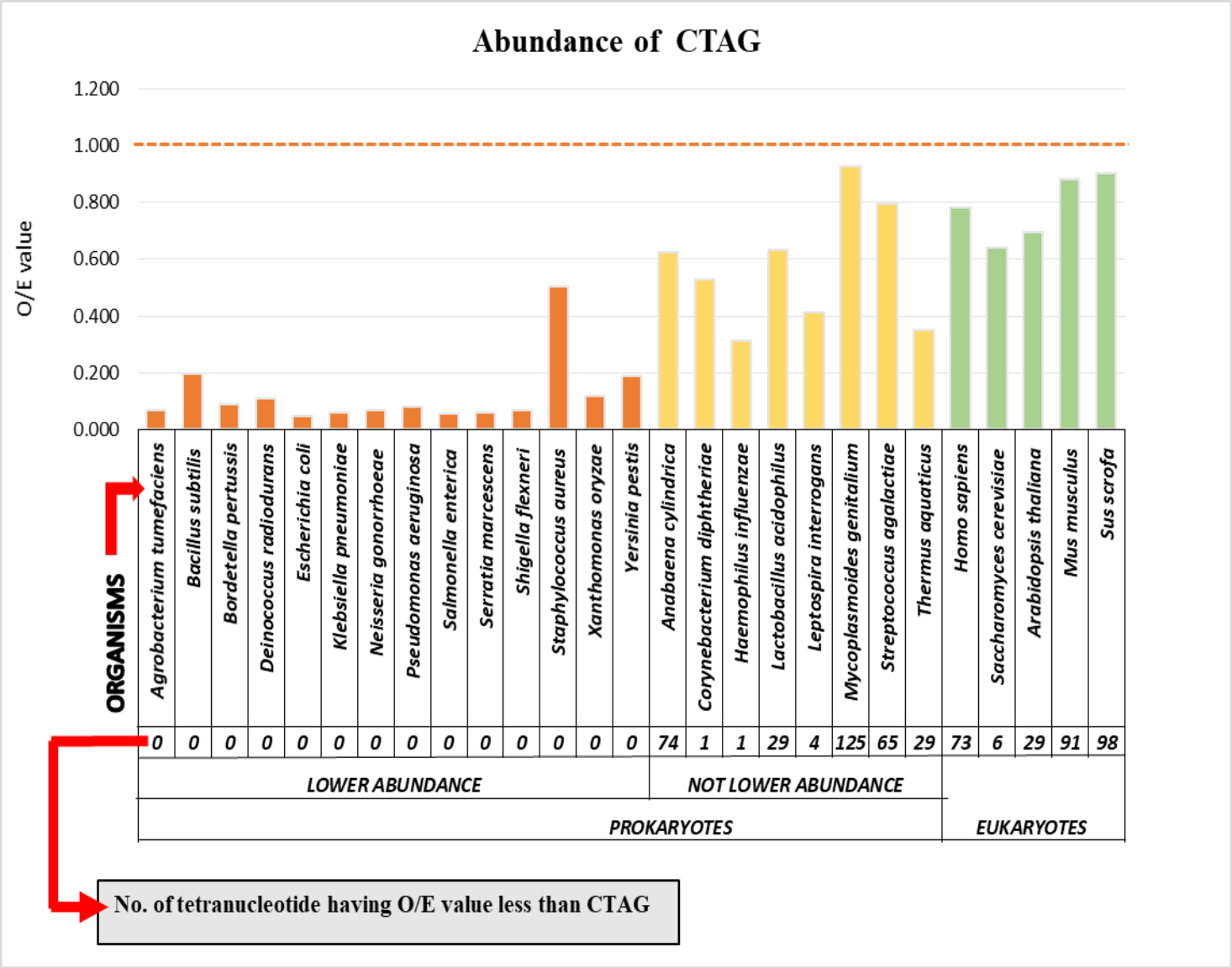
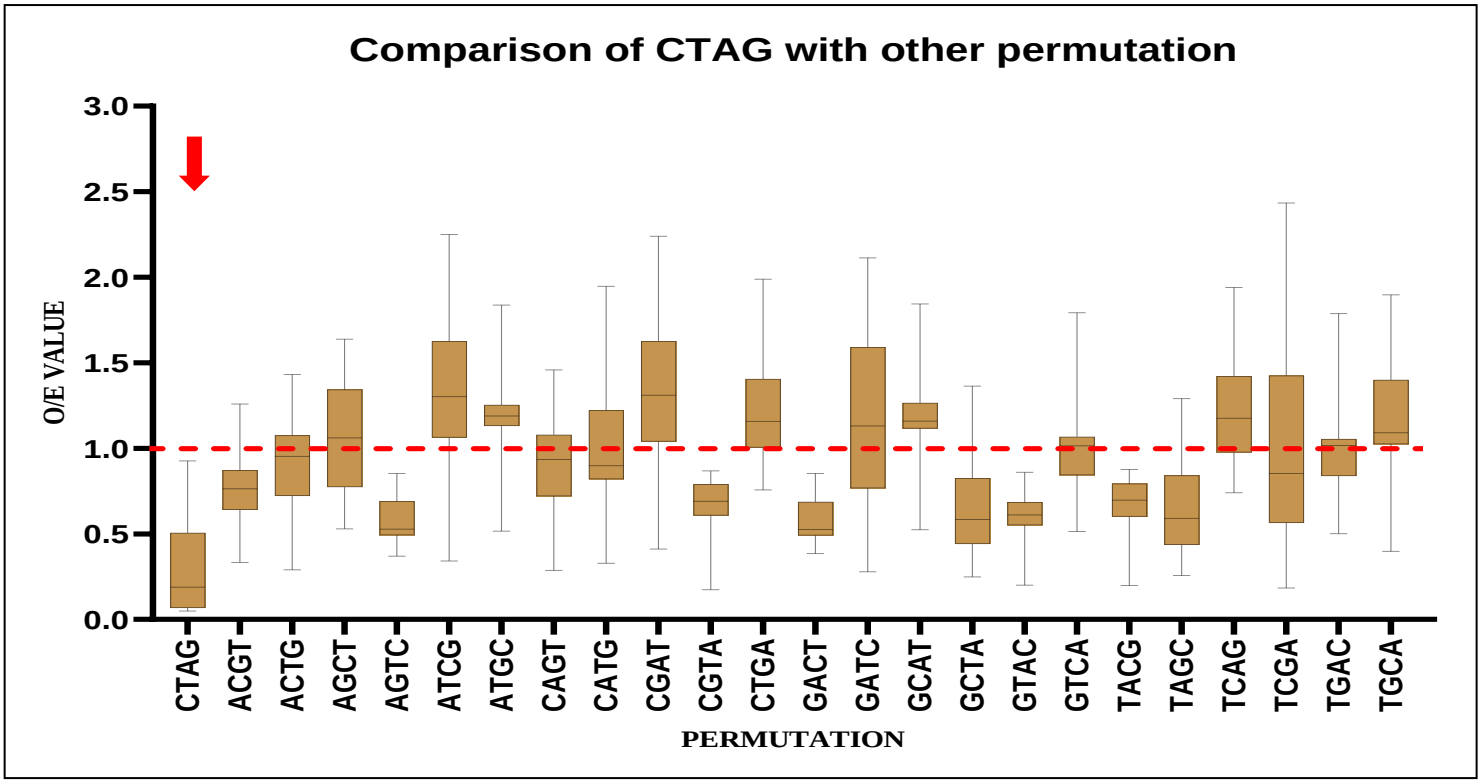


Table 3: O/E value of CTAG and their permutation in the bacterial genome

		PERMUTATIONS																						
BACTERIAL SPECIES	CTAG	ACGT	ACTG	AGCT	AGTC	ATCG	ATGC	CAGT	CATG	CGAT	CGTA	CTGA	GACT	GATC	GCAT	GCTA	GTAC	GTCA	TACG	TAGC	TCAG	TCGA	TGAC	TGCA
Agrobacterium tumefaciens	0.067	0.518	0.517	0.866	0.409	2.184	1.496	0.513	1.577	2.152	0.429	1.126	0.390	1.934	1.492	0.249	0.201	1.164	0.435	0.256	1.128	2.074	1.127	0.944
Anabaena cylindrica	0.806	0.407	1.076	1.130	0.811	0.779	0.905	1.079	0.585	0.757	0.459	1.178	0.786	0.722	0.895	1.160	0.608	0.926	0.465	1.169	1.176	0.383	0.924	1.103
Bacillus subtilis	0.195	0.568	0.803	1.405	0.554	1.153	1.141	0.791	1.032	1.141	0.608	1.422	0.566	1.131	1.137	0.470	0.490	1.040	0.600	0.468	1.415	0.688	1.044	1.050
Bordetella pertussis	0.088	0.872	0.679	1.061	0.531	2.014	1.677	0.699	1.947	2.006	0.735	1.023	0.513	1.739	1.682	0.400	0.722	1.082	0.727	0.396	0.988	2.254	1.076	1.365
Corynebacterium diphtheriae	0.529	0.880	0.893	1.149	0.706	1.441	1.149	0.935	1.222	1.457	0.743	0.925	0.688	1.414	1.154	0.759	0.645	0.918	0.741	0.776	0.931	1.425	0.934	1.167
Deinococcus radiodurans	0.148	1.259	1.128	1.343	0.827	1.172	1.131	1.103	1.239	1.114	0.773	1.989	0.838	0.692	1.116	0.431	0.798	1.792	0.846	0.389	1.940	1.851	1.787	1.432
Erwinia carotovora	0.202	0.884	1.040	0.814	0.528	1.545	1.201	1.046	0.928	1.546	0.819	1.456	0.529	1.131	1.204	0.758	0.617	1.079	0.809	0.772	1.465	0.886	1.081	1.073
Escherichia coli	0.049	0.803	1.128	0.736	0.518	1.344	1.200	1.130	0.842	1.339	0.791	1.345	0.526	1.055	1.197	0.585	0.664	1.015	0.779	0.586	1.360	0.853	1.026	1.091
Haemophilus influenzae	0.312	0.871	0.751	0.891	0.420	1.071	1.199	0.747	0.328	1.083	0.846	0.924	0.436	0.767	1.159	0.825	0.557	0.669	0.806	0.826	0.891	0.565	0.690	1.584
Klebsiella pneumoniae	0.059	0.693	0.817	1.123	0.492	1.627	1.132	0.806	0.893	1.627	0.634	1.429	0.507	1.491	1.142	0.630	0.581	1.052	0.652	0.598	1.423	1.074	1.017	1.024
Lactobacillus acidophilus	0.631	0.754	1.074	1.414	0.691	0.883	1.193	1.032	0.899	0.842	0.643	1.070	0.742	1.151	1.178	1.201	0.716	0.855	0.618	1.154	1.063	0.545	0.881	1.514
Leptospira interrogans	0.414	0.810	0.659	0.577	0.700	1.444	0.516	0.665	0.402	1.443	0.835	0.908	0.688	1.592	0.524	0.398	0.324	0.514	0.832	0.401	0.907	1.305	0.501	0.852
Mycoplasmoides genitalium	0.927	0.334	1.432	1.638	0.507	0.341	0.984	1.458	0.675	0.411	0.174	1.225	0.496	1.342	0.932	1.363	0.479	0.696	0.198	1.29	1.182	0.184	0.712	1.401
Neisseria gonorrhoeae	0.067	0.549	0.630	0.530	0.466	1.302	1.254	0.606	0.522	1.296	0.785	1.033	0.455	0.279	1.251	0.304	0.551	0.760	0.775	0.292	1.019	1.050	0.737	0.981
Pseudomonas aeruginosa	0.099	0.715	0.920	1.075	0.637	1.827	1.251	0.910	1.333	1.819	0.626	1.157	0.633	2.114	1.265	0.464	0.713	0.844	0.613	0.490	1.190	2.434	0.838	1.210
Salmonella enterica	0.055	0.813	0.991	0.752	0.516	1.401	1.171	0.980	0.865	1.420	0.869	1.295	0.526	1.010	1.143	0.700	0.686	1.018	0.877	0.695	1.284	0.752	1.032	0.851
Serratia marcescens	0.061	0.763	0.722	1.087	0.401	1.926	1.279	0.721	1.030	1.961	0.571	1.398	0.405	2.073	1.284	0.442	0.464	0.984	0.566	0.440	1.429	1.313	0.969	1.240
Shigella flexneri	0.070	0.823	1.172	0.773	0.535	1.296	1.190	1.180	0.850	1.311	0.807	1.412	0.542	1.017	1.178	0.585	0.672	1.029	0.793	0.591	1.441	0.819	1.042	1.074
Staphylococcus aureus	0.506	1.038	0.980	1.028	0.654	1.063	1.348	0.949	1.055	1.040	0.790	0.994	0.679	0.609	1.364	0.979	0.860	1.014	0.787	0.970	0.965	0.792	1.021	1.647
Streptococcus agalactiae	0.795	0.756	0.953	1.521	0.854	0.748	0.946	0.915	0.820	0.756	0.603	1.125	0.853	0.584	0.905	1.275	0.582	1.071	0.586	1.252	1.112	0.527	1.095	1.043
Thermus aquaticus	0.350	0.641	0.290	1.445	0.370	0.555	0.641	0.286	1.097	0.494	0.621	0.757	0.384	0.873	0.639	0.583	0.563	0.605	0.698	0.556	0.740	0.304	0.619	0.398
Xanthomonas oryzae	0.119	0.856	0.957	1.008	0.494	2.249	1.837	0.951	1.408	2.239	0.643	1.005	0.491	2.038	1.844	0.457	0.614	0.943	0.649	0.436	0.977	1.644	0.949	1.897
Yersinia pestis	0.188	0.711	1.206	0.708	0.526	1.134	1.134	1.241	0.894	1.145	0.690	1.406	0.519	1.011	1.139	0.819	0.610	1.068	0.689	0.841	1.422	0.655	1.055	1.045

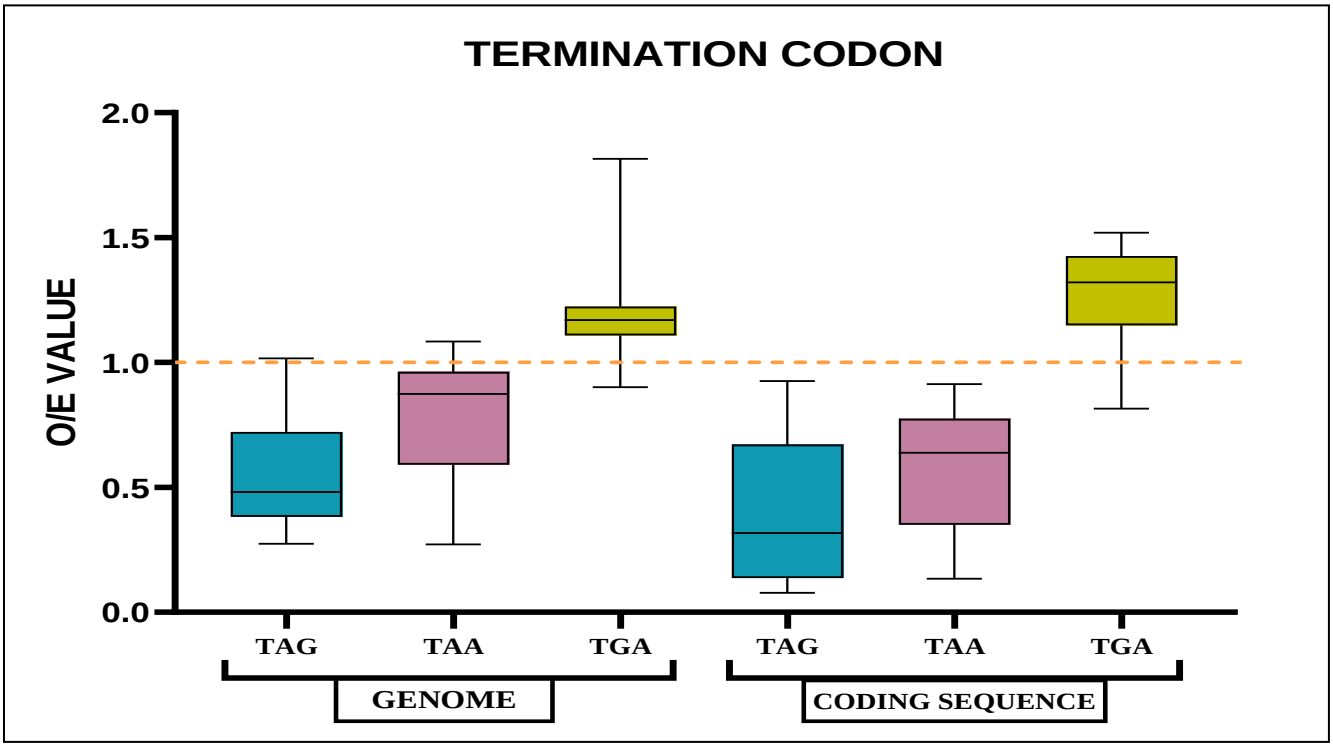
Fig. 2: Comparison of CTAG abundance with other tetranucleotides with same base composition



The CTAG consists of a submotif TAG, which is a termination codon. In order to investigate if presence of this submotif is contributing to the under-representation of CTAG, the abundance of all the three termination codons (TAA, TAG, and TGA) was determined for all of the studied bacterial genomes (Table 4). It was found that TAG is the least abundant termination codon as it has the lowest O/E value in all the bacterial genomes under the study. Further, their O/E value was determined in the coding sequence and at the termination site (trinucleotide functioning as the termination codon) of the protein coding genes. It was found that TAG was the least common of the three termination codons in coding sequences as well as at the termination codon sites (Fig. 3). The abundance of the three termination codons (TAG, TGA, and TAA) follows an order, viz. O/E value: TAG<TAA<TGA in the genomic sequence as well as the coding sequence. At termination codon sites, the order is O/E value: TAG<TGA<TAA. This is perfectly in agreement with the earlier reports on the abundance of termination codon in bacterial genome (Ho & Hurst, 2021). Interestingly, this pattern remains unaltered when these codons have a C nucleotide at their 5' end for genomics and coding sequence. In case of termination codon site, the O/E value exhibit a minor change in the order viz. O/E value: TAG<TAA<TGA.

Table 4: O/E value of three termination codons in the genome, coding sequence, and at the termination site

BACTERIAL SPECIES	TERMINATION CODON								
	GENOME			CONDING SEQUENCE			TERMINATION SITE		
	TAG	TAA	TGA	TAG	TAA	TGA	TAG	TAA	TGA
Agrobacterium tumefaciens	0.274	0.478	1.258	0.078	0.244	1.307	0.000	33.982	59.049
Anabaena cylindrica	0.871	0.901	1.053	0.926	0.693	1.099	13.205	19.238	11.424
Bacillus subtilis	0.385	0.748	1.283	0.319	0.590	1.469	8.182	27.953	13.280
Bordetella pertussis	0.390	0.308	1.167	0.085	0.134	1.138	27.640	35.265	68.554
Corynebacterium diphtheriae	0.630	0.699	1.125	0.463	0.574	1.221	21.170	41.432	11.929
Deinococcus radiodurans	0.459	0.589	1.815	0.090	0.216	1.520	11.449	59.622	69.471
Erwinia carotovora	0.494	0.910	1.234	0.316	0.687	1.429	9.013	37.613	19.987
Escherichia coli	0.382	0.995	1.172	0.255	0.782	1.405	4.428	41.599	20.317
Haemophilus influenzae	0.572	0.940	1.030	0.607	0.723	1.201	7.965	25.711	5.066
Klebsiella pneumoniae	0.406	1.042	1.219	0.204	0.829	1.441	7.913	57.279	25.411
Lactobacillus acidophilus	0.768	0.959	1.149	0.725	0.891	1.351	11.137	20.741	3.636
Leptospira interrogans	0.677	0.764	0.901	0.687	0.574	0.815	7.898	16.654	15.431
Mycoplasmoides genitalium	0.875	0.964	1.221	0.890	0.913	1.474	15.265	17.689	0.180
Neisseria gonorrhoeae	0.291	0.732	0.942	0.164	0.483	1.027	7.826	36.195	26.559
Pseudomonas aeruginosa	0.469	0.272	1.112	0.109	0.135	1.204	12.671	20.643	87.409
Salmonella enterica	0.428	1.084	1.140	0.280	0.862	1.335	6.896	42.276	21.035
Serratia marcescens	0.363	0.862	1.263	0.128	0.523	1.417	8.077	61.264	32.716
Shigella flexneri	0.379	0.980	1.190	0.248	0.775	1.415	4.884	43.055	18.772
Staphylococcus aureus	0.716	0.926	1.185	0.807	0.775	1.353	8.131	19.561	6.104
Streptococcus agalactiae	0.836	0.887	1.165	0.784	0.723	1.302	10.043	19.509	8.663
Thermus aquaticus	0.724	0.596	1.040	0.343	0.463	0.991	42.884	49.274	49.639
Xanthomonas oryzae	0.340	0.377	1.107	0.127	0.229	1.094	11.182	35.817	64.290
Yersinia pestis	0.521	1.016	1.224	0.382	0.765	1.445	8.780	31.590	17.575



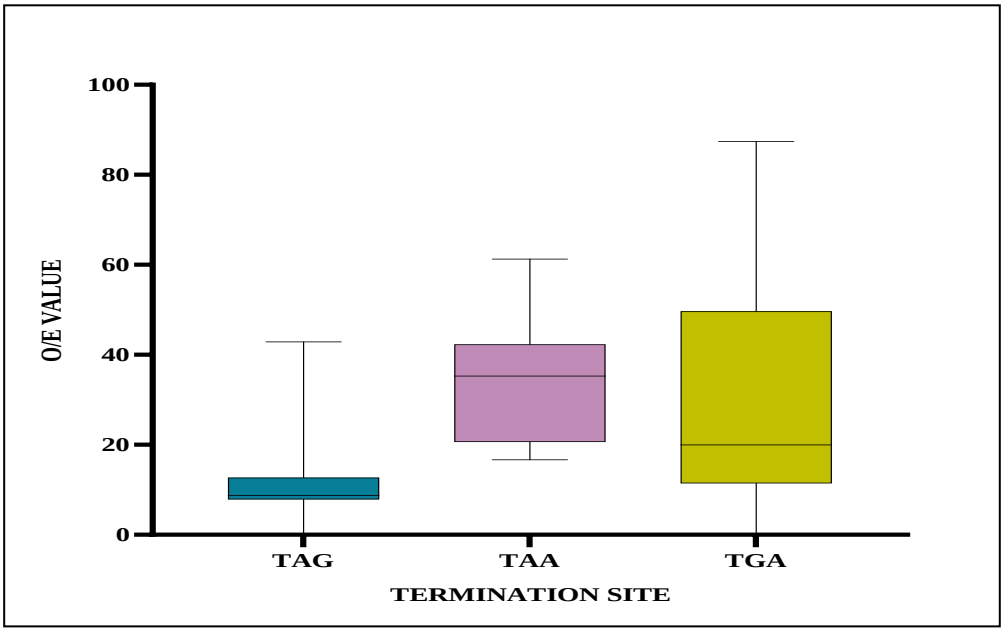


Fig. 3: Comparison of three termination codons (A) in the entire genome, coding sequence and (B) at the termination site.

5.3 Effect of 5' cytosine on abundance of TAG, TGA and TAA

Table 5: O/E value of Cytosine preceding termination codon on in the genome, coding sequence, and at the termination site

BACTERIAL SPECIES	Cytosine preceding termination codon								
	GENOME			CONDING SEQUENCE			TERMINATION SITE		
	CTAG	CTAA	CTGA	CTAG	CTAA	CTGA	CTAG	CTAA	CTGA
Agrobacterium tumefaciens	0.067	0.143	1.126	0.030	0.095	1.549	0.000	24.049	98.342
Anabaena cylindrica	0.806	0.977	1.178	0.816	0.697	1.207	7.215	14.652	8.773
Bacillus subtilis	0.195	0.441	1.422	0.172	0.353	1.582	4.713	15.463	12.241
Bordetella pertussis	0.088	0.095	1.023	0.064	0.051	1.535	43.058	18.141	119.360
Corynebacterium diphtheria	0.529	0.543	0.925	0.491	0.548	1.120	27.341	55.133	9.829
Deinococcus radiodurans	0.148	0.269	1.989	0.059	0.142	2.024	9.742	37.077	120.744
Erwinia carotovora	0.202	0.493	1.456	0.184	0.439	1.652	6.260	29.479	29.519
Escherichia coli	0.049	0.430	1.345	0.035	0.366	1.585	1.537	22.907	27.252
Haemophilus influenzae	0.312	0.821	0.924	0.281	0.510	1.064	4.014	17.424	2.341
Klebsiella pneumoniae	0.059	0.413	1.429	0.042	0.371	1.748	5.453	43.966	42.558
Lactobacillus acidophilus	0.631	1.000	1.070	0.620	1.010	1.264	7.610	15.526	2.367
Leptospira interrogans	0.414	0.849	0.908	0.444	0.669	0.831	2.075	14.065	13.079
Mycoplasmoides genitalium	0.927	1.174	1.225	0.961	1.148	1.417	14.798	24.102	0.000
Neisseria gonorrhoeae	0.067	0.260	1.033	0.044	0.193	1.211	3.521	16.932	41.436
Pseudomonas aeruginosa	0.099	0.111	1.157	0.055	0.067	1.748	16.778	15.491	165.061
Salmonella enterica	0.055	0.469	1.295	0.043	0.433	1.483	3.703	27.395	30.641
Serratia marcescens	0.061	0.313	1.398	0.034	0.227	1.811	5.480	35.265	53.147
Shigella flexneri	0.070	0.418	1.412	0.055	0.363	1.658	2.584	21.832	28.777
Staphylococcus aureus	0.506	0.866	0.994	0.503	0.650	1.131	5.145	12.904	1.543
Streptococcus agalactiae	0.795	0.988	1.125	0.773	0.843	1.113	11.337	15.836	5.125
Thermus aquaticus	0.350	0.638	0.757	0.328	0.629	1.014	55.578	79.255	36.403
Xanthomonas oryzae	0.119	0.197	1.005	0.073	0.133	1.170	9.233	29.698	87.823
Yersinia pestis	0.188	0.627	1.406	0.158	0.531	1.593	6.219	24.104	23.961

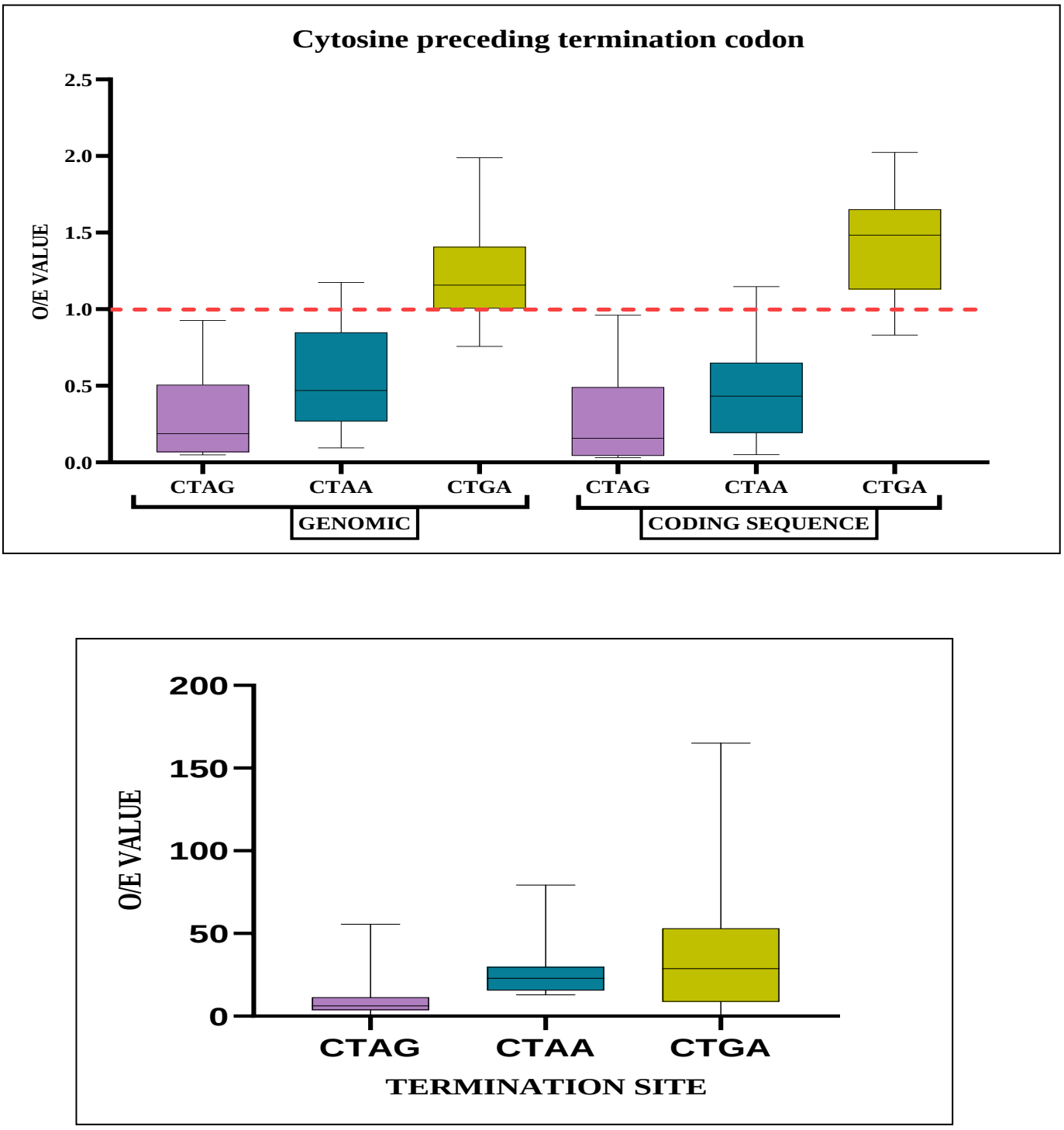


Fig. 4: Comparison of Cytosine preceding termination codons (A) in the genome, coding sequence, and (B) at the termination site.

In addition to TAG, a termination codon, the other part of CTAG is a preceding cytosine. To determine the effect of this preceding C (at 5'end of the termination codon) abundance of CTAG was compared with that of similar tetranucleotides consisting of the rest of the two termination codons, viz. CTAA and CTGA. The O/E values of CTAG was found to be lower than CTGA and CTAA in the genomic sequences of the bacteria (Table5). The abundance of the three tetranucleotides followed the order of their respective submotifs, O/E value: CTAG < CTAA < CTGA. An identical pattern of O/E values was seen when the three tetranucleotides were analysed in the coding sequences and at the termination sites (Fig4). CTAG is found to be less abundant in coding sequence than whole genome.

This result indicates that occurrence of a cytosine at the 5' end of the termination codons fails to change the pattern of their abundance. However, the cytosine at 5' end might have made the differences either stronger or weaker in a uniform fashion for all the three termination codons. To study any such effect, an index based on the ratio of O/E values of (CTGA+CTAA) and CTAG and the ratio of O/E value of (TGA+TAA) and TAG was used (Table6). This index value explained if O/E value of CTAG is greater or smaller than that of TAG that too in comparison with the respective values of the other two termination codons. The index values were found to be greater than 1.0, which means that TAG preceded by a cytosine is more strongly under-represented than TAG alone in the genome. In some of the genomes the effect was several old difference. Identical effect was observed in the coding sequences also. Similar through less pronounced results were obtained at the termination sites which clearly showed that TAG is the least abundant termination codon, yet occurrence of C at its 5' end makes it further suppressed.

Table 6: Ratio of the $\frac{\text{OBSERVED (CTAA+CTGA)}}{\frac{\text{EXPECTED (CTAA+CTGA)}}{\text{O/E (CTAG)}}}$ to the $\frac{\text{OBSERVED(TAA+TGA)}}{\frac{\text{EXPECTED (TAA+TGA)}}{\text{O/E (TAG)}}}$

THE RATIOS									
BACTERIAL SPECIES	$\frac{\text{OBSERVED (CTAA + CTGA)}}{\frac{\text{EXPECTED (CTAA + CTGA)}}{\text{O/E (CTAG)}}}$			$\frac{\text{OBSERVED(TAA + TGA)}}{\frac{\text{EXPECTED (TAA + TGA)}}{\text{O/E (TAG)}}}$			$\frac{\frac{\text{O/E(CTAA + CTGA)}}{\text{O/E(CTAG)}}}{\frac{\text{O/E(TAA + TGA)}}{\text{O/E(TAG)}}}$		
	GENOME	CONDING REGION	TERMINATION SITE	GENOME	CONDING REGION	TERMINATION SITE	GENOME	CONDING REGION	TERMINATION SITE
Agrobacterium tumefaciens	10.845	31.556	#DIV/0!	3.431	11.204	#DIV/0!	**3.161**	**2.817**	**#DIV/0!**
Anabaena cylindrica	1.309	1.097	1.715	1.103	0.918	1.227	**1.187**	**1.195**	**1.397**
Bacillus subtilis	4.444	5.168	2.984	2.550	3.051	2.636	**1.742**	**1.694**	**1.132**
Bordetella pertussis	8.233	16.505	2.015	2.285	9.567	2.093	**3.603**	**1.725**	**0.963**
Corynebacterium diphtheriae	1.413	1.740	1.130	1.473	1.989	1.211	**0.960**	**0.875**	**0.933**
Deinococcus radiodurans	9.603	23.852	9.567	3.074	12.146	5.784	**3.123**	**1.964**	**1.654**
Erwinia carotovora	4.880	5.735	4.712	2.175	3.369	3.178	**2.243**	**1.702**	**1.483**
Escherichia coli	18.318	28.189	16.341	2.838	4.310	6.955	**6.456**	**6.540**	**2.349**
Haemophilus influenzae	2.762	2.567	2.908	1.704	1.492	2.239	**1.621**	**1.721**	**1.299**
Klebsiella pneumoniae	16.792	27.322	7.916	2.814	5.764	4.943	**5.967**	**4.740**	**1.601**
Lactobacillus acidophilus	1.622	1.773	1.434	1.335	1.452	1.324	**1.215**	**1.221**	**1.083**
Leptospira interrogans	2.101	1.633	6.614	1.199	0.959	2.055	**1.752**	**1.702**	**3.219**
Mycoplasmoides genitalium	1.284	1.283	1.115	1.194	1.224	0.797	**1.075**	**1.048**	**1.399**
Neisseria gonorrhoeae	9.854	16.313	8.442	2.894	4.681	3.982	**3.405**	**3.485**	**2.120**
Pseudomonas aeruginosa	8.127	21.390	6.823	1.764	7.757	5.116	**4.607**	**2.757**	**1.334**
Salmonella enterica	16.313	22.608	7.856	2.603	3.961	4.524	**6.266**	**5.708**	**1.737**
Serratia marcescens	15.713	34.522	8.383	3.038	8.274	5.475	**5.173**	**4.173**	**1.531**
Shigella flexneri	13.175	18.490	9.814	2.871	4.438	6.290	**4.589**	**4.166**	**1.560**
Staphylococcus aureus	1.796	1.609	1.778	1.414	1.197	1.858	**1.271**	**1.344**	**0.957**
Streptococcus agalactiae	1.303	1.215	1.060	1.179	1.186	1.558	**1.105**	**1.024**	**0.681**
Thermus aquaticus	2.053	2.712	0.902	1.240	2.393	1.155	**1.655**	**1.133**	**0.781**
Xanthomonas oryzae	5.982	10.867	7.240	2.483	6.143	4.831	**2.409**	**1.769**	**1.499**
Yersinia pestis	5.323	6.570	3.865	2.142	2.850	2.836	**2.485**	**2.305**	**1.363**

5.4 Comparison of CTAG with their DTAGs

Table 7: O/E value of DTAG in the genome, coding sequence, and at the termination site

BACTERIAL SPECIES	O/E value of DTAG											
	GENOME				CONDING REGION				TERMINATION SITE			
	CTAG	GTAG	ATAG	TTAG	CTAG	GTAG	ATAG	TTAG	CTAG	GTAG	ATAG	TTAG
Agrobacterium tumefaciens	0.067	0.270	0.709	0.145	0.030	0.073	0.189	0.043	0.000	0.000	0.000	0.000
Anabaena cylindrica	0.806	0.939	0.732	1.008	0.816	0.985	0.614	1.256	7.215	12.396	12.309	18.422
Bacillus subtilis	0.195	0.354	0.512	0.427	0.172	0.325	0.295	0.448	4.713	8.300	12.929	6.036
Bordetella pertussis	0.088	0.652	0.759	0.098	0.064	0.086	0.184	0.027	43.058	22.327	26.130	8.126
Corynebacterium diphtheriae	0.529	0.839	0.587	0.548	0.491	0.618	0.261	0.441	27.341	20.662	17.023	18.809
Deinococcus radiodurans	0.148	0.851	0.485	0.270	0.059	0.109	0.127	0.075	9.742	12.077	15.179	9.931
Erwinia carotovora	0.202	0.561	0.733	0.489	0.184	0.341	0.318	0.423	6.260	10.160	12.678	7.001
Escherichia coli	0.049	0.523	0.538	0.426	0.035	0.364	0.233	0.388	1.537	5.956	6.454	3.813
Haemophilus influenzae	0.312	0.549	0.483	0.836	0.281	0.623	0.282	1.120	4.014	5.502	10.792	9.091
Klebsiella pneumoniae	0.059	0.491	0.760	0.404	0.042	0.246	0.257	0.311	5.453	8.858	10.068	7.796
Lactobacillus acidophilus	0.631	0.917	0.535	0.993	0.620	0.939	0.381	0.997	7.610	11.611	13.071	10.790
Leptospira interrogans	0.414	0.839	0.590	0.819	0.444	0.857	0.488	0.915	2.075	6.996	7.621	11.809
Mycoplasmoides genitalium	0.927	0.745	0.615	1.180	0.961	0.785	0.569	1.239	14.798	5.643	13.506	21.824
Neisseria gonorrhoeae	0.067	0.394	0.451	0.263	0.044	0.183	0.225	0.212	3.521	6.011	12.363	10.018
Pseudomonas aeruginosa	0.099	0.926	0.653	0.124	0.055	0.167	0.169	0.037	16.778	13.224	11.056	4.902
Salmonella enterica	0.055	0.517	0.694	0.467	0.043	0.354	0.326	0.406	3.703	8.378	8.557	7.079
Serratia marcescens	0.061	0.480	0.694	0.305	0.034	0.162	0.173	0.169	5.480	7.722	10.507	10.015
Shigella flexneri	0.070	0.509	0.523	0.420	0.055	0.357	0.223	0.357	2.584	6.607	5.970	4.403
Staphylococcus aureus	0.506	0.746	0.635	0.884	0.503	0.944	0.559	1.126	5.145	7.152	12.245	5.975
Streptococcus agalactiae	0.795	0.771	0.786	0.946	0.773	0.786	0.499	1.069	11.337	8.294	10.980	9.353
Thermus aquaticus	0.350	1.205	0.589	0.635	0.328	0.224	0.490	0.482	55.578	37.095	32.549	38.380
Xanthomonas oryzae	0.119	0.526	0.538	0.199	0.073	0.162	0.184	0.103	9.233	13.770	10.695	10.529
Yersinia pestis	0.188	0.514	0.725	0.625	0.158	0.380	0.357	0.610	6.219	8.552	10.951	9.132

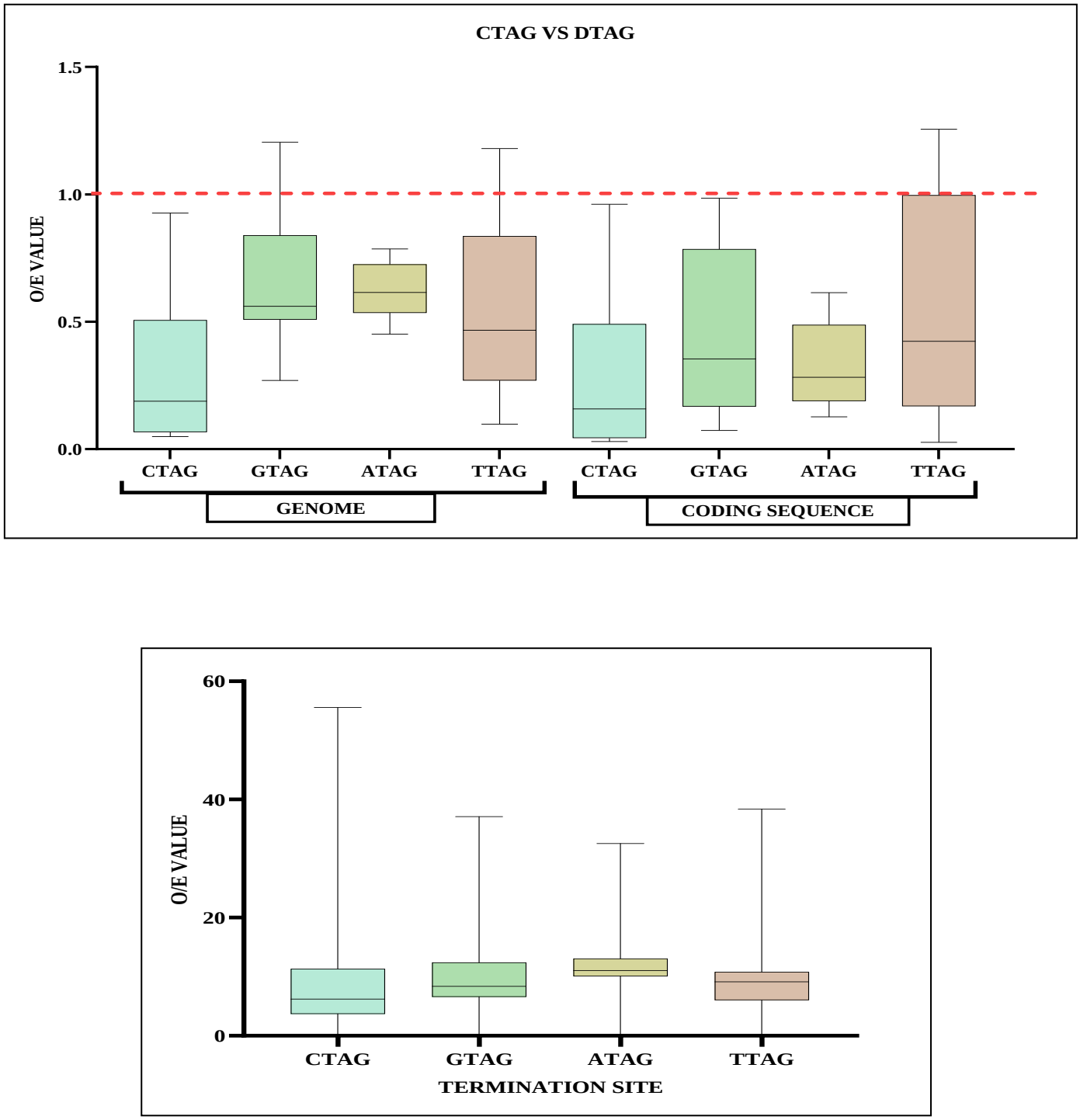


Fig. 5: Comparison of DTAG (A) in the entire genome, coding sequence and (B) at the termination site.

Table 8: Ratio of the $\frac{O/E\ (CTAA+CTGA)}{O/E\ (DTAA+DTGA)}$ to the $\frac{O/E\ (CTAG)}{O/E\ (DTAG)}$

BACTERIAL SPECIES	THE RATIO								
	$\frac{O/E(CTAA+CTGA)}{O/E(DTAA+DTGA)}$			$\frac{O/E(CTAG)}{O/E(DTAG)}$			$\frac{\frac{O/E(CTAA+CTGA)}{O/E(DTAA+DTGA)}}{\frac{O/E(CTAG)}{O/E(DTAG)}}$		
	GENOME	CONDING REGION	TERMINATION SITE	GENOME	CONDING REGION	TERMINATION SITE	GENOME	CONDING REGION	TERMINATION SITE
Agrobacterium tumefaciens	0.731	1.142	1.316	0.244	0.388	#DIV/0!	**2.997**	**2.940**	**#DIV/0!**
Anabaena cylindrica	1.103	1.202	0.764	0.925	0.885	0.546	**1.192**	**1.358**	**1.398**
Bacillus subtilis	0.917	1.065	0.672	0.508	0.540	0.576	**1.806**	**1.972**	**1.166**
Bordetella pertussis	0.758	1.312	1.324	0.226	0.753	1.558	**3.357**	**1.741**	**0.850**
Corynebacterium diphtheriae	0.805	0.996	1.217	0.840	1.068	1.291	**0.958**	**0.933**	**0.943**
Deinococcus radiodurans	0.935	1.323	1.223	0.301	0.656	0.851	**3.106**	**2.015**	**1.437**
Erwinia carotovora	0.909	1.057	1.024	0.408	0.584	0.695	**2.229**	**1.810**	**1.475**
Escherichia coli	0.819	0.948	0.810	0.128	0.138	0.347	**6.414**	**6.889**	**2.334**
Haemophilus influenzae	0.886	0.932	0.642	0.545	0.464	0.504	**1.626**	**2.009**	**1.274**
Klebsiella pneumoniae	0.815	0.956	1.046	0.145	0.207	0.689	**5.600**	**4.607**	**1.518**
Lactobacillus acidophilus	0.982	1.169	0.734	0.822	0.860	0.683	**1.195**	**1.359**	**1.074**
Leptospira interrogans	1.055	1.212	0.846	0.611	0.650	0.263	**1.726**	**1.864**	**3.221**
Mycoplasmoides genitalium	1.098	1.306	1.349	1.059	1.085	0.969	**1.036**	**1.204**	**1.391**
Neisseria gonorrhoeae	0.773	0.988	0.930	0.232	0.272	0.450	**3.335**	**3.637**	**2.067**
Pseudomonas aeruginosa	0.916	1.442	1.672	0.211	0.512	1.339	**4.351**	**2.816**	**1.248**
Salmonella enterica	0.793	0.903	0.917	0.129	0.155	0.537	**6.148**	**5.814**	**1.707**
Serratia marcescens	0.805	1.094	0.941	0.169	0.266	0.683	**4.775**	**4.107**	**1.378**
Shigella flexneri	0.843	0.982	0.819	0.185	0.223	0.529	**4.557**	**4.397**	**1.547**
Staphylococcus aureus	0.881	0.989	0.563	0.706	0.626	0.633	**1.247**	**1.581**	**0.890**
Streptococcus agalactiae	1.030	1.122	0.744	0.952	0.990	1.129	**1.082**	**1.133**	**0.659**
Thermus aquaticus	0.853	1.168	1.169	0.484	0.957	1.296	**1.762**	**1.221**	**0.902**
Xanthomonas oryzae	0.810	1.040	1.174	0.351	0.577	0.826	**2.307**	**1.803**	**1.422**
Yersinia pestis	0.908	1.033	0.978	0.360	0.414	0.708	**2.519**	**2.496**	**1.380**

Further insight into the role of C in under-representation of CTAG could be gained by studying effect of other three bases preceding the TAG. The comparison clearly showed that O/E value of CTAG was lowest when compared to any one of DTAGs (GTAG, ATAG and TTAG) in most of the bacterial genomes (Table 7). Similar were the results in coding regions and the termination sites (Fig. 5). Based on the counts of the three termination codon trinucleotides with or without C

preceding them, another index was used to study the effect of C at 5'end. In this index the CTAG abundance was evaluated against the DTAGs and compared with that of the other two termination codons, viz. CTAA & CTGA against DTAA & DTGA. The index value (Table8) indicated that CTAG shows stronger under-representation than the DTAGs in most of the bacterial genomes as well as their coding sequences. In nearly half of the bacteria, it was more than two-fold stronger effect. At termination site also most of the bacteria exhibited stronger under-representation than that of DTAGs. This result clearly showed that it is the combination of C and the following TAG that is causing CTAG under-representation.

5.5 Comparison of 64 trinucleotides among the selected bacteria genomes

The low abundance of TAG in comparison with the other two termination codons and this effect of under-representation is strengthened by occurrence of C at its 5' end. This observation prompted to study abundance of rest of all the possible trinucleotides, which are act as codons in open reading frame encoding amino acids and are 61 in number. When O/E value of all the possible 64 trinucleotides was compared among the studied bacterial genomes, it was found that in most of the bacterial genomes, TAG was one of the least abundant trinucleotide. However, the other two termination codon (TAA & TGA) were not strongly under-represented in the most of the bacteria. Their O/E value was higher than several other trinucleotides in most of the bacteria (Table 9). These result (Fig. 6) indicate that the termination codon TAG is not only less abundant than TAA and TGA, the other two termination codons but also less abundant in comparison with most of the other 61 trinucleotides. TAG role in under-representation of CTAG is further strengthened by a C preceding it in the sequence.

Fig. 6: Heat map representation of 64 trinucleotide present in the bacterial genomes.

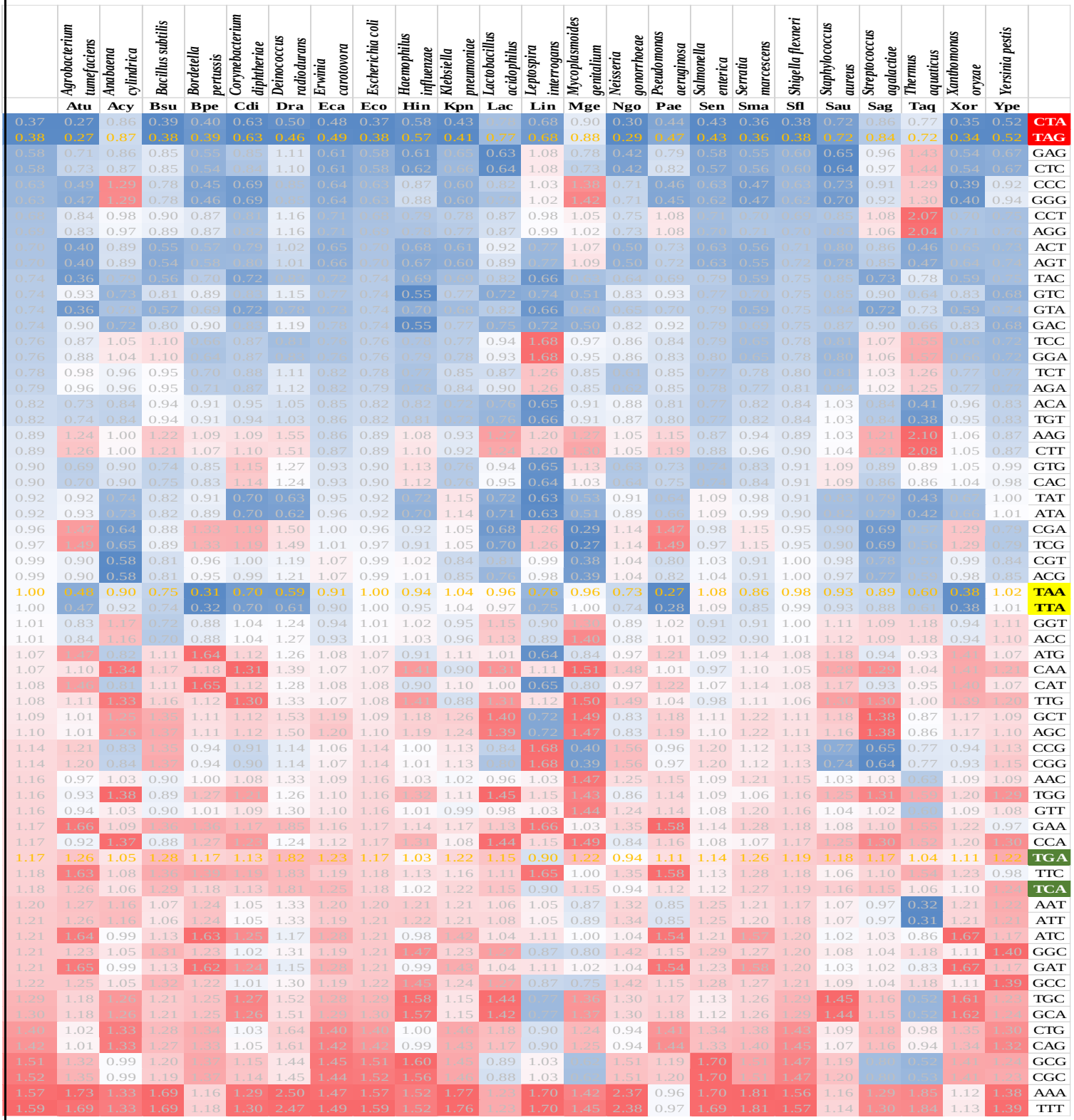
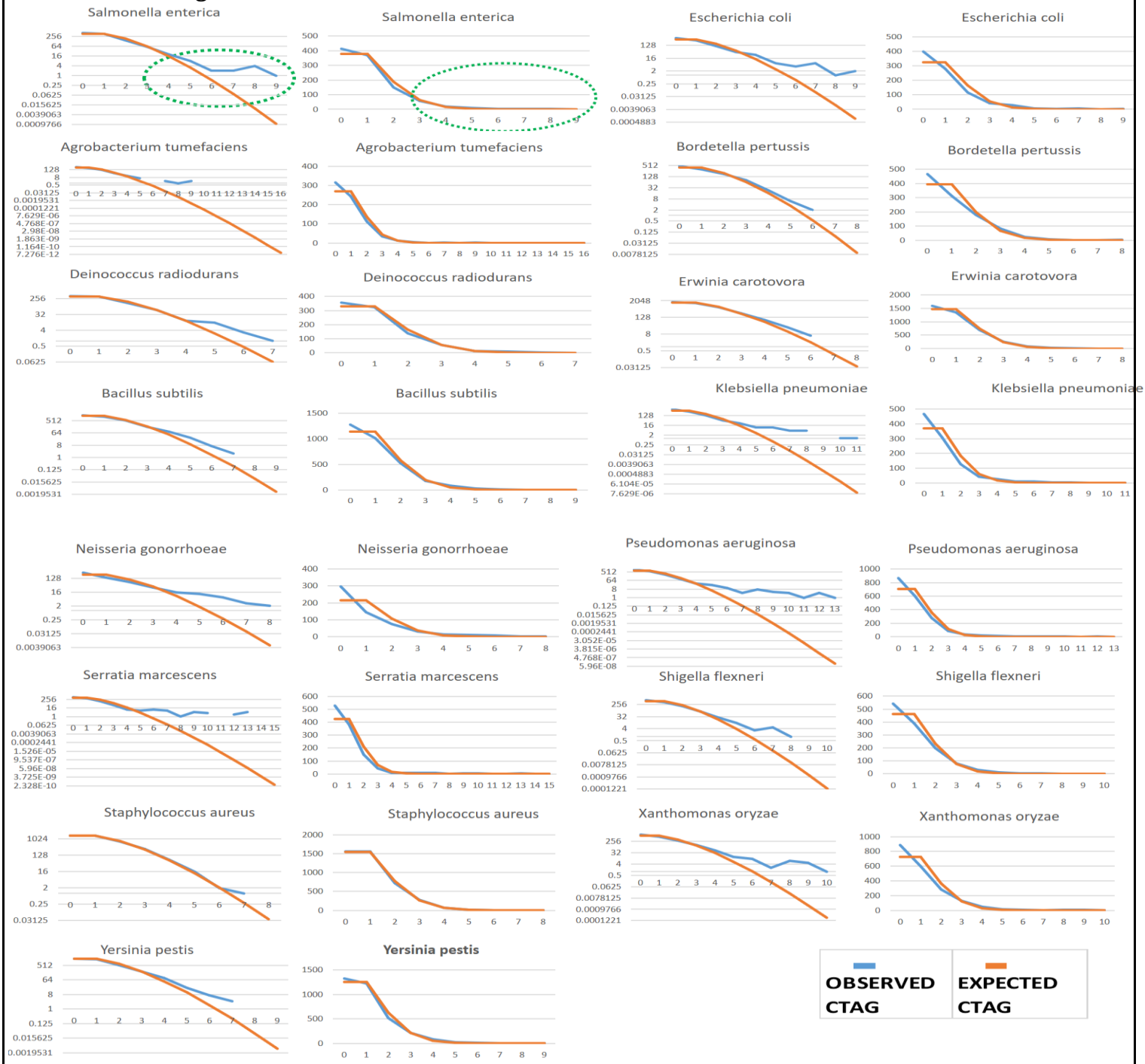


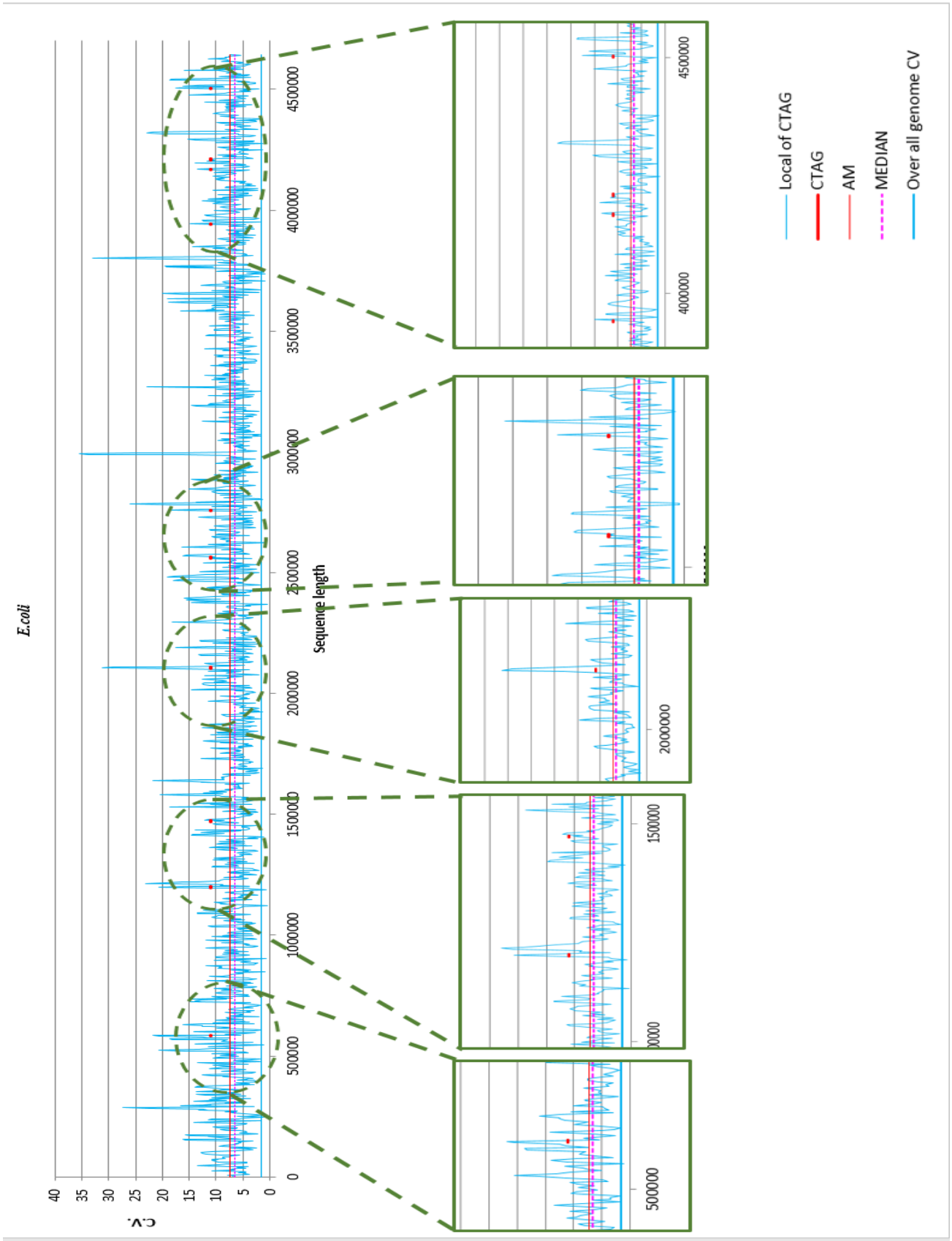
Table 9: Rank of termination codon in the comparison of other 61 trinucleotide

	RANK					
Bacterial species	**CTA**	**TAG**	**TAA**	**TTA**	**TGA**	**TCA**
Agrobacterium tumefaciens	1	2	9	8	48	50
Anabaena cylindrica	17	19	24	26	42	44
Bacillus subtilis	2	1	12	9	53	54
Bordetella pertussis	4	3	1	2	41	42
Corynebacterium diphtheriae	1	2	6	8	45	46
Deinococcus radiodurans	2	1	3	4	60	59
Erwinia carotovora	1	2	24	23	53	54
Escherichia coli	1	2	31	32	48	50
Haemophilus influenzae	4	3	27	28	40	37
Klebsiella pneumoniae	2	1	34	33	52	51
Lactobacillus acidophilus	13	12	33	35	48	47
Leptospira interrogans	12	11	18	17	30	27
Mycoplasma capricolum	26	24	30	34	44	43
Neisseria gonorrhoeae	2	1	15	17	34	36
Pseudomonas aeruginosa	3	6	1	2	39	40
Salmonella enterica	2	1	36	37	50	45
Serratia marcescens	1	2	24	23	51	54
Shigella flexneri	1	2	29	30	52	51
Staphylococcus aureus	5	4	25	26	56	51
Streptococcus agalactiae	18	13	21	20	53	49
Thermus aquaticus	27	23	17	19	40	42
Xanthomonas oryzae	2	1	3	4	39	37
Yersinia pestis	1	2	34	32	51	55

5.6 CTAG local over-representation in the 15 bacterial genomes with lowest O/E value of CTAG

Fig. 7: logarithmic and line graph representation of CTAG cluster in the least abundant bacterial genome




E.coli
C.V.
Sequence length
Local of CTAG
CTAG
AM
MEDIAN
Over all genome CV

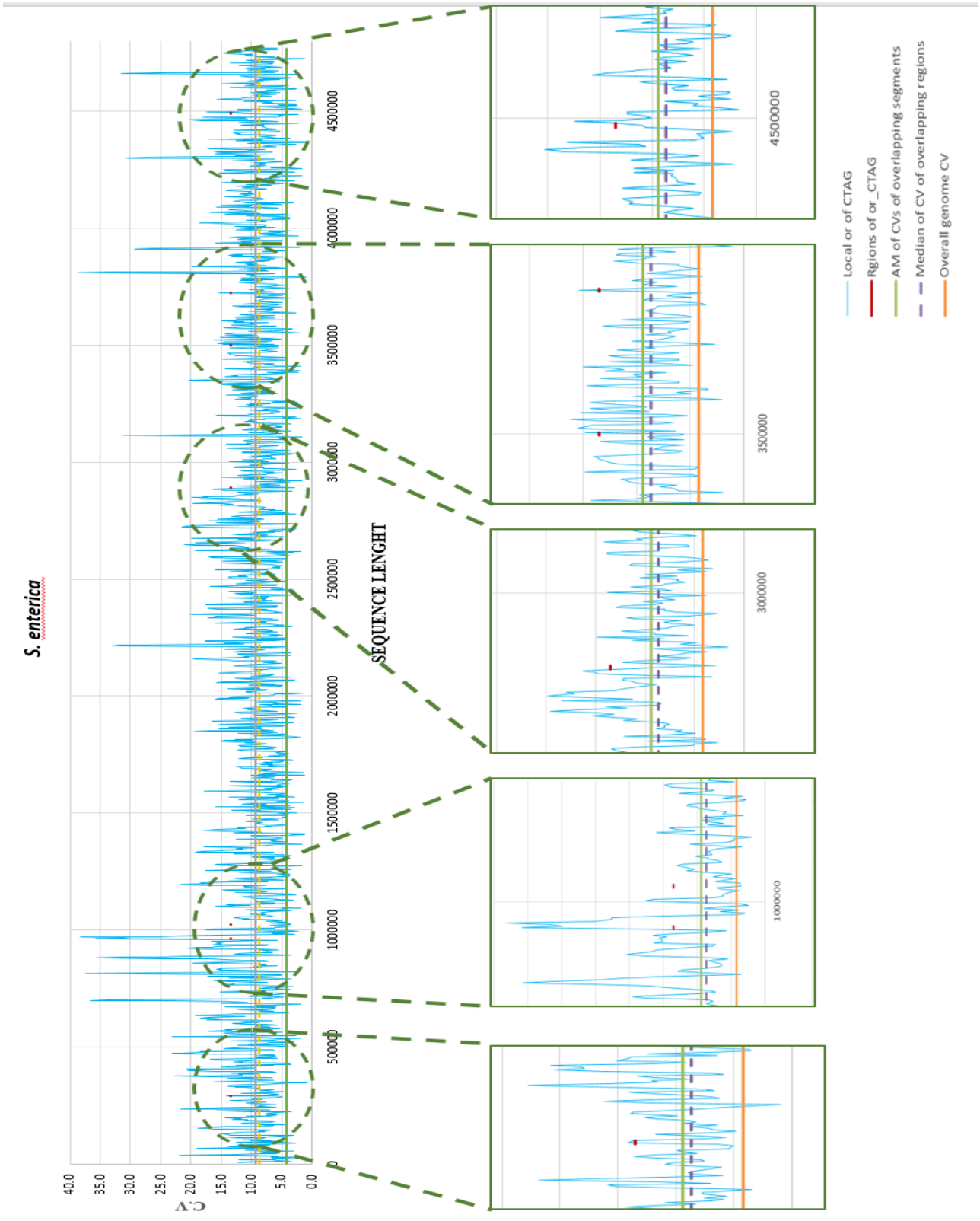


Fig. 8: Local over-representation of CTAG in the (A) *E. coli* and (B) *S. enterica*

The strong under-representation in the bacterial genomes prompted to study its distribution in the genomes. The positional analysis of CTAG motifs showed that in several regions of genome they are found in high local abundance. Analysis of genomes for distribution of CTAGs was based on determining their position in genomes, which were used to determine the spaces between successive CTAGs. Since CTAG occurrence in the genome is a rare event, it was expected to follow a Poisson distribution. As a result, the CTAG count in a unit length of genome was compared with the expected count based on its Poisson distribution. In most of the studied bacteria it was found that the CTAGs have local high occurrence in the genome at a number of regions which is distinctly higher than expected on the basis of Poisson distribution statistics (Fig7). The unexpected occurrence of CTAG clustering in small regions of genome points towards their biological role in bacteria which requires to be studied further. This requires to learn about the factors associated with local abundance of CTAG. In preliminary studies based on two bacterial genomes, viz. *E. coli* and *S. enterica*, it has been observed that most of the high abundance regions of CTAG overlap with sequence characteristics of high base heterogeneity (Fig8). The coefficient of variation of all the four base counts was used as a measure of base heterogeneity. The CV of a few thousand bases long regions was determined in the sliding window fashion. Further work is required to understand this association of base composition heterogeneity and local abundance of CATGs in bacterial genomes which may also help in understanding their biological role in the bacterial genomes.

CHAPTER 6

DISCUSSION

Genomic sequences exhibit certain characteristics that show their non-randomness. These patterns can be detected by analyzing the compositional characteristics of the sequences. Typical manifestation of non-random nature of genome sequences lies in its heterogeneity in terms of over-represented or under-represented motifs (Burge et al., 1992; Karlin et al., 1994). One of such example is a palindromic tetranucleotide motif, CTAG, which is under-represented in bacterial genomes. There are several reasons for the under-representation of the CTAG motif. The presence of CTAG motifs has been associated with structural defects or kinks in the DNA. These kinks may have detrimental effects on DNA structure elsewhere in the genome. Additionally, the vsr gene product, which is part of the very short patch repair system, is known to play a role in reducing the frequency of CTAG in certain bacterial genomes (Burge et al., 1992). One of the possible reasons for the under-representation of CTAG could be its TAG submotif, which is a termination codon. In order to evaluate this hypothesis experiments were designed to study abundance of CTAG and its submotifs and their comparison with other similar tetranucleotide.

It was observed that the O/E value of CTAG was < 1 for all the studied bacteria, indicating an under-representation. Out of the 23 bacterial genomes, 15 showed significant under-representation, with the CTAG sequence ranking last among the 256 possible tetranucleotides in relation to O/E values. The rest of the 8 bacteria also showed under-representation of CTAG but with higher O/E values compared to few to several other tetranucleotides. A similar under-

representation of CTAG was observed in eukaryotic genomes, which is in agreement with the earlier reports (Karlin et al., 1998)(Fig. 1).

Following this observation, CTAG abundance among the selected bacterial genomes was compared with other permutations of the same base composition. Distinctly, CTAG was the least abundant when compared to the rest of the 23 other tetranucleotides (Fig. 2). The CTAG sequence motif has trinucleotide TAG, a termination codon sequence, in it as a submotif. To investigate the role of TAG submotif in under-representation of CTAG, its abundance in the bacterial genomes was determined and compared with the other two termination codons, viz. TAA and TGA in three different regions, i.e., the entire genome, the coding sequences and at the termination sites of protein coding genes. It was observed that TAG had the lowest O/E value among the three termination codons, indicating its lower frequency throughout the genomes, coding sequences, and at the termination sites of protein-coding genes (Fig. 3). In earlier work also it has been reported that TAG is the least used termination codon (Ho & Hurst, 2022).Another interesting observation was that O/E value of TAG and TAA in coding region was lower than that in the whole genome, whereas converse is true for TGA.

Comparing the abundance of CTAG to that of similar tetranucleotides but with TAA and TGA instead of TAG (CTAA and CTGA) it was found that CTAG had lower O/E values than CTGA and CTAA. This result was in line with the previous result reiterating the importance of TAG in CTAG suppression in the genomes. Apparently it seemed that the abundance of the termination codons was unaffected by the presence of a cytosine at their 5' end. An index based on the ratio of the O/E values of (CTGA+CTAA)/(CTAG) and (TGA+TAA)/(CTAG) was used to further assess

the effect of the 5' cytosine on CTAG under-representation. The value of this index were > 1, which indicates that CTAG is underrepresented more significantly than TAG alone when there is a cytosine at the 5' end of the termination codon (Table 6). Further, O/E value of CTAG was compared to those of the other DTAGs (GTAG, ATAG, and TTAG), and it was found that CTAG had the lowest O/E value of any DTAG in the majority of bacterial genomes, coding regions, and termination sites. This observation suggested that the combination of C in addition to TAG led to a stronger under-representation (Fig. 5)

So fart the results showed that C and TAG, both contribute to the under-representation of CTAG in bacterial genomes. The result was based only on the basis of comparison of TAG with the other two termination codon trinucleotides. The O/E values of rest of the possible 61 trinucleotides were also found to be higher than TAG in most of the bacteria. However, there were few to several trinucleotides that had O/E value less than TAA and TGA (Fig. 6). Thought TAG, TAA and TGA, all three can function as termination codons, it is only TAG that has strong under-representation in bacterial genomes.

It is expected that trinucleotides functioning as termination codons may have lower abundance than rest of the 61 trinucleotides. But in bacteria TAG is the least preferred termination codon. In case it is preceded by a C, TAG results is CTAG tetranucleotide which is a palindrome. Thus CTAG presents a TAG on both the strands of DNA when part of CTAG and on both the strands the TAG is preceded by a C. It appears to have strong contribution to the under-representation of CTAG in bacterial genomes.

The study also explored the distribution of CTAG motifs in bacterial genomes and found that they are highly sparsely distributed in most of the genome while occur in high abundance in certain small regions of the genome. Based on Poisson distribution statistics, it was found that there was a higher than expected concentration of CTAGs in specific small regions of the genome, indicating that this clustering effect may have some biological significance. Further investigation revealed a connection between the high base heterogeneity characteristics of the sequence and the CTAG high abundance regions. (Fig. 7,8). More work on these two aspects may further reveal the significance of GTAC suppression in general and high local occurrence in the regions with highly uneven base composition.

CHAPTEER -7

CONCLUSION

In conclusion, the under-representation of the CTAG tetranucleotide motif in bacterial genomes is a significant genomic feature that has been studied in this work. The results indicate that TAG trinucleotide motif which is also a termination codon sequence, particularly when preceded by a cytosine (C) at its 5' end, is associated with its strong under-representation in bacterial genomes. This under-representation is not observed for the other two termination codons trinucleotides, TAA and TGA. The combination of the preceding C and TAG appears to be responsible for the CTAG under-representation, which is not simply due to its function as a termination codon.

The positional analysis of CTAG motifs in bacterial genomes also demonstrated that they frequently cluster in locations where their abundance exceeds the value expected from the Poisson distribution. These high-CTAG patches indicate overlap with high-base heterogeneity sequencing characteristics. The coefficient of variance of all four base counts was used to assess base heterogeneity. This shows that the number of CTAG motifs in particular regions of bacterial genomes and base composition heterogeneity may be connected.

These results raise some interesting questions about the role of CTAG under-representation in bacterial organisms. However, the exact biological function of this under-representation and its clustering in certain genomic regions is still unknown. There is a need for additional research into the factors associated with the local abundance of CTAG motifs and their potential implications for the organization and function of the bacterial genome.

Overall, this study contributes to our understanding of the diverse features and complexities of genomic DNA sequences in living organisms and emphasizes the importance of investigating sequence heterogeneities to gain insights into the genetic information and the related biological processes of both prokaryotic and eukaryotic organisms.

Document Information

Analyzed document	Sonia _thesis.pdf (D172382472)
Submitted	2023-07-26 08:40:00
Submitted by	Atul Kumar Upadhyay
Submitter email	atul.upadhyay@thapar.edu
Similarity	0%
Analysis address	atul.upadhyay.thapar@analysis.urkund.com

Sources included in the report

W URL: https://www.bioinformatics.org/sms2/dna_stats.html
Fetched: 2023-07-26 08:41:00 1

Entire Document

ABSTRACT DNA sequences in the genomes have several examples of inhomogeneities which point towards their non-random nature. One such case is under-representation of CTAG tetranucleotide in bacterial genomes. This work aims to study the reasons behind the under-representation of CTAG in bacterial genomes. A randomly selected set of 23 bacterial genomes, largely from diverse prokaryotes, showed that CTAG is under-represented (Observed / Expected value >1.0). In the majority of the selected genomes, CTAG had the lowest O/E value. It was found that TAG, a submotif of CTAG has one of the lowest O/E values in most of the bacteria. The under- representation effect is further enhanced by the occurrence of a cytosine at the 5' end of TAG which makes it CTAG and explains its under-representation. CTAG distribution in the genome is also very uneven. While most of the genome has very low occurrence, there are certain short regions where CTAG is found in relatively much higher abundance. Such regions coincide with the sequence with highly uneven base composition. This study demonstrates base composition as one of the factors causing highly heterogeneous distribution of CTAGs in the bacterial genomes and indicates that CTAGs play some important biologic role in bacteria. Key words: Under-representation, Inhomogeneity, CTAG, bacterial genome, termination codon. CHAPTER 1 INTRODUCTION The diversity of life, especially at cellular and molecular levels, is a remarkable feature among living organisms. The diversity in the living world is based on the genetic information of the living organisms. All living organisms harbor this information encoded in the sequence of DNA (RNA in the case of RNA viruses). The unit of genetic information is gene, a defined segment of DNA that encodes specific proteins. The genome of a cellular organism consists of hundreds to tens of thousands of genes, each encoding a unique protein. These proteins are responsible for most of the cellular functions. The structure of DNA can be described as two long polynucleotide chains bound by hydrogen bonds between their nucleic acid moiety and wrapped around each other to form a double helix. The backbone of DNA is made up of repeating deoxyribose sugar-phosphate units. Attached to each deoxyribose molecule is one of four possible bases, adenine (A), cytosine (C), guanine (G), and thymine (T). In DNA each nucleotide carries any one of the four bases. Thus, in turn, the polynucleotide chain consists of a sequence of bases in DNA, which is capable of storing genetic information. The structure of DNA, with its sugar-phosphate backbone and variable bases, plays a crucial role in storing, replicating and encoding genetic information in living organisms. This is how a chain of DNA forming chromosomes consists of a long sequence of bases that stores an enormous amount of genetic information in cells. Cellular organisms are classified into two different types, prokaryotes and eukaryotes. Prokaryotes, lacking a nucleus, possessed typically a single linear or circular chromosome. In case of eukaryotes, their genetic information is stored in multiple linear chromosomes which are present in membrane bound organelle called nucleus. Eukaryotic genomes are much larger in size when compared to prokaryotes. As a result, eukaryotic genomes have much higher number of gene and exhibit more complex organization.

REFERENCES

Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M., & Ikemura, T. (2006). Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene*, *365*, 27–34. https://doi.org/10.1016/j.gene.2005.09.040

Balaceanu, A., Buitrago, D., Walther, J., Hospital, A., Dans, P. D., & Orozco, M. (2019). Modulation of the helical properties of DNA: Next-to-nearest neighbour effects and beyond. *Nucleic Acids Research*, *47*(9), 4418–4430. https://doi.org/10.1093/nar/gkz255

Bhagwat, A. S., & McClelland, M. (1992). DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. *Nucleic Acids Research*, *20*(7), 1663–1668. https://doi.org/10.1093/nar/20.7.1663

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, *277*(5331), 1453–1462. https://doi.org/10.1126/science.277.5331.1453

Bohlin, J., & Skjerve, E. (2009). Examination of Genome Homogeneity in Prokaryotes Using Genomic Signatures. *PLoS ONE*, *4*(12), e8113. https://doi.org/10.1371/journal.pone.0008113

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten

things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Burge, C., Campbell, A. M., & Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, *89*(4), 1358–1362. https://doi.org/10.1073/pnas.89.4.1358

Cohen, N. R., Ross, C. A., Jain, S., Shapiro, R. S., Gutierrez, A., Belenky, P., Li, H., & Collins, J. J. (2016). A role for the bacterial GATC methylome in antibiotic stress survival. *Nature Genetics*, *48*(5), 581–586. https://doi.org/10.1038/ng.3530

Eyre-Walker, A. (1992). Evidence that both G + C rich and G + C poor isochores are replicated early and late in the cell cycle. *Nucleic Acids Research*, *20*(7), 1497–1501. https://doi.org/10.1093/nar/20.7.1497

Gaultney, R. A., Vincent, A. T., Lorioux, C., Coppée, J.-Y., Sismeiro, O., Varet, H., Legendre, R., Cockram, C. A., Veyrier, F. J., & Picardeau, M. (2020). 4-Methylcytosine DNA modification is critical for global epigenetic regulation and virulence in the human pathogen *Leptospira interrogans*. *Nucleic Acids Research*, *48*(21), 12102–12115. https://doi.org/10.1093/nar/gkaa966

Gelfand, M. S., & Koonin, E. V. (1997). Avoidance of palindromic words in bacterial and archaeal genomes: A close connection with restriction enzymes. *Nucleic Acids Research*, *25*(12), 2430–2439. https://doi.org/10.1093/nar/25.12.2430

Hämälä, T., Ning, W., Kuittinen, H., Aryamanesh, N., & Savolainen, O. (2022). Environmental response in gene expression and DNA methylation reveals factors influencing the adaptive potential of Arabidopsis lyrata. *ELife*, *11*, e83115. https://doi.org/10.7554/eLife.83115

Häring, D., & Kypr, J. (1999). Variations of the Mononucleotide and Short Oligonucleotide Distributions in the Genomes of Various Organisms. *Journal of Theoretical Biology*, *201*(2), 141–156. https://doi.org/10.1006/jtbi.1999.1019

Ho, A. T., & Hurst, L. D. (2021). Effective Population Size Predicts Local Rates but Not Local Mitigation of Read-through Errors. *Molecular Biology and Evolution*, *38*(1), 244–262. https://doi.org/10.1093/molbev/msaa210

Ho, A. T., & Hurst, L. D. (2022). Stop Codon Usage as a Window into Genome Evolution: Mutation, Selection, Biased Gene Conversion and the TAG Paradox. *Genome Biology and Evolution*, *14*(8), evac115. https://doi.org/10.1093/gbe/evac115

Humayun, M. Z., Zhang, Z., Butcher, A. M., Moshayedi, A., & Saier, M. H. (2017). Hopping into a hot seat: Role of DNA structural features on IS5-mediated gene activation and inactivation under stress. *PLOS ONE*, *12*(6), e0180156. https://doi.org/10.1371/journal.pone.0180156

Kandel, D., Matias, Y., Unger, R., & Winkler, P. (1996). Shuffling biological sequences. *Discrete Applied Mathematics*, *71*(1–3), 171–185. https://doi.org/10.1016/S0166-218X(97)81456-4

Karlin, S. (2005). Statistical signals in bioinformatics. *Proceedings of the National Academy of Sciences*, *102*(38), 13355–13362. https://doi.org/10.1073/pnas.0501804102

Karlin, S., Campbell, A. M., & Mrázek, J. (1998). COMPARATIVE DNA ANALYSIS ACROSS DIVERSE GENOMES. *Annual Review of Genetics*, *32*(1), 185–225. https://doi.org/10.1146/annurev.genet.32.1.185

Karlin, S., Ladunga, I., & Blaisdell, B. E. (1994). Heterogeneity of genomes: Measures and values. *Proceedings of the National Academy of Sciences*, *91*(26), 12837–12841. https://doi.org/10.1073/pnas.91.26.12837

Karlin, S., Mrázek, J., & Campbell, A. M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*, *179*(12), 3899–3913. https://doi.org/10.1128/jb.179.12.3899-3913.1997

Keil, K. P., & Lein, P. J. (2016). DNA methylation: A mechanism linking environmental chemical exposures to risk of autism spectrum disorders? *Environmental Epigenetics*, *2*(1), dvv012. https://doi.org/10.1093/eep/dvv012

Kirilov, K. T., Golshani, A., & Ivanov, I. G. (2013). Termination Codons and Stop Codon Context in Bacteria and Mammalian Mitochondria. *Biotechnology & Biotechnological Equipment*, *27*(4), 4018–4025. https://doi.org/10.5504/BBEQ.2013.0052

Korkmaz, G., Holm, M., Wiens, T., & Sanyal, S. (2014). Comprehensive Analysis of Stop Codon Usage in Bacteria and Its Correlation with Release Factor Abundance. *Journal of Biological Chemistry*, *289*(44), 30334–30342. https://doi.org/10.1074/jbc.M114.606632

Lee, J. Y., & Lee, T.-H. (2012). Effects of DNA Methylation on the Structure of Nucleosomes. *Journal of the American Chemical Society*, *134*(1), 173–175. https://doi.org/10.1021/ja210273w

Liu, S. L., Hessel, A., & Sanderson, K. E. (1993). The XbaI-BlnI-CeuI genomic cleavage map of *Salmonella typhimurium* LT2 determined by double digestion, end labelling, and pulsed-field gel electrophoresis. *Journal of Bacteriology*, *175*(13), 4104–4120. https://doi.org/10.1128/jb.175.13.4104-4120.1993

McVean, G. A. T., & Hurst, G. D. D. (2000). Evolutionary Lability of Context-Dependent Codon Bias in Bacteria. *Journal of Molecular Evolution*, *50*(3), 264–275. https://doi.org/10.1007/s002399910031

Nan, X., Ng, H.-H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., & Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, *393*(6683), 386–389. https://doi.org/10.1038/30764

Nekrutenko, A., & Li, W. H. (2000). Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome research, 10(12), 1986–1995. https://doi.org/10.1101/gr.10.12.1986

Oggenfuss, U., Badet, T., Wicker, T., Hartmann, F. E., Singh, N. K., Abraham, L., Karisto, P., Vonlanthen, T., Mundt, C., McDonald, B. A., & Croll, D. (2021). A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. *ELife*, *10*, e69249. https://doi.org/10.7554/eLife.69249

Povolotskaya, I. S., Kondrashov, F. A., Ledda, A., & Vlasov, P. K. (2012). Stop codons in bacteria are not selectively equivalent. *Biology Direct*, *7*(1), 30. https://doi.org/10.1186/1745-6150-7-30

Reisenauer, A., Kahng, L. S., McCollum, S., & Shapiro, L. (1999). Bacterial DNA Methylation: A Cell Cycle Regulator? *Journal of Bacteriology*, *181*(17), 5135–5139. https://doi.org/10.1128/JB.181.17.5135-5139.1999

Saitou, N. (2013). Eukaryote Genomes. In N. Saitou, *Introduction to Evolutionary Genomics* (Vol. 17, pp. 193–222). Springer London. https://doi.org/10.1007/978-1-4471-5304-7_8

Sivaraman, K., Seshasayee, A. S. N., Swaminathan, K., Muthukumaran, G., & Pennathur, G. (2005). Promoter addresses: Revelations from oligonucleotide profiling applied to the

Escherichia coli genome. *Theoretical Biology and Medical Modelling*, *2*(1), 20. https://doi.org/10.1186/1742-4682-2-20

Sobetzko, P., Jelonek, L., Strickert, M., Han, W., Goesmann, A., & Waldminghaus, T. (2016). DistAMo: A Web-Based Tool to Characterize DNA-Motif Distribution on Bacterial Chromosomes. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.00283

Stibitz, S. (1998). IS *481* and IS *1002* of *Bordetella pertussis* Create a 6-Base-Pair Duplication upon Insertion at a Consensus Target Site. *Journal of Bacteriology*, *180*(18), 4963–4966. https://doi.org/10.1128/JB.180.18.4963-4966.1998

Stibitz, S., & Garletts, T. L. (1992). Derivation of a physical map of the chromosome of Bordetella pertussis Tohama I. *Journal of Bacteriology*, *174*(23), 7770–7777. https://doi.org/10.1128/jb.174.23.7770-7777.1992

Sung, W., Ackerman, M. S., Gout, J.-F., Miller, S. F., Williams, E., Foster, P. L., & Lynch, M. (2015). Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments. *Molecular Biology and Evolution*, *32*(7), 1672–1683. https://doi.org/10.1093/molbev/msv055

Tang, L., Liu, W.-Q., Fang, X., Sun, Q., Zhu, S.-L., Wang, C.-X., Wang, X.-Y., Li, Y.-G., Zhu, D.-L., Sanderson, K. E., Johnston, R. N., Liu, G.-R., & Liu, S.-L. (2014). CTAG-Containing Cleavage Site Profiling to Delineate *Salmonella* into Natural Clusters. *PLoS ONE*, *9*(8), e103388. https://doi.org/10.1371/journal.pone.0103388

Tang, L., Mastriani, E., Zhou, Y.-J., Zhu, S., Fang, X., Liu, Y.-P., Liu, W.-Q., Li, Y.-G., Johnston, R. N., Guo, Z., Liu, G.-R., & Liu, S.-L. (2017). Differential degeneration of the ACTAGT sequence among *Salmonella*: A reflection of distinct nucleotide amelioration

patterns during bacterial divergence. *Scientific Reports*, *7*(1), 10985. https://doi.org/10.1038/s41598-017-11226-9

Tang, L., Zhu, S., Mastriani, E., Fang, X., Zhou, Y.-J., Li, Y.-G., Johnston, R. N., Guo, Z., Liu, G.-R., & Liu, S.-L. (2017). Conserved intergenic sequences revealed by CTAG-profiling in *Salmonella*: Thermodynamic modeling for function prediction. *Scientific Reports*, *7*(1), 43565. https://doi.org/10.1038/srep43565

Wei, Y., Wang, J., & Xia, X. (2016). Coevolution between Stop Codon Usage and Release Factors in Bacterial Species. *Molecular Biology and Evolution*, *33*(9), 2357–2367. https://doi.org/10.1093/molbev/msw107

Wong, T.-Y., Fernandes, S., Sankhon, N., Leong, P. P., Kuo, J., & Liu, J.-K. (2008). Role of Premature Stop Codons in Bacterial Evolution. *Journal of Bacteriology*, *190*(20), 6718–6725. https://doi.org/10.1128/JB.00682-08