# INVESTIGATION AND ANALYSIS OF CONTEXT DEPENDENT MUTATION RATES CONTRIBUTION TO SYNONYMOUS CODON USAGE VIA DEVELOPMENT OF CDMAP

by

David Logan Patton

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics & Computational Biology

Charlotte

2022

Approved By:

Dr. Way Sung

Dr. Cynthia Gibas

Dr. Anthony Fodor

Dr. Rebekah Rogers

Dr. Mariya Munir

**Abstract:**

David Logan Patton

Investigation and analysis of context dependent mutation rates contribution to synonymous codon usage via development of CDMAP

(Under the direction of Dr. Way Sung)

Mutations play a pivotal role as a driver of genome evolution in organisms across the tree of life. However, site-specific mutation rates appear to have a non-uniform mutation rate and vary across the genome. Currently, site-specific variation has not been well characterized or analyzed using a uniform treatment. To address the lack of uniform analysis and provide an open-source, non-proprietary means of in-depth analysis, we developed the Context Dependent Mutation Analysis Pipeline (CDMAP for short). CDMAP is a novel pipeline that provides an automated, end-to-end, reproducible pipeline that provides in-depth analysis of genome-wide and replichore specific mutation rates and nucleotide triplet usage in bacterial prokaryotes. Additionally, we utilize the CDMAP pipeline to investigate patterns influencing genomic nucleotide triplets and synonymous codon usage through the lens of genomic GC content and various spatiotemporal factors.

# DEDICATIONS

This dissertation is dedicated to the many, many friends, family, and loved ones I have had the honor of meeting and sharing unforgettable memories with. I dedicate this to my parents and grandparents, James. G. Lewis, Mary G. Lewis, Jefferey M. Patton, Rosemary Patton, Opal M. Potter, and Chris Potter for believing in my ability to achieve whatever goal I put my mind to, and to be the best person I could be no matter what adversity came my way. This thesis is also dedicated to my amazing wife, Shelvasha Burkes-Patton, who was my pillar through the highest peaks and the darkest valleys during thesis and encouraging me to continue on no matter how many times the odds felt insurmountable. This thesis is dedicated to my brothers Richard Patton, Vincent Simone, Martavias Cary, Sheterral Burkes, and Mikayah Hoskins, whom are the best siblings I could ever ask for and strive to be a positive part of their lives through my actions every day.

This thesis is dedicated to my extended family of the Burkes, Givens, Lewis, and Patton family. The love, support, and understanding of the importance of this milestone in my life that has driven me towards the finish line time and time again. Additionally, this work is dedicated to my closest friends who have been with me through the all the ups and downs throughout my undergraduate and graduate work: Victor Todd Williams, Joshua Clay, Benjamin Williamson, Jen Williamson, Zachary Torrence, Jeri Colbert, Kari Hosey, Joshua Taylor, Daniel Benton, Kristin Cowherd, and the many, many more people who never stopped believing in me to achieve this dream of mine.

In addition, this thesis is dedicated in loving memory of those who unfortunately whose light have burned bright and were extinguished too soon. I dedicate this doctoral thesis to Katlin Lafitte and James Kim, whom a day does not pass without feeling the impact of your loss. Last but not least, I dedicate this thesis to Dr. Alexander Y. Gordon, professor of mathematics at UNC

Charlotte. Your infectious enthusiasm, thirst for knowledge and drive to live, learn and pass on is an irreplaceable quality that left an everlasting impact.

# ACKNOWLEDGEMENTS

This Doctoral work is dedicated to the many, folks that believed in my academic ability and determination to Succeed. I dedicate this work to Dr. Way Sung, who took a chance on me as first doctoral student and never stopped emphasizing the importance of the significance and the story your research tells. I would like to additionally thank Dr. Anita Rana, and Nathan Tyler Turner and the other members of the Sung lab past and present as serving as the bedrock of colleagues and friends that I could seek advice, knowledge, and support from. I also wish to thank Dr. Rebekah Rogers and the past and current members of the Rogers lab along with Dr. Elizabeth Cooper and the past and current members of the Rogers lab for the collaboration, camaraderie, and many, many memorable journal club meetings which served as a forum of knowledge from a diverse range of viewpoints.

I would like to also take the time to thank my graduate program advisor both past and present Dr. Cynthia Gibas and Dr. Dennis Livesay for their support and encouragement. No matter the whatever snag came up academically, financially, or logistically, Dr. Gibas would tirelessly strive to ensure myself and other graduate students were being represented in a fair and equitable fashion. Dr. Dennis Livesay taught me it was not necessary to know everything, nor should it be expected of anyone. He taught me it's important to understand the question you're asking, why, and have the ability to discover and integrate the knowledge necessary to illuminate the answers. I would like to thank the members of my doctoral thesis committee Dr. Anthony Fodor and Dr. Mariya Munir for taking the time to acknowledge the doctoral work presented in this dissertation and judge objectively on the merits of my effort. I would also like to thank Dr. Lawrence Mays, chair of the UNCC Bioinformatics department, who served as a great instructor for which to serve as a teaching assistant for and provide advice from an outside perspective as both an educator and

friend. Finally, I wish to thank the unnamed mathematics tutor at Central Piedmont Community College, who worked tirelessly with me when I set out on this path to instill the fundamental understanding of mathematics that set me on the journey that has brought us here today.

# Table of Contents

## List of Tables

## List of Figures

## List of Abbreviations/Terms/Jargon

| | |
|---|---|
| **CDM** | Context Dependent Mutation |
| **CDMAP** | Context Dependent Mutation Analysis Package |
| **CDMAP-SOA** | CDMAP Single Organism Analysis pipeline |
| **CDMAP-MOA** | CDMAP Multi Organism Analysis pipeline |
| **CDVIS** | Context Dependent Visualizer |
| **MA** | Mutation Accumulation |
| **GWTC** | Genome Wide Triplet Count |
| **RWTC** | Replichore Wide Triplet Count |
| **GC3C** | Codon Triplet |
| **Ne** | Effective Population Size |
| **CUB** | Codon Usage Bias |
| **MMR** | Mismatch Repair |
| **ORI** | Origin of Replication |
| **TERM** | Terminus of Replication |
| **NXN/NXNY/YNXN** | Nucleotide Triplet where X, Y are mutable nucleotides |
| **Fold** | the class of codon triplet conferring the number of nucleotide triplet encoding configurations |
| **DNA** | Deoxyribonucleic Acid |
| **RNA** | Ribonucleic Acid |
| **NER** | Nucleotide Excision Repair |
| **BER** | Base Excision Repair |
| **SSBR** | Single Stranded Break Repair |
| **CUB** | Codon Usage Bias |
| **START** | START codon |
| **STOP** | STOP codon |
| **tRNA** | Transfer Ribonucleic Acid |
| **Kb/Mb** | Kilobase/Megabase |
| **N-mer** | Combinatorial string of nucleotides of length N |
| **CpG** | CpG island |
| **Ts** | Transition rate |
| **Tv** | Transversion rate |
| **GBK** | Genbank .GBK file |
| **FASTA** | .FASTA sequence file |
| **VCF** | Variant Call File |

# Chapter 1: Introduction & Background

## Overview:

Understanding the role mutations play is critical, serving a fundamental force in evolutionary processes as a primary driver for genetic variation in all organisms. Shedding light on the spectrum of spontaneous mutations and factors driving mutation rates allow us to understand how organisms evolve and adapt to changing environments. Mutations can vary in size, ranging from a point mutation at a single nucleotide to large scale insertion and deletion events spanning thousands of kilobases (1-5). Understanding the role of genetic variation within organisms gives us clues to understanding mechanisms that shape gene evolution (6). Spontaneous mutation events that occur within genes are capable of generating new functionality, which can be beneficial or deleterious. In order to understand how mutations, serve a foundational role in evolutionary principles, exanimating mutations with respect to numerical, spatial, and compositional factors over time can elucidate how genomic patterns drives site specific mutation rates.

Given the importance of the mutation process in understanding evolutionary processes, little is known about the variation in mutation rate across a genome. Prior work has shown that mutation rates are not uniform across a genome, which can impact how quickly genes evolve in different regions and within different contexts of the genome. One major type of variation in mutation rate can arise depending on the localized context of neighboring nucleotides, and these context-dependent mutation rates can vary by up to 75-fold depending on its immediate 5’ and 3’ neighbor (7). The overall goal of this dissertation is to build an automated, modular, rapid, and reproducible framework to generate context-dependent mutation rates across bacterial genomes.

Previously to date, there has not been a uniform framework of which to generate and easily compare spatiotemporal variation in mutation rates, motivating us to develop CDMAP, the Context Dependent Mutation Analysis Package.

The primary Innovation of CDMAP is that it is a novel pipeline planned to be an open source, non-proprietary analysis of context dependent mutations on a per chromosome, per replichore basis. In prior studies, research and analysis on genome wide variation in mutation rates has been done, however they were done primarily using in-house scripts that cannot be readily compared across studies. CDMAP is designed with "First party principles" in mind, meaning that we designed our pipeline with the end user in mind. We designed CDMAP to intake required files for analysis, with minor user input (such as the organisms name) and then provide a fully automated end-to-end analysis. We also designed our pipeline with an emphasis on reproducibility in mind, so that when a researcher uses our software to conduct analysis, using the same input files and parameters, they should receive the same output every time. One final innovation we built in mind with was having a modular framework, so that over time, as research needs demand can incrementally implement new features on top of an existing framework to expand breadth and depth of analysis over time. This framework will be applied to a wide range of bacteria harboring single and multiple chromosomes and will be used to determine whether mutation-rate variation across organisms is linked to various evolutionary parameters like genomic GC content and codon usage. In order to accomplish these tasks, we outlined below briefly 3 distinct aims that we wish to use in order to accomplish these goals.

## Objectives:

Our first objective is to develop an automated, rapid, bioinformatics framework to generate context-dependent mutation patterns in bacteria. This framework operates with minimal data input and can be modified to analyze multiple chromosomes and multiple species. In our second objective, we apply the developed algorithms to 20 prokaryotic organisms that have available mutation accumulation data. This data contains unbiased estimates of mutation rates from both wild-type and mismatch-deficient bacteria. This analysis will also generate statistical comparisons of context-dependent mutation rates within (intraspecific) and across (interspecific) species. In our final objective, we will examine the in-depth relationship of context-dependent mutation rates and codon usage and investigate whether context dependent mutational patterns exhibit a direct or inverse relationship with codon usage.

## Background:

### Bacterial Replication and Context-Dependent Mutations

A context-dependent mutation is a mutation that is affected by local sequence context, particularly the immediate upstream 5' and downstream 3' nucleotide. To understand the context-dependent mutation process, we must first understand how mutations arise during replication. In most cases, bacteria have circular chromosomes, and in rare cases bacteria harbor linearized chromosomes similar to eukaryotic organisms (e.g., *Borrelia burgdorferi (8) )*.Often there is only a single chromosome but, in some cases, there are multiple chromosomes. Each chromosome is replicated from a single origin of replication (referred to as the ORI) and replication ends at the terminus of replication (TERM). The ORI is located near the replication origin protein DnaA, a highly conserved protein that promotes the unwinding of DNA so polymerases can attach (9). From the ORI, DNA replication occurs bidirectionally forming two hemispherical segments of chromosomal DNA referred to as replichores. Within each replichore, DNA is synthesized on both the leading and lagging strand. Replication of the leading strand occurs in a continuous manner in the 5' to 3' direction, where each nucleotide base is continuously added to the strand until reaching the TERM. Meanwhile, the lagging strand begins replication shortly after the leading strand and replicates discontinuously by the use of 150-200bp fragments called Okazaki Fragments adding nucleotides

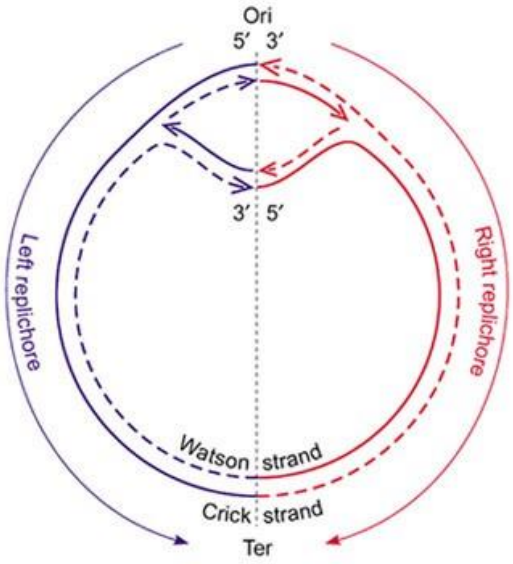



*Figure 1- An example image of replication with respect to leading and lagging strands of replication. In this example by Mackiewicz 2001, replication begins with respect to the ORI and replicates bidirectionally, where the left replichore replicates 5' to 3' and right replichore replicates 3' to 5'*

in the 5’ to 3’ direction (10). These okezaki fragments are then added incrementally on a per segment basis starting from the ORI until reaching the TERM. Due to the bidirectionality of synthesis away from the ORI (Fig. 1), and the double-stranded nature of DNA, the leading and lagging strands are synthesized in reverse. As a result, the leading strand synthesizes the “top strand” of left replichore and the discontinuous lagging strand synthesizes the “top strand” of the right replichore. This is critical to the analysis because all genomes are annotated with respect to the “top strand”. This is critical as the enzymes involved in synthesis of the top strand are asymmetrical with respect to each other in each respective replichore.

Sometimes bacteria have more than one chromosome or plasmids that replicate differently from the primary chromosome such as plasmids, megaplasmids, and Chromids. It is unclear how context-dependent mutation patterns differ with different mechanisms and timing of replication. Plasmids are among the most common extrachromosomal DNA structure harbored by bacteria (10, 11). These are most commonly found in circular bacterial organisms and are generally small (<5-10KB) circular genetic elements that have been shown to contain genes that are involved in virulence, pathogenicity, and antibiotic or heavy metal resistances (12, 13). These plasmids can move from bacterial cell to cell using horizontal gene transfer it’s not uncommon for a bacterial organism to harbor one or more plasmids within its genome (13). Megaplasmids are an in-between step between plasmids and Chromids, ranging in size from tens of kilobases to several hundred kilobases in size. Megaplasmids are shown to be a fusion of smaller compatible plasmids, that may also contain important genetic elements (12, 13). One such example would be the megaplasmids found in *Agrobacterium tumefaciens* plasmid pAtC58, which harbors a range of adaptations related to repair, resistance, and catabolism (12). The final auxiliary genomic structure present in bacterial organisms are secondary chromosomes, also referred to as Chromids (14). Chromids are larger

than megaplasmids and are speculated to have evolved from non-essential megaplasmids, since they harbor a plasmid-like replication origin proteins (14). One critical difference however is that over time chromids have come to harbor essential core genes necessary to survival within nature. Furthermore, chromids can be broken into ‘primary’ and ‘secondary’ chromids, where primary chromids are considered existential to the bacterial organism, while secondary chromids are not always necessary for survival (14).

## Mutations and Repair Mechanisms

Mutations play a fundamental role in driving the evolution of genome architecture and can be defined as the alteration of one or more nucleotide bases within the genomic sequence of an organism (15, 16). This broad classification can be further subdivided into specific types of genomic mutations. The simplest type of mutation is called a point mutation and is a mutation at a single nucleotide. This mutation can be either be a base substitution that replaces a single nucleotide with another nucleotide, or a single base-pair indel (insertion/deletion), which inserts or deletes a single nucleotide. More complex mutations include inversions, where a segment of one or more nucleotide bases have their order inverted in a genomic sequence and large-scale indels, which involve indels greater than a single nucleotide.

| Repair Pathway Mechanism: | Overview: |
|---|---|
| Mismatch Repair (MMR) | Evolutionarily conserved pathway; repairs base mismatches occurring from replication or indel loops; contributes 100x to replication fidelity. |
| Nucleotide Excision Repair | Repair pathway responsible for removing bulky lesions such as Cyclobutane Pyrimidine (CPD) dimers, Photoproducts (PP), errors caused by chemotherapeutics. |
| Base Excision Repair | Corrects abasic, alkyation, deamination, and oxidative single base damage. |
| Single Stranded Break Repair | Corrects unresolved breaks caused by replication failure occurring from oxidative damage, or defective TOP1 enzyme. |

*Table 1- Table of Bacterial Repair Mechanisms*

The genome-wide base-substitution rate has been shown to range from $1\times10^{-11}$ to $1\times10^{-8}$ mutations per site per generation (7, 17-23). However, an ever-growing body of work has observed that the point mutation rate can be heavily affected by neighboring nucleotides (7). Sung et al., found that site-specific mutation rates greatly vary depending upon the nucleotide, and varying the neighboring nucleotide can have a 75-fold impact on mutation rate (24, 25). This effect, coined as

Context-dependent Mutations, is a measure of the effect that neighboring nucleotides have on the rate of a point mutations. Within the scope of this dissertation, we will be primarily be focused on this specific category of mutations at a singular nucleotide base, and the influence of its neighboring nucleotides play in modifying the rate at that base. When discussing how context dependent mutations arise within the genome of an organism.

Mutations arise through endogenous and exogenous damage. Endogenous mutations primarily occur through oxidative and hydrolytic reactions within an organism (26). Exogenous mutations occur primarily from external influences such as various forms of radiation, chemical agents such as alkylizers, toxins, and environmental stressors such as extreme heat or cold (26). Both exogenous and endogenous mutators can produce a wide variety of mutations, from benign to lethal mutations, so organisms naturally have developed multiple different repair pathways to address these mutations.

There are several prominent and well characterized repair pathways that directly address mutations arising from replication error and mutagenesis. Mutations occurring due to oxidative stressors or deamination usually are repaired using the Base Excision (BER) pathway, though this pathway is often limited to single base repair, while multi-base mutation or replication errors require more robust pathways (26). Damage caused by chemotherapeutics or from UV radiation often will use the Nucleotide Excision Repair (NER) pathway to repair these types of errors, which specialize in removing bulky lesions that arise from photoproducts and Cyclobutene Pyrimidine Dimers (CPD) (26). The last repair mechanism worth mentioning within the scope of this dissertation is the Mismatch Repair pathway (MMR). MMR is an evolutionarily conserved post-replication pathway that contributes upwards of 100-fold to replication fidelity (7, 18, 20, 26). These repair mechanisms are highly conserved across organisms, and the absence or the failure of

these repair mechanisms can have a significant impact on the replication errors that are not repaired at a site. In addition to wild-type organisms, we will be evaluating context-dependent mutation patterns in repair deficient bacteria. These repair mechanisms govern the repair of various types of damage. If the absence or mutations affecting these repair pathways can cause significant elevations in mutation rate and can lead to fundamental changes in genome architecture and accelerate the evolutionary trajectory of features such as new virulence or antibiotic resistance mechanisms in bacteria.

## Mutation Accumulation and the Evolution of Mutation Rate

When investigating the evolution of mutation rates across organisms, we must have a basic grasp of the major evolutionary forces driving the fixation of mutations. The two primary mechanisms that drive the fixation of mutations are Natural Selection, and Genetic Drift. Natural selection is the classical mechanism associated with driving evolutionary change, in which one or more alleles increases in frequency, due to its ability to improve the fitness (ability to survive and reproduce in an environment) (18, 27). Selection operates to remove mutations that are deleterious, and a majority of mutations are deleterious. Genetic drift operates irrespective of fitness, and mutations are driven to fixation by random chance (18, 27). In order to investigate the full distribution of possible mutations, we try to examine organisms that accumulate mutations in the absence of natural selection. (25).

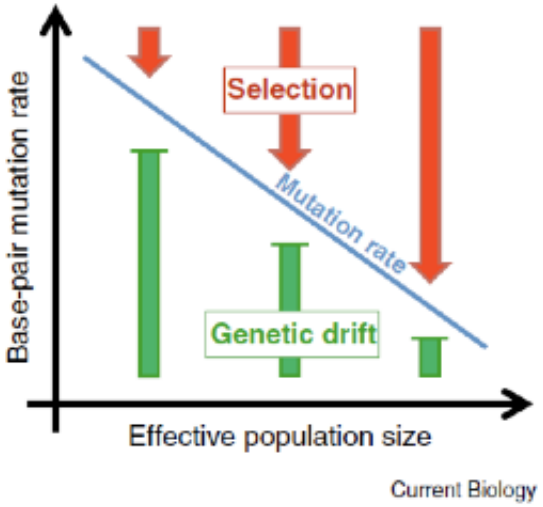



*Figure 2: Genetic Drift versus Selection –As effective population size (Ne) increases, that fixation of mutation driven less by genetic drift, and increasingly by natural selection. Illustration by Sniegowski & Raynes from Cu Biology 2013*

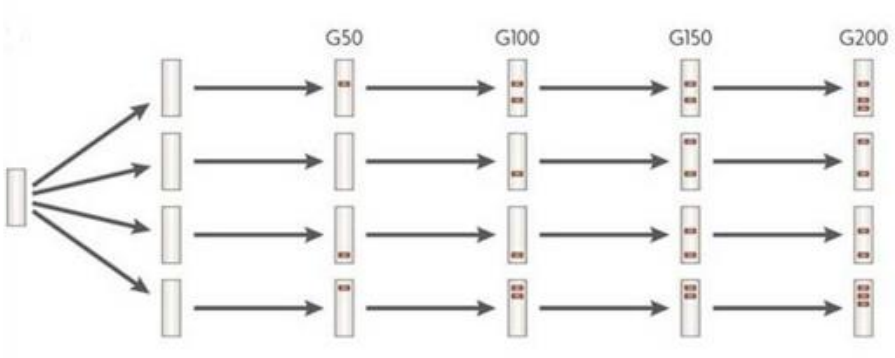



*Figure 3: Mutation accumulation and Bottlenecking – Mutation Accumulation occurs by taking a single ancestral organism and establishing multiple MA lines, which are repeatedly propagated by selecting a single offspring in a new environment repeatedly allowing an unbiased accumulation of mutations. Illustration by Baer et al., Nature Reviews: Genetics 2007.*

One such method commonly employed to study mutations in the near absence of natural selection is a Mutation Accumulation (MA) experiment. When carrying out an MA experiment, a single organism is taken and repeatedly bottlenecked. Bottlenecking takes the single offspring of the ancestral organism and transfers to a new, identical environment to propagate. The repeated process of bottlenecking across an MA experiment minimizes selection, and genetic drift ultimately drives the fate of new mutations in a nearly neutral fashion (25). After reaching a designated number of generations in the MA experiment, whole-genome sequencing can be used on each lineage of a MA to provide a comprehensive picture of where, what type, and how often spontaneous mutations occurred (7, 19-21, 25, 28, 29).

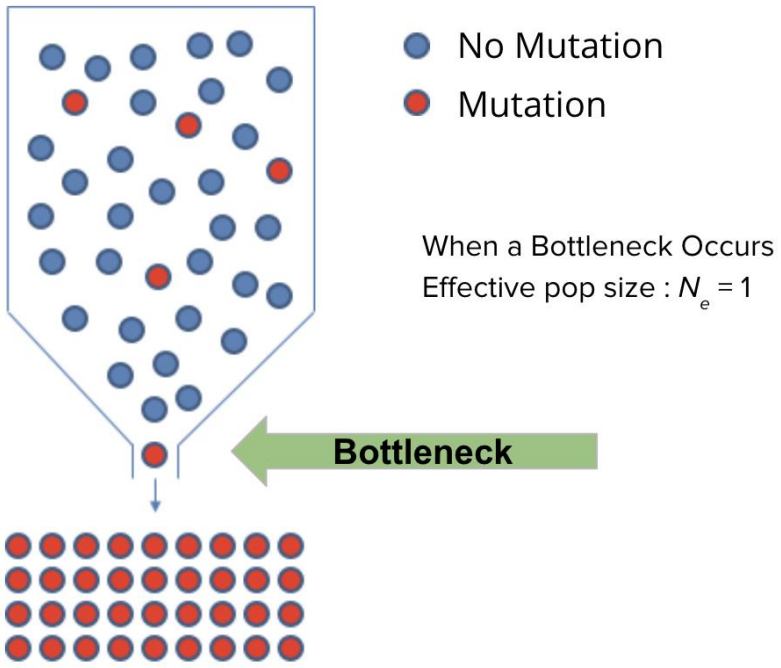



*Figure 4: Bottlenecking – Illustration of the Bottlenecking process courtesy of the Sung Lab. When an organism is bottleneck occurs, the effective population (Ne) effectively becomes one. As a result, the fate of mutational fixation is entirely determined by genetic drift.*

**Synonymous Codon Usage and Underlying Patterns of Contextual Bias**

Understanding the evolutionary role that mutations and their contextual patterns play on genomic architecture is fundamental in gaining deeper understanding how mutational adaptation affects organisms across the tree of life. One approach is observing and analyzing consequential effects at play in a genomic context. One way to accomplish this is by inspection of frequencies in

which nucleotide triplets occur within coding regions. For every amino acid, including both the initiator (START) and terminator (STOP) amino acids, one or more nucleotide triplet encodes a given amino acid. These nucleotide triplets are called Codons (30, 31). For a given codon $N_1N_2N_3$, where $N_i$ is any nucleotide at $i^{th}$ position, one or more of these nucleotides contribute to an amino acid's codon degeneracy, i.e., a mutation conferring a synonymous or nonsynonymous change in the amino acid expressed (32). Oftentimes, both $N_1$ and $N_2$ are more actively conserved than $N_3$ as a mutation in the former positions will induce a nonsynonymous mutation changing the amino acid encoded (32-34). This degeneracy in codon usage corresponds to the number of codons encoding a particular amino acid. For example, Alanine is a four-fold degenerate amino acid as it is encoded in the generalized form $GCN_3$, where the third nucleotide position can be any nucleotide. As alluded to in the prior sentence, Four-fold degenerate amino acids illustrate an interesting phenomenon, Where all synonymous mutations that encode them are dependent on $N_3$, while other fold degenerate amino acids either require multiple $N_i$ mutations to encode another amino acid, or are unable to use a subset of nucleotides in $N_3$ lest encoding a nonsynonymous mutation.

The corresponding frequency in which codons are utilized in fold degenerate amino acids is referred to as their Codon Usage Frequency, or Codon Usage for short (30, 31). The frequencies in codons are used in each fold degenerate amino acid are often unequally utilized, generating biases a given amino acids codon usage, or more succinctly called Codon Usage Bias (CUB). The prominence of these biases is influenced by both exogenous and endogenous factors in varying intensities (30, 32-40). One immediate factor influencing codon usage would be strand specific mutational biases, which relates to the asymmetric nature of replication between the leading and lagging strand (30-32). For a given codon on the leading strand 5'- $N_1N_2N_3$ -3' would be encoded 3'- $N_{3c}N_{2c}N_{1c}$ -5' on the lagging strand, where $N_{ic}$ represents the ith nucleotides complementary

nucleotide base on the lagging strand. In many cases this leads to the resulting translated lagging strand codon often will neither encode the same amino acid, or even the same class of fold degenerate amino acids.

Another endogenous factor that has shown to play a significant role affecting codon usage is the presence of genomic GC/AT content (30, 33, 38, 40). Organisms harboring a particular enrichment towards either extreme have shown evidence towards specific codons, nucleotides in specific positions, or even the strength of selection acting upon a given codon (33, 37). Another endogenous factor considered when analyzing patterns arising often in tandem is selection pressure acting on the abundance and efficiency of tRNA populations in an organism (30, 37, 40). Prior research suggests that selective forces act on tRNA isoacceptors maximizing expression of higher accuracy and efficiency is also driven by tertiary factors such as organism growth stage and minimizing rejection of incorrect tRNAs (30, 31, 37). Meanwhile, Exogenous factors have shown a less uniformly measured effect on amino acid and codon usage. Some factors such as exogenous damage from radiation have shown a consistent effect on biases in codon usage, while other forms of exogenous damage like nutrient poor environments, retroviruses, PH specificity of environment can contribute to differences in usage in specific circumstances (41).

## Datasets:

| Number | Organism | Gen. | Lines | Chromosomes | Mutations | GC Content | ORI (KB) |
|---|---|---|---|---|---|---|---|
| 1 | *Agrobacterium tumefaciens C58* | 5819 | 47 | 2 | 233 | 60.2 | 2765 |
| 2 | *Bacillus subtilis NCIB 3610* | 5077 | 50 | 1 | 350 | 43.65 | 4170 |
| 3 | *Bacillus subtilis NCIB 3610 (MMR-)* | 5077 | 50 | 1 | 5295 | 43.65 | 4170 |
| 4 | *Burkholderia cenocepacia HI2424* | 5554 | 47 | 3 | 130 | 67.28 | 69 |
| 5 | *Caulobacter crescentus NA1000* | 4284 | 44 | 1 | 259 | 67.62 | 3818 |
| 6 | *Colwellia psychrerythraea 34H* | 1078 | 84 | 1 | 400 | 38.95 | 4903 |
| 7 | *Deinococcus radiodurans BAA-816* | 5961 | 43 | 2 | 331 | 67.55 | 258 |
| 8 | *Escherichia coli K-12 MG1655* | 1682 | 46 | 1 | 1623 | 51.86 | 3644 |
| 9 | *Escherichia coli K-12 MG1655 (MMR -)* | 1682 | 46 | 1 | 231 | 51.86 | 3644 |
| 10 | *Kineococcus radiotolerans SRS30216* | 4724 | 44 | 1 | 280 | 74.4 | 522 |
| 11 | *Lactococcus lactis DSMZ20481* | 3973 | 63 | 1 | 813 | 36.58 | 1 |
| 12 | *Mesoplasma florum L1* | 2351 | 28 | 1 | 544 | 27.14 | 326 |
| 13 | *Mycobacterium smegmatis MC2 155* | 4900 | 49 | 1 | 856 | 67.7 | 3609 |
| 14 | *Rhodobacter sphaeroides ATCC 17025* | 4544 | 46 | 1 | 107 | 68.63 | 2 |
| 15 | *Ruegeria pomeryoi DSS-3* | 5386 | 47 | 1 | 147 | 64.71 | 1 |
| 16 | *Staphyloccus aureus ATCC 25923* | 2716 | 83 | 1 | 274 | 33.74 | 3 |
| 17 | *Staphyloccus epidermis ATCC 122228* | 7101 | 22 | 1 | 294 | 33.2 | 2419 |
| 18 | *Teredinibacter turnerae T7901* | 3025 | 42 | 1 | 779 | 51.72 | 4252 |
| 19 | *Vibrio fischeri ES114* | 5187 | 48 | 2 | 132 | 39.79 | 1 |
| 20 | *Vibrio fischeri ES114 (MMR -)* | 5187 | 48 | 2 | 2909 | 39.79 | 1 |

*Table 2- List of bacterial organisms analyzed with CDMAP*

Mutation accumulation datasets span the prokaryotic kingdom (7, 17-21, 24, 42). The organisms analyzed in Table 1 harbor a variety of different genomic architectures, the majority of which harbor a single circular chromosome such as *Escherichia coli* and *Bacillus subtilis.* In this

scenario, these organisms are the most straightforward to run an analyze. We also analyzed wildtype organisms harboring multiple genomic architecture elements outside of a single primary chromosome such as *Deinococcous radiodurans* and *Rhodobacter sphaeroides* to name a few. In the case of these two organisms such as *R. Sphaeroides* exhibits a primary circular chromosome with library of 5 different plasmids that vary from 13kb to over 877kb in size. Meanwhile, *D. Radiodurans* harbors a primary circular chromosome, a secondary chromosome of 400kb, a 177kb megaplasmid, and a 45kb plasmid. While these organisms both share similarly sized genomic elements, they are categorically different in classification and function.

Using CDMAP to analyze these organisms may illuminate genome-wide patterns that may provide insight into the similarities and differences of genomic architecture composition that govern these organisms. Additionally, we analyze several organisms that had MA lines carried out that either had MMR- repair knockouts, or in the case of *Mesoplasma Florum* have no MMR repair mechanism present in the entire genome. Analyzing the differences in MMR deficient organisms can possibly provide insight into how organisms in the absence of certain repair pathways can mitigate excessive elevation in site-specific mutation rates, as well as providing possible insight into how novel mechanisms relating to pathogenicity and antibiotic resistance arise in bacterial organisms, so we can adjust and adopt new targeting methods in future therapeutics.

## Prior Work Regarding Context-dependent Mutations

There are many spatiotemporal factors that can influence mutation rate. First, there is an increasing amount of evidence of context-dependent mutation processes. In prior work, MA experiments with *Mesoplasma florum*, *Escherichia coli*, and *Bacillus subtilis* revealed context-dependent mutation rates that different by 75-fold depending on the upstream and downstream nucleotides. Furthermore, it was shown that these mutations arose asymmetrically within each replichore (35). This study posited that asymmetrical context dependent mutations were caused by the ORI locus orientation with respect to the leading and lagging strand. Furthermore, studies in *E. coli* also revealed asymmetric patterns in replication between the leading and lagging strand, which was supported by sharp differences in GC content and nucleotide composition between the leading and lagging strand (25). Finally, within the 1000 genomes project for humans, Harris, et. al. found a prevalence of the 3mer triplet 5'-TCC-3' to 5'-TTC-3' being a central causative factor present in human skin cancers (43). Another study investigated how replication asymmetry within *Vibrio Cholerae* developed into the current representative pathogenic strain from a non-pathogenic strain (5). In this paper, the authors carried out MA experiments on non-pathogenic WT and MMR deficient strains of *V. Cholerae*, using varied media limited and antibiotic resistance laden environments. The primary results of their work found concentrated, elevations in mutation rate at the TERM of chromosome 2, and midpoint of chromosome 1. This was due in part to replication timing mechanisms that begin replication of the second chromosome after temporarily suspending replication in the primary chromosome. Understanding the effect of neighboring nucleotides in triplets that may have a predisposition to increased mutational pressures at key points in the replication process can possibly shed light into the how novel mutations that arise conveying pathogenicity or resistance in bacterial organisms occur.

One paper that investigated the role of flanking base composition, and mutation dynamics in the maize genome revealed illuminated several findings related to the role sequence context plays in driving mutation rates. Originally, they investigated the CpG effect driving mutation rates due to being well characterized as a driver of transition rate mutation deamination at methylated CpG sites (44). One finding from the paper showed a consistent G+T skew over C and A nucleotides, which was in agreement with a recent observation in human genomes (45).When they calculated GC and AT mutation rates with respect to immediate 5' and 3' flanking nucleotides, they found higher GC content across all sites, which couldn't be accounted for by the CpG effect (44). As a result, they investigated the role of cytosine deamination, and found a 2.1x increase in transition site mutation compared to other contexts, with the highest transition rates occurring at CG > CA and CG > TG sites. When they analyzed context specific sites in the absence of CpG sites, and analyzing whether a context was flanked by 0, 1, or 2 A+T nucleotides, they found significant differences in Transition / Transversion (Ts:Tv) ratios between sites flanked by 1 or 2 A+T nucleotides, supporting that flanking bases influence mutation rates. Finally, they compared GC > AT and AT > GC mutation rates and found that as regional A+T nucleotide content increases in the genome, that transition mutation rates of GC > AT increased, while AT > AT and GC > GC mutations remained relatively constant (44).

In another study done using *B. subtilis* that investigated the effect of sequence context of DNA polymerase error rates both in vitro and in vivo. They conducted MA experiments using WT and MMR- using MutSL knockouts. The results found a 60-fold increase in mutation rates, which was shown to be similar to rifampin resistant strains, averaging 15.5 generations as opposed to 909 generations per mutation (11). When investigating site-specific mutation rates, they found a 403-fold difference in mutation rates between the highest transition rate triplet, 5'-CCG-3' and the

lowest transition rate triplet 5’-AGT-3’after normalization with respect to the leading strand. Although transversions were rare in this study, they were able to conclude sequence context highly influenced transition mutation rates. In a study examining germline mutation rates in humans, they took 36 million singleton variants spanning 3560 whole genome sequences studying extremely rare variants (ERVs) showed heterogeneity of mutation rates were dependent on nucleotide contexts (46). They found when using ERVs to estimate germline mutation rates according to mutation type with respect to context mutations, they found effects of sequence context 3 bases up and downstream of the mutation from a given variant site (46). They found these 7mer context frames in agreement with prior work, and found several higher mutable 3mer triplets, 5mers, and a hypermutable 7mer NTT[A > T]AAA which had a 6-fold higher mutation rate than the generic A>T 1mer.

## Prior Work Regarding Codon Usage

Factors affecting Biases in codon and amino acid usage frequency are multifactorial, with most prior work often investigating one or two dimensions that could contribute to fluctuation in rates. For example, in one study, the authors chose to look at genome architecture variation contributions to CUB in 16 *S. aureus* phages (47). The investigators analyzed over 900 protein coding genes, filtering out genes with less than 50 codons analyzing the codon usage count relative to mutation rate. Analysis of the 16 phages showed AT enrichment at the $N_3$ codon position, and self-separating patterns of codon usage in more virulent strains of the phages (47). More interestingly, the investigators found a pattern of over representation of 19 codons, with 14 $N_3$ codons ending in a T, while the other 5 ended in A nucleotide (47). One study focused on the analyzing the fitness cost of synonymous variants of the GFP gene in *E. coli* strains (39). The authors found synonymous variation in the GFP conferred significant reduction in cell growth rate viability, and upon evolving suppressor strains that could mediate the toxicity of the variant GFP gene by mediating mRNA expression levels (39).

Several studies have investigated the relationship between synonymous codon usage bias (SCUB) and radioresistance levels in bacterial organisms (40). The authors found a tandem relationship in tRNA gene copy number and CUB, where 9 of the 16 species shared the similar tRNA copy numbers for each codon (40). More interestingly, they found that in more radioresistant bacteria GC enrichment in the $N_3$ codon position allowing increased thermic stability over AT $N_3$ codons (40). The authors further speculated this GC enrichment is crucial to prevention of exogenous damage from ionizing and UV radiation. Another similar study in Cyanobacterial genomes sought to specifically analyze relationships of GC/AT enrichment at $N_3$ nucleotides and codon usage patterns (38). The authors noted distinct groupings in relative synonymous codon

usage for AT rich and GC enriched organisms, citing nucleotide compositional constraints and habitats as drivers in codon usage patterns (38).

Several authors have attempted to analyze patterns of codon usage from a broader perspective. In one such study, authors analyzed 29 prokaryote families and found codon optimization relative to tRNA presence and gene expression level (33). An interesting takeaway from their work was fold-specific enrichment of $N_3$ nucleotides in codons, where 2-fold and 3-fold degenerate codons were enriched towards C nucleotides, and 4-fold degenerate codons being biased towards T nucleotides at $N_3$ nucleotides (33). Another notable takeaway found was a threshold in GC enrichment (approximately 40% GC rich) pointing towards increased selective pressure at $N_3$ nucleotides (33). Another cohort of researchers investigated how patterns in codon usage vary when examining bacterium in extremophilic environments, where they found a combination of lower CUB values in tandem with transcription-based adaptations induced by the selective pressures of high acidity, low PH environments (41).

# Chapter 2: Development of the CDMAP Pipeline

## Introduction

Parsing context-dependent mutation patterns is not an easy task. First, to analyze both upstream and downstream effects, each mutation has to be divided into multiple different replichores and categorized into 64 different bins. Second, characterizing the influence of context dependent mutations has to be broken down into a per-replichore, per-chromosome basis due to the asymmetry of replication from an ORI. Third, the ORI and TERM points have to be identified and mutations have to be reoriented to those points. Fourth, this process must be agnostic to input data and be developed within a uniform reproducible framework. While individual tools have been developed for each of these steps, there has not been an end-to-end automated process to easily analyze context-dependent mutation processes.

With these goals in mind, we began development of CDMAP incrementally, expanding features as we verified compatibility and consistency of results in each successive step. The assumptions made for the current version of CDMAP assumes that we are analyzing a prokaryotic organism, with a single ORI and TERM within a single circular chromosome, chromid, or plasmid. With those assumptions in mind, CDMAP takes the required input files and automatically partitions both replichores, calculating and recording mutation rates, triplet rates, codon usage, and upstream and downstream context dependent mutation influence for a single organism. CDMAP conveniently and easily organizes text and image output for postprocess analysis and downstream interspecies analysis. Pipeline validation was carried out using prior results from context mutation analysis in *B. subtilius, E. coli, and M. florum (7).*

# Methods

## *Data Input*

CDMAP requires several input files. The first file required would be a variant base call file, which contains the position of the variant, the original nucleotide, and the variant nucleotide. The next file required is the reference FASTA file, which contains the genomic sequence needed for mapping the variant nucleotide to the neighboring nucleotides. The final file needed is optional, though highly recommended for best results is a fully annotated GBK file from which the ORI and TERM positions are identified. In the absence of this GBK file, manual designation of the ORI and TERM can be used to execute the program.

## *Data Dependencies*

In terms of the actual pipeline, several package dependencies are required to run the CDMAP pipeline for single organism analysis and visualization. The first package used during analysis is the SeqINR, a bioinformatics toolkit that allows easy manipulation and partitioning of the reference FASTA file, along with containing other needed dependencies (48).One such dependency is OriLoc, a tool that was developed to determine replication origins in prokaryotic organisms using DNA skew, which helps identify nucleotide usage asymmetries to determine the ORI and term position using the GBK file (8). Another bioinformatics toolkit used in analysis is BiocManager, which is a collection of various bioinformatics analysis tools allowing easy manipulation of data objects at various stages of processing data. In BiocManager we use two specific packages to aid with analysis at different steps throughout the pipeline: pracma and genbankr. Pracma stands for ‘practical math’ and is itself a collection of different numerical analysis methods that we are implemented on the back end for partitioning data objects at various

stages, and aids in developing backend objects for downstream visualization. Genbankr is a tool primarily used for parsing and manipulation of genbank files into data objects within R to easily extract and coerce features from as needed. The last major dependency used in our pipeline is Lattice, a lightweight data visualization package that enables us to generate high quality visual heatmaps without excessive data transformation steps (49).

## Replication ORI and TERM Determination, Replichore Partitioning

In prior studies, results have shown asymmetrical context-specific mutation patterns with respect to the ORI and TERM, therefore accurate determination of the replication origin is a critical first step in our analysis (4, 7, 11, 18-21, 42, 43, 50). In previous work, replication origin was approximated based on the location of the replication origin protein DnaA or using the midpoint of the sequence. However, the true starting point of replication may vary by several hundred bases to several thousand Kb from the start position of DnaA (29).To provide the currently most accurate estimate of ORI and TERM position, we used the OriLoc R package to determine ORI based on nucleotide skew (8, 48).

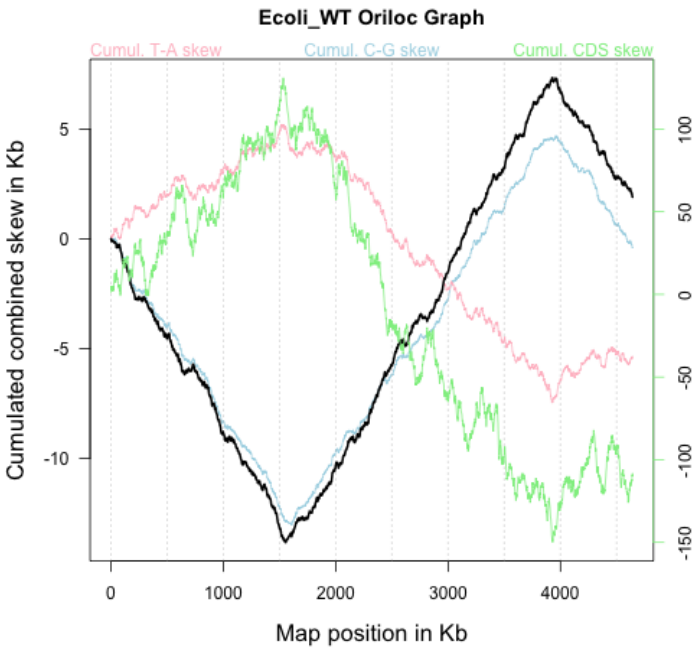
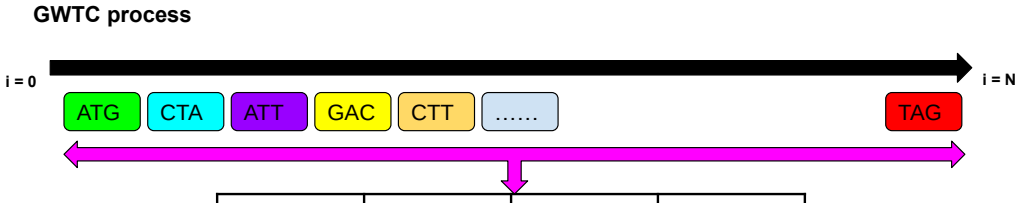



*Figure 5: Visual depiction of the oriloc, where a DNA walk is performed then transformed into a combined skew vs position in KB. The Supremum, or maximum skew position of the dataset, and the infimum, or the minimum skew position, make up the ORI and term respectively.*

Single organism analysis in CDMAP was carried out using MA datasets from prokaryotic organisms (4, 7, 11, 24, 36), most of which harbored a single chromosome. Some of the

prokaryotes harbored secondary genomic elements such as secondary chromosomes, chromids or plasmids. Some of the bacterial organisms we analyzed included both wildtype and MMR- strains. Having a large breadth of different prokaryotes along the tree of life allowed us to compare and contrast context-dependent mutation patterns across varying levels of replication and repair fidelity, nucleotide base composition, and genome sizes, allowing us to generate a comprehensive picture context mutation influence on context-dependent mutations.

## Nucleotide Frequency Determination and Codon Usage




| | | | |
|---|---|---|---|
| TTT | TGT | TCT | TAT |
| TGT | TGG | TCG | TAG - 1 |
| TTC | TGC | TCC | TAC |
| TTA | TGA | TCA | TAA |
| GTT | GGT | GCT | GAT |
| GTG | GGG | GCG | GGA |
| GTC | GGC | GCC | GAC - 1 |
| GTA | GGA | GCA | GAA |
| CTT | CGT | CCT - 1 | CAT |
| CTG | CGG | CCG | CAG |
| CTC | CGC | CCC | CCC |
| CTA - 1 | CGA | CCA | CAA |
| ATT -1 | AGT | ACT | AAT |
| ATG - 1 | AGG | ACG | AAG |
| ATC | AGC | ACC | AAC |
| ATA | AGA | ACA | AAA |

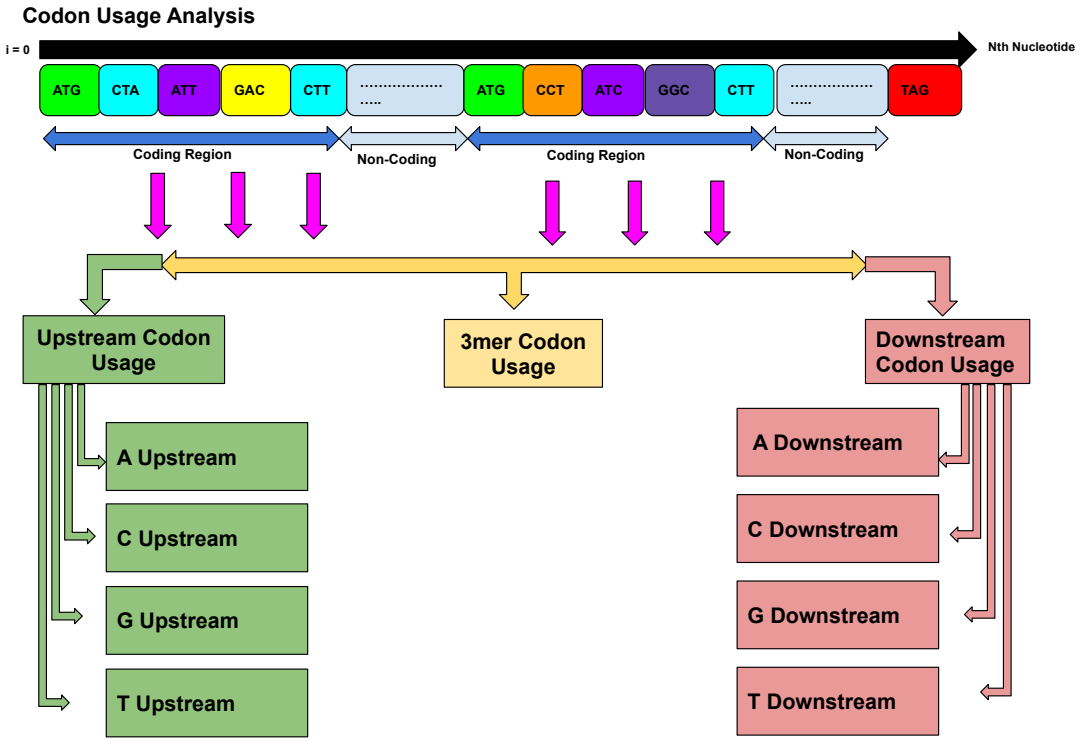



*Figure 6: GWTC and Codon Usage pipeline representation. The genome wide triplet count is performed by sequentially calculating for each triplet in the sequence and storing them in 16x4 matrix.. In a similar manner to GWTC, codon usage for triplets, upstream, and downstream calculations are performed by sequentially analyzing each triplet in the FASTA sequence contained in coding regions. In upstream and downstream calculations there is an additional step determining which upstream and downstream matrix to be stored in dependent of the upstream and downstream nucleotide.*

After each replichore has been partitioned and oriented with respect to the ORI and TERM, nucleotide triplet counts (both coding and non-coding) and codon usage within coding regions are generated. The observed counts for nucleotide usage and codon usage are done on a chromosome wide (Genome Wide Triplet Count, or GWTC) and a replichore specific basis (Replichore Wide Triplet Count, or RWTC) respectively. For coding regions, additional upstream and downstream nucleotides are captured for each codon in order to evaluate the context specific mutation patterns on codon usage bias.

Given the methodical construction taken in partitioning both the variant base call mutations and reference sequence into their respective replichores, we can examine resulting output in a variety of ways. By default, CDMAP analyzes a given chromosome bidirectionally beginning at

the ORI to TERM, however due to the flexibility of the existing framework, the end user can easily analyze site-specific mutation rates solely with respect to leading strand (also referred to as clockwise chromosome orientation) or the lagging strand (also referred to as counterclockwise chromosome orientation) if the user desires. A visual representation of this can be seen in figure 7.

**Site Specific Mutation Frequency & Rate Analysis**

CDMAP parses each subset of base-substitutions in each replichores for the position and the type of base substitution that occurred, then extracts the neighboring nucleotide information from the reference FASTA file. In order to calculate mutation rates and analyze individual contexts of each nucleotide triplet site (*K*), the triplet count for each nucleotide triplet site in each replichore and for each chromosome was calculated using the following formula:

$$K_{genome} = \frac{M_{Triplet}}{GWTC_{Triplet}} \qquad K_{replichore} = \frac{M_{Triplet}}{RWTC_{Triplet}}, \textit{ for Left, Right Replichore}$$

Where $K_{genome}$ and $K_{replichore}$ represent the mutation ratio for each nucleotide triplet site on a chromosome or replichore wide scale, $M_{codon}$ represents the number of observed base substitutions for a given nucleotide triplet site triplet, and $GWTC_{codon}$ and $RWTC_{codon}$ represent the genome wide triplet count and the replichore wide triplet count on a given replichore at a given nucleotide triplet site. To calculate the base substitution mutation rate, which forms the baseline rate for the majority of analysis, we used the following equation for chromosome and replichore specific mutation rates:

$$U_{bs} = \frac{M_{triplet}}{(GWTC_{triplet})(G)(N)} \qquad R_{bs} = \frac{M_{codon}}{(RWTC_{Triplet})(G)(N)}$$

Where $U_{bs}$ represents the chromosome and $R_{bs}$ represents the replichore specific base substitution rates for a given nucleotide triplet site for each generation. $M_{codon}$, $GWTC_{codon}$, and $RWTC_{codon}$, are the same as above, and *N* represents the number of MA lines, and G represents the number of generations occurred for a given organisms MA study.

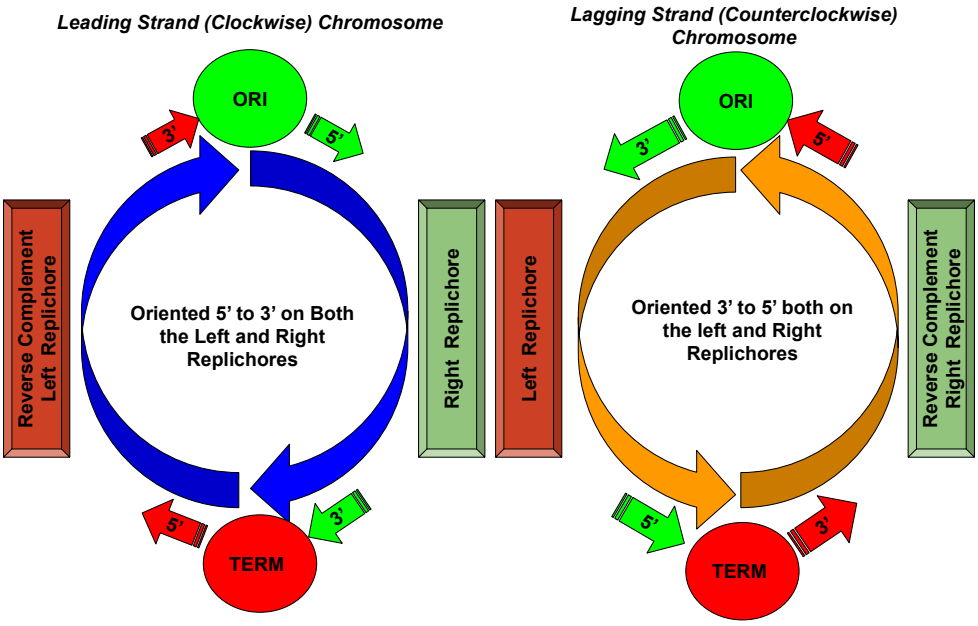



*Figure 7: Illustration of Clockwise and Counterclockwise Chromosome Orientation. The Clockwise chromosome corresponds to the normal top strand analysis where the look at replication 5'>3' with respect to the origin of replication. The Counterclockwise chromosome looks at the Lagging strand replication from 3'>5' to compare local sequence context variation on the leading and lagging strand.*

Once all observations about codon usage, nucleotide triplet frequency, and mutation rate calculation have been completed, CDMAP prepares all output for end user analysis. As a convenience for the user, CDMAP dynamically generates output repositories and subdirectories for each organism as results are generated, and then outputs them into the proper subdirectories for the ease of navigation. By default, all output datasets generated by CDMAP are generated both as output visualizations by lattice for immediate graphical interpretation, and as text files so

visualization or post processing may be carried out by the end user if they so desire (such as commonly used graphing programs (examples: ggplot or tableau).

**Visualization and Benchmarking**

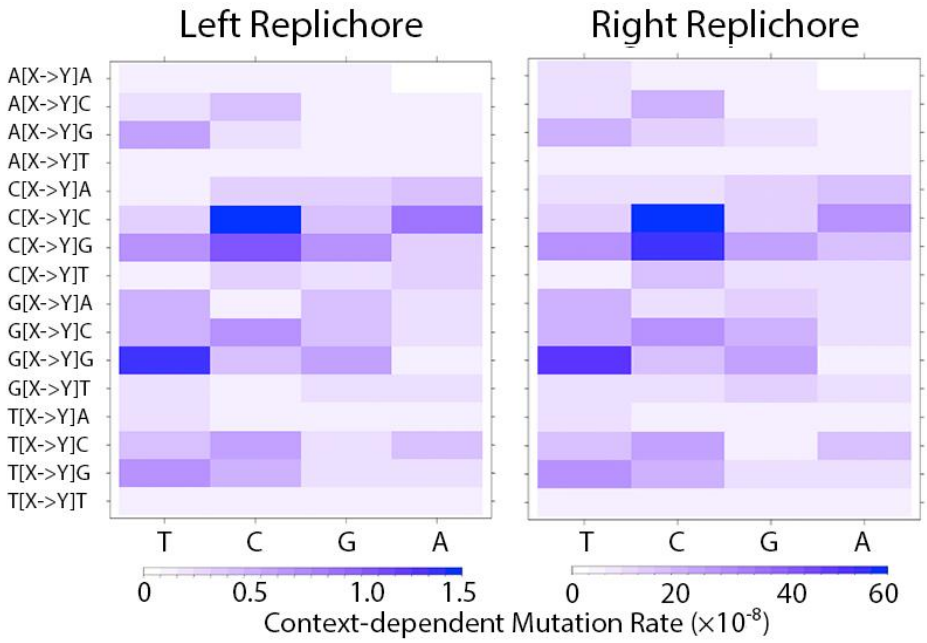



Figure 8- CDMAP Single Organism Analysis (SOA) output of *Bacillus subtilis* mismatch repair deficient MA lines. Context-dependent mutation rates shown for left replichore and right replichore. Each row repesents a triplet N[X->Y]N with N as the left and right neighboring nucleotide and X and Y repesenting the reference nucleotide and mutation respectively.

Complex patterns within large-scale data sets are often easier to identify using visualization tools. Relevant information about triplet frequency, variant distribution, and genome-wide and replichore-specific mutation rates are passed through Lattice and correlation between input files can be automatically formatted (Figure 8 and 10) (49). Throughout the process, CDMAP collects and outputs both CSV format spreadsheets and heatmaps in dynamically generated output directories that are categorized for easy navigation and downstream analyses.

CDMAP was benchmarked against 17 mutation accumulation datasets in prokaryotic organisms (7, 18-21, 24, 36, 42), which harbor a variety of different genomic architectural features. The majority of MA studies contain organisms with a singular, circular chromosome such as *Escherichia coli*, while others may have multiple genomic elements such as chromids and plasmids (Table S1), or may be deficient in repair enzymes such as mismatch repair. An example of replichore-specific output automatically generated using Lattice during the execution of CDMAP is shown in Figure 8. CDMAP can be installed and run on a standard laptop or desktop, with an approximate runtime of 90 minutes for an average size bacterial genome (~5mb), and benchmarked data for MA lines has been uploaded to CDVIS. CDMAP/CDVIS serves as a foundational framework for analysis of context-dependent mutations and spatiotemporal variation in mutation rate across multiple of organisms.

## Discussion

There are several potential issues and concerns regarding the input, output, and development of software. Currently, in Aim 1, CDMAP was developed to handle a single prokaryotic circular chromosome with a singular ORI and TERM. In future work, plans to expand CDMAP to analyze different genomic features and genomic architecture. For example, archaea and eukaryotes have multiple ORI points and the distance between ORI points would have to be accounted for in the analysis. One solution would be to develop a pre-processing mechanism that would allow us to find an 'optimal' replication origin point to act as an anchor for genome wide analysis in organisms with multiple ORIs. Additionally, in light of recent events, we would also like to consider adding functionality for analysis of viral organisms as well. Understanding the genome-wide patterns influencing factors such as how the context-dependent mutation process affects the evolution of viruses such as Covid-19 or other future threats may prove invaluable for

understanding and development of targeted therapeutics in the future. While early work on development has been promising on expansion of this feature, additional modular features incorporating phylogenetic methods would warrant additional consideration as well.

Another issue of CDMAP is that it relies on the quality of the input data. Experimental errors, sequencing errors, and statistical power in detecting mutations are all upstream issues that may influence the output data. An example would be extraction from an overgrown colony on an agar plate, whether due to user choice, or due to overgrowth from incorrect incubation times leading to selection operating within MA line. Another possible upstream problem could arise from an inadequate number of generations carried out in the MA experiment, leading to lack of accumulated mutations due to premature sequencing and termination of the experiment. Problems may also occur upstream due to sequencing errors due to lack of coverage depth leading inaccurate identification mutations. While analysis and mutation calling in MA data are generally straight forward due to high sequencing coverage, CDMAP may have more issues identifying context-dependent mutation patterns in lower-covered datasets or data from natural populations.

# Chapter 3: Multi-Organism Analysis of Context Dependent Mutation Rates Across a Diverse Array of Bacterium

## Introduction

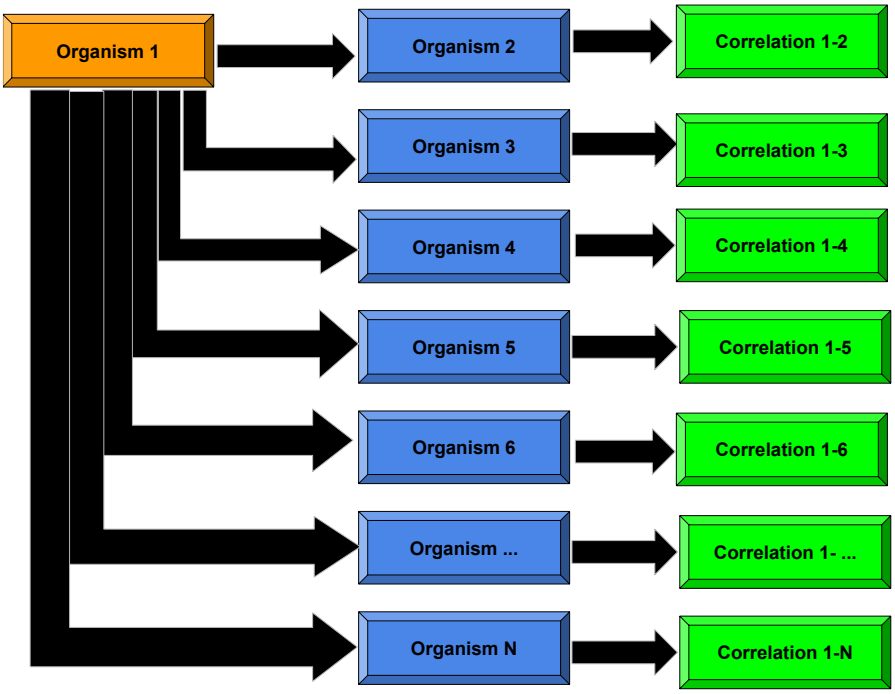



*Figure 9: CDMAP interspecies analysis. During a one-to-many interspecies analysis, CDMAP sequentially takes each organism and compares it to every other organism within the repository of analyzed organisms using Pearson's correlation coefficient. This analysis can be done carried out using a variety of lenticular parameter such as Ne, Chromosome size, and GC content.*

After establishing a solid framework to analyze the role of genome-wide patterns and the effect on site specific context mutation rates, we wanted to develop a framework to analyze and compare genome-wide patterns across a large breadth of bacterial organisms in a reproducible manner. Understanding whether these patterns are local to a given organism due to unique mechanistic properties they exhibit in replication and repair machinery, or the observed pattern are clade-specific, kingdom-specific, or even are a fundamental rule driving context specific mutation rates across all organisms. With this philosophical approach in mind, we wanted to build a similarly modular, flexible framework to allow the end user to easily carry out a one-to-many

analyses for a collection of bacterial organisms (and later on other prokaryotes and eventually eukaryotes). The interspecies analysis framework is built with a similar philosophy that we adopted when developing the Single organism pipeline, where a user may analyze relationships under varying parameters among a collection of organisms.

## Methods

### *Required Input*

CDMAP was developed in later versions with the flexibility of interspecies analysis in mind, so prior to the user running any steps of the analysis, CDMAP dynamically creates a repository of requisite information needed downstream for this portion of analysis and conveniently stores it for later use. By default, genome-wide and replichore-specific output datasets of every organism ran by the end user are stored for interspecies analysis. When the user runs the multi-organism analysis pipeline, CDMAP makes compares organisms on a chromosome-wide, replichore specific basis, including analysis based on leading or lagging strand orientation and includes analysis of upstream and downstream nucleotide effects. Upon initialization of the multi-organism analysis portion of the pipeline, CDMAP dynamically generates the genomic GC content % for each organism from the provided reference FASTA files piped to the repository from the single organism analysis pipeline.

## *Data Generation Methodology and Sample Output*

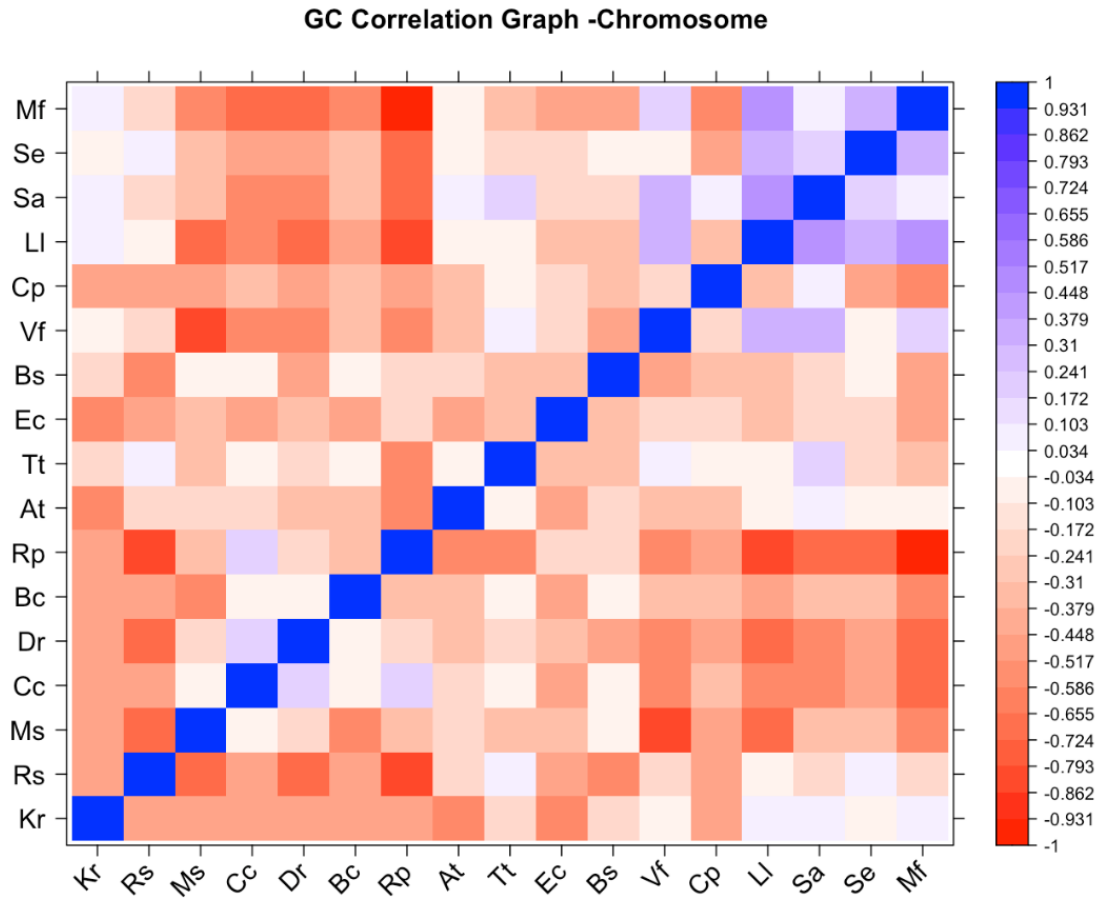



*Figure 10 - CDMAP One-To-Many Correlation Analysis. For each organism in the single organism output repository, CDMAP conducts a one-to-many analysis generating a Pearson's product moment correlation value with respect to all possible combination of organisms. Next, each comparison is oriented with respect to genomic GC content, from the most GC rich (bottom left) to AT rich (top right). In this figure, an example of a one-to-many analysis with respect to the chromosome has been conducted.*

After the context-dependent mutation patterns for each organism have been dynamically loaded into memory, CDMAP performs a one-to-many correlation analyses sequentially with each organism using Pearson's correlation coefficient. Genome-wide and replichore-specific comparisons are performed. Once correlation patterns are generated across all possible combination of organisms, CDMAP then orients the relationship coefficients of each organism

according to GC content and then records the genome-wide, replichore-specific, upstream and downstream interspecies correlations, p-values, and test statistics in both high quality visualizations using lattice, and as text output for external post processing or visualization if the user so desires using another visualization software of their choosing.

As Shown in figure 10, we have taken a subset of organisms that have had the single organism analysis conducted using CDMAP and ran an all against all correlation of each organism's primary chromosome with respect to GC content. In this visualization, columns are ranked from AT rich (top) to GC- rich (bottom) and rows are oriented GC rich (left) to AT-rich (right). The 1:1 organism comparison is color coordinated relative to its Pearson's coefficient, as indicated by the heatmap legend. In figure 10, the upper rightmost portion of the graph shows the 1:1 comparisons of all organisms harboring the most AT rich chromosome, which displays an increase in correlation among site-specific

| Organism Name | Abbreviation | 2letter code |
|---|---|---|
| Agrobacterium tumefaciens C58 | A. tumefaciens | At |
| Bacillus subtilis 3610 | B. subtilis | Bs |
| Burkholderia cenocepacia HI2424 | B. cenocepacia | Bc |
| Caulobacter Crescentus NA1000 | C. Crescentus | Cc |
| Colweillia psychrerythraea 34H | C. psychrerythraea | Cp |
| Deinococcus radiodurans R1 | D. radiodurans | Dr |
| Escherichia coli K12 MG1655 | E. coli | Ec |
| Kineococcus radiotolerans SRS30216 | K. radiotolerans | Kr |
| Lactococcus lactis DSMZ20481 | L. lactis | Ll |
| Mesoplasma florum L1 | M. florum | Mf |
| Mycobacterium Smegmatis MC2 155 | M. Smegmatis | Ms |
| Rhodobacter sphaeroides ATCC 17025 | R. sphaeroides | Rs |
| Rugeria Pomeryoi DSS3 | R. Pomeryoi | Rp |
| Staphyloccus aureus ATCC 25923 | S. aureus | Sa |
| Staphyloccus epidermidis ATCC 12228 | S. epidermidis | Se |
| Teredinibacter turnerae T7901 | T. turnerae | Tt |
| Vibrio fischeri ES114 | V. fischeri | Vf |

*Table 3- Name, Abbreviation and 2 letter codes of bacterial organisms analyzed with CDMAP*

mutation rates in comparison to the correlation coefficients of GC rich organism site-specific mutation rates.

# Results

## Discussion Order of Results

Given the breadth and depth of analysis conducted using CDMAP's multi organism analysis pipeline, discussion of results will be structured in the following manner: First we will be looking at the one-to-many analysis of organisms with respect to their genomic NXN triplets, with respect to their primary chromosome and each replichore, where N represents neighbor nucleotides held constant, while X represents a nucleotide N mutating to another nucleotide. Next, we will discuss the analysis of genomic triplets NXN, and the influence of context mutation rate patterns given an upstream nucleotide Y, i.e. YNXN with respect each chromosome and replichore. Finally, in a similar manner we will discuss analysis of NXN genomic triplets, and the influence of context mutation rate patterns given a downstream nucleotide Y, i.e., NXNY.

**Across a spectrum of prokaryotic organisms, do we see patterns in genomic triplets NXN corresponding to GC content % across organisms?**

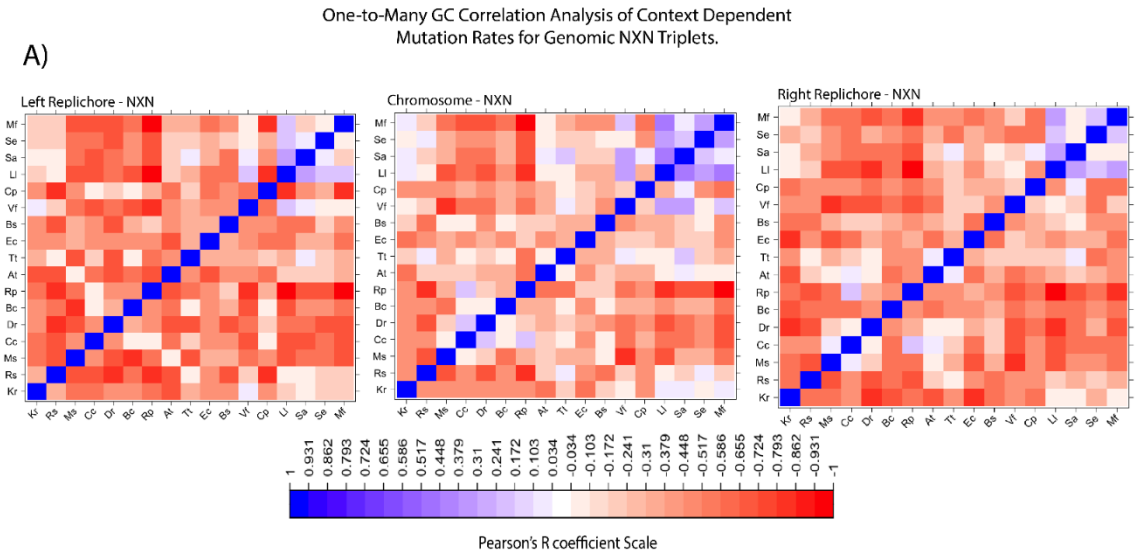



*Figure 11 - One-to-Many Analysis of Bacterial organisms. Context dependent mutation rates for each organism were correlated using Pearson's product moment correlation with each cell representing an individual correlation. Each matrix from left to right represents: Left Replichore NXN triplet Correlation, Chromosome NXN correlation, Right Replichore NXN Correlation.*

The first area of investigation we wanted to explore was to determine if patterns existed baseline with genomic NXN, where X is a point mutation occurring in a nucleotide triplet on a per-chromosome, per replichore basis. We first examined whether correlations of context mutation rate patterns would be influenced being on the forward 5' to 3' versus being on the reverse 3' to 5' strand of replication. We found no evidence of asymmetry on a chromosome wide or replichore specific basis in context mutation rate pattern correlations. Additionally, there were no discernable differences in patterns between the left and right replichore on the forward 5' to 3' or reverse 3' to 5' strand of replication for genomic NXN triplets. Correlations only appear to be diluted upon examination of NXN replichore. However, it should be noted that L. lactis maintains the strongest positively correlated mutation rates regardless of strand orientation or replichore.

Next, we investigated the relationships among the most GC rich organisms and whether organisms exhibited any correlation in context mutation rate patterns. Generally speaking, organisms oriented on the GC rich end of the dataset exhibited no correlation to being moderately negative correlated context mutation rate patterns with one exception. In the case of C. crescentus, we found moderately strong correlations with both D. radiodurans ($r = .4651$, $p = 0.000107$) and R. pomeryoi ($r = .4651$, $p = 0.000129$). Upon examination of C. crescentus with respect to each organism, it appears the primary N[A]N nucleotide is the primary driver of mutational correlation between each organism. In the relationship C. crescentus to R. pomeryoi, the strongest relationships are driven particularly by T[A]T, T[A]G, and T[A]C context mutations, while the relationship between C. crescentus and D. radiodurans appears to be driven by A[A]C, C[A]C, G[A]C, and G[T]C context mutations. It is interesting to note that in both instances of mutational correlation, that we see unique anchoring nucleotides (a left neighbor nucleotide T in the former, and right C neighbor nucleotide in the latter) appearing to be the singular relationship between all highly mutable sites in each contextual dinucleotide scenario.

After examination of GC rich organisms, we turned our attention to AT rich organisms, as the most AT rich organisms presented a pronounced set of correlations in context mutation rate patterns amongst each other. Each positively correlated AT rich organism correlation highlighted in blue from figure 11 exhibited a moderately to nearly strongly correlated Pearson's r value, and was statistically significant as shown in table 5**.**

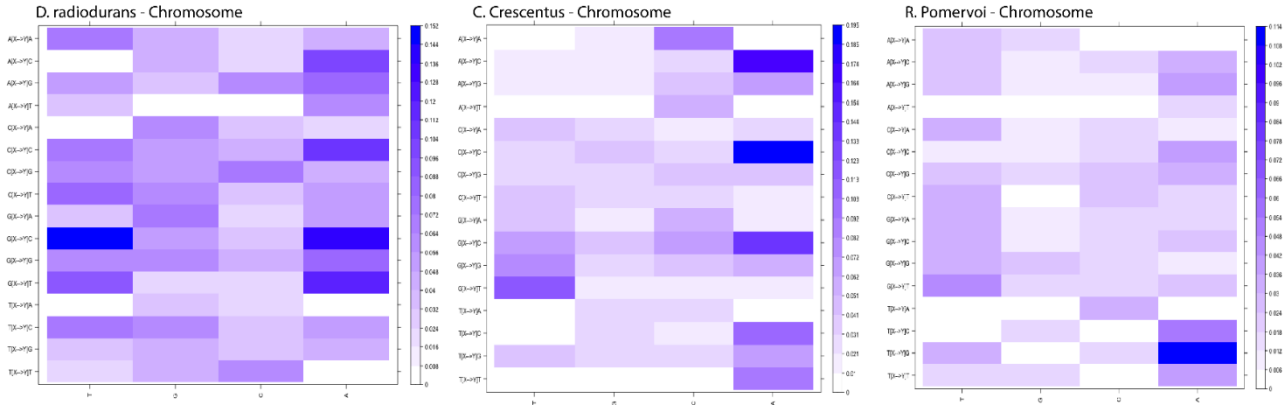



*Figure 12- Chromosome Context Dependent Mutation Rates in GC rich Organisms: site specific mutation rates for C. crescentus, D. radiodurans, and R.Pomeryoi.*

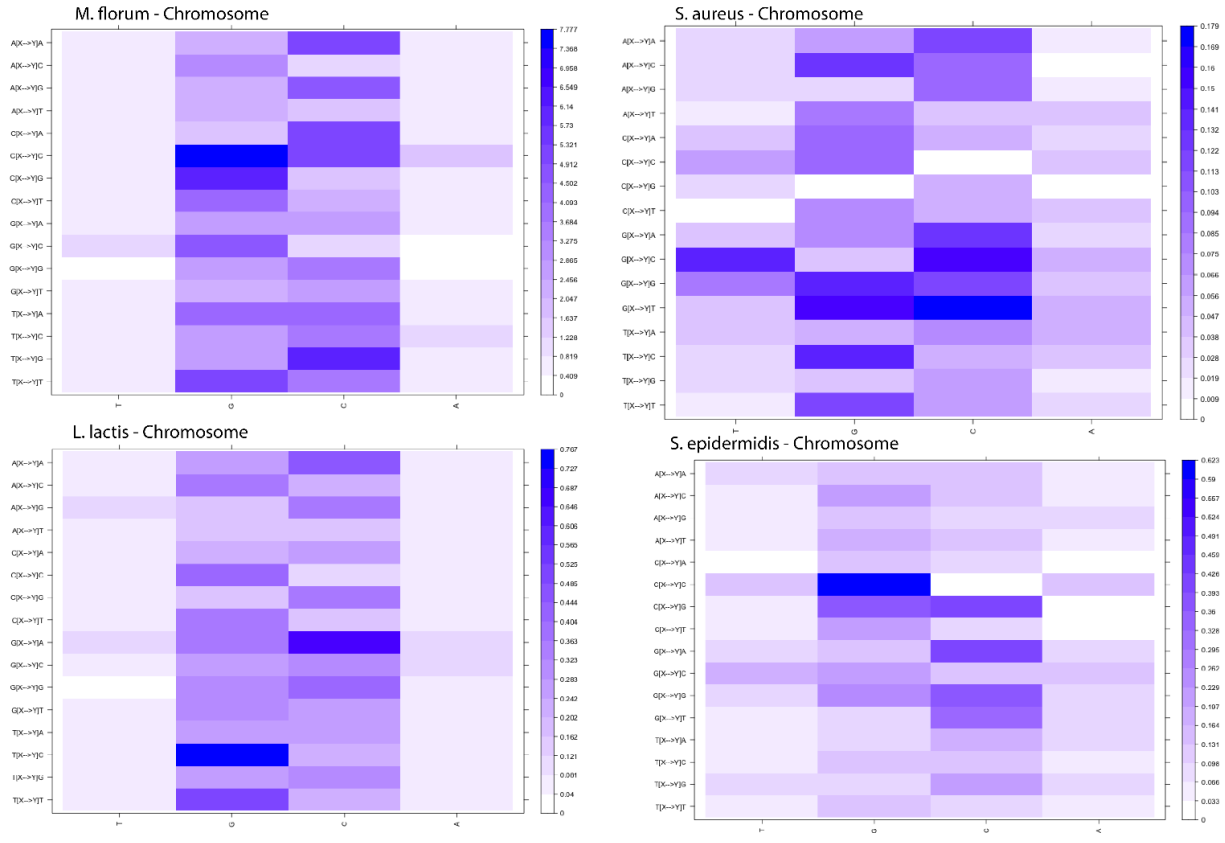



*Figure 13 - Chromosome Site Specific Context Dependent mutation Rates of the Most AT rich Organisms within the CDMAP one- to many analysis. Mutation rates were scaled to 1E-8*

**AT-Rich Correlations**

| Organism 1 | Organism 2 | Pearson's R | P-Value |
|---|---|---|---|
| L. lactis | S. aureus | 0.688433765 | 3.24E-10 |
| L. lactis | M. florum | 0.665113872 | 2.04E-09 |
| L. lactis | S. epidermidis | 0.588309108 | 3.18E-07 |
| M. florum | S. epidermidis | 0.56726887 | 1.02E-06 |
| S. Aureus | S. epidermidis | 0.453536056 | 0.0001674 |
| V. fischeri | S. aureus | 0.573151491 | 7.43E-07 |
| V. fischeri | M. florum | 0.44627313 | 0.0002189 |

*Table 4 – Pearson's correlation coefficients and p-values for each positively correlated AT rich organism (chromosome) from Figure 11 above.*

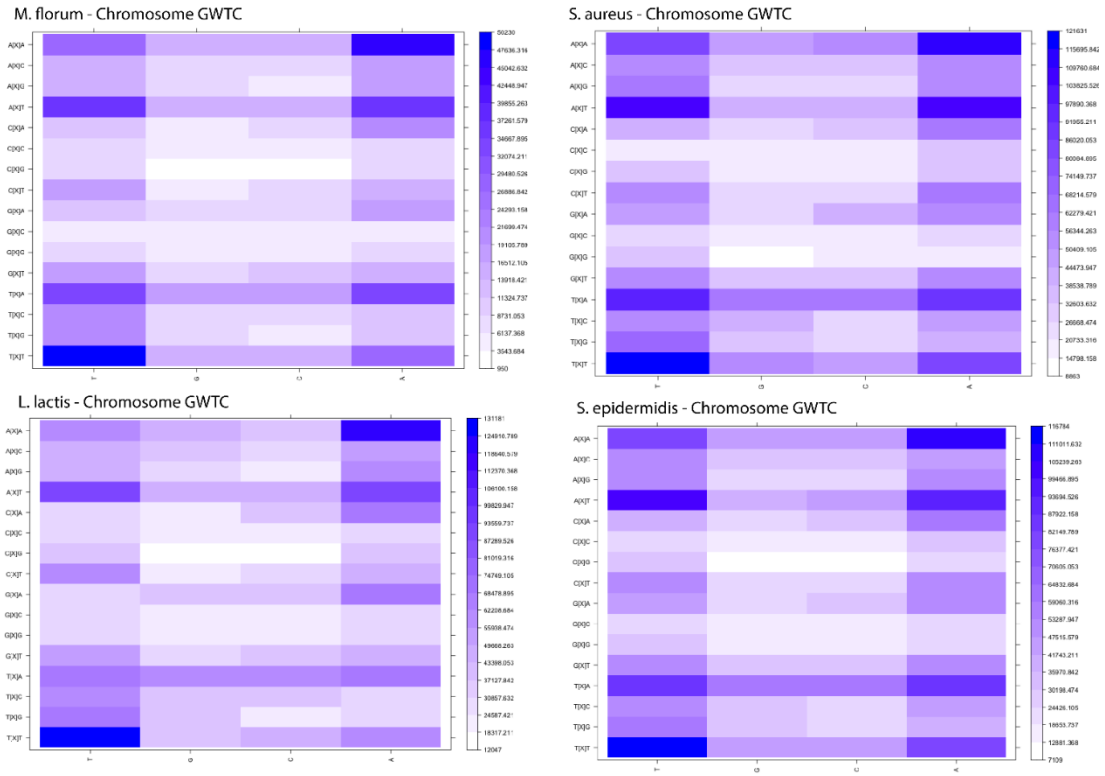
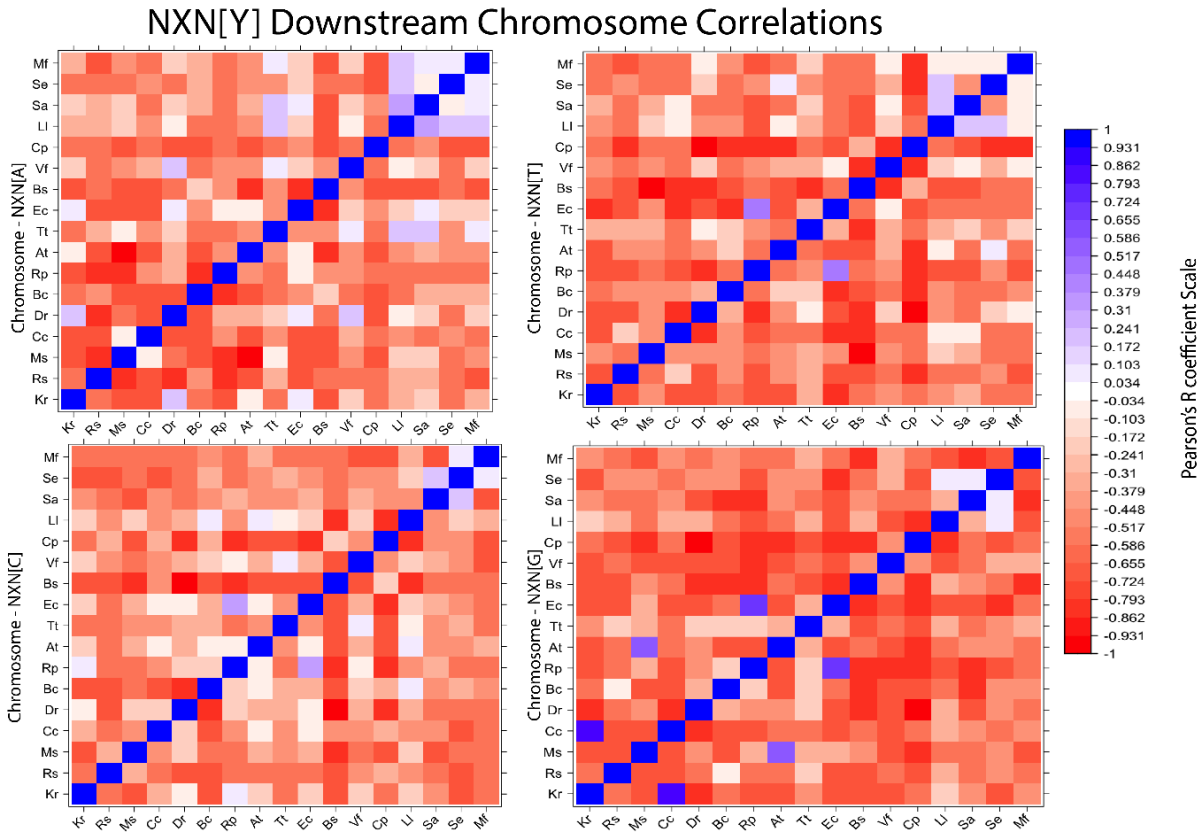



*Figure 14 - Genome Wide Triplet counts for all 64 NXN genomic triplets within the chromosome.*

Immediately upon inspection of the context dependent mutation rates of AT rich organisms in figure 13, it was evident the lion's share of mutations occurred with respect to N[G]N and N[C]N, which was not wholly unexpected given the genomic AT content of these organisms. Interestingly enough, N[G]N and N[C]N context dependent mutation rates, and N[T]N and N[A]N appeared visually inverse from one another. The second majorly pronounced visual comparison was the relative activity of which may serve as the driver of correlation among each organism's context mutation rates were the 4-fold amino acids Threonine (Thr) and Alanine (Ala), and the 6-fold amino acid of Arginine (Arg).

For the 6-fold amino acid Arg, we found Arg genomic triplets to be highly mutable 3 of the 4 most AT-rich organisms as shown in figure 14. C[G]N sites, in particular C[G]C and C[G]G were found exhibiting the hottest areas of mutation for Arg while A[G]A maintained the lowest mutation rate for Arg sites. Meanwhile for 4-fold amino acids, Ala genomic triplets stood out the most among 4-fold amino acids, particularly for S. aureus. Further examinations showed G[C]N mutation rates were active amongst all AT rich organisms. In the instance of Thr genomic triplets we found A[C]A and A[C]G triplets to be the most mutationally active triplets, while A[C]T appeared to have the least mutable Thr Triplet.

**Across a spectrum of prokaryotic organisms, do we see downstream patterns in genomic triplets NXNY corresponding to GC content % across organisms?**



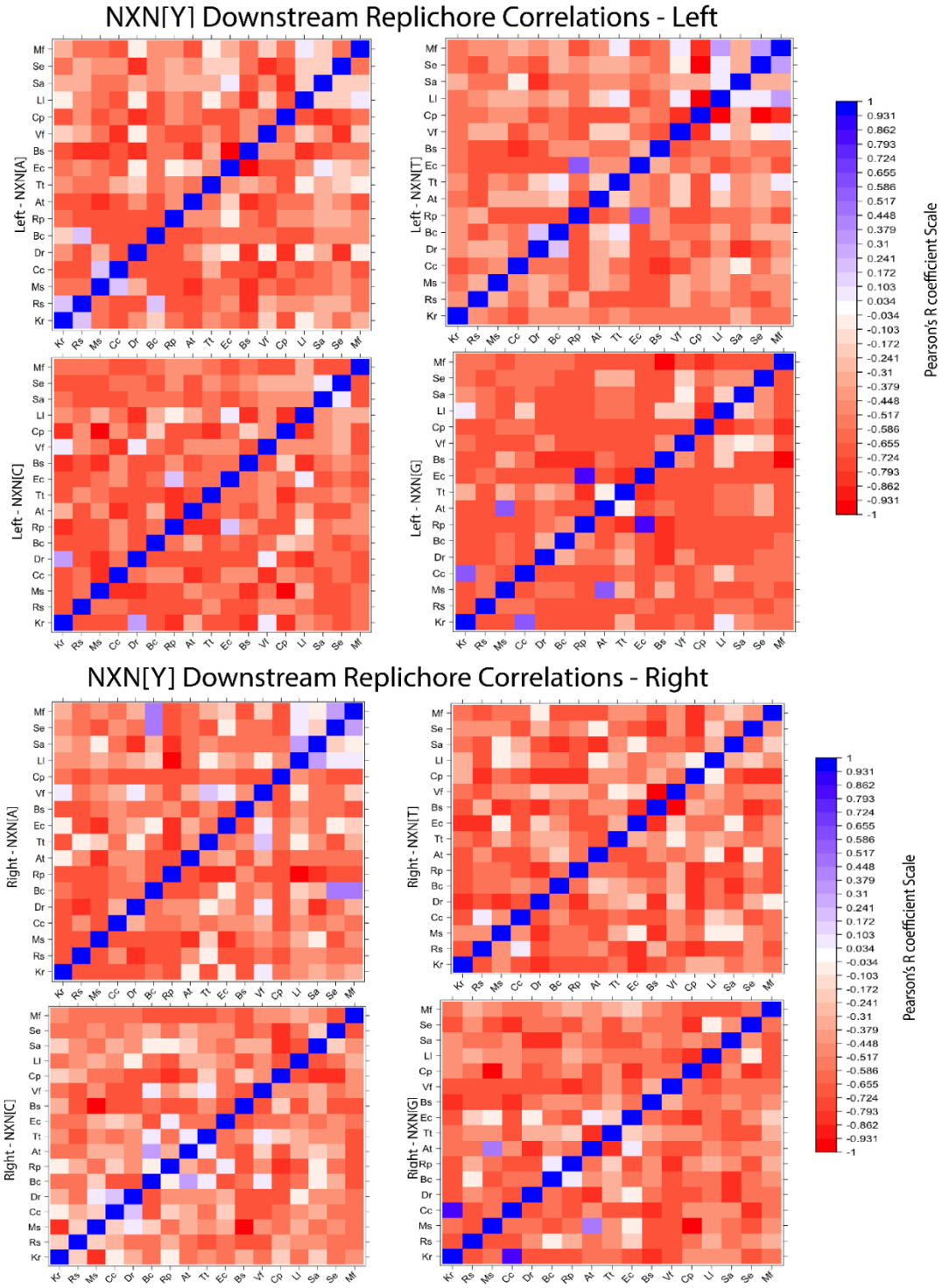



*Figure 15 - NXN[Y] CDMAP one-to-many downstream analysis of genomic triplets with respect to the chromosome and each replichore. Each Matrix represents the chromosomal or replichore-specific analysis of context dependent mutation rates of genomic NXN triplets with respect to a downstream nucleotide Y (indicated by the label on the matrix). Each cell represents the Pearson's product moment correlation between the 64 context dependent rates of two Bacterial organisms similar to the genomic NXN correlation in Figure N.*

When examining the downstream NXNY correlations of bacterial organisms, where Y represents a downstream nucleotide, followed a similar process to analysis in prior sections in terms of identifying relationships between organisms. In terms of potential asymmetry between the 5' to 3' forward strand and 3' to 5' reverse strand of replication, no asymmetry in correlations were found with respect to the chromosome or either replichore. Overall analysis of NXNY genomic triplet downstream overall yielded very sparse positively correlated context dependent mutation rates, and quite often yielded moderately negative correlations for most organisms when compared against the spectrum of other bacterial organisms with a few notable exceptions on the GC rich end of the spectrum.

Upon inspection of downstream NXN[G] correlations in figure 15, we see three distinctly positive correlations in context dependent mutation rates among a sea of negatively correlated mutation rates: E. coli and R. pomeryoi, K. radiotolerans and C. crescentus and A. tumefaciens

**NXN[G] Positive Correlations**

| Organism 1 | Organism 2 | Chr Pearson's R | Chr P-Value | Left Pearson's R | Left P-Value | Right Pearson's R | Right P-Value |
|---|---|---|---|---|---|---|---|
| E. coli | R. pomeryoi | 0.904557833 | 1.23E-24 | 0.93049213 | 7.59E-06 | -0.073659862 | 0.562959407 |
| C. crescentus | K. radiotolerans | 0.933998393 | 2.08E-29 | 0.824786712 | 5.40E-17 | 0.938976669 | 1.97E-30 |
| A. tumefaciens | M. smegmatis | 0.811953192 | 3.94E-16 | 0.819550124 | 1.24E-16 | 0.720604551 | 1.90E-11 |

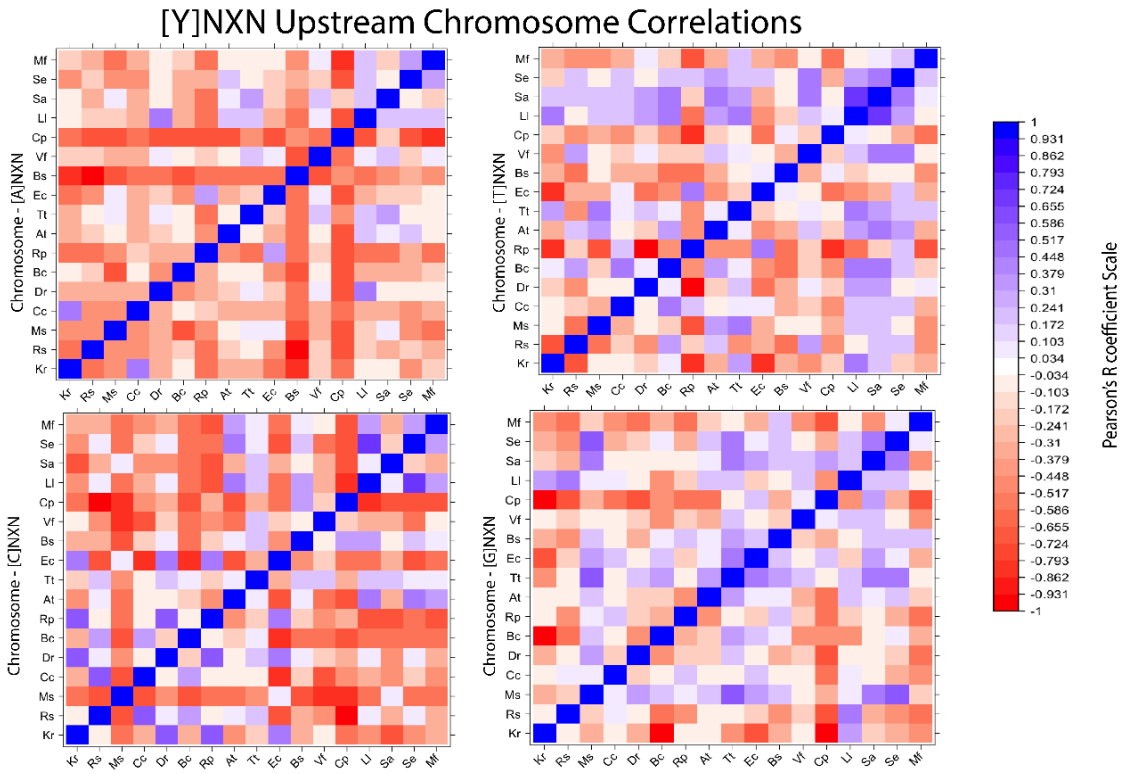
*Table 6- Table of NXNY Strongly correlated GC rich organisms. For each organism, the replichore specific and chromosome Pearson's product moment correlation and p-value are listed, showcasing the contribution of each replichores context mutation rate patterns contribute to the corresponding chromosomal relationship.*

with M. smegmatis. For each relationship denoted on the downstream chromosome, we can easily track the strength of the correlation back to it's given replichore in the following table:

In table 6 we see the breakdown of each relationship with respect to the chromosome, and the individual replichore. We observe a unique scenario where each possibility driving context mutation rate patterns occurs: E. coli and R. pomeryoi is entirely driven by relationships occurring in the left replichore, A. tumefaciens and M. smegmatis's contextual patterns being driven more so by the left replichore (though there is still a strong relationship in the right replichore) and K. radiotolerans and C. crescentus's relationship strength nearly mirrors itself between replichores.

The other peculiar behavior occurred in the most AT rich organisms, appearing to be sensitive to the downstream nucleotide Y that succeeds the genomic NXN triplet. For example, in the downstream chromosome NXN[A], L. lactis appears to preserve some degree of mutational relationship with other AT rich organisms, meanwhile if we change Y to NXN[T] not only does L. lactis's relationships invert to a negative correlation, we also see the most AT rich organism, M. florum, ceases to share any relationship with the other most AT rich organisms. Despite the apparent sensitivity NXNY context mutation rate patterns, we are still able to discern the root replichore where those relationships originated, such as the correlation between S. aureus and L. lactis in NXN[A] appears to originate from the right replichore. Meanwhile when examining individual replichores, we see several correlations occurring with respect to specific nucleotides Y (NXN[A] – Right and NXN[T] - left) that appear to show correlations in context mutation rates, that appear to vanish when scaling up to their respective NXNY chromosome correlation. This negation may be explained by the strong dissimilarity in the opposing replichore that gives the appearance at the chromosome level that a given relationship exists.

**Across a spectrum of prokaryotic organisms, do we see upstream patterns in genomic triplets YNXN corresponding to GC content % across organisms?**



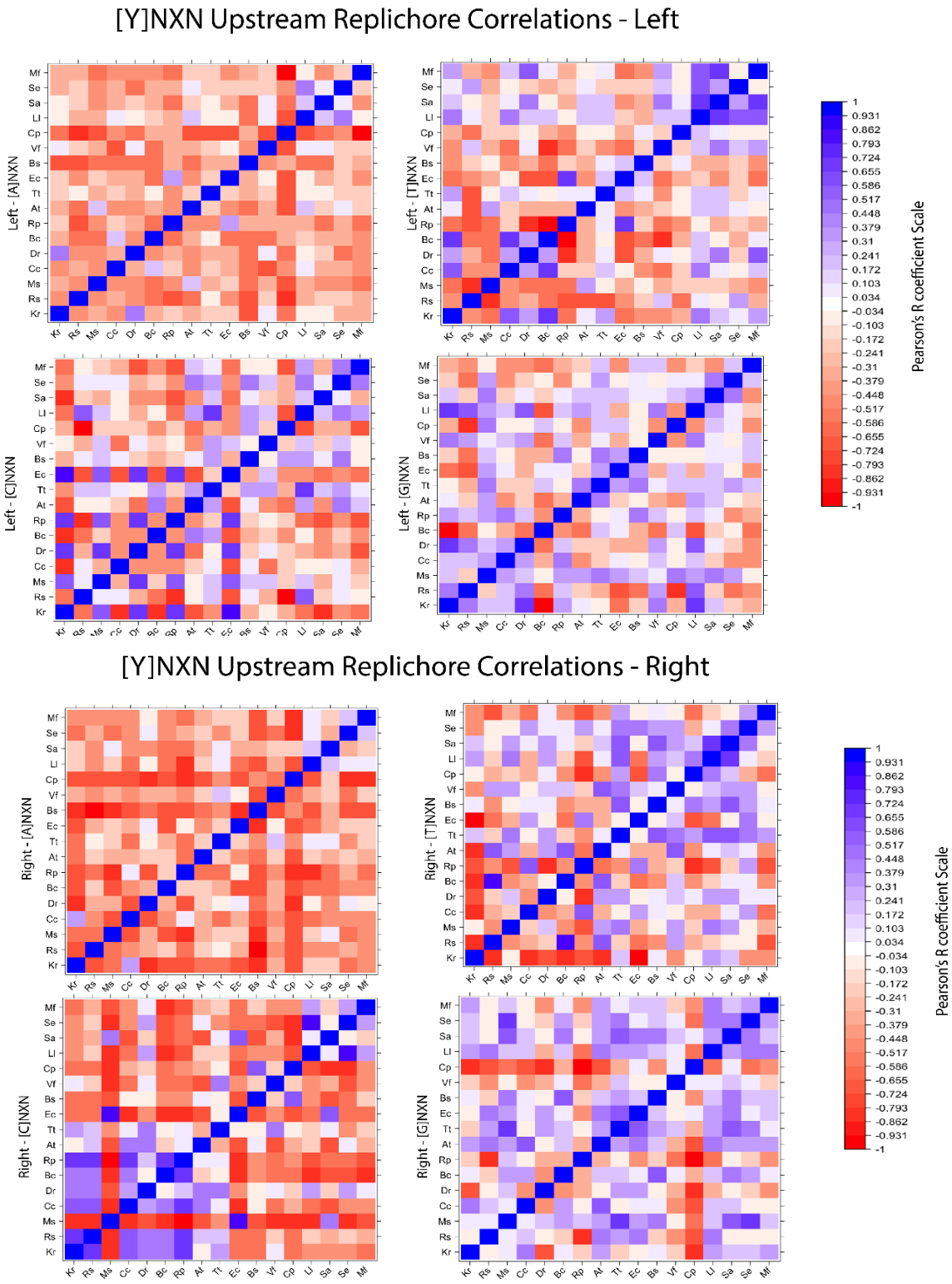



*Figure 16 – [Y]NXN CDMAP one-to-many upstream analysis of genomic triplets with respect to the chromosome and each replichore. Each matrix represents the chromosomal or replichore-specific analysis of context dependent mutation rates of genomic NXN triplets with respect to an upstream nucleotide Y (indicated by the label on the matrix). Each cell represents the Pearson's product moment correlation between the 64 context dependent rates of two Bacterial organisms similarly to the genomic NXN correlation in Figure 11.*

Our final area of examination was to determine given an upstream nucleotide composition YNXN where Y is upstream nucleotide of a genomic triplet NXN. We conducted our analysis in a similar manner to prior sections in terms of how we approached chromosomes and replichores. Similarly, both the analysis of NXN and NXNY downstream correlations, we found no discernable difference in relationships of context dependent mutation rates between the 5' to 3' forward and the 3' to 5' reverse strands of replication both with respect to the chromosome and their corresponding replichores. However, as seen in figure 16 there are distinctly pronounced effects on correlations with respect to given chromosome and its corresponding replichores.

To begin our dissection of the multitude of fluctuations of influence on YNXN genomic triplets, we decided to observe how Y influenced NXN genomic triplets on a chromosome wide level. When observing the [A]NXN triplets, for the most part with a few small, noteworthy relationships, we largely see patterns that align with base NXN genomic triplets. In terms of uniform patterns observed, we see that both C. psychrerythraea and B. subtilis exhibit moderate to strong negative correlations with most organisms. In terms of observed positive correlations of context dependent mutation rates we observe several interesting moderate to strong relationships as listed in table 7 below:

**[A]NXN GC- Correlations**

| Organism 1 | Organism 2 | Pearson's R | P-Value |
|---|---|---|---|
| **A. Tumefaciens** | **L. lactis** | 0.532215602 | 6.00E-06 |
| **A. Tumefaciens** | **M. florum** | 0.514329703 | 0.0086554 |
| **A. Tumefaciens** | **S. aureus** | 0.41436633 | 0.0006641 |
| **C. crescentus** | **K. radiotolerans** | 0.725943762 | 1.14E-11 |
| **D. Radiodurans** | **L. lactis** | 0.658014886 | 3.46E-09 |
| **D. Radiodurans** | **V. fischeri** | 0.437153993 | 0.000304 |
| **E. coli** | **R. pomeryoi** | 0.594550373 | 2.22E-07 |
| **T. turnerae** | **S. aureus** | 0.580021383 | 5.09E-07 |
| **T. turnerae** | **L. lactis** | 0.563767419 | 1.23E-06 |
| **T. turnerae** | **D. Radiodurans** | 0.444390678 | 0.0002345 |
| **V. fischeri** | **S. aureus** | 0.564494938 | 7.94E-06 |
| **V. fischeri** | **T. turnerae** | 0.476290152 | 6.94E-05 |

*Table 6 - Table of Pearson's product moment correlations and p-value for [A]NXN correlations of context dependent mutation rates between organisms.*

The primary observations that seem to be prevalent is there appears to be a polarization effect occurring with the relationships of more "GC neutral" (meaning they are neither biased towards the middle, rather than the AT or GC rich end of the spectrum) begin to correlate with more AT rich organisms, such as A. tumefaciens and T. turnerae. In the case of A. tumefaciens showed a pronounced uptick in positively correlated context mutation rates between multiple AT rich organisms. While T. turnerae appeared to show increased patterns in context mutation rates towards both S. aureus and L. lactis, T. turnerae also showed an increase in context mutation rate patterns with D. radiodurans, which is more GC rich organism. In terms of shifts in context mutation rate patterns in NXN[A] GC rich organisms, we observed some rather interesting patterns of note. The first of which was a rather strong positive relationship in correlation between K. radiotolerans and C. crescentus ($r = .72$, $p = 1.14E-11$). Additionally, despite D. radiodurans being a more GC rich organism within the dataset, we observed increases in context mutation rate patterns with both L. lactis and V. fischeri, both of which are more oriented on the AT rich end of the organisms in the dataset.

Next, we observed the influence of how an upstream C nucleotide influences context dependent mutation rates among bacterial organisms. The immediate takeaway is a moderate to strong polarization of patterns of context dependent mutation rates with respect to genomic GC content towards the most AT and GC rich organisms on both a chromosome wide and replichore specific level. In the table 8 below outlines sites exhibiting moderate to major correlations in context dependent mutation rate patterns:

| [C]NXN Correlations - Organism 1 | Chromosome Organism 2 | AT/GC/Neutral/Outlier | Pearson's R | P-Value |
|---|---|---|---|---|
| A. Tumefaciens | S. epidermidis | AT | 0.779065679 | 3.44E-14 |
| A. Tumefaciens | L. lactis | AT | 0.757543934 | 4.32E-13 |
| A. Tumefaciens | M. florum | AT | 0.644261427 | 9.25E-09 |
| B. cenocepacia | R. sphaeroides | GC | 0.692866961 | 0.012968576 |
| B. cenocepacia | C. crescentus | GC | 0.680125959 | 0.023297534 |
| B. Subtilis | L. lactis | AT | 0.652491734 | 5.16E-09 |
| B. Subtilis | C. psychrerythraea | AT | 0.649026876 | 6.61E-09 |
| C. crescentus | R. sphaeroides | GC | 0.7966761 | 3.48E-15 |
| D. radiodurans | K. radiotolerans | GC | 0.840575486 | 3.71E-18 |
| E. coli | K. radiotolerans | GC | 0.764332054 | 2.00E-13 |
| E. coli | D. radiodurans | GC | 0.715360029 | 3.10E-11 |
| E. coli | R. pomeryoi | Neutral | 0.730406045 | 7.40E-12 |
| L. lactis | S. epidermidis | AT | 0.919961737 | 6.66E-27 |
| L. lactis | M. florum | AT | 0.674291464 | 1.01E-09 |
| R. pomeryoi | K. radiotolerans | GC | 0.833370515 | 1.30E-17 |
| R. pomeryoi | D. radiodurans | GC | 0.803273758 | 1.39E-15 |
| S. epidermidis | M. florum | AT | 0.652155216 | 5.29E-09 |
| T. turnerae | L. lactis | AT | 0.641824553 | 1.10E-08 |
| T. turnerae | V. fischeri | AT | 0.596759352 | 1.95E-07 |
| T. turnerae | S. aureus | AT | 0.577533575 | 5.84E-07 |
| T. turnerae | D. radiodurans | GC | 0.618045278 | 5.29E-08 |
| T. turnerae | R. sphaeroides | GC | 0.58207357 | 4.54E-07 |
| T. turnerae | B. subtilis | Neutral | 0.572509326 | 7.70E-07 |

*Table 7 - Table of Pearson's product moment correlations and p-value for [C]NXN correlations of context dependent mutation rates between organisms. Organisms are categorized with respect to polarizing to the AT rich, GC rich, Neutral (meaning it correlated neither with an AT rich or GC rich oriented organism), or Outlier (meaning an AT rich organism correlated with a GC rich organism).*

When examining the one-to-many analysis of [C]NXN correlations, it appears that the dataset polarized down almost directly down the middle, and either favored in either correlating towards more AT rich organisms, or more GC rich organisms. In table 8 we can several patterns among context dependent mutation rate relationship polarization. In the case of AT rich Organisms when a C is aligned upstream of a given genomic Triplet, A. tumefaciens and B. Subtilis context mutation patterns have moderate to strongly correlated relationships with AT rich organisms. Though L. Lactis is an AT rich organism, when C is upstream, its context mutation patterns exhibit

moderate to strongly correlated relationships with other AT rich organisms. In the case of GC rich organisms, when a C is aligned upstream of a genomic triplet, Both B. cenocepacia and E. coli's mutation rate patterns appear to gravitate towards more GC rich organisms. Meanwhile, GC rich organisms such as R. pomeryoi, C. crescentus and D. Radiodurans context mutation patterns demonstrate an increasingly correlated relationship with other GC rich organisms.

Upon further analysis of [C]NXN with respect to each individual replichore, we found even more profound intensifications of the previously mentioned polarization effect. When examining GC rich organisms, we see apparent and pronounced intensifications of relationships for the GC rich end of the spectrum in both replichores. In the Right replichore, with the sole exception of M. smegmatis, exhibit strongly correlated patterns in all of the most GC rich organisms, and this intensification can even be seen in more GC neutral organisms such as A. tumefaciens, and T. turnerae. Meanwhile in the Left replichore, there is an apparent, albeit less uniform intensification of patterns in context mutation rates in GC rich organisms. It appears the polarization effect of E. coli to more GC rich organisms appears to originate exclusively from the left replichore given its lack of correlation to other GC rich organisms in the right replichore. It is interesting to note that while E. coli exhibits a strong positive correlation with M. smegmatis in the right replichore, and a moderately positive correlation in the left replichore, any semblance of a positive correlation vanishes at the chromosome level.

On the other end of the Spectrum with AT rich organisms, we found that correlations among the most AT rich organisms followed similar patterns, albeit with minor differences in L. lactis's relationships. upon examination of each given replichore and chromosome, [C]NXN mutation rates of L. lactis and S. epidermidis are strongly correlated with respect to both the chromosome and individual replichores. Meanwhile with respect to L. lactis and S. aureus, we see

a dilution effect, where in the left replichore we observe a weak to moderate correlation, but when paired with the right replichore (which exhibits no correlation between them) we see any relationship between said organisms on a chromosome level.

| [G]NXN Chromosome | | | | |
|---|---|---|---|---|
| **Organism 1** | **Organism 2** | **AT/GC/Neutral/Outlier** | **Pearson's R** | **P-Value** |
| **A. tumefaciens** | **S. epidermidis** | **AT** | **0.614891967** | **6.46E-08** |
| **A. tumefaciens** | **L. lactis** | **AT** | **0.660624581** | **2.85E-09** |
| **A. tumefaciens** | **M. smegmatis** | **GC** | **0.637223046** | **1.50E-08** |
| **A. tumefaciens** | **T. turnerae** | **Neutral** | **0.700255204** | **1.19E-10** |
| **B. cenocepacia** | **M. smegmatis** | **GC** | **0.61559293** | **6.18E-08** |
| **B. cenocepacia** | **E. coli** | **Neutral** | **0.628944582** | **1.98E-07** |
| **B. cenocepacia** | **T. turnerae** | **Neutral** | **0.705606837** | **7.48E-11** |
| **B. subtilis** | **M. smegmatis** | **GC** | **0.653535587** | **4.79E-09** |
| **C. psychrerythraea** | **S. aureus** | **AT** | **0.709577112** | **5.25E-11** |
| **D. radiodurans** | **M. smegmatis** | **GC** | **0.684374083** | **4.52E-10** |
| **E. coli** | **C. psychrerythraea** | **AT** | **0.645139557** | **8.70E-09** |
| **E. coli** | **S. aureus** | **AT** | **0.723137027** | **1.50E-11** |
| **E. coli** | **C. crescentus** | **GC** | **0.613479774** | **1.27E-09** |
| **E. coli** | **M. smegmatis** | **GC** | **0.676568197** | **4.73E-10** |
| **E. coli** | **B. subtilis** | **Neutral** | **0.647399307** | **7.42E-09** |
| **K. radiotolerans** | **L. lactis** | **Outlier** | **0.723901113** | **1.39E-11** |
| **L. lactis** | **S. aureus** | **AT** | **0.617924758** | **5.33E-08** |
| **L. lactis** | **S. epidermidis** | **AT** | **0.62310525** | **3.83E-08** |
| **R. pomeryoi** | **C. crescentus** | **GC** | **0.625546599** | **3.27E-08** |
| **R. pomeryoi** | **B. subtilis** | **Neutral** | **0.618021106** | **5.30E-08** |
| **R. sphaeroides** | **L. lactis** | **Outlier** | **0.796109118** | **3.76E-15** |
| **S. aureus** | **S. epidermidis** | **AT** | **0.75079876** | **9.05E-13** |
| **S. aureus** | **M. smegmatis** | **Outlier** | **0.779366923** | **3.31E-14** |
| **S. epidermidis** | **M. smegmatis** | **Outlier** | **0.809587866** | **5.59E-16** |
| **T. turnerae** | **V. fischeri** | **AT** | **0.613144792** | **7.21E-08** |
| **T. turnerae** | **L. lactis** | **AT** | **0.632602701** | **2.05E-08** |
| **T. turnerae** | **S. epidermidis** | **AT** | **0.760604533** | **3.06E-13** |
| **T. turnerae** | **S. aureus** | **AT** | **0.789052715** | **9.63E-15** |
| **T. turnerae** | **D. radiodurans** | **GC** | **0.612920411** | **7.31E-08** |
| **T. turnerae** | **M. smegmatis** | **GC** | **0.844754493** | **1.74E-18** |
| **T. turnerae** | **B. subtilis** | **Neutral** | **0.724117512** | **1.36E-11** |
| **T. turnerae** | **E. coli** | **Neutral** | **0.798571453** | **1.13E-06** |
| **V. fischeri** | **S. aureus** | **AT** | **0.64573657** | **8.34E-09** |
| **V. fischeri** | **L. lactis** | **AT** | **0.668662124** | **1.56E-09** |

*Table 8 - Table of Pearson's product moment correlations and p-value for [G]NXN correlations of context dependent mutation rates between organisms. Organisms are categorized with respect to polarizing to the AT rich, GC rich, Neutral (meaning it correlated neither with an AT rich or GC rich oriented organism), or Outlier (meaning an AT rich organism correlated with a GC rich organism).*

When we shift gears and examine [G]NXN upstream nucleotide triplets, we similarly see a general trend of increased mutational relationships with GC neutral organisms orienting towards either GC rich or AT rich organisms. However, we see the introduction of outliers on both the GC and AT rich end of the spectrum in the form of L. lactis and M. smegmatis both generating strongly correlated statistically significant relationships with multiple organisms opposite of their position on the genomic GC content spectrum (as shown in table 9).

On the chromosome level we can see patterns in context dependent mutation rates aligning with specific AT and GC rich organisms. The majority of strongly correlated, significant, GC rich [G]NXN context mutation rate patterns seemed to anchor themselves to M. smegmatis, while AT rich [G]NXN context patterns seem to be distributed between L. lactis, S. aureus, and S. epidermidis. When looking at how GC neutral organisms oriented themselves with respect to GC and AT rich organisms, A. Tumefaciens and T. turnerae context mutation rate patterns appeared biased towards AT rich organisms, while E. coli appeared to balance itself between both AT and GC rich organisms. However, it should be noted that all 3 GC neutral organisms did exhibit patterns across the genomic GC content spectrum.

After chromosomal analysis, we dissected context dependent mutation rate patterns on a per replichore basis and found a distinct discretization effect existing on each replichores. When we analyze patterns on the right replichore, on a surface level we clearly see strong patterns of context dependent mutation rates across the board irrespective of GC content, however we see several distinct patterns emerge. First off, AT rich organisms clearly, are exhibiting strong positive correlations across the board irrespective of GC content. The second clear relationship we see is that GC neutral organisms are exhibiting moderate to strong correlations with other GC neutral organisms. Finally, and most interestingly, we see GC rich organisms seem to have correlations

amongst other GC rich organisms, but in a much sparser density than AT rich or GC neutral organisms.

When we turn our analysis to the left replichore we appear to observe a nearly full inversion of context mutation rate patterns. Where we previously saw sparsely populated patterns of context mutation rates, we now see patterns reminiscent of [C]NXN GC rich correlations, albeit to a slightly less intense degree. When observing GC neutral organisms, we see for the most part similar patterns of context dependent mutation, albeit slightly less intense than observed in the right replichore. Most interestingly we find [G]NXN context dependent mutation rates of AT right organisms which were strongly correlated in the right replichore experience little to no positively correlated mutation rates in the left replichore. Most interestingly when referenced back to the chromosome level context mutation rate patterns, the only patterns that remain visible are among the GC neutral organisms, while the replichore specific AT rich mutation rate patterns appear to be greatly reduced in strength, while GC rich mutations are diluted to the point of no longer showing any visible signs on a chromosome wide level.

**[T]NXN Correlations**

| Organism 1 | Organism 2 | AT/GC/Outlier | Pearson's R | P-Value |
|---|---|---|---|---|
| A. tumefaciens | S. epidermidis | AT | 0.679523957 | 6.68E-10 |
| A. tumefaciens | S. aureus | AT | 0.729180612 | 8.35E-12 |
| A. tumefaciens | L. lactis | AT | 0.660891466 | 2.80E-09 |
| A. tumefaciens | V. fischeri | AT | 0.589230149 | 3.02E-07 |
| A. tumefaciens | R. sphaeroides | GC | 0.617858707 | 5.36E-08 |
| A. tumefaciens | M. smegmatis | GC | 0.668753093 | 1.55E-09 |
| B. cenocepacia | R. sphaeroides | GC | 0.666277934 | 1.87E-09 |
| B. cenocepacia | A. tumefaciens | GC | 0.666246452 | 1.87E-09 |
| B. cenocepacia | S. epidermidis | Outlier | 0.634635352 | 1.79E-08 |
| B. cenocepacia | S. aureus | Outlier | 0.739866997 | 2.86E-12 |
| B. cenocepacia | L. lactis | Outlier | 0.75919044 | 3.59E-13 |
| C. crescentus | R. pomeryoi | GC | 0.587395287 | 3.36E-07 |
| C. crescentus | B. cenocepacia | GC | 0.752495537 | 7.53E-13 |
| C. crescentus | S. epidermidis | Outlier | 0.602885024 | 1.35E-07 |
| C. crescentus | S. aureus | Outlier | 0.622159933 | 4.07E-08 |
| C. crescentus | L. lactis | Outlier | 0.615346776 | 6.28E-08 |
| D. radiodurans | T. turnerae | GC | 0.610708545 | 8.38E-08 |
| D. radiodurans | S. aureus | Outlier | 0.669038842 | 1.51E-09 |
| D. radiodurans | L. lactis | Outlier | 0.687266727 | 3.57E-10 |
| E. coli | S. epidermidis | AT | 0.592618877 | 7.89E-08 |
| L. lactis | S. epidermidis | AT | 0.68851052 | 3.22E-10 |
| L. lactis | S. aureus | AT | 0.871679248 | 7.21E-21 |
| L. lactis | K. radiotolerans | Outlier | 0.764955745 | 1.86E-13 |
| L. lactis | M. smegmatis | Outlier | 0.621225717 | 4.32E-08 |
| R. pomeryoi | E. coli | GC | 0.735046865 | 1.24E-16 |
| R. pomeryoi | S. epidermidis | Outlier | 0.582086351 | 4.53E-07 |
| S. aureus | K. radiotolerans | Outlier | 0.607750299 | 1.01E-07 |
| S. aureus | R. sphaeroides | Outlier | 0.62239791 | 4.01E-08 |
| S. aureus | M. smegmatis | Outlier | 0.59342886 | 2.37E-07 |
| S. epidermidis | S. aureus | AT | 0.7857174 | 1.48E-14 |
| S. epidermidis | M. florum | AT | 0.611014624 | 8.23E-08 |
| S. epidermidis | R. sphaeroides | Outlier | 0.634229251 | 1.84E-08 |
| T. turnerae | M. florum | AT | 0.599564079 | 1.65E-07 |
| T. turnerae | S. epidermidis | AT | 0.630781196 | 2.31E-08 |
| T. turnerae | S. aureus | AT | 0.707306238 | 6.43E-11 |
| T. turnerae | L. lactis | AT | 0.771998105 | 8.13E-14 |
| T. turnerae | K. radiotolerans | GC | 0.660565957 | 2.86E-09 |
| T. turnerae | M. smegmatis | GC | 0.745283206 | 1.63E-12 |
| V. fischeri | S. epidermidis | AT | 0.72652052 | 1.08E-11 |
| V. fischeri | S. aureus | AT | 0.733892369 | 5.24E-12 |
| V. fischeri | L. lactis | AT | 0.602554427 | 1.38E-07 |
| V. fischeri | R. sphaeroides | Outlier | 0.653175518 | 4.92E-09 |

*Table 9 - Table of Pearson's product moment correlations and p-value for [T]NXN correlations of context dependent mutation rates between organisms. Organisms are categorized with respect to polarizing to the AT rich, GC rich, Neutral (meaning it correlated neither with an AT rich or GC rich oriented organism), or Outlier (meaning an AT rich organism correlated with a GC rich organism).*

Our final upstream effect analyzed was the influence of T upstream nucleotides on [T]NXN triplets with respect to the chromosome and each corresponding replichore. We see a clear pattern of context mutation rates among organisms across the spectrum of Genomic GC content to polarize towards the AT rich end of the spectrum, specifically with respect to L. lactis, S. aureus, and S. epidermidis. Among these three organisms, all organisms exhibit a moderate to strongly correlated pattern of context mutation rates, except for B. subtilis and C. psychrerythraea with one or more these AT rich organisms. Among GC neutral organisms, we saw nearly all of their context dependent mutation rate patterns bias towards AT rich organisms, with the exception of B. subtilis, which exhibited no strong correlations across the [T]NXN organism dataset as shown in table 11:

| Organism | # of GC biased rate patterns | # of AT biased rate patterns |
|---|---|---|
| **A. tumefaciens** | **2** | **4** |
| **B. subtilis** | **0** | **0** |
| **E. coli** | **1** | **1** |
| **R. pomeryoi** | **0** | **2** |
| **T. turnerae** | **2** | **4** |

*Table 10 - [T]NXN frequency of GC neutral organisms correlating with a more AT rich or GC rich organism*

It is interesting to note that though S. aureus, S. epidermidis, and L. lactis exhibited increased patterns of context mutation rates, also at the same time appeared to demonstrate a diminished strength in patterns of context mutation rates with most AT rich organism in the dataset, M. florum.

When we further compartmentalize our analysis with respect to the individual replichores, we see more discretized patterns in mutation rates between organisms. Starting with the Left replichore, we see the concentration between more GC rich organisms largely concentrated here. We see strong positive correlations between K radiotolerans and C. crescentus, D. radiodurans, and B. cenocepacia, along with strong positive correlations between the latter three amongst each other. Also, towards the GC end of the spectrum, we see that R. pomeryoi exhibits strong negative correlations with both B. cenocepacia, and D. radiodurans. Meanwhile if we shift our focus

towards the AT rich end of the left replichore, we see a strongly positive correlated amongst the most AT rich organisms, particularly in the case of L. lactis with other AT rich organisms. While we also see AT rich organisms correlating across the GC content spectrum in the left replichore, it appears to be more interspersed than the uniform behavior observed in the right replichore.

When we shift our focus to the right replichore, we see marked similarity and differences to between the corresponding left replichore. In terms of the most AT rich organisms, we still see multiple strong positive correlations in context mutation rates amongst them, albeit not quite as uniform with L. lactis. However, what we do see is increase patterns of context mutation rates amongst both GC neutral and AT biased organisms. When observing patterns of mutation rates amongst GC neutral organisms, we see that the only positively correlated context mutation rate pattern appears to be E. coli with R. pomeryoi (though this relationship appears to exist on both replichores). When analyzing the GC Rich end of organisms, the only strongly positive correlated pattern of mutation rates existed between R. sphaeroides and B. cenocepacia, while other patterns between mutation rates on seen on the left were did not appear to carry over to the right replichore. The most interesting observation however appeared to be with K. radiotolerans exhibiting near uniformly strong negative correlations in patterns of context dependent mutation rates with both GC rich organisms and GC neutral organisms except for M. smegmatis, which had no correlation, and T. turnerae, which exhibited a moderately correlated positive relationship.

# Discussion

## GC & AT Rich Genomic NXN Triplets

For the correlation in mutation rates among GC rich organisms, one possible explanation for this relationship could be explained due to the fact of the environmental diversity that C. crescentus can be found in. Given both C. crescentus and R. pomeryoi may can both be found in sea water environments may explain their relationship in mutational patterns(51, 52). Meanwhile given D. radiodurans extremophile adaptations for surviving Ionizing radiation may have some overlap with C. crescentus's oligotrophic adaptations in nutritionally limited environments (51, 53). Additionally, there were no discernable differences in patterns between the left and right replichore on the forward 5' to 3' or reverse 3' to 5' strand of replication for genomic NXN triplets. Correlations only appear to be diluted upon examination of NXN replichore. However, it should be noted that L. lactis maintains the strongest positively correlated mutation rates regardless of strand orientation or replichore.

For the correlation in mutation rates amongst AT rich organisms, we found pronounced activity in Arg, Ala, and Thr amino acids. In Arg, we found that, C[G]X triplets, in particular C[G]C and C[G]G exhibited some of the highest context dependent mutation rates among Arg genomic triplets. Upon comparison of Mutation rates to GWTC, we found that in general, C[G]C and C[G]G triplets were the least encoded triplets, meanwhile A[G]A was found to be the most highly encoded NXN Arg triplet, which also corresponded with the lowest mutation rate among Arg sites. One may argue based on this interpretation that C[G]X Arg triplets are not favored for encoding genomic Arg amino acids in favor of the less mutable A[G]A triplet in AT rich organisms.

In 4-fold amino acid sites, Threonine exhibited highly mutable A[C]A and A[C]G triplets, however held some interesting insights when comparing A[C]X genomic triplet usage. When

comparing context mutation rates and GWTC between AT rich organisms, despite A[C]A being highly mutable, it was also highly encoded for as a Thr triplet. Meanwhile in the case of the other highly mutable triplet A[C]G, was found to have the lowest genomic triplet usage of A[C]X sites on average. In a similar manner to Threonine, when examining Alanine, we found G[C]A and G[C]G to be the most highly mutable G[C]X triplets. Upon an identical method of comparison used in Thr, we found despite G[C]A being highly mutable, it was one of the most highly encoded Ala NXN genomic triplets, and similarly G[C]G was the lowest encoded NXN genomic Ala triplet on average. Based on these observations, one may infer the following: "In a AT rich organism given the option to encode a given purine A or G to express a genomic Thr or Ala amino acid AT rich organisms appear to prefer to encode using an A purine for a right neighbor nucleotide."

## NXNY Downstream Triplets

After examining NXN genomic triplets, we wanted to examine the contribution of upstream and downstream effects a nucleotide Y would influence patterns of context mutation rates between prokaryotic organisms across the genomic GC spectrum. Generally, we found sparse positively correlated relationships amongst organisms with a few notable exceptions. When looking at GC rich organisms, we found 3 distinctly strong relationships between K. radiotolerans and C. crescentus, A. tumefaciens and M. smegmatis, and E. coli and R. pomeryoi. For each of these relationships, we saw both symmetric and asymmetric contributions for the left and right replichore responsible for influencing the relationship between their patterns of context dependent mutation rates. When turning our attention to AT rich organisms, we found noticeable sensitivity to the downstream Y nucleotide among AT rich organisms. We observed that by simply changing which nucleotide Y sat downstream positively correlated mutation rates would invert negatively

or even vanish entirely. When retracing patterns from the chromosome level to the individual replichores, we saw that though on some NXNY replichores positively correlated relationships may exist, the corresponding replichore paired with it may exhibit an equal or stronger negative relationship, thereby negating any trace of a positive relationship at the chromosome level.

## Downstream YNXN Triplets

The final portion of the analysis we investigated the influence of a given upstream nucleotide Y on genomic NXN triplets with respect to the individual chromosome and its corresponding replichores. The general trend we found with YNXN genomic triplets was a polarization in patterns of context dependent mutation rates biased towards more AT or GC rich organisms, and these patterns were observed in varying and degrees of intensity with respect to a given nucleotide Y with respect to the chromosome and the replichore. On a chromosome wide level, we found that if Y is a G or C upstream nucleotide, we observed that patterns of context dependent mutation rates between organisms will polarize towards AT/GC rich organisms, where A. tumefaciens and B. subtilis tend to exhibit correlations towards AT rich organisms more often, and T. turnerae and E. coli tend to correlate more frequently with GC rich organisms. Meanwhile, if T is an upstream nucleotide, organisms across the genomic GC content range tend to exhibit increased positive correlations in context dependent mutation rate patterns with AT rich organisms, specifically with respect to L. lactis, S. epidermidis, and S. aureus. Moreover, we can infer a pattern of replichore specific bias in relationships of context dependent mutation rates between organisms. Patterns biasing towards GC rich organisms appear to generally prefer the composition of CNXN for each replichore, but in in other YNXN configurations tend to favor the left replichore.

Meanwhile patterns biasing towards AT rich organisms appear to favor TNXN generally, but for other YNXN configurations bias tend to favor the right replichore.

When analyzing YNXN patterns with respect to the individual replichores, The ANXN upstream nucleotide context mutation rates appeared to exhibit minimal influence overall with respect to both the right and left replichore. When observing CNXN upstream context mutation rates with respect to the left and right, we see that patterns are more biased towards GC rich organisms on both the left and right replichore, with the right replichore showing particularly pronounced strong positive correlations amongst the most GC rich organisms. Meanwhile patterns biased with respect to AT rich organisms appears to be mor evenly distributed between each replichore. GNXN upstream context dependent mutation rates show a bias in patterns of context mutation rates in more GC rich organisms, while the right replichore tended to favor increased patterns in context mutation rates among AT rich organisms. GNXN GC neutral patterns of context mutation rates appeared to be for the most part replichore agnostic. Finally, in the case of TNXN upstream context mutation rates, more concentrated for AT and GC biased context mutation rate patterns were exhibited in the left replichore, while in the right replichore we saw correlations biasing AT rich organisms more broadly across the genomic GC spectrum, while we saw a noticeable diminish in GC rich biased patterns, and even a trend towards negatively correlated context mutation rate patterns across the genomic GC spectrum with K. radiotolerans.

# Chapter 4: Analysis of Synonymous Codon Triplet Usage and Mutation Rate Relationships Across a Diverse Array of Bacterium

## Introduction

The final major area of analysis we decided to explore was designed to be an extension of our work in aim 2. In the previous chapter, we discovered pronounced and distinct patterns between the relationships of genomic NXN triplets and mutation rates. These patterns were found to have sensitivity with respect to the strand of replication encoded upon, concentration of genomic GC content, and spatially with respect to a given replichore. We wanted to further expand our analysis to the subclass of genomic triplets responsible for synonymous codon usage in four-fold degenerate amino acids. We are specifically interested in four-fold degenerate codon usage sites, due to the fact the third nucleotide position in a codon triplet regardless of nucleotide will still encode the same amino acid (30). Ideally, the properties should extend to this class of triplets as well, however we cannot assume this to be the case.

In this chapter, we will take a similar systematic approach to analyzing, dissecting and interpretation of the potential observed patterns in codon usage and mutation rate. First off, we will need to determine if any pattern between codon usage and mutation rates exists, irrespective of any prior parameters at the most surface level. Next, we will delve further by looking at similar parameters previously affecting genomic triplet usage, like GC content and the native strand of replication a codon is observed on. If these patterns or other novel patterns corresponding to nucleotide composition in gene coding regions exist, they may exhibit similarly pronounced and distinct relationships affecting synonymous codon usage in an organism relative to the context dependent mutation rate for a given triplet. These patterns may have consequential effects from a biological perspective, i.e., an organism utilizing a specific 4-fold amino acid triplet with a higher

mutation rate may gain the genomic instability needed relative to its given neighbor nucleotides to induce a nonsynonymous mutation in subsequent generations to gain new adaptations affecting the overall fitness of the organism (such as a new mechanism for pathogenicity, or antibiotic resistance in bacterial prokaryotes). Based on our prior observations seen in Aim 2 with mutation rates and genomic triplets, we hypothesize that organisms ideally have an inverse relationship with their synonymous codon usage relative to their mutation rate, i.e., the higher observed count of a given synonymous codon triplet, its corresponding context dependent mutation rate will decrease.

## Methods

### Required Input and Tools

To conduct the requisite analysis needed in this chapter, a few required files, packages and software were necessary to accomplish this task. The input files needed are all generated from CDMAP's Single Organism Analysis (SOA) pipeline, which with minimal concatenation processing of the files can accomplish our analyses. The resulting concatenated input table contains the relevant codon usage, mutation rate, and genomic triplet information for ease of visualization and downstream analysis. To establish whether if any relationship exists between codon usage exists between codon usage and mutation rates, we will accomplish this using Pearson's product moment correlation. We will assume that codon usage is independent of each other and will analyze each organism with respect to all four-fold degenerate amino acid sites on both the forward and reverse strand of replication, then similarly we will conduct this analysis in the same manner with respect to each four-fold degenerate amino acid. This will be conducted using stripped down versions of the CDMAP Multi-Organism Analysis (MOA) pipeline to streamline calculation relative to the number of sites.

For the next part of our analysis, we will determine if synonymous codon usage exhibits a dependent relationship to mutation rate, and the factors that contribute to how sensitive the dependency between usage and rates are affected if they exist via regression analysis. Similarly, like above and in aim 2, we will first examine all four-fold degenerate codon triplets among all organisms exhibits a relationship with respect to the forward or reverse strand of replication. Next, we will subdivide our analysis to 4-fold synonymous codon triplets with respect to the native strand of replication to determine if patterns are influenced by GC content, and if those patterns vary based on the strand of replication. We opted to visualize these regressions easily within the Tableau data visualization software, however there are any number of other software packages easily capable of accomplishing this task.

## Analysis Methodology

To analyze patterns of codon usage in our diverse bacterial dataset, we had to determine the best approach that provides a uniform analytical treatment. At the most surface level when analyzing with respect to the chromosome, we realized that unlike analysis of NXN genomic triplets, directionality must be considered when we analyze synonymous codon usage. As outlined in table 12 we can clearly see that translation of a set codon triplets for an amino acid do not translate one-to-one between strands of replication.

In the case of amino acid encoding, no four-fold degenerate set of codon triplets encoding a given amino acid on the 5'-NNN-3' forward strand will encode the same amino acid. Meanwhile, even translation of a set of codon triplets rarely yields the complementary set will all translate into the same class of fold degenerate codon triplets. Ergo, to preserve as much biological information as possible, we opted to

| Amino Acid | 5'-NNN-3' | 3'-NNN-5' | Fold/AA Change |
|---|---|---|---|
| Valine | GTA | TAC | 2 fold - Tyr |
| | GTC | GAC | 2 fold - Asp |
| | GTG | CAC | 2 fold - His |
| | GTT | AAC | 2 fold - Asn |
| Threonine | ACA | TGT | 2 fold - Cys |
| | ACC | GGT | 4 fold - Gly |
| | ACG | CGT | 4 fold - Arg |
| | ACT | AGT | 6 fold- Ser |
| Alanine | GCA | TGA | 3 fold- STOP |
| | GCC | GGC | 4 fold - Gly |
| | GCG | CGC | 4 fold - Arg |
| | GCT | AGC | 6 fold - Ser |
| Glycine | GGA | TCC | 6 fold - Ser |
| | GGC | GCC | 4 fold - Ala |
| | GGG | CCC | 4 fold - Pro |
| | GGT | ACC | 4 fold - Thr |
| Arginine | CGA | TCG | 4 fold - Ser |
| | CGC | GCG | 4 fold - Ala |
| | CGG | CCG | 4 fold - Pro |
| | CGT | ACG | 4 fold - Pro |

*Table 11- 4-fold degenerate codon triplets in their 5-NNN-3' configuration, and their respective 3'-NNN-5' codon triplet on the reverse strand of replication, along with the respective amino acid and fold change*

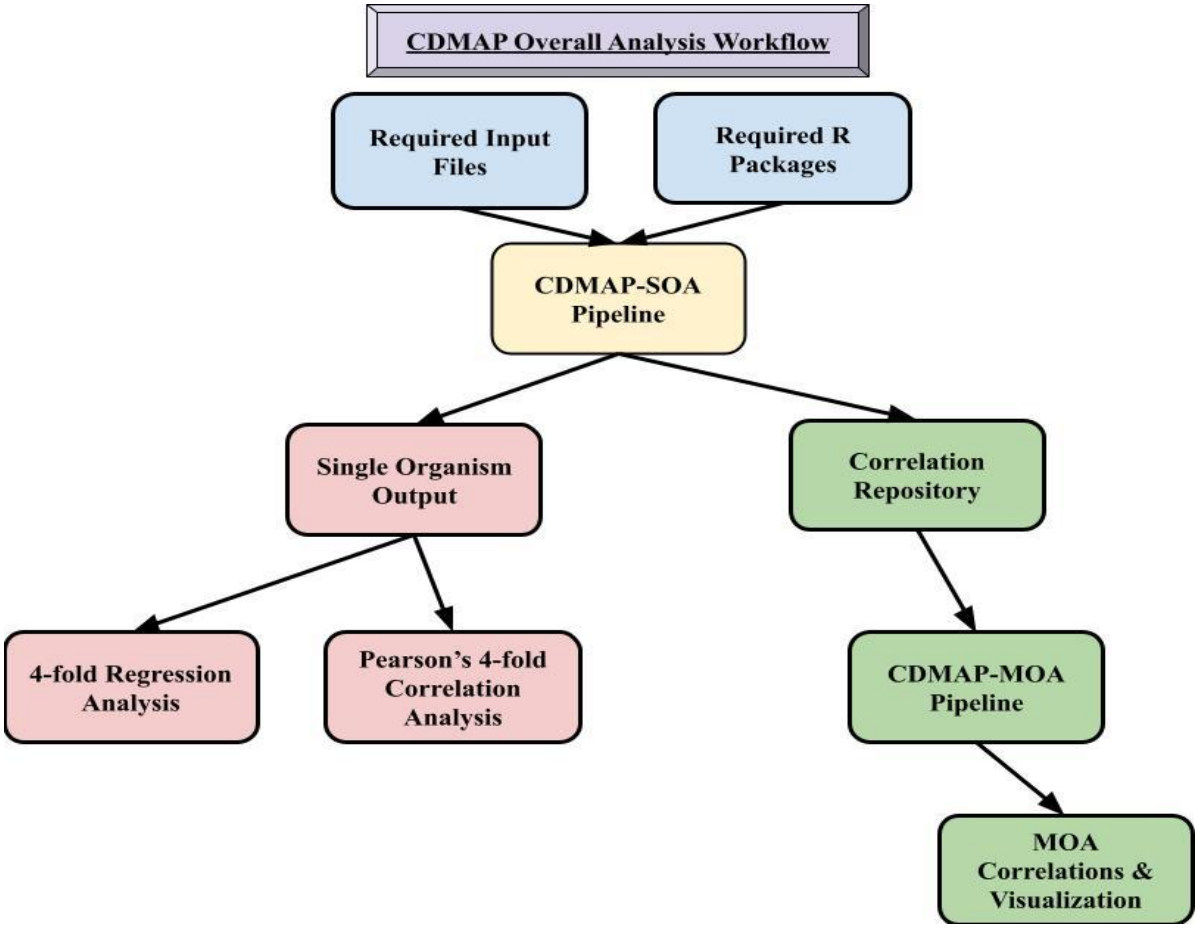



*Figure 17 - Simplified Workflow diagram of CDMAP downstream analysis. Steps outlined in green were automated downstream analysis steps outlined and discussed in Aim 2. Downstream analysis carried out in Aim 3 was carried out utilizing output and methods outlined in pink.*

separate codon usage with respect to their native strand of replication. Given the potential loss of information that could be introduced by attempting to translate codon usage between strands of replication, we opted at least for the scope of the dissertation to analyze synonymous codon mutation rates with respect to the chromosome, and each native strand of replication we identified gene coding regions.

For regression analysis, concatenated output files generated from CDMAP and were uploaded into the tableau environment. Log-log plots were then generated for each set of codon triplets that encode each 4-fold degenerate amino acid and the associated regression and p-value

was generated for both the forward and reverse strand of replication. Based on evidence in genomic NXN triplets, we condensed our scope of analysis to the most AT and GC rich organisms.

# Results

## Discussion Order of Results

As we alluded to in the prior sections, we shall be approaching our results in the following manner: First, we shall analyze whether any relationship between codon usage and mutation rate exists, independent of each other, then analyze the strand of replication's contribution to the sensitivity of these effects. Next, we shall determine whether codon usage exhibits a dependency on mutation rate for their relationship, and whether other independent factors such as replication strand, and genomic GC content effect the sensitivity of this relationship.

### Do four-fold degenerate amino acids exhibit a relationship between codon usage and mutation rate and the factors effecting the strength of their relationship?

| Organism Name | (+) Correlation Coefficient | (+) P-Value | (-) Correlation Coefficient | (-) P-Value |
|---|---|---|---|---|
| Agrobacterium tumefaciens C58 | -0.459061656 | 0.041745263 | 0.015878422 | 0.94702598 |
| Bacillus subtilis 3610 | -0.065059874 | 0.785229507 | -0.204760803 | 0.386498629 |
| Burkholderia cenocepacia HI2424 | -0.357783913 | 0.121424756 | -0.521164693 | 0.018453565 |
| Caulobacter crescentus NA1000 | -0.021750504 | 0.9274781 | -0.619164254 | 0.003602612 |
| Colwellia psychrerythraea 34H | -0.636405563 | 0.002554528 | -0.603806166 | 0.004815324 |
| Deinococcus radiodurans | -0.270974233 | 0.247850519 | -0.298969288 | 0.200380258 |
| Escherichia coli MG1655 | 0.144170919 | 0.544236828 | 0.064047865 | 0.788497064 |
| Kineococcus radiotolerans SRS30216 | -0.399841457 | 0.08068515 | -0.537621887 | 0.014492551 |
| Lactococcus lactis DSMZ20481 | -0.173529285 | 0.464378267 | -0.051466252 | 0.829388335 |
| Mesoplasma florum | -0.14247372 | 0.54903603 | -0.126848169 | 0.594091631 |
| Mycobacterium smegmatis MC2_155 | 0.542371989 | 0.013486688 | 0.285409386 | 0.222550374 |
| Rhodobacter sphaeroides ATCC17025 | -0.151285019 | 0.524330333 | -0.417963172 | 0.066680424 |
| Rugeria pomeryoi DSS3 | -0.59231769 | 0.005927661 | -0.434057047 | 0.055850744 |
| Staphylococcus aureus ATCC 25923 | -0.392490206 | 0.086948059 | -0.438073456 | 0.053369766 |
| Staphylococcus epidermidis ATCC12228 | -0.397507212 | 0.082636377 | -0.431050562 | 0.057764554 |
| Teredinibacter turnerae T7901 | 0.487546531 | 0.029216737 | 0.510653239 | 0.021406061 |
| Vibrio fischeri ES114 | -0.648581132 | 0.001979109 | -0.416473524 | 0.067757321 |

*Table 12 – Pearson's product moment correlation of each organism with respect to the 5' to 3' forward (+) strand of replication and the 3' to 5' reverse (-) strand of replication. Each entry in the table is colored whether we observe a positively correlated (yellow) or negatively correlated (red) relationship between codon usage and mutation rate. For each observed positive or negatively correlated relationship, statistically significant ($p < 0.05$) relationships are highlighted in green.*

We wanted to determine first if any relationship between codon usage and mutation rate existed on one or both strands of replication with respect to all four-fold degenerate codon triplets. Immediately, we saw most organisms exhibited some form of correlation between codon usage and mutation rate. The exceptions to this were L. lactis, M. florum, R. sphaeroides, and B. subtilis on either strand of replication. The next interesting observation we found was though we observed an abundance of organisms that though they demonstrated weak to moderately strong correlations, we found that most either exhibit basically no statistical significance (such as the case of C. crescentus (+) and B. subtilis (-) or were borderline not statistically significant (as in the case of S. aureus and S. epidermidis on both strands of replication). Of the 24 sites exhibiting some form of positively or negatively correlated relationship between their codon usage and mutation rates in our dataset of organisms, only 11 (45%) of those sites on either strand of replication were shown to be statistically significant.

Of the 11 Statistically significant correlated sites, 6 (54%) of them belong to organisms that are more biased GC rich end of the spectrum. Of those sites, just over half of GC biased organisms exhibited a moderate to a moderately strong, statistically significant inverse relationship between their codon usage and mutation rate. Meanwhile, we notice that both T. turnerae and M. smegmatis, organisms that favor more GC rich genomes, are the only organisms that exhibit a positive correlation on not only one, but both strands of replication within the entire dataset. For each strand of replication, we notice that they exhibit a moderate positively correlated relationship their usage and rates, except for M. smegmatis (-), which also corresponded to a non-statistically significant correlation. These datapoints lead us to the viewpoint that there is a demonstrable relationship occurring between codon usage and mutation rate, and in GC biased organisms, appear

to share moderately correlated relationships, while AT rich organisms on average appear to more uncommonly share a relationship under the assumption that their usage and rates are independent of one another.

Next, we decided to peer deeper and see if we could tease out the source of these rather strong correlations and the contribution each amino acid was playing for each organism as shown in table 13.

| Forward (+) Organism Amino Acids | Gly Corr | Gly Pval | Ala Corr | Ala Pval | Pro Corr | Pro Pval | Val Corr | Val Pval | Thr Corr | Thr Pval |
|---|---|---|---|---|---|---|---|---|---|---|
| Agrobacterium_tumefaciens_C58 | 0.3927 | 0.6073 | 0.3160 | 0.6840 | -0.0192 | 0.9808 | -0.8828 | 0.1172 | -0.5898 | 0.4102 |
| Bacillus_subtilis_3610 | 0.1628 | 0.8372 | -0.5190 | 0.4810 | -0.0743 | 0.9257 | -0.8252 | 0.1748 | -0.9361 | 0.0639 |
| Burkholderia_cenocepacia_HI2424 | -0.7458 | 0.2542 | -0.0401 | 0.9599 | 0.0289 | 0.9711 | -0.6722 | 0.3278 | -0.5031 | 0.4969 |
| Caulobacter_crescentus_NA1000 | 0.6298 | 0.3702 | 0.3733 | 0.6267 | 0.2111 | 0.7889 | 0.2937 | 0.7063 | -0.9128 | 0.0872 |
| Colwellia_psychrerythraea_34H | -0.2684 | 0.7316 | -0.1556 | 0.8444 | -0.9413 | 0.0587 | -0.4767 | 0.5233 | 0.5449 | 0.4551 |
| Deinococcus_radiodurans | -0.1481 | 0.8519 | 0.8427 | 0.1573 | 0.7294 | 0.2706 | 0.5271 | 0.4729 | -0.2955 | 0.7045 |
| Escherichia_coli_MG1655 | -0.9389 | 0.0611 | -0.5714 | 0.4286 | 0.4596 | 0.5404 | 0.3579 | 0.6421 | 0.8133 | 0.1867 |
| Kineococcus_radiotolerans_SRS30216 | 0.5789 | 0.4211 | -0.4934 | 0.5066 | -0.0787 | 0.9213 | -0.6266 | 0.3734 | -0.4644 | 0.5356 |
| Lactococcus_Lactis_DSMZ20481 | 0.1088 | 0.8912 | -0.1541 | 0.8459 | -0.2138 | 0.7862 | -0.7030 | 0.2970 | -0.8434 | 0.1566 |
| Mesoplasma_florum | -0.6058 | 0.3942 | -0.1907 | 0.8093 | 0.5769 | 0.4231 | -0.7932 | 0.2068 | -0.6287 | 0.3713 |
| Mycobacterium_smegmatis_MC2_155 | -0.0431 | 0.9569 | 0.4660 | 0.5340 | 0.5657 | 0.4343 | 0.8000 | 0.2000 | 0.8129 | 0.1871 |
| Rhodobacter_sphaeroides_ATCC17025 | -0.8189 | 0.1811 | -0.9278 | 0.0722 | 0.0582 | 0.9418 | NA | NA | -0.5338 | 0.4662 |
| Ruegeria_pomeroyi_DSS_3 | -0.1751 | 0.8249 | 0.9721 | 0.0279 | -0.4716 | 0.5284 | 0.0858 | 0.9142 | -0.9399 | 0.0601 |
| Staphylococcus_aureus_ATCC_25923 | 0.0719 | 0.9281 | 0.2043 | 0.7957 | -0.4223 | 0.5777 | -0.7817 | 0.2183 | -0.1659 | 0.8341 |
| Staphylococcus_epidermidis_ATCC12228 | -0.9495 | 0.0505 | 0.4114 | 0.5886 | -0.3427 | 0.6573 | -0.5620 | 0.4380 | 0.9657 | 0.0343 |
| Teredinibacter_turnerae_T7901 | 0.5115 | 0.4885 | 0.0823 | 0.9177 | -0.0489 | 0.9511 | 0.2580 | 0.7420 | 0.7536 | 0.2464 |
| Vibrio_fischeri_ES114 | -0.8749 | 0.1251 | 0.5997 | 0.4003 | -0.7659 | 0.2341 | -0.9077 | 0.0923 | -0.0564 | 0.9436 |

| Reverse (-) Organism Amino Acids | Gly Corr | Gly Pval | Ala Corr | Ala Pval | Pro Corr | Pro Pval | Val Corr | Val Pval | Thr Corr | Thr Pval |
|---|---|---|---|---|---|---|---|---|---|---|
| Agrobacterium_tumefaciens_C58_ChrCircular | 0.365125 | 0.634875 | -0.0439 | 0.956102 | 0.211284 | 0.788716 | -0.0734 | 0.9266 | -0.28258 | 0.717424 |
| Bacillus_subtilis_3610 | -0.14506 | 0.854941 | 0.376384 | 0.623616 | 0.324008 | 0.675992 | -0.40773 | 0.592267 | -0.41934 | 0.580662 |
| Burkholderia_cenocepacia_HI2424_Chr1 | -0.69278 | 0.307215 | -0.52784 | 0.472161 | 0.237214 | 0.762786 | -0.61083 | 0.389172 | -0.99023 | 0.009771 |
| Caulobacter_crescentus_NA1000 | 0.966457 | 0.033543 | 0.4734 | 0.5266 | 0.842525 | 0.157475 | 0.512336 | 0.487664 | -0.99056 | 0.009437 |
| Colwellia_psychrerythraea_34H | -0.60399 | 0.396013 | 0.900828 | 0.099172 | -0.78106 | 0.218941 | 0.55004 | 0.44996 | 0.241575 | 0.758425 |
| Deinococcus_radiodurans | -0.0593 | 0.940704 | -0.22469 | 0.77531 | 0.956711 | 0.043289 | 0.719204 | 0.280796 | -0.33899 | 0.661005 |
| Escherichia_coli_MG1655 | 0.947613 | 0.052387 | 0.758845 | 0.241155 | -0.1678 | 0.832202 | 0.011871 | 0.988129 | 0.463019 | 0.536981 |
| Kineococcus_radiotolerans_SRS30216 | 0.470286 | 0.529714 | -0.49844 | 0.501562 | -0.91858 | 0.081416 | -0.5467 | 0.453295 | -0.50158 | 0.49842 |
| Lactococcus_Lactis_DSMZ20481 | 0.789067 | 0.210933 | -0.21872 | 0.781276 | 0.126523 | 0.873477 | -0.0874 | 0.912602 | -0.78955 | 0.210451 |
| Mesoplasma_florum | -0.27714 | 0.722856 | -0.93751 | 0.06249 | -0.9236 | 0.076398 | -0.38903 | 0.610967 | -0.43609 | 0.563912 |
| Mycobacterium_smegmatis_MC2_155 | 0.292444 | 0.707556 | 0.813077 | 0.186923 | 0.833249 | 0.166751 | 0.543898 | 0.456102 | -0.96803 | 0.031965 |
| Rhodobacter_sphaeroides_ATCC17025 | 0.402778 | 0.597222 | -0.69786 | 0.302141 | -0.94741 | 0.052588 | -0.91865 | 0.081348 | -0.56306 | 0.436936 |
| Ruegeria_pomeroyi_DSS_3 | -0.83977 | 0.160227 | -0.52275 | 0.477249 | -0.03025 | 0.969745 | -0.8594 | 0.1406 | -0.34309 | 0.656914 |
| Staphylococcus_aureus_ATCC_25923 | -0.88687 | 0.113129 | 0.361055 | 0.638945 | -0.75865 | 0.241348 | 0.676018 | 0.323982 | 0.729598 | 0.270402 |
| Staphylococcus_epidermidis_ATCC12228 | -0.83417 | 0.165834 | -0.76886 | 0.231139 | -0.96272 | 0.037281 | -0.9872 | 0.012796 | -0.21512 | 0.784885 |
| Teredinibacter_turnerae_T7901 | 0.613556 | 0.386444 | 0.331464 | 0.668536 | -0.63071 | 0.369294 | 0.644851 | 0.355149 | 0.869766 | 0.130234 |
| Vibrio_fischeri_ES114_Chr1 | -0.0573 | 0.942702 | -0.08041 | 0.919586 | -0.88624 | 0.113757 | 0.195574 | 0.804426 | 0.479938 | 0.520062 |

*Table 13 – Pearson's product moment correlation of each organism with respect to each four-fold degenerate amino acid. Each organism specific four-fold degenerate site was separated into their respective 5' to 3' forward (+) strand of replication and the 3' to 5' reverse (-) strand of replication. Each entry in the table is colored whether we observe a positively correlated (yellow) or negatively correlated (red) relationship between codon usage and mutation rate. For each observed positive or negatively correlated relationship, statistically significant ($p < 0.05$) relationships are highlighted in green.*

When dissecting each organism's relationship with respect to the individual amino acid relative to their native replication strand, we observe several interesting patterns. The first immediate observation is that in the forward strand of replication, only two individual amino acids

among all organisms (R. Pomeryoi – Alanine, and S. epidermidis – Valine) exhibited statistically significant direct correlations, due in part to both having a Pearson's $r > 0.95$. After combing over the regression analysis, we wanted to delve into potential patterns occurring on between four-fold degenerate codon usage and mutation rates on a per amino acid basis, to discern any deeper level of insight into pre-existing patterns observed. We decided to correlate codon triplet usage and site-specific mutation rates on a per amino acid basis for all four-fold degenerate organisms within our database using Pearson's product moment correlation. In table 14 we find very interesting case, where in the other previously statistically significant organisms on the forward strand of replication do not exhibit an individual relationship between codon usage and mutation rate that is statistically significant. However, what is prevalent amongst them all are amino acids on their respective forward strand of replication are polarized either in the direct or inversely related relationships that are moderate or strongly correlated, rarely exhibiting no correlation.

Meanwhile, on the reverse strand of replication, though we see the largely ubiquitous correlation between mutation rate and codon usage like in the forward strand, we see a larger smattering of those sites observed to be statistically significant. Of the 6 statistically significant correlations in usage and rates, we found that two-thirds of these relationships belonged to GC rich organisms, while the other two sites were contained within one AT rich organism (S. epidermidis, Proline and Valine). Like the forward strand, we also see all statistically significant sites harboring a Pearson's $r > 0.95$, while the remainder of non-statistically significant correlations exhibited a similar behavior of moderate to strongly correlated relationships skewed either directly or inversely. Given the observations with respect to each strand of replication and individual amino acid, there appears to be evidence supporting a clear relationship between codon usage, however it appears they are merely not independent of each other. Therefore, we find it necessary to

investigate the relationship between codon usage and mutation rate where codon usage is dependent upon the site-specific context mutation rates and vice versa for a given organism.

**For four-fold degenerate amino acids, do we see a dependent relationship of codon usage given its context dependent mutation rate with respect to a given strand of replication?**

| (+) Regression Coefficient | (+) P-Value | (-) Regression Coefficient | (-) P-Value |
|---|---|---|---|
| **0.232326** | **< 0.0001** | **0.166383** | **< 0.0001** |

*Table 15 - Regression Coefficients and statistical significance of all four-fold codon triplet families with respect to the forward (+) and reverse (-) strand of replication.*

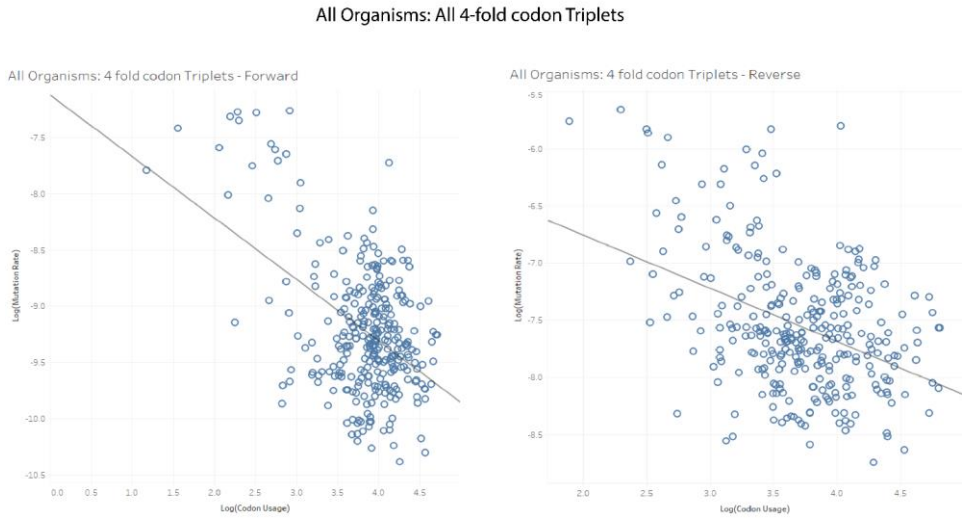



*Figure 18 – Linear regression of All organism four-fold degenerate codon triplets among all organisms within the Bacterial dataset analyzed by CDMAP. Each relationship between codon usage and mutation rate on the forward (+) strand of replication (left figure) and the reverse (-) strand of replication (right figure).*

After our initial investigation of any relationship between codon usage and mutation rates as factors independent of each other, we saw evidence pointing towards a relationship on a per organism, per individual synonymous four-fold amino acid basis. In the following sections, we are investigating whether codon triplet usage for four-fold degenerate sites harbors a dependent relationship with their respective site-specific context dependent mutation rates. Similar to our previous avenues of analysis, we approached this question in the same level-by-level method of dissection. First as illustrated in figure 18 and table 14 we can see the log transformed regression of codon usage and mutation rate for each strand of replication. At the broadest level of analysis, we can observe some evidence lending credence to codon usage dependency on mutation rate with respect to the forward strand. Although we see weak regression for the forward strand, and borderline no relationship for reverse strand of replication.

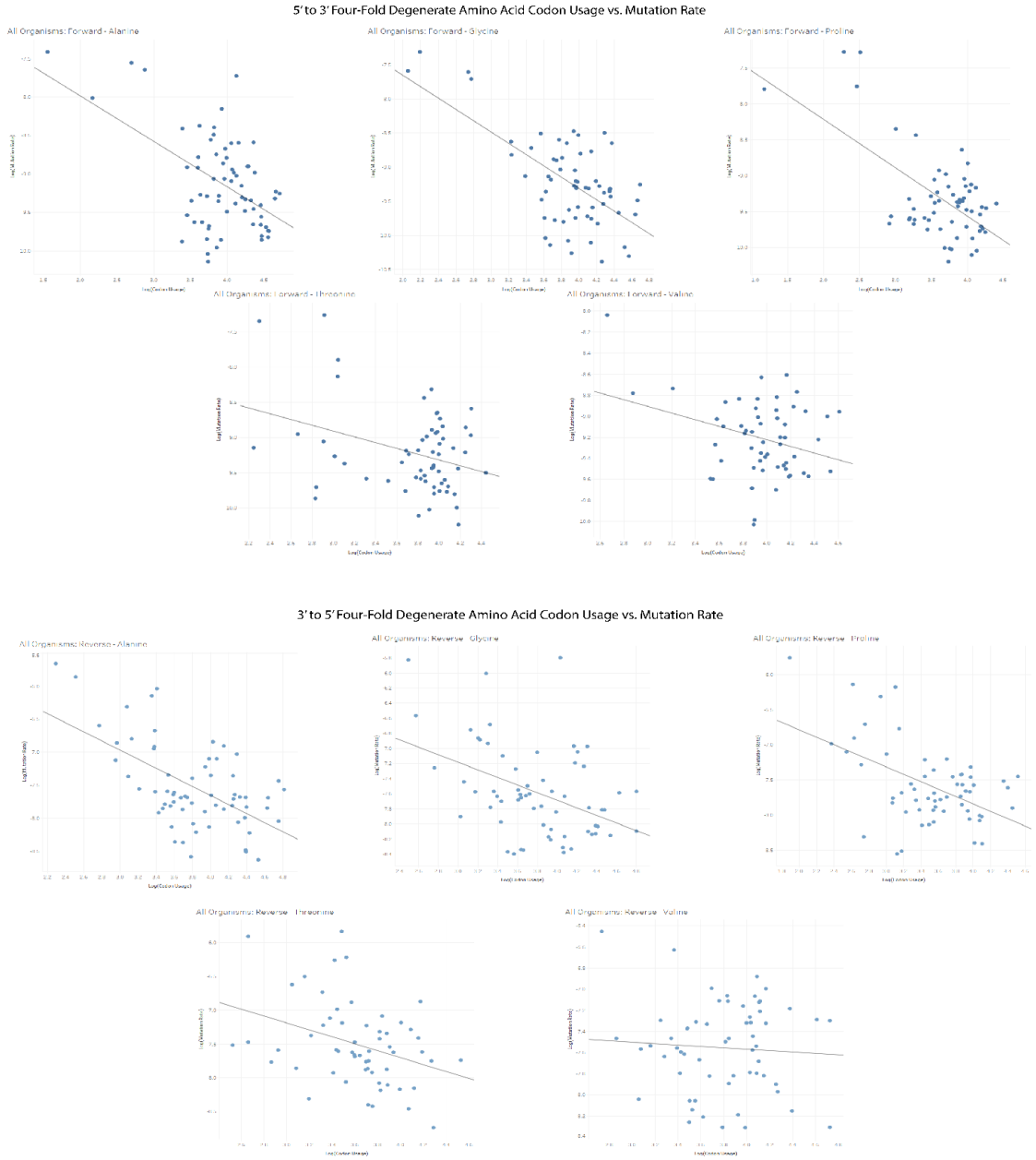



*Figure 19 - Codon usage versus Mutation rate regressions for all four-fold degenerate amino acids. Each amino acid has been broken down into the corresponding forward (+, top figure) or reverse (-, bottom figure) strand of replication.*

Under normal circumstances, one may argue the merit of further dissection and interpretation of results. However, as we saw both in the second aim, and earlier in our analysis that both patterns of both genomic triplet and codon triplet usage may be obfuscated or seemingly

vanish when observing from a given scope, and four-fold degenerate codon triplets harbor a highly varied nucleotide triplet composition. Given the generic shape of our regression datapoints in both the forward and reverse strand of replication, the possibility exists these underlying patterns being present, and merely necessitated delving deeper into the data. Therefore, in order to tease out possible relationships, we subdivided our dataset on each strand of replication into their respective four-fold degenerate amino acid representations as shown in figure 19 and their corresponding coefficients provided in table 16:

| Amino Acid | (+) Regression Coefficient | (+) P-Value | (-) Regression Coefficient | (-) P-Value |
|---|---|---|---|---|
| **Alanine** | **0.271956** | **< 0.0001** | **0.338428** | **< 0.0001** |
| **Glycine** | **0.416689** | **< 0.0001** | **0.182795** | **0.0004741** |
| **Proline** | **0.382659** | **< 0.0001** | **0.254959** | **< 0.0001** |
| **Threonine** | **0.117489** | **0.0068478** | **0.12425** | **0.0066526** |
| **Valine** | **0.097819** | **0.0189299** | **0.0041097** | **0.62354** |

*Table 15 - Regression coefficients and P-value of four-fold degenerate amino acids for all organisms in Figure 19. Each amino acids codon usage and mutation rate were regressed with respect to the individual forward (+) and reverse (-) strand of replication.*

Upon partitioning organism's codon triplets in the dataset with respect to their individual amino acid, we immediately notice an emergence of an inverse relationship of codon usage being dependent upon mutation rate. When observing four-fold degenerate codon triplets on the forward strand, we can clearly see evidence of this relationship in usage and rates particularly in Alanine, Proline, and Glycine. Each of these four-fold sites harbors a weak (or moderate in the case of Glycine) relationship showing clear statistical significance, while Threonine and Valine do not show signs of a pronounced relationship. On the 3' to 5' reverse strand of replication we observe a more subtle representations of patterns in codon usage and mutation rate observed on the forward strand.

There are several notable observations when comparing differences in patterns between replication strands. First off, we note a substantial drop in relationship in both Glycine and Proline, where Glycine was observed to be the strongest relationship on the forward strand of replication, now shows little to no relationship on the reverse strand of replication. Meanwhile, in the case of Proline where we observed previously a borderline moderately strong relationship between usage and rates, has dropped to a solidly weakly expressed relationship between usage and rates. Interestingly enough, Alanine represented the strongest relationship between codon usage in mutation rate on the reverse strand, mainly due in part to having a similarly expressed relationship in usage and rates seen on the forward and reverse strand. Similar to the forward strand of replication, both Threonine and Valine expressed little to no relationship in patterns codon usage and mutation rate.

When comparing individual amino acid relationships back to the comparisons of all sites with respect to a given strand of replication, we can observe an interesting phenomenon occurring with respect to each individual strand of replication. On the forward strand of replication, we see a loose clustering occurs roughly around the point of (-9.5,4) which corresponds to roughly a mutation rate of $3.16 \times 10^{-9}$ and a codon usage count of approximately 10,000. Meanwhile, on the reverse strand of replication we observe a similar loose clustering around (-7.75, 3.75) which corresponds to a mutation rate of $3.16 \times 10^{-8}$ and a codon usage count of approximately 5,623, a nearly two-fold drop in codon usage within one order of magnitude difference in mutation rate. This clustering event may serve as another signpost in support of the inverse relationship between codon usage and mutation rate. However, other factors such as genomic GC content may be playing a subtle, albeit obfuscated contribution to patterns of codon usage and context mutation rates, which we shall explore in the next section.

## Are patterns of codon usage and mutation rate being influenced with respect to a given concentration of genomic GC content?


*Figure 20 – Linear regression of All AT rich (left column) and GC rich (right column) four-fold degenerate codon triplets among all organisms within the Bacterial dataset analyzed by CDMAP. Each relationship is represented relative to the forward strand (top row) and reverse strand (bottom row) of replication.*

| Organisms | (+) Regression Coefficient | (+) P-Value | (-) Regression Coefficient | (-) P-Value |
|---|---|---|---|---|
| **AT Rich** | **0.565279** | **< 0.0001** | **0.313148** | **< 0.0001** |
| **GC Rich** | **0.0009908** | **0.767079** | **0.0008295** | **0.785214** |

*Table 17- Regression coefficients and statistical significance of AT and GC rich regressions of codon usage and mutation rate with respect to the forward (+) and reverse (-) strand of replication.*

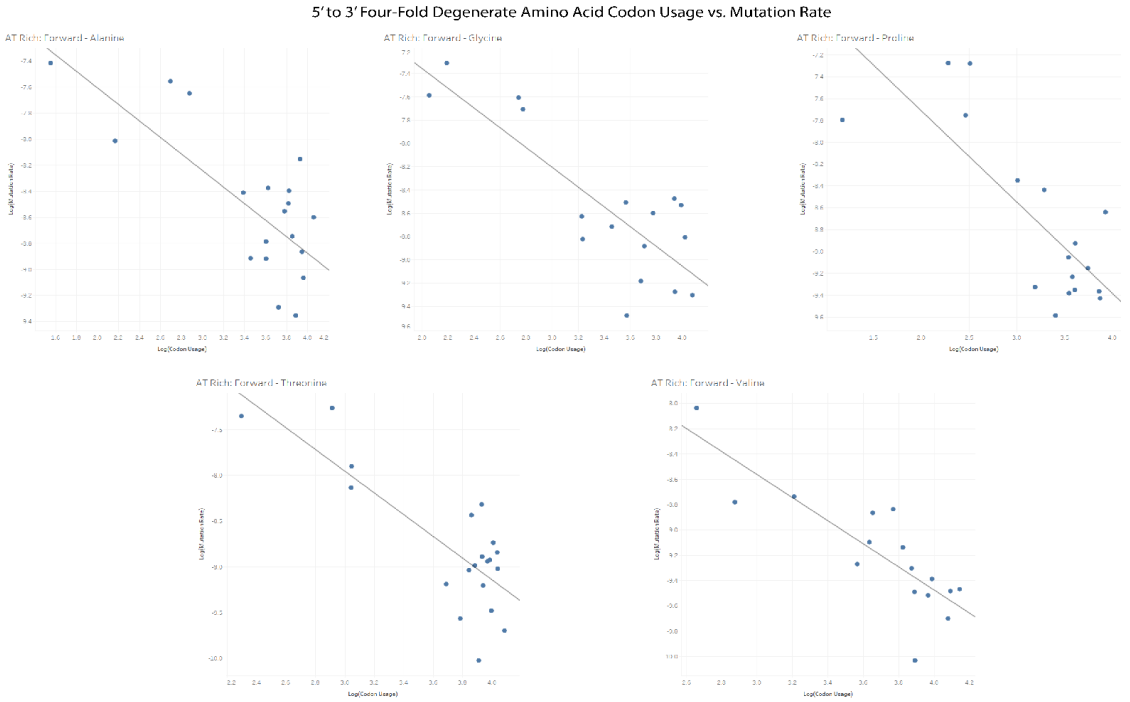



*Figure 21 - Codon usage versus Mutation rate regressions for all four-fold degenerate amino acids for the most AT rich organisms on the forward strand of replication with respect to their individual amino acid.*

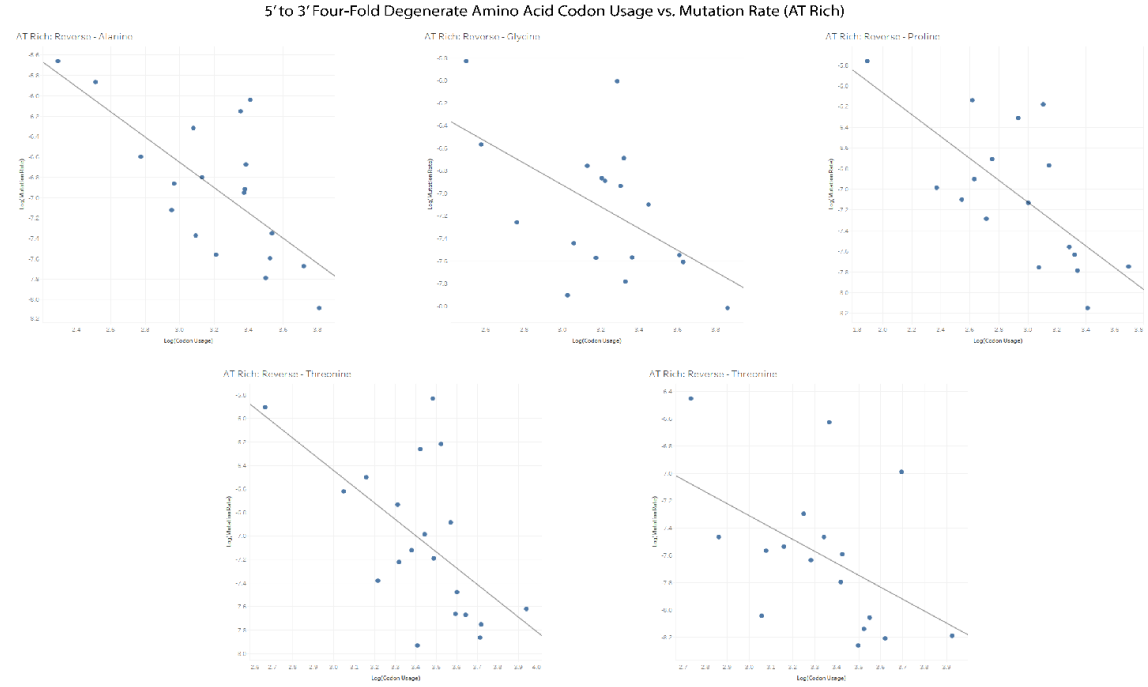



*Figure 22 - Codon usage versus Mutation rate regressions for all four-fold degenerate amino acids for the most AT rich organisms on the reverse strand of replication with respect to their individual amino acid.*

In our previous section, we investigated the possible relationship of codon triplet usage being dependent upon their corresponding site-specific context dependent mutation rate. When sub-partitioning our usage and rates with respect to their native strand of replication, and further down to a per amino acid basis we found varying levels evidence supporting a dependent relationship with respect to each strand of replication. These patterns were partially obscured at the organism level but became more prevalent when observed on a per amino acid basis. In aim 2, we found relationships between genomic NXN triplets and mutation rate that gravitated to either the genomic GC or AT rich biased end of our dataset. Additionally, we saw these patterns obfuscated when observing patterns at higher order scope, only teased out when delving deeper with respect to each organism. In this section we investigate codon triplet usage's dependent relationship in the most AT and GC rich organisms in our dataset, using a similar methodology to prior sections. Immediately upon inspection of AT rich and GC rich organisms on the forward and reverse strand of replication we can observe stark contrasting differences in their relationships. AT rich organisms on both strands of replication demonstrate a clear, statistically significant relationship showcasing their codon triplet usage being dependent upon their site-specific context dependent mutation rates. The main distinction between replication strands of AT rich organisms mainly boils down to the strength, where we observe a moderately expressed ( $r^2 = 0.565279, p < 0.0001$) on the forward strand, while on the reverse strand we observe a weakly expressed ( $r^2 = 0.313148, p < 0.0001$) relationship their usage and rates. Meanwhile, when we shift our focus to GC rich organisms, we observe nearly no relationship on either strand of replication ( $r^2 < 0.001, p > 0.75$). This result is of particular interest, given when analyzing correlation relationships for our collection of organisms, the majority of statistically significant correlations were found to be in GC Biased organisms. Given the stark contrasts between AT and

GC rich organism's patterns between their codon usage and mutation rate, and similar to previous sections we decided to further tease out the root of these patterns my observing these contributions on a per-amino acid basis.

| | | AT Rich | | |
|---|---|---|---|---|
| **Amino Acid** | **(+) Regression Coefficient** | **(+) P-Value** | **(-) Regression Coefficient** | **(-) P-Value** |
| **Alanine** | **0.54287** | **0.0002111** | **0.489676** | **0.0008524** |
| **Glycine** | **0.676732** | **< 0.0001** | **0.293385** | **0.020247** |
| **Proline** | **0.613534** | **0.0002** | **0.479153** | **0.0020755** |
| **Threonine** | **0.632271** | **< 0.0001** | **0.340288** | **0.0069366** |
| **Valine** | **0.699635** | **< 0.0001** | **0.238034** | **0.0469474** |

*Table 18 - Regression coefficients and P-value of four-fold degenerate amino acids for AT rich organisms in Figure 21 & 22. Each amino acids codon usage and mutation rate were regressed with respect to the individual forward (+)*

When we separate out AT rich organism's four-fold degenerate patterns of codon usage and mutation rate into their individual amino acids in figures 21 & 22 we see a further distinction of patterns as outlined in table 18. When examining the forward strand of replication, we see clearly represented, statistically significant moderate to moderate-strong relationships between each amino acids pattern of codon usage and mutation rates. Meanwhile on the reverse strand of replication, we see similarly to other patterns observed in previous sections more tempered version of weak to moderate strength relationships compared to the forward strand of replication. These patterns observed with respect to the individual strand of replication were expected and agree with our larger surface level analysis of patterns of codon triplets and mutation rates. However, upon closer inspection, we see subtle nuances in the pattern strength on an individual amino acid basis. We observe that the patterns in usage and rates for Alanine codon triplets were largely independent of replication orientation, while we saw larger variances in other four-fold degenerate amino acids. We notice that for Threonine, Valine, and Glycine which all harbor moderate-strong regression coefficients on the forward strand of replication, seemingly invert from the three strongest relationships between codon usage and mutation rates, to the three weakest relationships. This

could suggest these four-fold degenerate amino acids are more highly subject to influences from replication machinery (direct continuous replication versus Okazaki fragments.)

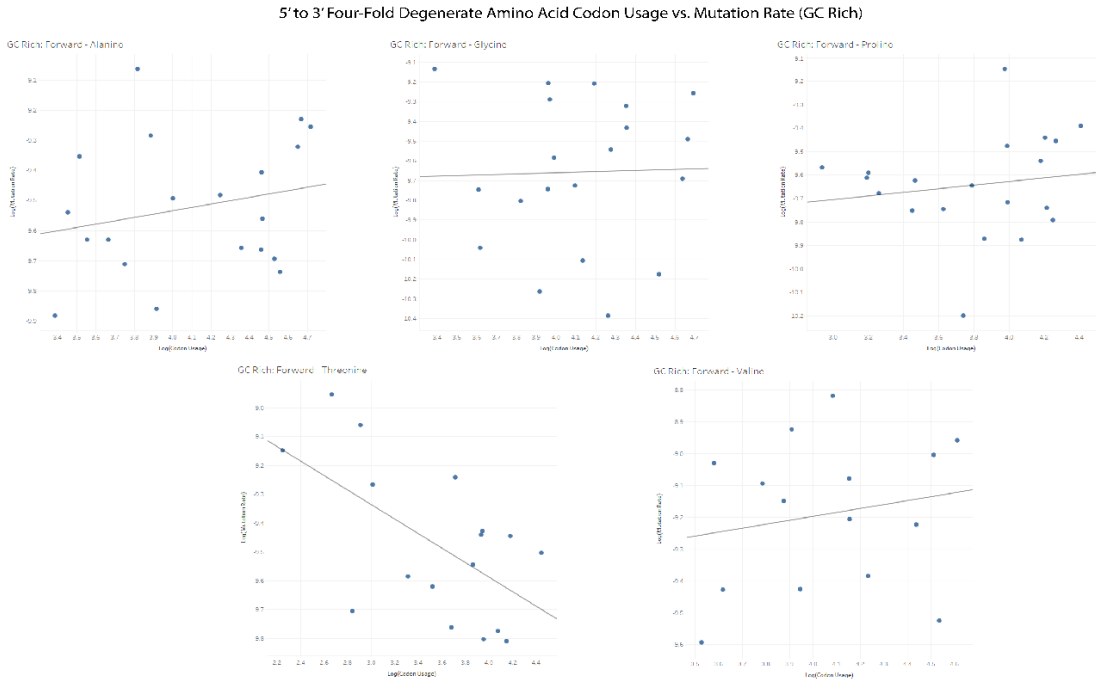



*Figure 23- Codon usage versus Mutation rate regressions for all four-fold degenerate amino acids for the most GC rich organisms on the forward strand of replication with respect to their individual amino acid.*

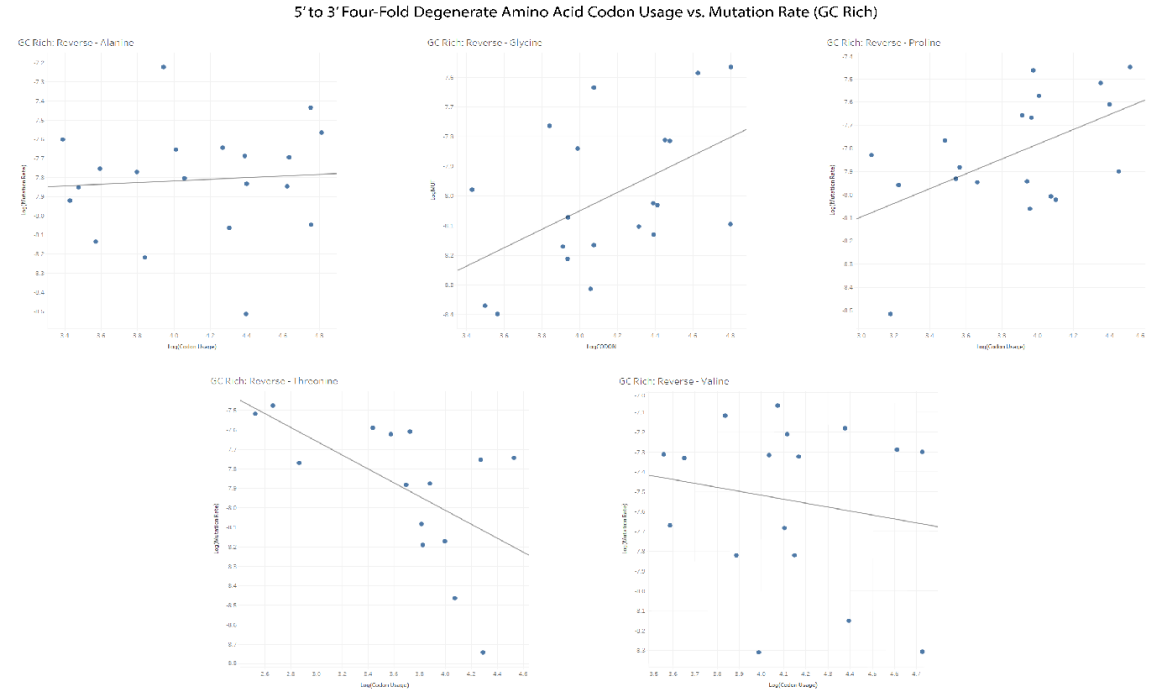



*Figure 24 - Codon usage versus Mutation rate regressions for all four-fold degenerate amino acids for the most GC rich organisms on the reverse strand of replication with respect to their individual amino acid.*

When we shift the scope of our analysis over to the GC rich end of organisms, a very different picture arises when analyzing codon usage and mutation rate relationships shown by figures 23 & 24 and table 19. On the whole, we observe very disparate relationships relative to AT

| Amino Acid | (+) Regression Coefficient | (+) P-Value | (-) Regression Coefficient | (-) P-Value |
|---|---|---|---|---|
| **Alanine** | **0.0525285** | **0.331052** | **0.005129** | **0.764143** |
| **Glycine** | **0.0007193** | **0.910635** | **0.249078** | **0.0250796** |
| **Proline** | **0.0228611** | **0.524569** | **0.282188** | **0.0192616** |
| **Threonine** | **0.344439** | **0.0132618** | **0.329457** | **0.0252533** |
| **Valine** | **0.0352854** | **0.502597** | **0.0310703** | **0.49857** |

*Table 18 - Regression coefficients and P-value of four-fold degenerate amino acids for GC rich organisms in Figure 23 & 24. Each amino acids codon usage and mutation rate were regressed with respect to the individual forward (+) and reverse (-) strand of replication.*

rich organisms, but even among individual amino acids. On the whole, we found weak to no relationship between codon usage and mutation rate with respect to either strand of replication. The most interesting and stable of these relationships observed appeared to be Threonine on both strands of replication, where it held a nearly identical regression and statistical significance. We found interestingly enough a positively expressed, albeit weak relationship on the reverse strand of replication, which aligns with earlier observations in prior sections of GC biased organisms exhibiting statistically significant correlations on the reverse strand of replication, and also harboring multiple positive strongly correlated relationships in both proline and glycine. Additionally, a biological basis for this positive relationship could be due in part to the large GC nucleotide bias in Glycine (GGX) and Proline (CCX) that may contribute to this phenomenon.

In summary, from our observations above, upon subdividing our dataset into the most AT and GC rich organisms, we observe distinct patterns with respect to codon usage and mutation

rates on each respective strand. We have observed that given a direction of replication, said relationship tends to appear more often the forward 5’ to 3’ replication strand, if this relationship also appears on the 3’ to 5’ strand of replication, it often appears in a reduced capacity, and even rarely to an equal degree of intensity. When observing differences between the most AT rich and GC rich organisms, we observed a clear moderate to moderately strong inverse relationships between in codon usage and mutation rate in the most AT rich organisms. Meanwhile, when we look at the most GC rich organisms, we see mixed weak relationships, or often no relationships between the usage and rates. Our evidence leads us to take the viewpoint that patterns of codon usage and mutation rates are influenced by genomic GC content, and demonstrate a strong case for an inverse relationship between codon usage and mutation rate.

## Discussion

In our final research aim, we sought to demonstrate possible relationships between synonymous codon triplet usage and mutation rates across our set of bacterial organisms. This investigation was motivated by patterns established in our prior aim of research when investigating patterns in genomic NXN triplets and site-specific context dependent mutation rates. The interest in analyzing four-fold degenerate codon triplets was due to their unique nucleotide triplet architecture NNX, where any change in X induces a synonymous amino acid mutation. We opted to apply a similar methodology used in Aim 2, where we analyzed codon usage patterns on an organism wide level, then with respect to factors such as their native strand of replication and genomic GC content. In order to preserve as much biological information as possible, we took a conservative approach and only analyzed four-fold degenerate codon triplet usage relative to its native strand of replication. This was due in part that fold-degenerate information cannot easily be preserved within the same amino acid encoded, let alone the same fold degenerate class of amino acid.

When approaching our analysis, we made no prior assumptions about the nature of the relationship that codon usage and site-specific context dependent usage rates shared, if any. Therefore, we opted to use Pearson's product moment correlation for each organism's codon triplet usage, and their respective site-specific context dependent mutation rates. Upon analysis we found a near ubiquitous inverse correlation between codon usage and mutation rate, though we saw just under half of these organism specific sites exhibit statistical significance. Of these strand specific organism wide sites, the majority of them were found to be GC biased with the majority of those sites residing on the reverse strand of replication, while roughly equal representations of AT rich, GC rich, and neutral organisms were found in equal measure on the forward strand of replication.

We then separated our Pearson's correlation analysis down to individual amino acid basis, where we interestingly found statistically significant patterns appear and vanish in nearly equal measures. When analyzing sites on a per amino acid basis, the majority of strongly correlated statistically significant sites belonged to GC biased organisms, with all significant AT rich amino acid correlations belonging to S. epidermidis. Outside of the statistically significant sites, we observed a near ubiquitous polarization of positively and negatively correlated patterns of codon usage rates and context dependent mutation rates. We determined that if a significant relationship between codon usage and mutation rates exists, it was possible there are other obfuscating factors that may be concealing the nature of their relationship. We progressed to the notion of the possibility of codon usage having dependent relationship given its mutation rate and decided to investigate this possibility via regression analysis.

When we analyzed all organisms with respect to their native strand of replication and observed a teaser of evidence on the forward strand of replication, while we saw little evidence supporting a dependent relationship on the reverse strand of replication. However, in prior aims figured it would be necessary to descend and sub-partition organisms further to tease out these possible patterns. So, in our next step, we separated our regression analysis into each organisms four-fold amino acid specific regression. Upon doing so we revealed patterns lending towards an inverse relationship between codon usage and mutation rate, more directly in the forward strand of replication, and to a reduced effect in the reverse strand of replication. We then pondered if, similarly to genomic triplet and mutation rate patterns that genomic GC content could affect the sensitivity of these patterns.

We followed through on this notion by taking codon triplet usage and context dependent mutation rates and analyzing the most AT and GC rich organisms with respect to their native strand

of replication. Upon doing so we found clear, distinct contributions from organismal GC/AT content affecting the dependent relationship between codon triplet usage and their respective mutation rates. In the most AT rich organisms, we observed inverse relationships between usage and rates on each strand of replication, and similarly to previous avenues of analysis found this relationship was more pronounced on the forward strand of replication versus the reverse strand of replication. When turning our focus to GC rich organisms, we puzzlingly found essentially no evidence in GC rich organisms codon usage rates being dependent on their respective mutation rates. For our final level of analysis, we opted to observe how these patterns in GC and AT rich organisms interacted on a per amino-acid basis.

When Analyzing AT rich organisms on a per amino acid basis, we only found a confirmation and intensification of regression relationships observed on the organism specific level. Each amino acid demonstrated some form of inverse relationship in their codon usage patterns and mutation rates, with subtle and nuanced differences between the forward and reverse strand of replication. When Analyzing GC rich organisms on a per amino acid basis, we found mixed and inconsistent patterns between codon usage and mutation rate, where the only notable patterns existed within Threonine on both strands, and Glycine and Proline on the reverse strand of replication. These analyses from a myriad of viewpoints lead us to the following conclusion: "We observe an inverse relationship between codon usage and mutation rate, whereas codon usage increases at four-fold degenerate amino acids, their corresponding mutation rate decreases. Additionally, genomic AT content exhibits a fundamental role influencing and intensifying this relationship on the forward strand of replication."

# Chapter 5: Conclusions and Future Work

## Conclusions

In this dissertation thesis, we began by wanting to address the problem of providing a consistent, reproducible method of analysis for context dependent mutation rates. Prior literature had shown that genetic variation in organisms across the tree of life is non-uniform, and understanding the fundamental factors shaping this primary evolutionary force can play a pivotal role in applications such as conferring novel adaptations within an organism, to shaping predictions in antibiotic resistance or virulence in pathogenic organisms. To this date, we have not observed a standard, benchmarkable method was not widely available or accessible that could be a reliable baseline for analysis of mutation rate analysis. As a result, CDMAP was developed to address the research need for an open source, reproducible software package that can be modularly built upon as research demands warranted. We accomplished our goals during development by create a streamlined platform of analysis where a user with minimal input required a consistent method of mutation rate analysis that provided an extensive breadth and depth of different scopes to analyze a given organisms context dependent mutation rates at all possible nucleotide triplet combinations.

Our next objective in this dissertation work was to build upon the framework of analyzing mutational patterns to develop a methodology to reliably analyze a dynamically scaled collection of organisms to observe whether we could discern possible overarching patterns influencing their behavior. Therefore, we developed the CDMAP-MOA pipeline to directly interface with the CDMAP-SOA pipeline to provide an all-by-all comparison of organisms analyzed by a given end user. We analyzed a total of 17 wildtype organisms, 3 of which additionally included MMR-mutational variants. We found in general that AT-rich organisms harbored significant relationships

in context dependent mutation rates amongst each other. Meanwhile, we found more generally factors such as replichore specificity, or specific upstream and downstream nucleotides showed a significant influence on the strength of relationships observed between the various organisms within the collection.

In our final analysis, we further expanded our biological analysis to understand the possible factors influencing contextual mutation patterns in prokaryotes by analyzing synonymous codon usage at four-fold degenerate amino acid sites. Additional considerations had to be made in this step due to the directionality of replication playing a fundamental factor in the preservation of biological information, therefore analysis was further separated into forward and reverse strands of replication. Then in a similar manner to our previous objective, methodically dissected the relationship between four-fold synonymous codon usage and mutation rates. We found strong evidence for an inverse relationship between codon usage an mutation rates, which appeared particularly prevalent in AT rich organisms, suggesting genomic AT exerts a direct influence upon this phenomenon.

## Future Work

### *The relationship of Context dependent mutation rates and Ne*

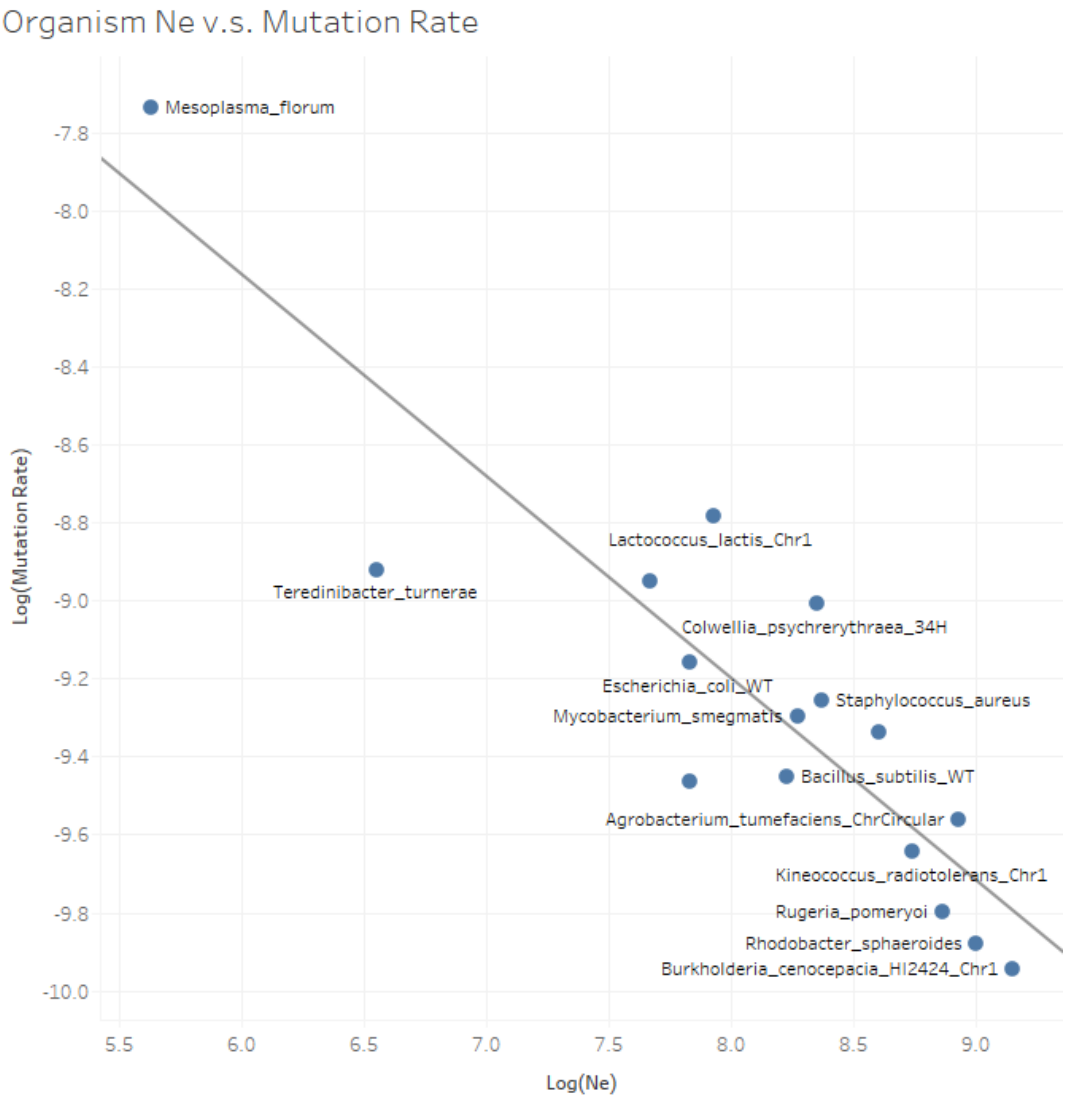



*Figure 25 – Effective population versus Mutation rate regressions for all four-fold degenerate amino acids for the most GC rich organisms on the reverse strand of replication with respect to their individual amino acid.*

At this current juncture, there are two interesting paths forward for continuation of the work outlined in this dissertation work. The first direction would be a further investigation into the role effective population size plays influencing the behavior of context dependent mutation rates. We currently hypothesize that the effective population size of an organism is a limiting factor for

regulation of genome wide contextual patterns. The motivation for this hypothesis comes from the relationship arises from earlier when discussing the power of selection and drift with respect to Ne. When using mutation accumulation to examine the range of spontaneous mutations occurring in an organism, we do this with the notion of limiting selective forces acting upon an organism by effectively making Ne = 1 (18, 24, 35). As a result, if selective forces are capable of purifying deleterious genomic patterns from a given population, then conversely if Ne is limited, then selective forces may be unable to select for repair pathways and mechanisms capable of removing damaging contextual patterns.

As of this writing, the investigation is ongoing, with our focus primarily on improving the number of samples that we have effective population size data. Though we have a limited dataset from which to draw observations from, it is interesting to note among the organisms analyzed using CDMAP in figure 24 that we have reliable Ne data for, we can there is a clearly defined significant inverse relationship between context dependent mutation rates and Ne exhibited. One possibility for this is AT rich organisms harbor a lower Ne tend to have a higher mutation rate than GC rich organisms, suggesting the repair mechanisms that mutation rates become stronger proportionally to Ne. However, increasing the number of organisms analyzed via CDMAP both among prokaryotes, and to eukaryotes and archaea along with reliable effective population estimation would be necessary to draw any definitive conclusions between this relationship, and the scope of its reach across the tree of life.

**The Spatiotemporal Relationship of Context Dependent Mutation Rates and Replication Distance From The ORI and TERM**

Our second possible direction to expand upon this dissertation work would be our understanding of how spatiotemporal forces affect replication fidelity in prokaryotic organisms, investigation of replication in early replication regions (closer to the ORI) and late replication regions (closer to the TERM) differ in site-specific mutation rates. Prior unpublished work has shown mutation rate elevation in late replicating regions of *A. tumefaciens*, driven by excess G/C to A/T transition mutations, leading to several biological consequence (54). One observation from these results could lead to an increased drive in mutational burden in the absence of selection, causing an overall reduction in fitness within the organism. Another possible consequence could

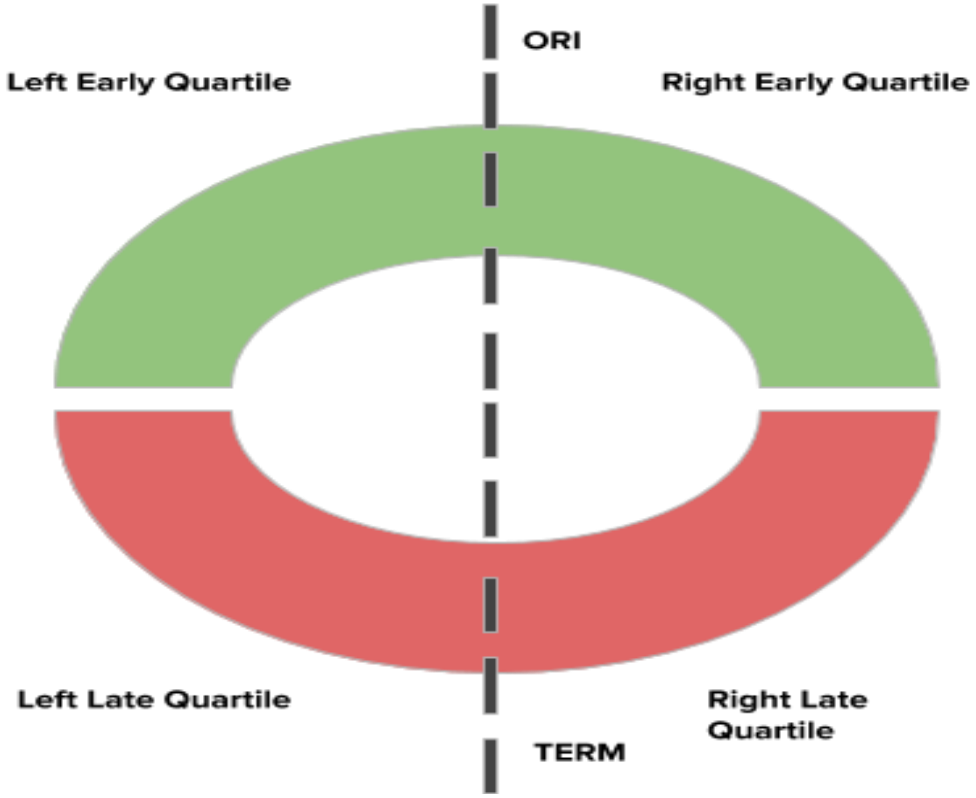



*Figure 26 - CDMAP Quartile Analysis. In order to extend analysis to investigate late replication region within organisms, replichores previously constructed by CDMAP in existing steps would be used to create hemispherical partitions representing the early and late replication regions within a given organism.*

involve possible genomic architecture rearrangements within the chromosome by moving larger,

essential genes to either earlier in the replication process on the chromosome, or onto secondary chromosomes and chromids in the case of multi-chromosomal organisms.

The process for tackling analysis of late replication analysis in CDMAP would ideally be fairly straightforward. Since the pipeline already creates hemispherical partitions for chromosome wide, and replichore-specific analyses with respect to the ORI, then extending our framework into a quartile-based partition and constructing and constructing an early (H_early) and late (H_late) replication hemisphere for analysis as shown in figure 25. Afterwards, downstream steps would merely need to be modified to incorporate H_early and H_late into the pre-existing framework of analysis. As a result, we would be able to investigate genome-wide and replichore specific contextual patterns across a wide variety of prokaryotic organisms in both early and late replication regions and expand the scope of our investigation within our current framework. While being straightforward, we would need to devote a non-trivial amount of time to ensure a new fundamental feature of analysis in CDMAP would meet the same quality standards held previously to CDMAP-SOA and CDMAP-MOA.

**REFERENCES:**


1. C. F. Baer, M. M. Miyamoto, D. R. Denver, Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet* **8**, 619-631 (2007).
2. I. Gordo, L. Perfeito, A. Sousa, Fitness Effects of Mutations in Bacteria. *Journal of Molecular Microbiology and Biotechnology* **21**, 20-35 (2011).
3. K. Heilbron, M. Toll-Riera, M. Kojadinovic, R. C. MacLean, Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* **197**, 981-990 (2014).
4. N. Keith *et al.*, High mutational rates of large-scale duplication and deletion in Daphnia pulex. *Genome Res* **26**, 60-69 (2016).
5. W. Wei *et al.*, Mutation Landscape of Base Substitutions, Duplications, and Deletions in the Representative Current Cholera Pandemic Strain. *Genome Biol Evol* **10**, 2072-2085 (2018).
6. C. J. Shaw, J. R. Lupski, Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* **13 Spec No 1**, R57-64 (2004).
7. W. Sung *et al.*, Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol Biol Evol* **32**, 1672-1683 (2015).
8. J. R. L. A. C. Frank, Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**, 560-561 (2000).
9. D. Trojanowski, J. Holowka, J. Zakrzewska-Czerwinska, Where and When Bacterial Chromosome Replication Starts: A Single Cell Perspective. *Front Microbiol* **9**, 2819 (2018).
10. N. V. Sernova, M. S. Gelfand, Identification of replication origins in prokaryotic genomes. *Brief Bioinform* **9**, 376-391 (2008).
11. J. W. Schroeder, W. G. Hirst, G. A. Szewczyk, L. A. Simmons, The Effect of Local Sequence Context on Mutational Bias of Genes Encoded on the Leading and Lagging Strands. *Curr Biol* **26**, 692-697 (2016).
12. T. G. Platt, E. R. Morton, I. S. Barton, J. D. Bever, C. Fuqua, Ecological dynamics and complex interactions of Agrobacterium megaplasmids. *Front Plant Sci* **5**, 635 (2014).
13. D. P. Jinshui Zheng, Lifang Ruan, and Ming Sun, Evolution and dynamics of megaplasmids with sizes larger than 100kb in the Bacillus Cereus group. *BMC Evolutionary Biology* 10.1186/1471-2148-13-262 (2013).
14. F. Fournes, M. E. Val, O. Skovgaard, D. Mazel, Replicate Once Per Cell Cycle: Replication Control of Secondary Chromosomes. *Front Microbiol* **9**, 1833 (2018).
15. A. Eyre-Walker, P. D. Keightley, The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**, 610-618 (2007).
16. A. F. Agrawal, M. C. Whitlock, Mutation Load: The Fitness of Individuals in Populations Where Deleterious Alleles Are Abundant. *Annual Review of Ecology, Evolution, and Systematics* **43**, 115-135 (2012).
17. W. Sung, Unpublished Agrobacterium proposal. *unpublished* (2015).
18. M. Lynch *et al.*, Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* **17**, 704-714 (2016).

19. H. Long *et al.*, Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient Pseudomonas fluorescens ATCC948. *Genome Biol Evol* **7**, 262-271 (2014).
20. S. Kucukyildirim *et al.*, The Rate and Spectrum of Spontaneous Mutations in Mycobacterium smegmatis, a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair Pathway. *G3 (Bethesda)* **6**, 2157-2163 (2016).
21. M. M. Dillon, W. Sung, M. Lynch, V. S. Cooper, The Rate and Molecular Spectrum of Spontaneous Mutations in the GC-Rich Multichromosome Genome of Burkholderia cenocepacia. *Genetics* **200**, 935-946 (2015).
22. A. B. Sahakyan, S. Balasubramanian, Single genome retrieval of context-dependent variability in mutation rates for human germline. *BMC Genomics* **18**, 81 (2017).
23. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Mol Biol Evol* **33**, 268-280 (2016).
24. W. Sung, M. S. Ackerman, S. F. Miller, T. G. Doak, M. Lynch, Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A* **109**, 18488-18492 (2012).
25. H. Lee, E. Popodi, H. Tang, P. L. Foster, Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A* **109**, E2774-2783 (2012).
26. N. Chatterjee, G. C. Walker, Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* **58**, 235-263 (2017).
27. A. Eyre-Walker, M. Woolfit, T. Phelps, The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**, 891-900 (2006).
28. W. S. Michael Lynch*†, Krystalynne Morris‡, Nicole Coffey*, Christian R. Landry§¶, Erik B. Dopman§, W. Joseph Dickinson , Kazufusa Okamoto‡, Shilpa Kulkarni‡, Daniel L. Hartl†§, and W. Kelley Thomas‡, A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**, 9722-9727 (April 22, 2008).
29. D. L. Halligan, P. D. Keightley, Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annual Review of Ecology, Evolution, and Systematics* **40**, 151-172 (2009).
30. M. D. Ermolaeva, Synonymous Codon Usage In Bacteria.
31. A. Iriarte, G. Lamolle, H. Musto, Codon Usage Bias: An Endless Tale. *J Mol Evol* **89**, 589-593 (2021).
32. X. X. Ma *et al.*, Comparative genomic analysis for nucleotide, codon, and amino acid usage patterns of mycoplasmas. *J Basic Microbiol* **58**, 425-439 (2018).
33. J. L. Lopez, M. J. Lozano, M. L. Fabre, A. Lagares, Codon Usage Optimization in the Prokaryotic Tree of Life: How Synonymous Codons Are Differentially Selected in Sequence Domains with Different Expression Levels and Degrees of Conservation. *mBio* **11** (2020).
34. J. L. de Oliveira *et al.*, Inferring Adaptive Codon Preference to Understand Sources of Selection Shaping Codon Usage Bias. *Mol Biol Evol* **38**, 3247-3266 (2021).
35. N. Galtier *et al.*, Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. *Mol Biol Evol* **35**, 1092-1103 (2018).

36. M. A. Gilchrist, W. C. Chen, P. Shah, C. L. Landerer, R. Zaretzki, Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. *Genome Biol Evol* **7**, 1559-1579 (2015).
37. P. M. Sharp, L. R. Emery, K. Zeng, Forces that influence the evolution of codon bias. *Philos Trans R Soc Lond B Biol Sci* **365**, 1203-1212 (2010).
38. R. Prabha, D. P. Singh, S. Sinha, K. Ahmad, A. Rai, Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. *Mar Genomics* **32**, 31-39 (2017).
39. J. B. Pragya Mittala, Julie Stephena, Joshua B. Plotkinb, and Grzegorz Kudlaa, Codon usage influences fitness through RNA toxicity. *Proc Natl Acad Sci U S A* **115**, E9753 (2018).
40. M. Dilucca, A. Pavlopoulou, A. G. Georgakilas, A. Giansanti, Codon usage bias in radioresistant bacteria. *Gene* **742**, 144554 (2020).
41. A. Hart, M. P. Cortes, M. Latorre, S. Martinez, Codon usage bias reveals genomic adaptations to environmental conditions in an acidophilic consortium. *PLoS One* **13**, e0195869 (2018).
42. T. T. K. M. Lynch, Estimate of the genomic mutation rate deleterious to overall fitness in E. coli. *Nature* **381**, 694-696 (1996).
43. K. Harris, Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci U S A* **112**, 3439-3444 (2015).
44. B. R. Morton, I. V. Bi, M. D. McMullen, B. S. Gaut, Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics* **172**, 569-577 (2006).
45. J. O. Elizabeth Louie, Jacek Majewski, Nucleotide Frequency Variation Across Human Genes. *Genome Research* (2003).
46. J. Carlson *et al.*, Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nat Commun* **9**, 3753 (2018).
47. K. Sau, S. K. Gupta, S. Sau, T. C. Ghosh, Synonymous codon usage bias in 16 Staphylococcus aureus phages: implication in phage therapy. *Virus Res* **113**, 123-131 (2005).
48. D. Charif, J. R. Lobry, "SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis" in Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, U. Bastolla, M. Porto, H. E. Roman, M. Vendruscolo, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007), 10.1007/978-3-540-35306-5_10, pp. 207-232.
49. D. Sarkar, *Lattice: Multivariate Data Visualization with R* (Springer, 2008).
50. N. Kono, M. Tomita, K. Arakawa, Accelerated Laboratory Evolution Reveals the Influence of Replication on the GC Skew in Escherichia coli. *Genome Biol Evol* **10**, 3110-3117 (2018).
51. P. D. JEANNE L. STOVE POINDEXTER, and GERMAINE COHEN-BAZIRE, Ph.D., THE FINE STRUCTURE OF STALKED BACTERIA BELONGING TO THE FAMILY CAULOBACTERACEAE. *The Journal of Cell Biology* (1964).
52. A. B. Mary Ann Moran, Jose´ M. Gonza´ lez, John F. Heidelberg, William B. Whitman, Ronald P. Kiene, James R. Henriksen, Gary M. King, Robert Belas, Clay Fuqua, Lauren Brinkac, Matt Lewis, Shivani Johri, Bruce Weaver, Grace Pai, Jonathan A. Eisen, Elisha

Rahe, Wade M. Sheldon, Wenying Ye, Todd R. Miller, Jane Carlton, David A. Rasko, Ian T. Paulsen, Qinghu Ren, Sean C. Daugherty, Robert T. Deboy, Robert J. Dodson, A. Scott Durkin, Ramana Madupu, William C. Nelson, Steven A. Sullivan, M. J. Rosovitz, Daniel H. Haft, Jeremy Selengut & Naomi Ward, Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment. *Nature* **432**.

53. K. S. Makarova *et al.*, Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev* **65**, 44-79 (2001).

54. Anonymous, <agro.pdf>.