Final Year Project

Mutation Bias in Pathogenic Bacteria: A computational Approach to Explore Mononucleotide Sequence Repeats as a Hotspot of Genetic Mutation

James Watson

Department of Biology and Biochemistry (Natural Sciences) | University of Bath | Claverton Down, BA2 7AY, United Kingdom | jw3528@bath.ac.uk

May 2024

Word-count: 5992

TABLE OF CONTENTS

LAY SUMMARY

Every living organism undergoes mutations, which are changes in the nucleotide sequence within DNA. While many mutations have no effect, some can influence an organism's ability to thrive under different environmental conditions. In other words, some mutations give rise to a phenotypic advantage which heightens the likelihood of an organism's survival, thereby increasing the chance of that mutation being passed on upon replication (for bacteria) or reproduction. This is the process of natural selection, which drives evolution. There is a long-standing belief that these types of mutations causing evolution occur randomly. However, there is now increasing evidence showing that certain types of mutations and specific regions of DNA can mutate much more frequently (mutational hotspots) or, alternatively, mutate much less (mutational coldspots) than the genomic average. These are called mutational biases. If mutation within these biases can give rise to a phenotypic advantage and improve the organism's likelihood of survival and reproduction, then this long-standing belief of evolution being driven by random mutations would be challenged. In this study, we examine a specific nucleotide sequence within the DNA of pathogenic bacteria, which is known to contain a mutational hotspot. Consequently, by assessing the potential that mutation within this hotspot has to influence the survival of pathogenic bacteria, we shed light on whether this mutational hotspot influences bacterial evolution. Indeed, we found some indications that imply these hotspots have the potential to influence evolution of pathogenic bacteria. However, these findings are rudimentary and further work is required to verify these claims. Additionally, we also suggest these sequences of DNA may become useful tools when predicting future evolution paths of pathogenic bacteria. In turn, this would help us develop strategies to block harmful strains of pathogens from emerging and infecting us.

ABSTRACT

Mutational biases can introduce fluctuations in mutation rates across a genome. As such, certain types of mutations and specific genomic positions often exhibit higher tendencies to produce functional consequences that may, in turn, alter the spectrum of cell fitness. Recent studies continue to underscore the role mutational biases can play in influencing adaptive trajectories of bacteria, and hence, the ability to understand the mechanisms underlying these biases is critical to enhance our abilities to predict evolutionary outcomes. However, it is unclear if the presence of hotspot biases are influenced by selection. In this study, we utilise a fully computational approach involving R scripting to examine a genomic sequence motif known to boast a mutational hotspot. Specifically, we targeted mononucleotide sequence repeats of guanine preceding thymine (GnT) with respect to a pathogenic lab (PAO1) and clinical isolate strain (B136-33) of *Pseudomonas aeruginos*a. In turn, we have attempted to assess the potential these motifs have to induce functional consequences upon their mutation, and shed light on whether these hotspots can synergise with

selection. Our findings not only imply that GnT hotspots have the potential to alter cell fitness, but further, appear to be subject to purifying selection across various functional gene classes. Conversely, we observe a lack of significant GnT motif depletion for genes within the antibiotic resistance and susceptibility (AMR) class. Consequently, we conclude that if further studies confirm GnT hotspot motifs can indeed alter cell fitness within the AMR class, then these motifs may be found in high enough frequencies for their utilisation in forecasting evolutionary trajectories of AMR genes. Thus, if successful, we pave the way for the development of strategies to mitigate the harmful adaptive outcomes pathogenic bacteria exhibit across evolution.

INTRODUCTION

The ever-emerging evidence underscoring the importance of mutational biases has fundamentally challenged the traditional view of mutations as being a stochastic and directionless driving force in evolution (Horton and Taylor, 2023). Implication wise, this suggests genetic mutation, the fundamental underlying's of evolutionary adaptation, may not be an entirely random force (Horton and Taylor, 2023). Instead, it has become evident that mutations are subject to strong preferences, in which certain types of mutations and specific genomic positions are more prone to change than others (Horton and Taylor, 2023; Cherry, 2023; Horton *et al*., 2021; Monroe *et al.,* 2022). Further, these biases are not uniformly distributed across the genome (Horton *et al*., 2021), where certain loci, such as mononucleotide sequence repeats, have a tendency to boast these biases (Horton and Taylor, 2023; Cherry, 2023). Consequently, the existence of biases induces fluctuations in mutation rates across a genome, thereby resulting in mutational hotspots, where mutagenicity at a certain genomic position is significantly heightened compared the genomic average, and mutational coldspots, where mutation rate is supressed compared to the genomic average (Horton and Taylor, 2023). Mutation bias has shown to play a crucial role in shaping the adaptive landscape of bacteria, influencing their evolutionary trajectories in both laboratory and clinical environments (Horton *et al.,* 2021; Moxon *et al.,* 2006; Cano *et al*., 2023; Payne *et al*., 2019; Rokyta *et al*., 2005; MacLean *et al*., 2010; Couce *et al*., 2015; Sackman *et al*., 2017; Stoltzfus *et al*., 2017; Storz et al*.,* 2019) . Hence, much attention has been brought into the realm of mutational bias, as understanding the mechanisms and environmental factors that underpin them is pivotal in utilising these biases for predicting evolutionary outcomes in pathogenic bacteria (Lässig *et al*., 2017; Cano *et al*., 2022; Franke *et al*., 2011; Stern *et al*., 2009).

On the notion of mutational coldspots, recent studies in organisms such as *Arabidopsis thaliana* demonstrate that within functionally critical regions of the genome, mutations have a tendency to be suppressed (Monroe *et al*., 2022). In turn, this would imply a bias that reduces the incidence of potentially deleterious

mutations, thereby challenging the long-standing beliefs of mutation as an entirely random force in evolution. Furthermore, a study by Horton *et al*. (2021) in *Pseudomonas fluorescens* underscored the impact mutational hotspots could have on repeatable evolutionary dynamics. Through building and breaking a mutational hotspot through silent genetic changes, Horton found adaptation was reliant on this hotspot. Specifically, this work revealed that local genetic sequence boasted the ability to steer and ensure consistent evolutionary trajectories across populations, thereby underscoring the previously overlooked importance of mutational biases. In addition, Horton and Taylor (2023) along with Cherry (2023) have emphasised the role of mononucleotide sequence repeats as a hotspot of genetic mutation. During DNA replication, these repeats often induce DNA polymerase slippage and unfaith base-pairing, thereby significantly heightening mutation rate (Horton and Taylor, 2023; Zhou *et al*., 2004).

In particular, recent work by Cherry (2023) has underscored the significance of guanine (G) mononucleotide sequence repeats as hotspots of genetic mutation. These repeats, often referred to as 'G tracks' or 'G runs', were found to significantly heighten the mutagenicity of thymine (T) to guanine base substitutions to rates vastly exceeding the genomic average when the G track precedes thymine. In discussions with Horton, we learned of his yet-to-be-published research that builds upon Cherry's findings by examining the mutagenicity associated with varying lengths of guanine tracks preceding thymine (e.g., GGGT (G3T), GGGGGT (G4T), GGGGGGT (G5T)). We have coined the term 'GnT motifs' to describe these sequences, where 'n' indicates the length of the G track. With permission from Horton, his preliminary results suggest that the mutagenic potential of the 3' T → G base substitution within GnT motifs varies according to the length of the G run and the specific nucleotides flanking the motif. These insights are concisely presented in Figure 1.

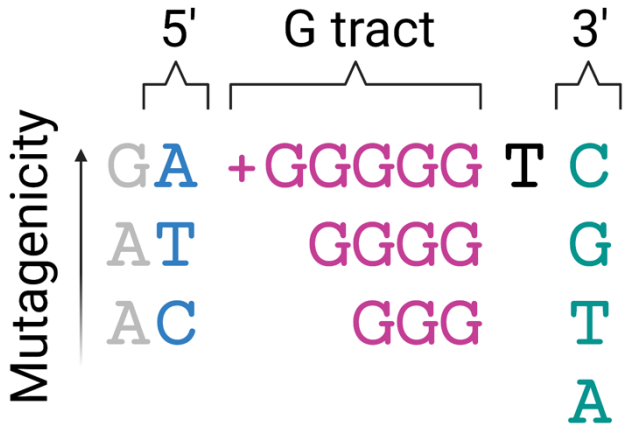


Figure 1 | Effect of the monoguanine repeat length and nucleotides flanking the GnT motif on the mutagenicity of the 3' terminal thymine within the GnT motif.

Figure reproduced with permission from Horton et al.

Armed with this knowledge, our research has taken a computational dive into the genomic sequences of pathogenic bacteria. Through leveraging data from the National Center for Biotechnology Information (NCBI) and employing R scripts, we have meticulously analysed GnT motifs as hotspots of genetic mutation within a laboratory (PAO1) and clinical isolate (B136-33) strain of *Pseudomonas aeruginos*a. We have, in light of their abundance and context, assessed GnT hotspot effects on coding regions of PAO1, their implication for protein tertiary structure, and evaluated their distribution across various functional genes classes. While hotspot motifs have shown to have a potent influence on evolution under some adaptive scenarios (Horton *et al.,* 2021), it is unclear if the presence of GnT hotspot motifs are influenced by selection. Subsequently, we have attempted to shed light on the possible selection pressures that may operate on GnT hotspot motifs by examining the potential they have to induce functional consequences which, in turn, would alter cell fitness. Primed with these foundations, our findings along with subsequent future studies will expand our understanding of how local genomic sequence, particularly GnT motifs, can drastically alter mutation rate, thereby paving the way to determine if this unique genetic architecture can underpin evolutionary dynamics in pathogenic bacteria.

Hence, our approach not only examines the potential synergy between GnT hotspot motifs and selection, but also opens avenues for the development of strategies to combat pathogenic bacteria by anticipating and potentially manipulating their harmful evolutionary trajectories.

MATERIALS AND METHODS

The complete R script for this research paper can be found here: https://github.com/jameswatsonn1/FYP-Mutation-Bias-in-Pathogenic-Bacteria.git

BACTERIAL DATA

All genomic sequences (FASTA) and annotation feature (GFF) datasets were harvested from the NCBI database and imported into R. Genomes utilised consisted of ASM676v1 (*Pseudomonas aeruginosa* (strain: PAO1)) and ASM35950v1 (*Pseudomonas aeruginosa* (strain: B136-33)). For ease of computational analysis, genomic and genomic coding sequences from FASTA file formats were read as a single string using the '*readDNAStringSet*' function from the '*Biostrings*' package.

CALCULATING ABUNDANCE OF MONONUCLEOTIDE SEQUENCE REPEATS AND GnT MOTIFS

In our analysis, we quantified the abundance of mononucleotide sequence repeats within the PAO1 genome, specifically targeting motifs from di-nucleotide to pentanucleotides (i.e., XX to XXXXX, where X denotes one nucleotide, either adenine (A), cytosine (C), guanine (G), or thymine (T)). Utilising the '*gregexpr*' function across a predefined list of motifs, these motifs were counted by identifying their starting positions within the sequence string and excluding any non-match indicators. The results for each base were then stored in separate vectors using the '*sapply*' function. These results were then aggregated into a single data frame to provide a suitable format for plotting. For comparative visualisation, we employed the '*ggplot2*' package to generate a bar plot to compare the frequency of differing lengths of mononucleotide sequence repeats for each base.

We then focused on the prevalence of GnT motifs, targeting trinucleotide to pentanucleotide runs of G and considering all combinations of nucleotides flanking the GnT motif. Utilising the '*gregexpr*' function, we counted the occurrence of these GnT motifs for each genome, and visually presented the data as separate bar plots for triplet, quadruplet, and pentaplet GnT motifs using the '*ggplot2*' package.

EXTRACTING MUTATED AMINO ACID SEQUENCE

Our next steps were to investigate the consequence of the T → G base substitution within the most mutable GnT motif, AGGGGGTC, on amino acid sequence.

The first step involved identifying the locus tags of genes containing AGGGGGTC motifs. To achieve this, we began utilising the '*matchPattern*' function from the '*Biostrings*' package to identify all positions of

AGGGGGTC motifs within the genomic sequence string. We then used the '*Granges*' function from the '*GenomicRanges*' package to determine the genomic range for each position of the motif, accounting for the entire run of the motif, before applying the '*findOverlaps*' function from *'GenomicRanges*' to see if the motif genomic range overlaps with any genes from the GFF data. Finally, using the '*mcols*' function from *'GenomicRanges*', we extracted all locus tags and gene names (if provided) containing this motif. Then, using data available from NCBI, we filtered out only coding genes and split these into two classes based on whether they occurred on the leading or lagging stand.

The next step was to determine the type and position of amino acid mutation induced upon a T → G base substitution within these genes identified, focusing on the leading strand genes to begin with. To achieve this, we first built the function *'translateSequence*' to translate nucleotide sequence into amino acid sequence, which ensures correct positioning of codons. Then, using '*substr*', '*gsub*', and our function '*translateSequence*', we extract the original nucleotide sequence of the predefined list of genes, introduce the T → G mutation within the AGGGGGTC motif, followed by translation of the original and mutated nucleotide sequences to the original and mutated amino acid sequences, respectively. Finally, we utilise an additional function we built, '*findAminoAcidMutations*', to determine the type of amino acid substitution and the position within the amino acid sequence which this occurs. To repeat this for lagging strand genes, we utilised a slightly modified approach that includes an additional function we built to handle lagging strand sequences, namely '*reverseComplementSequence*', which reverses the nucleotide sequence of these genes and replaces each nucleotide with its complement (A ↔ T; G ↔ C). The script conducts these operations in the following order: Introduce mutation in nucleotide sequence → reverse complement → translate to amino acid sequence.

In this way, we not only obtain the type and position of amino acid mutation as a result of the T → G substitution within AGGGGGTC motifs, but additionally, extract the full mutated amino acid sequence for every gene containing the motif, allowing for further study into GnT mutagenicity effect on protein structure.

PROTEIN STRUCUTRE DETERMINATION

Primed with these foundations, we were keen to explore the effect of the highly mutable AGGGGGTC motif on protein tertiary structure, specifically, the locational consequence of the amino acid mutation. To achieve this, the mutated amino acid sequence encoded via each gene originally containing a AGGGGGTC motif was imported into the SWISS-MODEL from the *Swiss Institute of Bioinformatics* (Waterhouse *et al*, 2018;

Bienert *et al*., 2017; Guex *et al*., 2009; Studer *et al*., 2020; Bertoni *et al*., 2017), a fully automated online tool for protein structure homology modelling, available from the *Expasy* web server (https://www.expasy.org/). Primed with these structures along with the type and position of the mutation, the location of the mutation with respect to secondary structure elements was examined through manual inspection (figure 3). To determine the significance of the results from figure 3, we utilised a statistical bootstrap technique over 100,000 iterations to generate an expected distribution of randomly sampled outcomes. Specifically, for each iteration, we randomly sampled 30 outcomes based on the predefined probabilities (weightings). We then determined the proportion of bootstrap samples that exhibit mutation counts as more extreme than the observed counts, thereby yielding P-values.

DISTRIBUTION OF GnT MOTIFS ACORSS FUNCTIONAL GENE CLASSES

Finally, we wanted to determine whether GnT motifs were acting under pressure from selection. To accomplish this, we examined of the distribution of GnT motifs across particular functional gene classes, specifically targeting G4T motifs (i.e., GGGGT) with all combinations of flanking nucleotides considered. The most mutable of the GnT motifs, G5T, were not considered since their lack of occurrences across the genome introduce uncertainty and inaccuracy when drawing conclusions surrounding their distribution. For each gene class of interest, we extracted all genes belonging to a class from *PseudoCAP* classification from the *Pseudomonas Genome Databank* (https://www.pseudomonas.com/) and imported them into R as a CSV file. Utilising previous methodology, genes containing G4T motifs across the entire genome were identified through their locus tags. Subsequently, we compared these G4T-containg genes with the list of genes from the class of interest, and using the '*intersect*' function, we returned and stored the count of those genes appearing in both datasets. We then divided the result by the total number of genes in the class we were investigating to generate the probability of a gene within that class of containing a G4T motif. Upon repeating for each gene class of interest, along with creating a suitable data frame of the results, a bar chart was plotted for comparative visualisation through '*ggplot2*' (figure 4A/5A).

In order to verify the significance of these results from figure 4A/5A, we employed a 'bootstrap' technique. Specifically, our script utilises the *Monte Carlo Simulation*. This involves a three-step process: 1) Randomly select 311 genes (the mean average of total genes across our functional classes of interest) without replacement from the annotation feature file (GFF) along with their genomic ranges. 2) Extract the nucleotide sequence of the genes from the sequence string, before concatenating these sequences into a single string to form a 'dummy' dataset. 3) Count the occurrences of G4T motifs within this sequence utilising the same technique previously used. This whole operation was then repeated for 10,000 iterations,

and an average number of G4T motifs across the iterations was calculated to form a null distribution. All values within this distribution were then divided by 311 and multiplied by 100 to obtain the probabilities of finding a G4T motif within a gene. We then represented this distribution through a histogram (figure 4B/5B) annotated with a normal density line. In addition, we also plotted the probability of a gene within each class of interest of containing a G4T motif, as calculated from figure 4A/5A. Our final step was to conduct one-sided Z-tests to determine if the results from figure 4A/5A were statistically significant. For ease of computational analysis, the null distribution was converted from percentage format to decimals before Z-tests were carried out.

RESULTS

ABUNDANCE OF MONONUCLEOTIDE SEQUENCE REPEATS AND GnT MOTIFS

Our first dive into the exploration of mutation bias involved quantifying the abundance of various types of mononucleotide sequence repeats within the PAO1 genome. Specifically, we began targeting dinucleotide to pentanucleotide runs of G, C, A, and T, before moving onto known motifs to boast a mutational hotspot. These hotspot motifs included trinucleotide to pentanucleotide runs of G preceding a T, with all combinations of flanking nucleotides considered (i.e., $X_1(G)_nTX_2$, where $X_{1,2}$ represent any nucleotide (although $X_1 \neq G$ to preserve the G run) and n corresponds to the length of the G track). Herein, the collated group of these hotspot motifs are abbreviated to GnT. These findings are presented in figure 2 below. Figure 2A informs us that homopolymeric tracks of C and G occur in much higher frequencies throughout the PAO1 genome than A and T tracks. Diving into hotspot motifs, specifically G4T motifs (i.e., XGGGGTX), we observe approximately 2800 cases of these motifs (figure 2C). In addition, we also find that the nucleotides flanking GnT hotspot motifs are more frequently C and G compared to A and T (figure 2B, C, D). To contextualise these results, we calculated the frequency of each nucleotide within the PAO1 genome. As expected, we observe C/G content to be significantly higher than A and T (C = 33.6%; G = 33.0%; A = 16.9%; T = 16.6%), and hence, this is likely what's driving the increased occurrence of C and G homopolymeic tracks and C/G flanking hotspot motifs. Due to the abundant frequency of GnT motifs, this represents the high potential for mutational hotspots within this genome, thereby making PAO1, and *Pseudomonas aeruginosa* more generally, ideal candidates to study GnT motifs as hotspots of mutation. While this data shows GnT hotspot motifs are common, genome-wide resolution is likely insufficient to quantify if selection operates on these motifs since we do not expect selection to operate on mutational hotspots equally throughout the genome. For-example, certain gene classes, such as essential genes, act

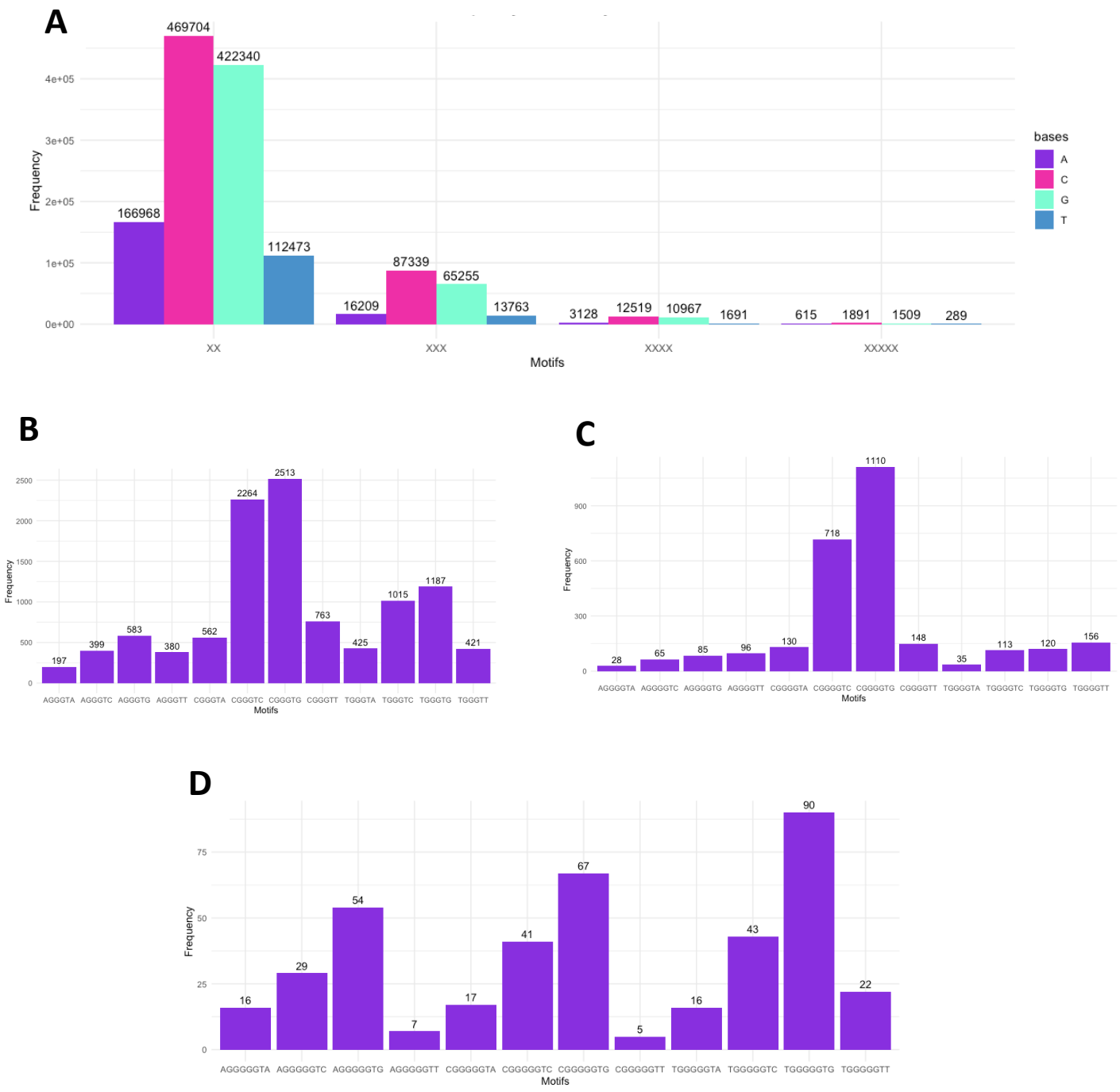


Figure 2 | Abundance of mononucleotide sequence repeats within coding regions of PAO1. |(A) Frequency comparison of mononucleotide repeats for each of the bases, A, C, G and T. These repeats consist of motifs XX, XXX, XXXX and XXXXX, where X denotes a particular base through the colour scheme as shown: Purple (A); pink (C); cyan (G); blue (T) **| (B, C, D)** Frequency comparison of GnT hotspot motifs. **| (B)** Abundance of G3T motifs (XGGGTX), with all combinations of flanking nucleotides considered. **|(C)** Abundance of G4T motifs (XGGGGTX), with all combinations of flanking nucleotides considered. **| (D)** Abundance of G5T motifs (XGGGGGTX), with all combinations of flanking nucleotides considered.

under stronger pressures from selection compared to non-essential genes (Monroe *et al*., 2022). Therefore, to explore the potential selection pressures operating on hotspot motifs, we seek to explore regions of the genome where the hotspot T → G mutation at GnT motifs will have functional consequences that have the potential to alter cell fitness. Subsequently, we investigate GnT motifs with respect to coding regions of genomes and the effect of their hotspot mutation on protein tertiary structure.

GnT MOTIFS AND PROTEIN TERTAIRY STRUCUTRE

Having established GnT motifs are common within the PAO1 genome, we then wanted to determine whether hotspot mutations at the most mutable of GnT motifs, $A(G)_5TC$, would induce any functional consequences regarding protein tertiary structure. To achieve this, we began by identifying all coding genes containing $A(G)_5TC$ motifs. Subsequently, for each of these genes, we calculated the mutated amino acid sequence that is encoded upon the highly mutable T → G base substitution within the $A(G)_5TC$ motif. We found a total of 30 coding genes boasting $A(G)_5TC$ motifs across PAO1, of which 16 occurred on the leading strand and 14 on the lagging strand. We find that when $A(G)_5TC$ motifs are found within the coding sequence, they can only encode a finite number of amino acids, which frequently, leads to conserved amino acid mutations across the various encoding genes containing $A(G)_5TC$ motifs. For-example, of the 16 leading strand genes, 15 experience a valine → glycine mutation in their encoded amino acid sequences upon the T → G base substitution within the $A(G)_5TC$ nucleotide motif. The remaining gene, however, exhibit a synonymous amino acid mutation. In addition, of the 14 lagging strand genes, 10 experience a threonine → proline mutation, where the remaining 4 present an aspartate → alanine mutation in their encoded amino acid sequences upon the T → G base mutation. Indeed, this consistency would make GnT mutation highly predictable. However, we anticipate this consistency is likely representative of genomic organisation (i.e., restricted combinations of codon alignments across the motif), rather than pressures influenced by selection itself. Primed with these foundations, we then wanted to explore the distribution of these amino acid mutations with respect to protein tertiary structure. Utilising the SWISS-MODEL from the *Swiss Institute of Bioinformatics*, a fully automated online tool for protein structure homology modelling, we inputted our mutated amino acid sequences into this server to obtain protein tertiary structures that have been encoded via genes that underwent a T → G base substitution within their $A(G)_5TC$ motif. A select few of these structures are presented in figure 3 below (see appendices for all 30 structures). It's worth noting that all structures generated have high sequence identities and global model quality estimates (GMQE), and hence, we are confident in our structures ability to provide accurate representations. Through manual examination of our 30 structures, we observe 7 amino acid mutations occur in alpha helices, 4 in beta sheets, and the remaining within disordered or loop regions. Utilising a statistical bootstrap technique over 100,000 iterations,

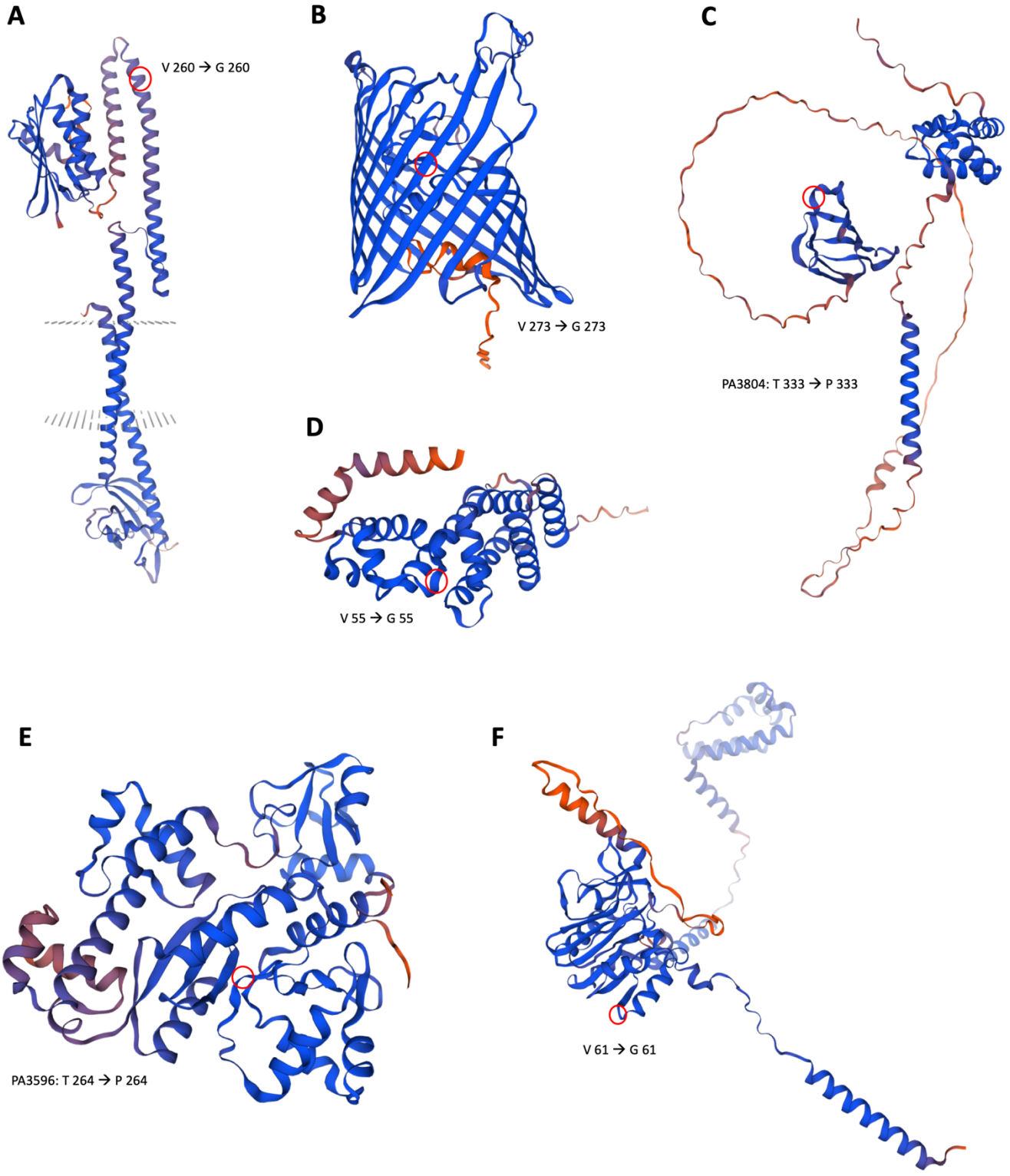


Figure 3 | SWISS-MODEL structure prediction of proteins encoded via PAO1 genes that are susceptible to T → G base mutations, of which are derived from the highly mutable AGGGGGTC nucleotide motif. Circled in red accompanied by the respective label indicates the locational consequence of amino acid substitution as a result of the T → G base mutation **| (A) PhoQ primed with a V260 → G260 mutation.** *Template for structure prediction: AlphaFold DB model of PHOQ_PSEAE (gene: phoQ, organism Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.78%, global model quality estimate (GMQE): 0.84.* **| (B) PA1974 primed with a V273 → G273 mutation**. *Template for structure prediction: Uncharacterized protein. AlphaFold DB model of A6V6J6_PSEA7 (gene: A6V6J6_PSEA7, organism: Pseudomonas aeruginosa, strain: PA7). Sequence identity: 96.24%, GMQE: 0.9.*

Figure 3 continued | (C) PA3804 primed with a T333 → P333 mutation. *Template for structure prediction: HTH cro/C1-type domain-containing protein. AlphaFold DB model of Q9HXJ3_PSEAE (gene: Q9HXJ3_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.71%, GMQE: 0.73.* **| (D) PA0828 primed with a V55 → G55 mutation.** *Template for structure prediction: Probable transcriptional regulator. AlphaFold DB model of Q9I5B1_PSEAE (gene: Q9I5B1_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.56%, GMQE: 0.86.* **| (E) PA3596 primed with a T264 → P264 mutation.** *Template for structure prediction: Probable methylated-DNA--protein-cysteine methyltransferase. AlphaFold DB model of Q9HY30_PSEAE (gene: Q9HY30_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.72%, GMQE: 0.88.* **| (F) PA4321 primed with a V61 → G61 mutation.** *Template for structure prediction: DUF4350 domain-containing protein. AlphaFold DB model of Q9HW80_PSEAE (gene: Q9HW80_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.76%, GMQE: 0.84.*

---

we find 4/30 mutations occurring in beta sheets and 19/30 in disordered/loop regions to both be statistically significant at the 2.5% and 5% significance levels (Bootstrap-test; $P_{values} = 0.0126; 0.00270$, respectively), where 7/30 occurring in alpha helices was a non-significant depletion (Bootstrap-test; $P_{value} = 0.432$). Hence, G5T-coordinated amino acid mutations have a tendency to be enriched in disordered/loop regions and depleted in beta sheets, where there is no bias for coordination in alpha helices. However, the validity of our bootstrap technique is questionable, since we assume each secondary structure element occurs in equal frequency across our structures. In turn, since we commonly observe less beta sheets and an abundance of disordered/loop regions within our structures, then this may trivially explain the distribution of mutations. Hence, we cannot conclude on the distribution of G5T-coordinated amino acid mutations with respect to secondary structure elements and selection effects. However, if our bootstrap findings are verified, then it's likely the increased occurrence of mutations within discorded/loop regions is consistent with the notion that, in general, mutations have less tendency to occur in alpha helices and beta sheets in order to conserve structural stability (Klosterman *et al*., 2006). Regardless, our analysis indicates that the location of the amino acid mutation is not solely confined to a particular structural domain or secondary structure element. In addition, these mutations occur within a wide diversity of proteins, including a range of sizes, functions, and cellular localisations. For-example, we observe G5T coordinated mutations to occur in proteins including transcriptional regulators, membrane proteins, transporters, a diversity of enzymes classes, and many more (see appendix). In turn, these results indicate the high potential G5T-coordinated amino acid mutations have to yield a myriad of functional consequences that may alter cell fitness.

However, while we have established the role G5T motifs play in protein structure, it remains unclear whether hotspots motifs are localised within particular functional gene classes which, in turn, would further enrich our understanding of the potential selection pressures operating on GnT motifs.

DISTRIBUTION OF GnT MOTIFS ACROSS FUNCTIONAL GENE CLASSES

Our final goal was to determine if GnT motifs are subject to pressure from selection. Our null hypothesis is that the frequency of GnT motifs shall be evenly distributed across the genome. Alternatively, if selection is acting upon GnT motifs, then we would expect particular gene classes to either be enriched or depleted in such motifs. To test this hypothesis, we examined the distribution of G4T motifs with all combinations of flanking nucleotides considered across various gene classes of PAO1, including two-component systems, transcriptional regulators, motility and attachment, cell wall/LPS/capsule, transport of small molecules, antibiotic resistance and susceptibility (AMR), and membrane proteins. Consequently, we extracted all genes belonging to each of these classes from the *Pseudomonas* database and calculated the probability of a gene within each of these classes of containing a G4T motif. These results are presented in figure 4A below. To contextualise these results, we first generated our null expected by randomly selecting 311 genes and calculating the frequency of G4T motifs within this group. Consequently, we could yield the expected chance of finding a G4T motif within a particular gene. The process was repeated 10,000 times to obtain the null distribution, represented as a histogram (figure 4B) and annotated with data from figure 4A. Interestingly, we observe all our functional gene classes of interest to contain less G4T motifs compared to the null average (figure 4B) in PAO1. Further, we find that two-component systems (Z-test; $p = 0.000263$), transcriptional regulators (Z-test; $p = 8.94e\text{-}08$), motility and attachment (Z-test; $p = 0.000105$), cell wall/LPS/capsule (Z-test; $p = 0.0173$), transport of small molecules (Z-test; $p = 0.0381$), and membrane proteins (Z-test; $p = 0.00140$) are all significantly depleted in G4T motifs compared to the null distribution at the 5% significance level, where AMR (Z-test; $p = 0.124$) was found to not differ significantly from the null distribution at the 2.5% significance level. Interestingly, at the 2.5% significance level, the transport of small molecules class shifts to become a non-significant depletion in G4T motifs compared to the null distribution. Hence, these results imply that in PAO1, G4T motifs are under pressure from purifying selection when found in two-component systems, transcriptional regulators, motility and attachment, cell wall/LPS/capsule, and membrane protein gene classes. Conversely, we find no evidence of purifying selection operating on G4T motifs in the AMR class, and it remains unclear if G4T motifs undergo pressure from selection in the transport of small molecules class.

Primed with these foundations, we then sought to explore how changes in environmental selection pressures would impact the distribution of GnT motifs across functional gene classes. Subsequently, we applied this same methodology as detailed above to a clinical isolate strain of *Pseudomonas aeruginosa,* B136-33 – a community-acquired hypervirulent pathogen inducing severe diarrhoea, sepsis, and enteritis in healthy infants (Lo *et al*., 2018). These results are presented in figure 5 below.

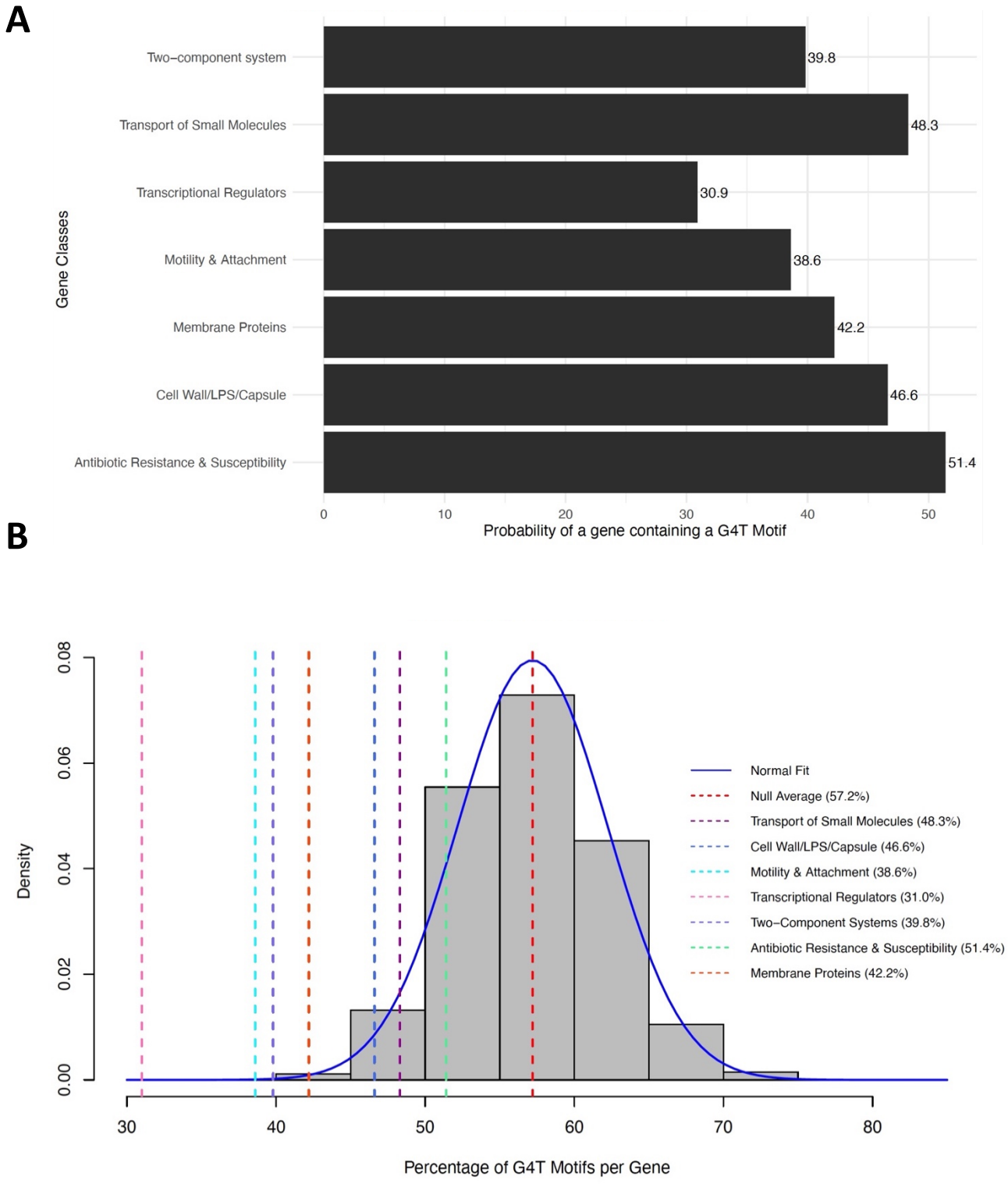


Figure 4 | (A) Probability of a gene within a particular functional class of containing a G4T motif in PAO1. These motifs can include any one of the following: A(G)$_4$TC, A(G)$_4$TG, A(G)$_4$TT, A(G)$_4$TA, T(G)$_4$TC, T(G)$_4$TG, T(G)$_4$TT, T(G)$_4$TA, C(G)$_4$TC, C(G)$_4$TG, C(G)$_4$TT, C(G)$_4$TA. **|(B) Null probability distribution of a gene containing a G4T motif in PAO1.** Annotated with the colour scheme as shown is the data from figure 4A to represent the depletion of G4T motifs in functional gene classes compared to the null average (red).

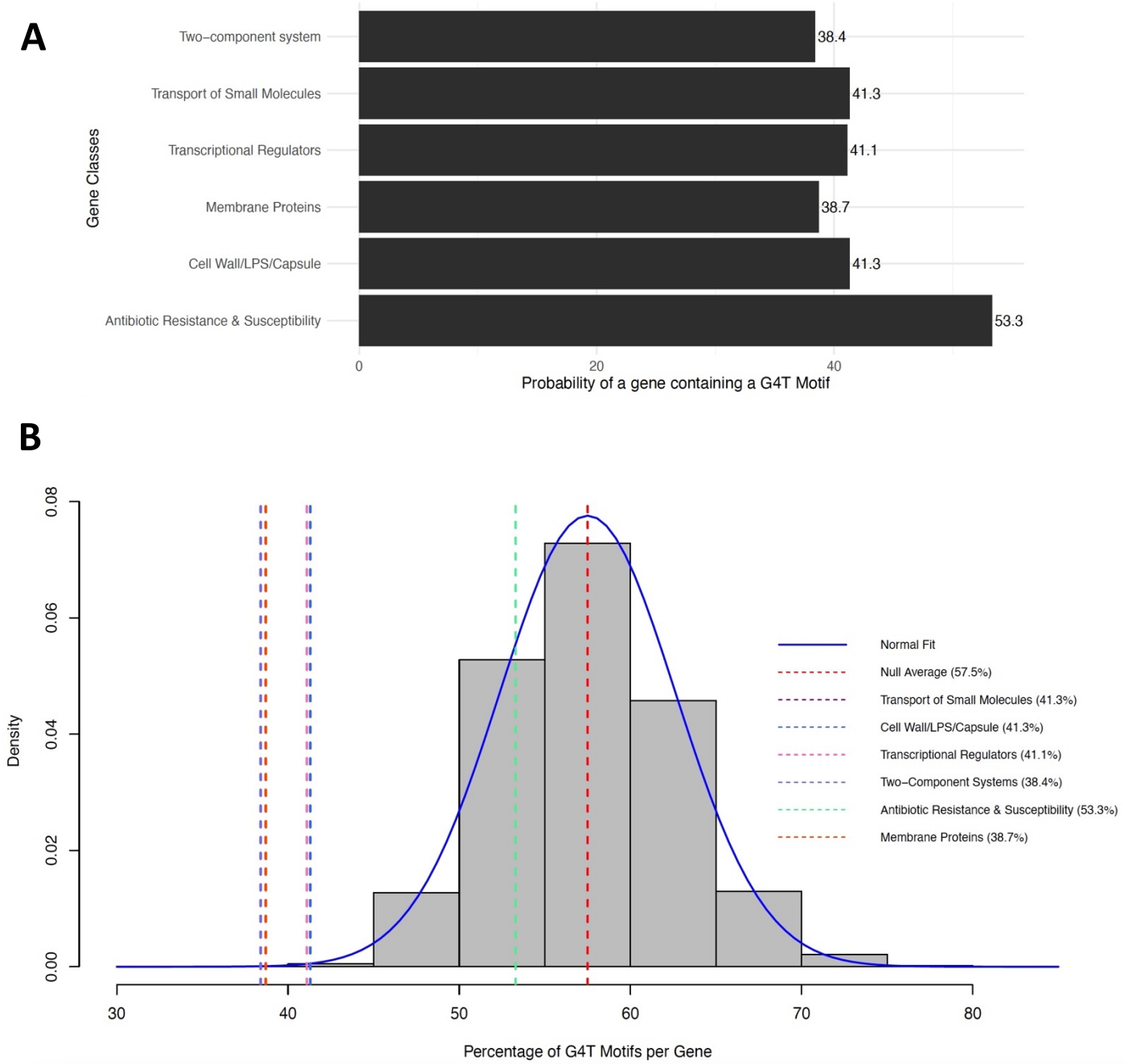


Figure 5 | (A) Probability of a gene within a particular functional class of containing a G4T motif in B136-33. These motifs can include any one of the following: $A(G)_4TC$, $A(G)_4TG$, $A(G)_4TT$, $A(G)_4TA$, $T(G)_4TC$, $T(G)_4TG$, $T(G)_4TT$, $T(G)_4TA$, $C(G)_4TC$, $C(G)_4TG$, $C(G)_4TT$, $C(G)_4TA$. **|(B) Null probability distribution of a gene containing a G4T motif in B136-33.** Annotated with the colour scheme as shown is the data from figure 5A to represent the depletion of G4T motifs in functional gene classes compared to the null average (red).

For B136-33, we observe a similar trend where all functional gene classes of interest are depleted compared to the null average. In this case, however, all gene classes of interest except AMR are significantly depleted in G4T motifs compared to the null distribution at the 2.5% significance level (two-component systems: Z-

test; p = 0.000102; transport of small molecules & cell wall/LPS/capsule: Z-test; p = 0.000814; transcriptional regulators: Z-test; p = 0.000712; membrane proteins: Z-test; p = 0.000128). Unfortunately, we were unable to analyse the motility and attachment class in B136-33 due to a lack of *PseudoCAP* functional classifications for this class. Further, as observed with PAO1, the AMR class experiences a non-significant depletion in G4T motifs at the 2.5% significance level (Z-test; p = 0.207), therefore further implying the lack of pressure G4T motifs face from selection within this class. Intriguingly, since the remaining classes of interest (two-component systems, transport of small molecules, cell wall/LPS/capsule, transcriptional regulators, and membrane proteins) are significantly depleted in G4T motifs to similar levels (figure 5), then there may be a requirement to equally distribute G4T motifs across these functional gene classes in B136-33.

DISCUSSION

In this study, we have attempted to shed light on the potential selection pressures that operate on GnT hotspot motifs. In addition to their abundance across the PAO1 genome, we have considered the effect of GnT hotspots on coding regions of PAO1, protein tertiary structure, and evaluated their distribution across various functional genes classes for a pathogenic lab (PAO1) and clinical isolate (B136-33) strain of *Pseudomonas aeruginosa.*

With respect to coding regions and protein structure within PAO1, we found G5T-coordinated amino acid mutations to occur in a wide diversity of proteins, including a range of sizes, functions, and cellular localisations. In addition, of all genes containing the highly mutable G5T motif, $A(G)_5TC$, we only observe one synonymous amino acid mutation upon the T → G base substitution within this hotspot motif. Thus, the notion that G5T-cooridnated mutations are not confined to a particular protein class and that nearly all genes containing this hotspot would experience a non-synonymous mutation, in turn, may underscore the potential G5T motifs have to coordinate a myriad of functional consequences and alter cell fitness.
Regarding secondary structure elements, we observe G5T-coordinated amino acid mutations to be significantly depleted in beta sheets, enriched in disordered/loop regions, and have no significant bias towards alpha helices. However, the validity of our statistical bootstrap technique to verify this significance is questionable, since we assume each secondary structure element occurs in equal frequency across our structures. In reality, since we commonly observe less abundance of beta sheets, this may trivially explain their depletion in mutations. Alternativity, the lack of beta sheets we observe may already be a consequence of selection, on average, inhibiting beta sheets in protein structures encoded via G5T-containing genes. However, if our bootstrap outcomes are verified, then the substantially higher rates of mutation we observe

within disordered/loop regions is likely consistent with the notion that, in general, mutations have a tendency to avoid alpha helices and beta sheets in order to maintain structural stability (Klosterman *et al*., 2006). Hence, as of current, we cannot conclude on the distribution of G5T-coordinated mutation with respect to protein tertiary structure and selection effects. We therefore propose examination of G4T-coordinated mutations with respect to protein tertiary structure should be considered as a next step. However, the computational strain associated with generating the vast abundance of protein structures that G4T-containing genes encode for would be challenging.

Regardless, our analysis indicates that G5T-coorindated amino acid mutations are not solely confined to a particular structural domain or secondary structure element. Thus, although we observe no common amino acid substitution hotspot, the findings that GnT hotspots are both frequent and can be found throughout coding sequences would therefore imply mutation within these motifs would have the potential to yield a myriad of functional consequences. However, due to the current limit of computational approaches, we remain to explore whether GnT-coordinated mutations could interfere with specific protein-protein or protein-ligand interactions that may, in turn, affect functionality. Hence, our view is that mutagenic studies involving crystal and cryo-EM structures of proteins encoded via genes containing GnT motifs is essential to assess the possible protein interactions that GnT hotspots may interfere with, thereby yielding further insight into whether GnT hotspots have the potential to alter cell fitness. Consequently, in an attempt to address the potential selection pressures that GnT motifs may be subject to, we dived more generally into the less mutable, but more frequently occurring G4T hotspot motifs and assessed their distribution across various functional gene classes.

In PAO1, we found signals of purifying selection significantly depleting G4T motifs in the following gene classes: two-component systems, transcriptional regulators, motility and attachment, cell wall/LPS/capsule, transport of small molecules, and membrane proteins. Intriguingly, although G4T motifs were depleted in the AMR class, this was found to be a non-significant depletion, and hence, it would appear G4T motifs are not subject to purifying selection pressures when found in this class. Further, we were unable to determine if G4T motifs are under pressure from purifying selection in the transport of small molecules class due to discrepancies in P values between the 2.5% and 5% significance levels. Similarly, for our clinical isolate strain, B136-33, we observed strong signals of purifying selection depleting G4T motifs in all gene classes of interest except the AMR class. Interestingly, since G4T motifs were depleted to similar levels across these classes, then one may assume a shift to a hostile environment has driven selection to be rigorous in depleting G4T motifs to a somewhat homeostatic level between numerous functional gene classes. However, further work would be required to determine as to why this might be the case. In general, evidence for the role selection may play here is lacking. Additionally, since our lab and clinical isolate strains have only diverged

recently in the evolutionary timeline, then it's likely the 'consistent' level of depletion observed in B136-33 is insignificant compared to the wider-spread depletion levels in PAO1. Hence, there is limited evidence to suggest purifying selection has driven numerous functional gene classes to be depleted to similar levels in B136-33.

Regardless, the notion of significant G4T depletion across an array of functional gene classes in both lab and clinical isolate strains would therefore imply these hotspot motifs have some functional consequences that would reduce cell fitness, and hence, purifying selection may ensure the potential deleterious G4T-mediated mutations are regulated in these functional gene classes. This is likely consistent with the notion that selection would attempt to filter out mononucleotide sequence repeats due to their costly effect on mRNA stability, their introduction of frame-shift mutations, and thus, reducing the fidelity of translation (Pernitzsch *et al*., 2014). Intriguingly, however, since purifying selection does not appear to act on G4T motifs in the AMR class in both our lab and clinical isolate strains of *Pseudomonas aeruginosa* (PAO1 and B136-33, respectively), then it's likely these hotspot motifs have limited functional consequences within this class. Consequently, this implies G4T motifs may not be suitable tools in forecasting evolutionary trajectories of AMR class genes. Conversely, if these motifs do boast some functional consequence, then this lack of depletion may mean these hotspot motifs are found in high enough frequencies that they can be utilised for predicting evolutionary outcomes of AMR genes. At this junction, it's critical to note that the lack of significant G4T depletion within the AMR class may be a product of hotspot enrichment in certain genes balanced with depletion within other genes of this class. This is a mechanism pathogenic bacteria have shown to utilise in order to unlock phase variation (Moxon *et al*., 2006; Moxon *et al*., 1994). That is, the co-existence of highly mutable 'contingency' genes alongside genes resilient to mutations, thereby allowing bacteria to explore adaptive outcomes in response to environmental changes while preserving essential functions that ensure fitness (Moxon *et al*., 1994). Thus, we don't ignore the possibility that certain genes within the AMR class may be enriched in G4T motifs to potentially facilitate phase variation, and hence, we urge next steps should seek to determine if such genes within the AMR class exist. In turn, these genes would be essential targets for evolutionary forecasting with respect to antibiotic resistance.

Finally, we acknowledge the possibility that the lack of G4T depletion within the AMR class may be due to lower expression levels of these genes, which overall, would reduce the costly effects of mononucleotide sequence repeats on translation for this class (Pernitzsch *et al*., 2014). Thus, this might alleviate the extent of the purifying selection pressures operating on G4T motifs within the AMR class.

Due to the lack of *PseudoCAP* functional classification for the motility and attachment class in B136-33, we missed a critical opportunity to analyse G4T distribution within this class. Since clinical isolate strains have likely evolved to evade invade immune response through suppression of genes within this class (Chaban *et al*., 2015), then we were unable to determine if G4T motifs have the potential to play any roles in immune

invasion. Finally, it's worth noting the findings that G4T motifs are significantly depleted in nearly all our functional gene classes of interest does raise an interesting question as to where hotspot motifs are predominately localised to. Thus, this is an area we seek to explore in future studies. Regardless, this G4T depletion will make it much more challenging to utilise such hotspot motifs for evolutionary forecasting if these motifs are to boast any functional consequences.

While we may have uncovered some signals of selection operating on GnT hotspot motifs, verifying and stitching these findings into the world of predicting adaptive phenotypes becomes highly challenging (Lind *et al*., 2019; Lässig *et al*., 2017). To do so, we require to know the fitness effects of specific mutations (Lind *et al*., 2019). Hence, next steps should seek to determine the potential functional consequences associated with various GnT motifs. To achieve this, we believe in-lab studies utilising a similar, but modified approach to Horton *et al*. (2021) should be conducted, where a targeted GnT hotspot is removed while preserving protein sequence, and the bacteria is then exposed to various environmental pressures that drive selection. Parallel to this, knock-in of the same GnT motif should be applied to a different strain at the same relative genomic position. Ultimately, the type of loss or gain function, if any, can be assessed with respect to the selection pressure in operation. Consequently, repetition across various functional gene classes would yield a wider view of the potential cell fitness effects and selection pressures that GnT hotspot motifs operate under. In turn, if GnT motifs are found to significantly alter cell fitness, then as described from before, these hotspots may provide pivotal value when predicting evolutionary trajectories of AMR class genes. This revelation echoes the findings of Cano *et al*. (2023), who also emphasize the importance of mutational bias in improving evolutionary forecasting, particularly for adaptation of infectious agents. On the contrary, since we found G4T motifs to be significantly depleted across an array of other functional gene classes, then utilising such hotspot motifs for predicting adaptive outcomes in these classes becomes more challenging.

Ultimately, further work as highlighted needs to be conducted to determine if GnT hotspot motifs have the potential to influence adaptation and synergise with selection. Indeed, there is an ever-increasing body of evidence documenting that specific mutational bias's influence genetic changes, which in turn, shape the spectrum of adaptation (Horton and Taylor, 2023; Horton *et al*., 2021; Cano *et al*., 2023; Payne *et al*., 2019; Rokyta *et al*., 2005; MacLean *et al*., 2010; Couce *et al*., 2015; Sackman *et al*., 2017; Stoltzfus *et al*., 2017; Storz et al*.,* 2019). Subsequently, this has major implications in the realm of predicting evolutionary outcomes (Lässig *et al*., 2017; Cano *et al*., 2022; Franke *et al*., 2011; Stern *et al*., 2009). Whether or not GnT motifs can exhibit this behaviour and coordinate evolutionary trajectories in pathogenic bacteria remains to be determined. Regardless, what we do provide in this paper is foundations for further studies to build on. We found G4T motifs are subject to purifying selection when located in numerous functional gene classes

within a lab and clinical isolate strain of *Pseudomonas aeruginosa*. Additionally, we emphasise the potential G4T motifs may provide to aid the evolutionary forecasting of AMR genes. Hence, subject to further investigations, we conclude by highlighting the potential G4T motifs may have to be useful tools in mitigating harmful antibiotic resistant strains of bacteria from developing.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr Tiffany Taylor for her continued support and guidance throughout this project. In addition, I would like to extend my gratitude to Dr James Horton, who oversaw my data collection and was pivotal to the success of this work.

REFERENCES

1. Horton, J.S., Taylor, T.B., 2023. Mutation bias and adaptation in bacteria. *Microbiology (Reading, England)*, 169(11). Available from: https://doi.org/10.1099/mic.0.001404. [Accessed 5 Feb 2024].

2. Cherry, J.L., 2023. T Residues Preceded by Runs of G Are Hotspots of T→G Mutation in Bacteria. *Genome biology and evolution*, 15(6). Available from: https://doi.org/10.1093/gbe/evad087. [Accessed 15 Mar 2023].

3. Horton, J. S., Flanagan, L.M., Jackson, R.W., Priest, N.K., & Taylor, T. B., 2021. A mutational hotspot that determines highly repeatable evolution can be built and broken by silent genetic changes. *Nature communications*, 12(1). Available from: https://doi.org/10.1038/s41467-021-26286-9. [Accessed 5 Feb 2024].

4. Monroe, J.G., Srikant, T., Carbonell-Bejerano, P. *et al.*, 2022. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602**, 101–105. Available from: https://doi.org/10.1038/s41586-021-04269-6. [Accessed 6 Apr 2024].

5. Moxon, R., Bayliss, C., & Hood, D., 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual review of genetics*, *40*, 307–333. Available from: https://doi.org/10.1146/annurev.genet.40.110405.090442. [Accessed 6 Apr 2024].

6. Lo, Y. L., Chen, C. L., Shen, L., Chen, Y. C., Wang, Y. H., Lee, C. C., Wang, L. C., Chuang, C. H., Janapatla, R. P., Chiu, C. H. and Chang, H. Y., 2018. Characterization of the role of global regulator FliA in the pathophysiology of Pseudomonas aeruginosa infection. *Research in microbiology* [Online], *169*(3). Available from: https://doi.org/10.1016/j.resmic.2018.02.001. [Accessed 25 Apr 2024].

7. Lind, P.A., Libby, E., Herzog, J. and Rainey, P.B., 2019. Predicting mutational routes to new adaptive phenotypes. *eLife* [Online], **8.** Available from: https://doi.org/10.7554/eLife.38822. [Accessed 28 Apr 2024].

8. Lässig, M., Mustonen, V. & Walczak, A., 2017. Predicting evolution. *Nat Ecol Evol* [Online], **1**(0077). Available from: https://doi.org/10.1038/s41559-017-0077. [Accessed 28 Apr 2024].

9. Cano, A.V., Gitschlag, B.L., Rozhoňová, H., Stoltzfus, A., McCandlish, D.M. and Payne, J.L., 2023. Mutation bias and the predictability of evolution. *Phil. Trans. R. Soc. B* [Online], 378. Available from: https://doi.org/10.1098/rstb.2022.0055. [Accessed 3 May 2024].

10. Cano, A.V., Rozhoňová, H., Stoltzfus, A., Payne, J.L., 2022. Mutation bias shapes the spectrum of adaptive substitutions. *Proc. Natl. Acad. Sci.* [Online], 119(7). Available from: https://doi.org/10.1073/pnas.2119720119. [Accessed 14 Apr 2024].

11. Payne, J.L., Menardo, F., Trauner, A., Borrell, S. and Gygli. S.M. et al., 2019. Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis. *PLOS Biology* [Online], 17(5). Available from: https://doi.org/10.1371/journal.pbio.3000265. [Accessed 3 May 2024].

12. Rokyta, D., Joyce, P., Caudle, S. *et al.,* 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat Genet* [Online], **37**. Available from: https://doi.org/10.1038/ng1535. [Accessed 1 May 2024].

13. MacLean, R. C., Perron, G. G. and Gardner, A., 2010. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in *Pseudomonas aeruginosa*. *Genetics* [Online], 186(4). Available from: https://doi.org/10.1534/genetics.110.123083. [Accessed 26 Apr 2024].

14. Couce, A., Rodríguez-Rojas, A. and Blázquez, J., 2015. Bypass of genetic constraints during mutator evolution to antibiotic resistance. *Proceedings. Biological sciences* [Online], 282(1804). Available from: https://doi.org/10.1098/rspb.2014.2698. [Accessed 28 Apr 2024].

15. Sackman, A. M., McGee, L. W., Morrison, A. J., Pierce, J., Anisman, J., Hamilton, H., Sanderbeck, S., Newman, C. and Rokyta, D. R., 2017. Mutation-Driven Parallel Evolution during Viral Adaptation. *Molecular biology and evolution* [Online], 34(12). Available from: https://doi.org/10.1093/molbev/msx257. [Accessed 15 Apr 2024].

16. Stoltzfus, A. and McCandlish, D. M., 2017. Mutational Biases Influence Parallel Adaptation. *Molecular biology and evolution* [Online], 34(9). Available from: https://doi.org/10.1093/molbev/msx180. [Accessed 16 Apr 2024].

17. Storz, J. F., Natarajan, C., Signore, A. V., Witt, C. C., McCandlish, D. M. and Stoltzfus, A., 2019. The role of mutation bias in adaptive molecular evolution: insights from convergent changes in protein function. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* [Online], 374(1777). Available from: https://doi.org/10.1098/rstb.2018.0238. [Accessed 16 Apr 2024].

18. Franke, J., Klözer, A., de Visser, J. A. and Krug, J., 2011. Evolutionary accessibility of mutational pathways. *PLoS computational biology* [Online], 7(8). Available from: https://doi.org/10.1371/journal.pcbi.1002134. [Accessed 10 Apr 2024].

19. Stern, D. L. and Orgogozo, V., 2009. Is genetic evolution predictable? *Science (New York, N.Y.)* [Online], 323(5915). Available from: https://doi.org/10.1126/science.1158997. [Accessed 25 Mar 2024].

20. Moxon, E.R., Rainey, P.B., Nowak, M.A. and Lenski, R.E., 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Current Biology* [Online], 4(1). Available from: https://doi.org/10.1016/s0960-9822(00)00005-1. [Accessed 9 May 2024].

21. Chaban, B., Hughes, H. V. and Beeby, M., 2015. The flagellum in bacterial pathogens: For motility and a whole lot more. *Seminars in cell & developmental biology* [Online], *46*. Available from: https://doi.org/10.1016/j.semcdb.2015.10.032. [Accessed 8 May 2024].

22. Zhou, Y., Bizzaro, J. W. and Marx, K. A., 2004. Homopolymer tract length dependent enrichments in functional regions of 27 eukaryotes and their novel dependence on the organism DNA (G+C)% composition. *BMC genomics* [Online], *5*(95). Available from: https://doi.org/10.1186/1471-2164-5-95. [Accessed 2 May 2024].

23. Pernitzsch, S. R., Tirier, S. M., Beier, D. and Sharma, C. M., 2014. A variable homopolymeric G-repeat defines small RNA-mediated posttranscriptional regulation of a chemotaxis receptor in Helicobacter pylori. *Proceedings of the National Academy of Sciences of the United States of America* [Online], *111*(4). Available from: https://doi.org/10.1073/pnas.1315152111. [Accessed 11 May 2024].

24. Klosterman, P.S., Uzilov, A.V., Bendaña, Y.R. *et al.,* 2006. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* [Online], **7**(428). Available from: https://doi.org/10.1186/1471-2105-7-428. [Accessed 11 May 2024].

25. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* [Online], *46*(W1). Available from: https://doi.org/10.1093/nar/gky427/ [Accessed 25 March 2024].

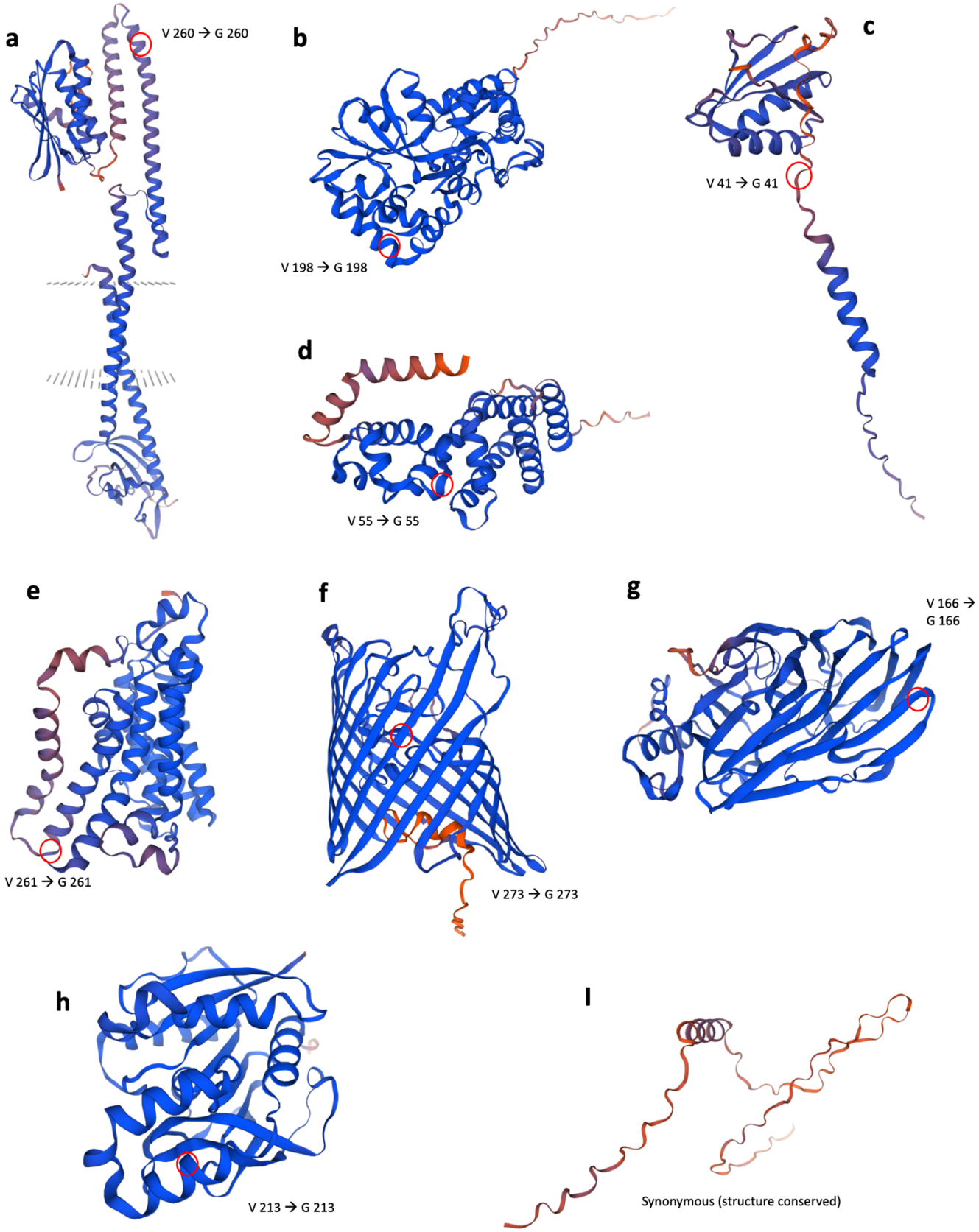
26. Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L. and Schwede, T., 2017. The SWISS-MODEL Repository-new features and functionality. *Nucleic acids research* [Online], *45*(D1). Available from: https://doi.org/10.1093/nar/gkw1132. [Accessed 25 March 2024].

27. Guex, N., Peitsch, M. C. and Schwede, T., 2009. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* [Online], *30 Suppl 1*. Available from: https://doi.org/10.1002/elps.200900140. [Accessed 25 March 2024].

28. Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J. and Schwede, T., 2020. QMEANDisCo-distance constraints applied on model quality estimation. *Bioinformatics (Oxford, England)*, *36*(6). Available from: https://doi.org/10.1093/bioinformatics/btz828. [Accessed 25 March 2024].

29. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. and Schwede, T., 2017. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Scientific reports* [Online], *7*(1). Available from: https://doi.org/10.1038/s41598-017-09654-8. [Accessed 25 March 2024].

APPENDIX 1


a
V 260 → G 260
b
V 198 → G 198
c
V 41 → G 41
d
V 55 → G 55
e
V 261 → G 261
f
V 273 → G 273
g
V 166 → G 166
h
V 213 → G 213
I
Synonymous (structure conserved)

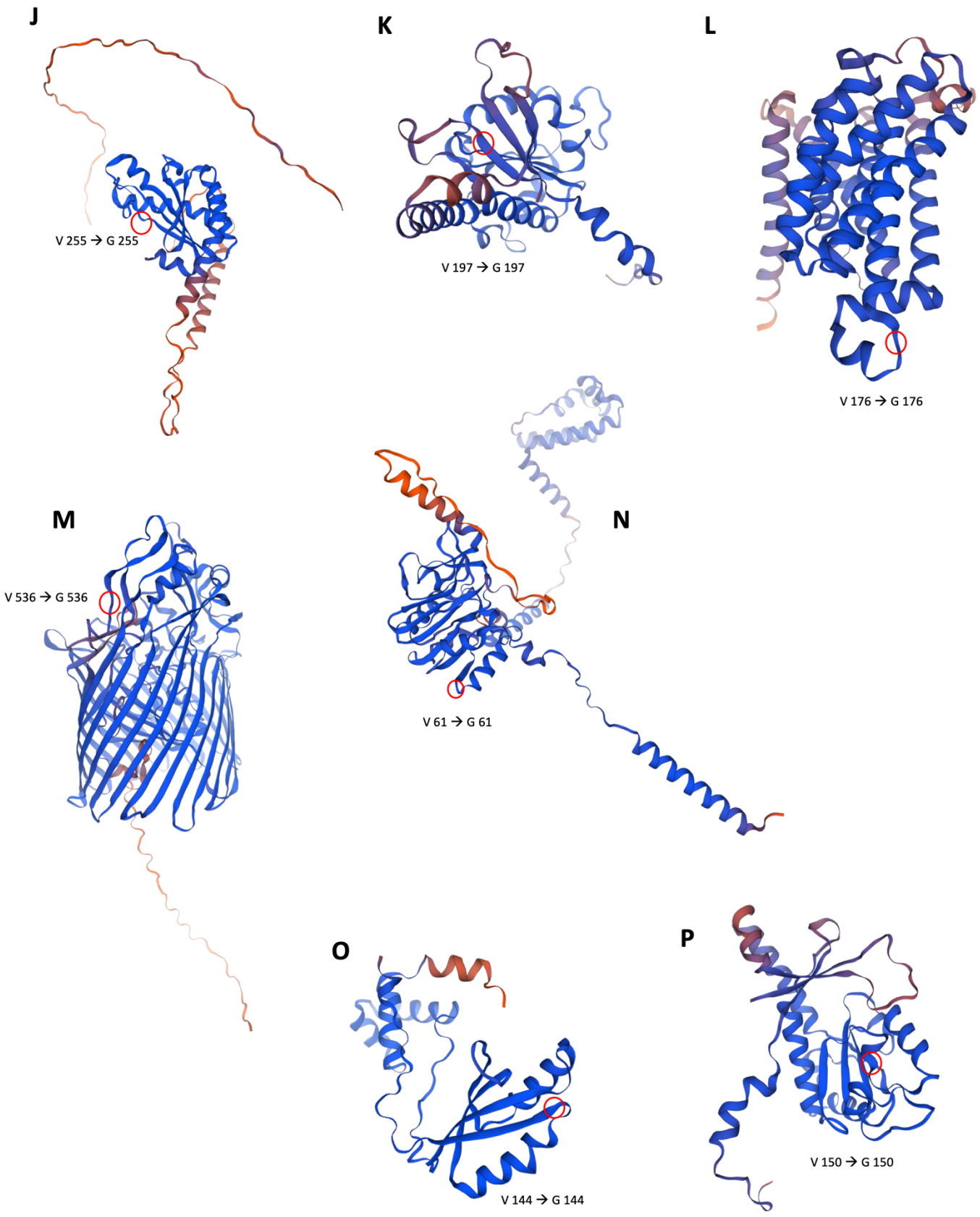
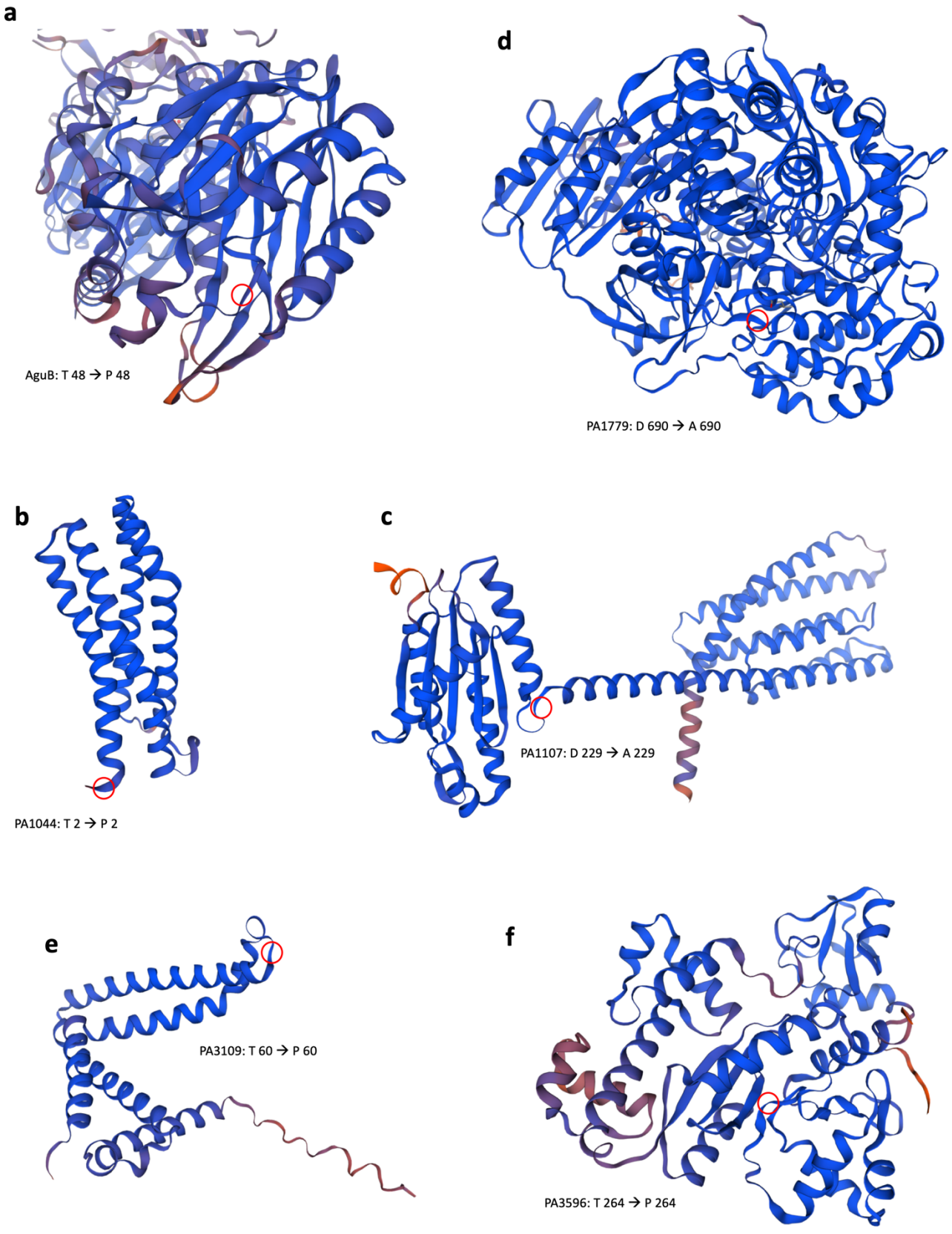
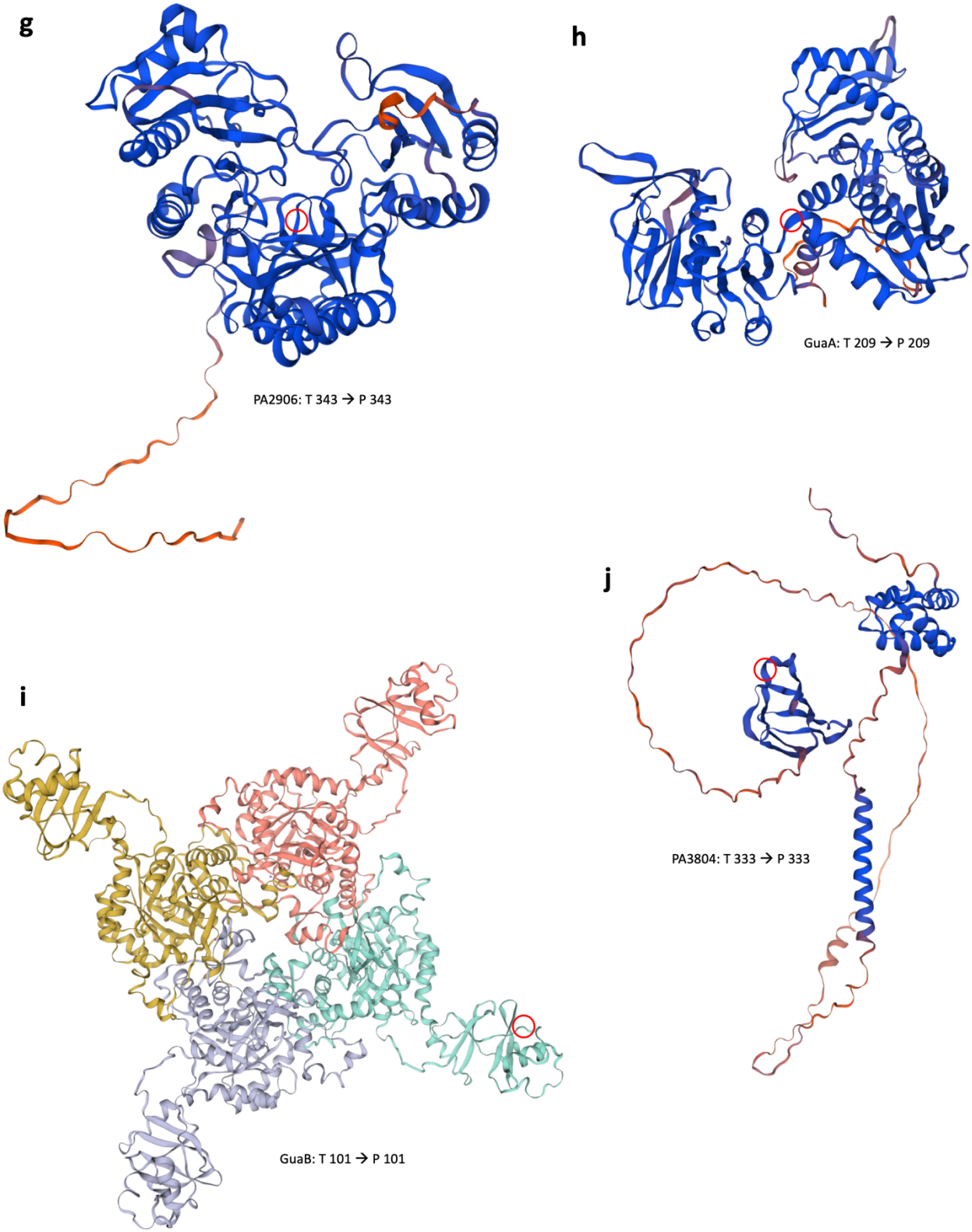

J
V 255 → G 255
K
V 197 → G 197
L
V 176 → G 176
M
V 536 → G 536
N
V 61 → G 61
O
V 144 → G 144
P
V 150 → G 150

Figure A1 | SWISS-MODEL structure prediction of proteins encoded via PA01 leading strand genes that are susceptible to T → G base mutations, of which are derived from the highly mutable AGGGGGTC nucleotide motif. Circled in red accompanied by the respective label indicates the locational consequence of amino acid substitution as a result of the T → G base mutation **| (a) phoQ primed with a V260 → G260 mutation.** (*Template for structure prediction: AlphaFold DB model of PHOQ_PSEAE (gene: phoQ, organism Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.78%, GMQE: insert here).* **| (b) PA0222 primed with a V198 → G198 mutation.** (*Template for structure prediction: Gamma-aminobutyric acid-binding protein. AlphaFold DB model of GABBP_PSEAE (gene: GABBP_PSEAE, organism: Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.72%, GMQE: 0.93).* **| (c) tolR primed with a V41 → G41 mutation.** (*Template for structure prediction: Tol-Pal system protein TolR. AlphaFold DB model of TOLR_PSEAE (gene: tolR, organism: Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.32%, GMQE: 0.78).* **| (d) PA0828 primed with a V55 → G55 mutation.** (*Template for structure prediction: Probable transcriptional regulator. AlphaFold DB model of Q9I5B1_PSEAE (gene: Q9I5B1_PSEAE, organism: Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.56%, GMQE: 0.86).* **| (e) PA0138 primed with a V261 → G261 mutation.** (*Template for structure prediction: Putative glucose ABC transporter permease protein TsgC13. AlphaFold DB model of A0A653B2C9_PSEOL (gene: tsgC, organism: Pseudomonas oleovorans). Sequence identity: 83.77%, GMQE: Insert here).* **| (f) PA1974 primed with a V273 → G273 mutation**. (*Template for structure prediction: Uncharacterized protein. AlphaFold DB model of A6V6J6_PSEA7 (gene: A6V6J6_PSEA7, organism: Pseudomonas aeruginosa, strain: PA7). Sequence identity: 96.24%, GMQE: 0.9).* **| (g) PA1205 primed with a V166 → G166 mutation.** (*Template for structure prediction: Pirin family protein. AlphaFold DB model of A0A221L1C9_PSEAI (gene: A0A221L1C9_PSEAI, organism: Pseudomonas aeruginosa). Sequence identity: 98.73%, GMQE: 0.94).* **| (h) PA1742 primed with a V213 → G213 mutation**. *Template for structure prediction: Amidotransferase. AlphaFold DB model of A0A0Q4E9A0_9PSED (gene: A0A0Q4E9A0_9PSED, organism: Pseudomonas sp Leaf15). Sequence identity: 78.33%, GMQE: 0.96.* **| (I) PA2485 showing a synonymous mutation.** *Template for structure prediction: Uncharacterized protein. AlphaFold DB model of Q9I0Z7_PSEAE (gene: Q9I0Z7_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 100%, GMQE: 0.52.* **| (J) PA1048 primed with a V255 → G255 mutation.** *Template for structure prediction: Probable outer membrane protein. AlphaFold DB model of Q9I4S6_PSEAE (gene: Q9I4S6_PSEAE, organism: Pseudomonas aeruginosa, strain: PA01). Sequence identity: 99.66%, GMQE: 0.73.* **| (K) nuoB primed with a V197 → G197 mutation.** *Template for structure prediction: NADH-quinone oxidoreductase subunit B. AlphaFold DB model of NUOB_PSEAE (gene: nuoB, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.56%, GMQE: 0.88.* **| (L) PA3837 primed with a V176 → G176 mutation.** *Template for structure prediction: Probable permease of ABC transporter. AlphaFold DB model of Q9HXG7_PSEAE (gene: Q9HXG7_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.66%, GMQE: 0.84.* **| (M) PA4675 primed with a V536 → G536 mutation.** *Template for structure prediction: Probable TonB-dependent receptor. AlphaFold DB model of Q9HVC0_PSEAE (gene: Q9HVC0_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.87%, GMQE: 0.91.* **| (N) PA4321 primed with a V61 → G61 mutation.** *Template for structure prediction: DUF4350 domain-containing protein. AlphaFold DB model of Q9HW80_PSEAE (gene: Q9HW80_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.76%, GMQE: 0.84.* **| (O) PA3965 primed with a V144 → G144 mutation.** *Template for structure prediction: Probable transcriptional regulator. AlphaFold DB model of Q9HX51_PSEAE (gene: Q9HX51_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.41%, GMQE: 0.93.* **| (P) nuoB primed with a V150 → G150 mutation.** *Template for structure prediction: NADH-quinone oxidoreductase subunit B. AlphaFold DB model of NUOB_PSEAE (gene: nuoB, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.56%, GMQE: 0.88.*

APPENDIX 2


a
AguB: T 48 → P 48
d
PA1779: D 690 → A 690
b
PA1044: T 2 → P 2
c
PA1107: D 229 → A 229
e
PA3109: T 60 → P 60
f
PA3596: T 264 → P 264


g
PA2906: T 343 → P 343
h
GuaA: T 209 → P 209
i
GuaB: T 101 → P 101
j
PA3804: T 333 → P 333

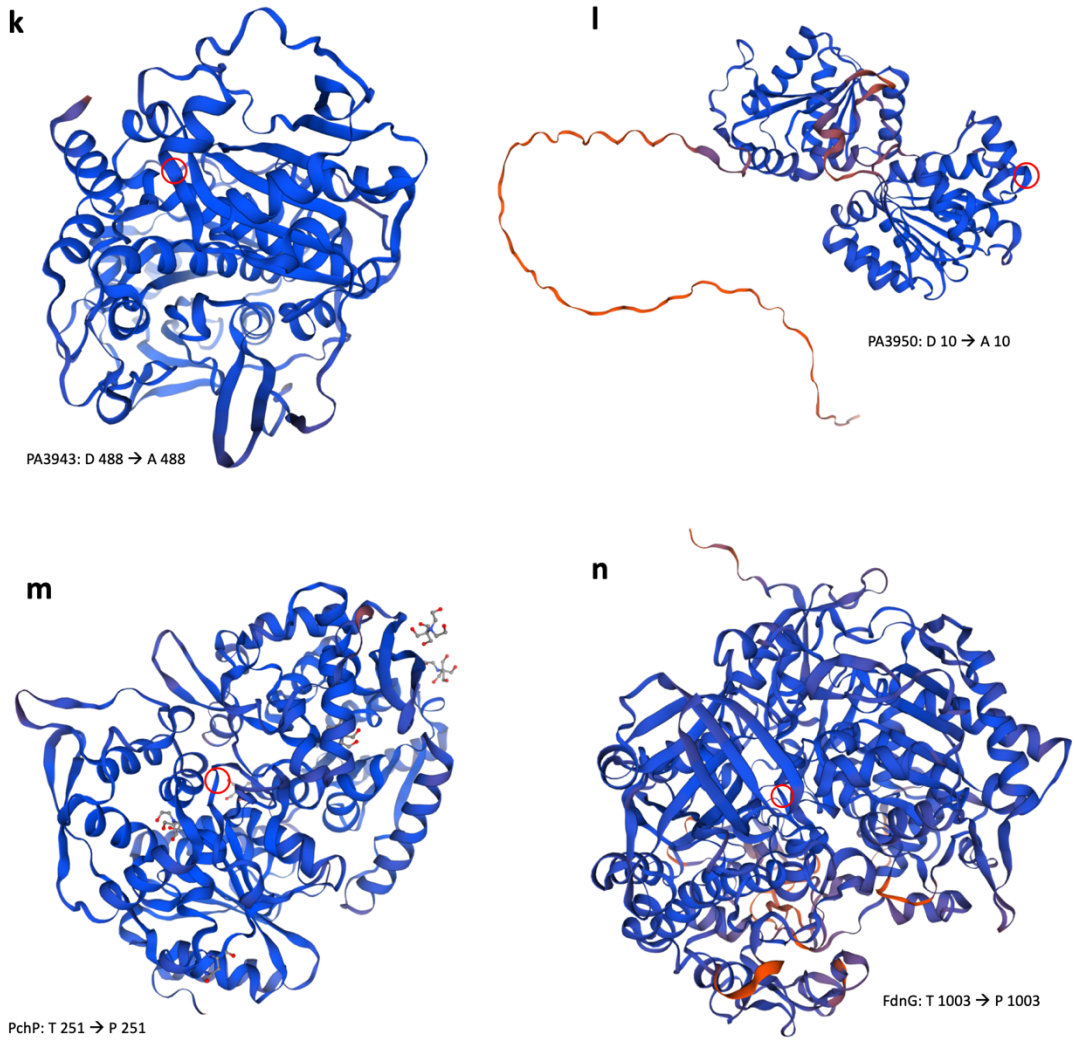


Figure A2 | SWISS-MODEL structure prediction of proteins encoded via PAO1 complementary strand genes that are susceptible to T → G base mutations, of which are derived from the highly mutable AGGGGGTC nucleotide motif. Circled in red accompanied by the respective label indicates the locational consequence of amino acid substitution as a result of the T → G base mutation **| (a) One of eight subunits of homo-octameric AguB primed with a T48 → P48 mutation.** *Template for structure prediction: N-carbamoylputrescine amidohydrolase. Crystal structure of Medicago truncatula N-carbamoylputrescine amidohydrolase (MtCPA) in complex* with N-(dihydroxymethyl)putrescine. Sequence identity: 63.41%, GMQE: 0.89. **| (b) PA1044 primed with a T2 → P2 mutation.** *Template for structure prediction: UniProtKB entry unknown, most likely obsolete. AlphaFold DB model of A0A083UKX2 (gene: unknown, organism: unknown). Sequence identity: 70.78%, GMQE: 0.95.* **| (c) PA1107 primed with a D229 → A229 mutation.** *Template for structure prediction: diguanylate cyclase. AlphaFold DB model of Q9I4M8_PSEAE (gene: Q9I4M8_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.75%, GMQE: 0.93.* **| (d) PA1779 primed with a D690 → A690 mutation.** *Template for structure prediction: Assimilatory nitrate reductase. AlphaFold DB model of Q9I2W3_PSEAE (gene: Q9I2W3_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.89%, GMQE: 0.95.* **| (e) PA3109 primed with a T60 → P60 mutation.** *Template for structure prediction: Colicin V production protein. AlphaFold DB model of Q9HZB0_PSEAE (gene: Q9HZB0_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.44%, GMQE: 0.88.* **| (f) PA3596 primed with a T264 → P264 mutation.** *Template for structure prediction: Probable methylated-DNA--protein-cysteine methyltransferase. AlphaFold DB model of Q9HY30_PSEAE (gene: Q9HY30_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.72%, GMQE: 0.88.* **| (g) PA2906 primed with a T343 → P343 mutation.** *Template for structure prediction: Precorrin-3B synthase. AlphaFold DB model of A0A6A9JLR3_PSEAI (gene: cobG, organism: Pseudomonas aeruginosa). Sequence identity: 98.15%, GMQE: 0.89.*

Figure 6 continued| (h) GuaA primed with a T209 → P209 mutation. *Template for structure prediction: GMP synthase [glutamine-hydrolyzing]. AlphaFold DB model of GUAA_PSEAE (gene: guaA, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.81%, GMQE: 0.91.* **| (i) GuaB primed with a T101 → P101 mutation. NB: mutation is only shown in one subunit of the homo-tetramer structure.** *Template for structure prediction: Inosine-5'-monophosphate dehydrogenase from Vibrio cholerae complexed with IMP and mycophenolic acid. Sequence identity: 64.26%, GMQE: 0.83.* **| (j) PA3804 primed with a T333 → P333 mutation.** *Template for structure prediction: HTH cro/C1-type domain-containing protein. AlphaFold DB model of Q9HXJ3_PSEAE (gene: Q9HXJ3_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.71%, GMQE: 0.73.* **| (k) PA3943 primed with a D488 → A488 mutation.** *Template for structure prediction: Nitroreductase family protein. AlphaFold DB model of A0A2R3IMX8_PSEAI (gene: A0A2R3IMX8_PSEAI, organism: Pseudomonas aeruginosa). Sequence identity: 93.03%, GMQE: 0.96*. **| (l) PA3950 primed with a D10 → A10 mutation.** *Template for structure prediction: RNA helicase. AlphaFold DB model of Q9HX66_PSEAE (gene: Q9HX66_PSEAE, organism: Pseudomonas aeruginosa, strain: PAO1). Sequence identity: 99.78%, GMQE: 0.85.* **| (m) PchP primed with a T251 → P251 mutation.** *Template for structure prediction: Pseudomonas Aeruginosa Phosphorylcholine Phosphatase (monoclinic form). Sequence identity: 99.69%, GMQE: 0.91.* **| (n) FdnG primed with a T1003 → P1003 mutation.** *Template for structure prediction: Formate dehydrogenase major subunit from e. coli. Sequence identity: 66.70%, GMQE: 0.85.*