# Effects of CRISPR-Cas System Presence on Lateral Gene Transfer Rates in Bacteria

Athena Zambelis, Utkarsh J. Dang, G. Brian Golding

Department of Biology, McMaster University, Hamilton, Ontario, L8S4K1 Canada

**Abstract**

An article by Gophna et al., (2015) defines a trade-off hypothesis between gene acquisition and CRISPR immunity. This hypothesis states that there is a trade-off between the adaptive immunity imparted by the CRISPR-Cas system against phage invasions and the ability of an organism to acquire new and potentially beneficial genetic material via LGT. The authors eventually conclude that on evolutionary timescales, the inhibitory effect of CRISPRs on LGT is not supported by evidence. We propose a better methodology by which the trade-off hypothesis can be tested. An R package called markophylo was used to measure rates of LGT among closely related bacteria with and without a CRISPR-Cas system, by using maximum likelihood estimates to measure unique gene insertion and deletion rates for each group. An R package called indelmiss was also used to find the value of a parameter for missing data, which measured whether species with a CRISPR-Cas system have "missing" genomic information due to the inhibition of LGT. For *Helicobacter* species with CRISPRs, the gene insertion rate was estimated to be 0.493101 (se = 0.01251881), while for the species without CRISPRs it was found to be higher at 1.872357 (se = 0.05224986) and 1.872357 (se = 0.02024776) for *Edwardsiella* and *Caulobacter* respectively. Similarly, for *Shewanella* species with CRISPRs, the gene insertion rate was estimated to be 0.5049471 (se = 0.01277100), while for taxa without CRISPRs it was estimated to be higher at 54.20031 (se = 1.402581) and 3.021449 (se = 0.07744549) for *Idiomarina* and *Marinobacter* respectively. This data represents a significant difference in LGT rates between the groups and supports the hypothesis that CRISPR function hinders rates of LGT.

# 1 Introduction

## 1.1 Lateral Gene Transfer

### 1.1.1 Mechanisms:

Lateral gene transfer (LGT) involves the transfer of genetic information between distantly related organisms. This mechanism acts in contrast to the vertical transfer of DNA from a parent cell to its descendant. The transfer of genetic information via LGT can occur through three mechanisms: transformation, transduction, and conjugation.

Transformation involves the uptake and incorporation of naked DNA from the environment [1]. Of the naturally transformable bacteria, many species become competent during the course of their lifecycle, while others remain in a competent state indefinitely [1].

The transformation efficiency may be enhanced by the presence of specific uptake sequences, which are frequently found in DNA exchanged between more closely related organisms [1].

Transduction describes the process whereby a bacteriophage acts as the vector for injection of DNA into a recipient cell [2]. The bacteriophage first replicates within a donor organism and then encapsulates the host DNA either adjacent to the phage attachment site (specialized transduction) or randomly across the genome (generalized transduction) [2]. Subsequent infection of a recipient organism by the bacteriophage then allows the transfer of the encapsulated donor DNA. With each transduction event, the amount of

DNA transferred is limited by the size of a phage capsid and limited to organisms that have suitable surface receptors recognized by the bacteriophage [2]. With both transduction and transformation, the donor and recipient cells are not required to be connected in time or space. Transformation is more sensitive as it relies on the sustained integrity of naked DNA in the environment. Conversely, transduction provides additional benefits whereby proteins encoded by the phage are able to mediate direct entry of phage DNA into the recipient cytoplasm, degrade host restriction endonucleases to prevent DNA damage, and mediate the integration of DNA into the chromosomes [1].

Conjugation is a contact-dependent mechanism of DNA transfer between the donor and recipient cells. It involves transfer of either a self-transmissible conjugative plasmid, mobilization of a non-self-transmissible plasmid containing an origin of conjugal transfer, or cointegration of two different circular plasmids that have undergone a fusion event [3].

The introduction of DNA into a recipient cell does not ensure successful gene transfer unless the sequences are stably maintained and expressed. In order for DNA to be assimilated into the bacterial genome from the cytoplasm, it must persist as an episome under positive selection forces to avoid stochastic loss [1]. Alternatively, the DNA can be incorporated directly into the recipient genome via bacteriophage integrases or transposase machinery [1].

A deletion bias in bacterial genomes serves to eliminate genes that fail to provide meaningful functionality. The small sizes of bacterial genomes imply that most of the DNA obtained through LGT fails to be maintained over a long-term basis [2]. In essence, LGT functions in allowing bacterial genomes to sample rather than accumulate sequences, since gene acquisition must be balanced with gene loss. This being said, not all lateral genes turn over rapidly; for instance, in the $\gamma$-proteobacteria, there are genes that have persisted vertically for a long period of time and have even become typical of the group [2].

### 1.1.2 Detecting LGT Rates:

There are several methods commonly used to measure rates of gene acquisition via LGT. One is to measure the fraction of recently acquired genes inferred on the basis of dinucleotide composition. Bacteria have base compositions, patterns of codon usage, and frequencies of di- and tri-nucleotides that are typical of a single species [4]. Therefore, it is possible to resolve which genes were obtained via LGT if they have a GC content uncharacteristic of the host species. However, this method can be unreliable, since the degree of differentiation between host and non-host GC content is not always clear. Finding the fraction of prophage genes in a genome may also be used to measure rates of LGT, and can be accomplished using PhageFinder [4]. This technique may cause shortfalls due to the limited availability of prophage genomes in current databases. A better method to identify the number of genes obtained by vertical versus horizontal transmission is to use parsimony methods. This is accomplished by selecting all sets of orthologous genes from the clusters of orthologous groups of proteins (COG) database, and then reconstructing for each gene the most parsimonious evolutionary phylogeny [5]. However, by using maximum likelihood techniques instead, many of the limitations of parsimony methods can be dealt with. Some of these limitations include the lack of consideration given to both the history of the gene families and the branch lengths of the organismal topology [5].

## 1.2 CRISPR-Cas System

### 1.2.1 Mechanism:

CRISPR (clustered, regularly interspaced, short, palindromic repeats)-Cas (CRISPR-associated proteins) systems in bacteria and archaea provide a form of heritable adaptive immunity against invading mobile genetic elements (MGEs). The CRISPR locus is composed of arrays of partially palindromic repeats ranging from 21-48 bp in length. These repeats are interspersed by 26-72 bp of variable spacer sequence derived from invading MGEs [6]. There are three main CRISPR-Cas system types, and multiple subtypes (I-A to F, II-A and B, and III-A and B) [7]. Depending on the type of system present in the organism, several Cas genes are encoded within or adjacent to the CRISPR locus.

There are three steps involved in the mechanistic action of CRISPR-Cas systems. Adaptation is the first step, and involves the formation of a complex by Cas1 and Cas2 proteins that excises protospacer DNA from invading elements and integrates them into the CRISPR array [8]. Following this is the transcription of the CRISPR locus to yield a single precursor RNA that is processed within the repeat regions into units of CRISPR RNAs (crRNAs) by a complex of Cas proteins known as CASCADE [8]. Interference is the final step, during which the crRNAs function as guides by recognizing the complementary protospacer sequence within the invading MGE from which the spacer was derived [8]. A type-specific complex of Cas proteins then degrade the invading DNA or RNA [8].

### 1.2.2 Biotechnological Applications:

Research relating to CRISPR function and evolution has peaked in recent years in response to the use of these systems in biotechnology. The bacterial Cas9 endonuclease is able to assume all of the functions of the CASCADE complex, and can therefore carry out crRNA maturation and biogenesis, as well as interference without support from additional proteins [9]. Cas9 has been shown to efficiently carry out genome modification in eukaryotes in a site-specific manner without detectable damage at known off-target regions [9]. This allows for genome editing of model organisms to produce key mutations without the need for more time consuming and expensive methodology. For instance, the microinjection of mouse embryos with Cas9 mRNA and single guide RNAs was shown to induce on-target mutations that were transmissible to offspring [9].

#### 1.2.3 Experimental Evidence:

In conflict with the immunity advantages imparted by the CRISPR-Cas system, there are negative selection pressures against CRISPR maintenance that may explain why these systems are present in less than half of the bacteria sequenced despite their ability to be laterally transferred [6]. Some of these pressures occur in response to the need for bacteria to maintain and replicate the DNA associated with these systems, which is an expensive and error-prone process [10]. In addition, autoimmunity caused by the incorporation of protospacers from the host genome may result in lethal genomic rearrangements from the activation of recombination mechanisms in response to chromosome breakage [10].

In 2007, the first experimental evidence of CRISPR-mediated immunity emerged with the isolation of phage resistant *Streptococcus thermophilus* cells [11]. These cells demonstrated that novel acquisition of phage sequences into a CRISPR locus after infection was the cause of a phage resistance phenotype. It was later shown that this immune response is also highly efficient, with the yield of cognate phage dropping by up to 5 orders of magnitude after spacer integration into the CRISPR locus [12]. However, recent evidence suggests that the CRISPR-Cas system is not only effective in preventing phage invasion, but also has the potential to block transfer of bacterial plasmids [6]. For instance, the conjugation efficiency of a plasmid into *Staphylococcus epidermidis* was reduced by greater than $10^4$-fold when a spacer derived from the plasmid was introduced into the bacterial CRISPR-Cas system [6]. In further support, the majority of spacer sequences have been shown to target bacteriophage genomes, though many match plasmids and other MGEs, as well as chromosomal regions of bacteria and archaea.

### 1.3 Specific Aims

Since transduction and conjugation - two of the three mechanisms by which LGT functions - are hindered by CRISPR function, it is hypothesized that lower rates of LGT will be observed in species with active CRISPR-Cas systems compared to closely related species without them.

An article by Gophna et al., (2015) defines a trade-off hypothesis between gene acquisition and CRISPR immunity. This hypothesis states that there is a trade-off between the adaptive immunity imparted by the CRISPR-Cas system against phage invasions and the ability of an organism to acquire new and potentially beneficial genetic material via LGT. The authors eventually conclude that on evolutionary timescales, the inhibitory effect of CRISPRs on LGT is not supported by evidence. Instead they make the conclusion that the trade-off hypothesis functions on a population level to increase the fitness of individual organisms and produce a more genetically diverse local population of species. This is a surprising observation given the high degree of experimental and conceptual support for LGT inhibition by CRISPR-mediated immunity outlined previously. It is possible that the methodology used by the authors can explain their unexpected results. Since Gophna et al. chose to consider the trade-off hypothesis over a long evolutionary timescale, the overall effects of CRISPRs on LGT may appear null if the functionality of the system is inconsistent over time. The CRISPR-Cas system has been shown to undergo rapid deletion and expansion events in response to different selective pressures. Under strong selective pressure for virulence or antibiotic resistance, bacterial pathogens have been shown to lose CRISPR function to allow for LGT. For example, when a spacer targeting a capsule gene encoding an essential virulence factor was provided to *Streptococcus pneumoniae*, the bacteria sometimes lost CRISPR function, acquired the capsule genes, and mounted a successful infection in mice [13]. This problem is made worse when the techniques used by the authors to measure rates of LGT lack sensitivity and statistical power. Thus, it is possible that the conclusions made by Gophna et al., are a product of their method falling short and not a true reflection of the impacts of the CRISPR-Cas system on LGT rates.

We propose a better methodology by which the trade-off hypothesis can be tested. An R package called indelmiss will be used to measure rates of LGT among closely related proteobacteria with and without a CRISPR-Cas system [14]. This package uses maximum likelihood estimates to measure genome-wide gene insertion and deletion (indel) rates [14]. This model will also find the value of a parameter for missing data, which will measure whether species with a CRISPR-Cas system have "missing" genomic information due to the inhibition of LGT. By using closely related species, we can consider short timescales that control for and limit any inconsistencies in the functionality of the CRISPR-Cas system. Finally, the use of maximum likelihood techniques will increase the sensitivity of LGT rate measurements and the statistical power of the results [14]. In order to conclude that the trade-off hypothesis functions on an evolutionary level, the missing data parameters provided for the species with CRISPRs will need to demonstrate a statistically significant increase compared to the missing data values generated by species without functional systems. In addition to this, the consistency of the results will be explored using different methodologies. Markophylo is an R package that fits maximum likelihood models on phylogenetic trees for the analysis of gene family presence/absence data [15]. This package will be used to estimate unique gene insertion and deletion rates for taxa on the basis of CRISPR presence or absence. These rates can then be used to infer rates of LGT in the taxa under consideration. To conclude that the presence of a CRISPR-Cas system hinders rates of LGT, the gene insertion and deletion rates will need to be significantly lower for the species with CRISPRs compared to the species without.

## 1.4 Evolution of Relevant Bacteria

### 1.4.1 Proteobacteria:

A number of factors were considered in choosing suitable species to be used throughout this study. Congeneric *Helicobacter* species were chosen as the primary proteobacteria under investigation, and provide 17 of the 25 genomes collected and analysed. *Helicobacter* species are opportunistic pathogens of the mammalian gastrointestinal tract and liver [16]. The most widely studied species within this genus is *H. pylori*, which colonizes the stomachs of more than 50% of the human population [16]. These organisms are considered pathogenic due to their association with the development of peptic ulcers, stomach cancer, and chronic gastritis [16].

The significant genetic variability of *H. pylori* is one of its identifying characteristics. Nearly every human infected with *H. pylori* harbours their own unique strain of the organism, since the species undergoes frequent genetic alterations driven by a high mutation rate and intraspecific recombination [16]. This allelic diversity in both housekeeping and virulence genes is believed to contribute to host adaptation [16]. In addition to this, *H. pylori* cells are naturally competent throughout their lifecycle, causing them to have increased expected rates of LGT [2]. This allows for missing data parameters to be exaggerated and more easily identifiable if transformation rates are hindered by the presence of CRISPR-Cas systems. Another marked feature considered in choosing *H. pylori* as the primary organism under study is its high variation in the number and types of CRISPR-Cas systems encoded by the different strains. This allows for the exploration of how the different CRISPR systems impact rates of LGT, as opposed to simply cataloguing their presence or absence.

Since the *Helicobacter* species collected are very closely related in evolutionary time and include genomes with and without CRISPR-Cas systems, two outgroups were chosen as more distant standards of comparison. Both of these standards represent groups that have no known presence of CRISPRs in any of their congeneric species. These standards include species from the genera of *Edwardsiella* and *Caulobacter*.

Similar to *Helicobacter*, *Edwardsiella* species are occasionally opportunistic pathogens of humans, though they are found more often in fish where they colonize the intestines and cause enteric septicemia [17]. Though the scientific literature exploring *Edwardsiella* evolution is limited, the similarities in virulence and niche with *Helicobacter* suggest a comparable requirement for frequent recombination events and high genetic diversity to allow for host specification [17].

Species that are obligate intracellular parasites or symbiotes of a host were not included in this study. This was done to prevent the genome size reduction associated with these relationships from increasing the missing data proportions calculated by indelmiss in comparison to extracellular bacteria. Despite the extremely limited evolutionary knowledge of *Caulobacter*, it was chosen as a standard due to its extracellular persistence in fresh water lakes and streams, and its classification as an Alphaproteobacterium closely related to *Helicobacter* [18]. Although *Caulobacter* is considered a model organism for studying the unique process of asymmetric cell division, the genes involved in its cell cycle control are conserved across most Alphaproteobacteria species [18]. This being said, *Caulobacter* possesses few genes that are specific and unique to its species, making it a good standard for comparison with *Helicobacter* [18].

### 1.4.2 Firmicutes:

Congeneric *Streptococcus pyogenes* subspecies were chosen as the primary firmicutes under investigation, and provide 17 of the 27 genomes collected and analysed. Similar to *Helicobacter* species, *S. pyogenes* are opportunistic human pathogens, though they often infect the respiratory tract and are only found in an estimated 15-20% of people [19]. Acute *S. pyogenes* infections can present as scarlet fever, pharyngitis, cellulitis, and impetigo [19]. Due to their strong prevalence as human pathogens, several closely-related *S. pyogenes* subspecies have undergone complete genome sequencing for use in scientific research. Unlike *Helicobacter* species, the genomes of *S. pyogenes* are not known to undergo large recombination events that may skew the missing data proportions being calculated in this study.

Two out-group species were once again chosen as standard for comparison against the *S. pyogenes* subspecies under investigation. Both of these standards represent groups that have no known presence of CRISPRs in any of their congeneric species. These standards include species from the genera of *Lactococcus* and *Pediococcus*.

*Lactococcus* species were initially included in the genus *Streptococcus* due to their immense genetic similarities and close evolutionary relatedness [20]. These bacteria are primarily used in the dairy industry for the production of cheese and milk, due to their classification as homofermenters [21]. They are often grown with other lactic acid bacteria, including *Streptococcus* species in mixed-strain starter cultures due to their metabolic similarities [21].

*Pediococcus* species belong to the order of Lactobacillales with *Lactococcus* and *Streptococcus*, denoting the close evolutionary relationship between all three species. This organism is a lactic acid bacteria used in the food industry for the fermentation of cabbage into sauerkraut, and as a beneficial microbe in the production of cheese and yoghurt [22].

### 1.4.3 Gammaproteobacteria:

Congeneric *Shewanella* species were chosen as the primary gammaproteobacteria under investigation, and provide 15 of the 20 genomes collected and analysed. *Shewanella* belongs to the proteobacteria phylum, and is included in this analysis as a comparison to the proteobacteria group outlined above. This genus was chosen due to its stark ecological contrasts to that of *Helicobacter*. For instance, unlike many of the known proteobacteria genera, *Shewanella* species are rarely pathogenic against humans [23]. They are widely distributed in soil and water, though have been found to colonize decaying tissues as well [23]. In addition, unlike *Helicobacter* species, the genomes of *Shewanella* are not known to undergo large recombination events that may skew the missing data proportions being calculated in this study.

Two out-group genera were once again chosen as standards for comparison against the *Shewanella* species under investigation. Both of these standards represent groups that have no known presence of CRISPRs in any of their congeneric species. These standards include species from both *Idiomarina* and *Marinobacter*. Both of these genera were included in the analysis due to their close evolutionary relationship with *Shewanella*, and their classification as gammaproteobacteria. In addition, both of these genera have ecological similarities to *Shewanella*, which limits the number of unique genes they are likely to possess. Neither of these groups is pathogenic, and both reside exclusively in marine environments [24] [25]. *Idiomarina* species reside predominantly near deep sea hydrothermal vents, and can withstand high temperatures and pressures [25]. *Marinobacter* remain motile in a variety of marine environments, though are often found near oil refineries [24]. Due to their use of hydrocarbons in energy production, they have been studied as a potential source of bioremediation [24]. As such, neither of these organisms have an extensive evolutionary literature available.

# 2 Methodology

The complete genome coding sequences of 25 proteobacteria were obtained as GenBank files from NCBI (Table 11). Of these genomes, 9 belong to *Helicobacter* species with CRISPR-Cas systems present (obtained on 07/02/2015). The remaining genomes, 4 of which belong to *Edwardsiella* species (obtained on 09/21/2015), 4 to *Caulobacter* species (obtained on 09/21/2015), and 8 to *Helicobacter* species (obtained on 07/02/2015), did not code for CRISPR-Cas systems. The CRISPR database (http://crispr.u-psud.fr/crispr/ accessed on 07/02/2015) was used to identify groups of closely related proteobacteria species suspected to be without any CRISPR-Cas systems. Since Cas1 and Cas2 proteins are present in all CRISPR-Cas system types and are required for spacer integration, an active CRISPR-Cas system was deemed to be present if both of these proteins in addition to a CRISPR array were encoded in the genome. Since many Cas genes are often annotated as hypothetical proteins by NCBI, *Helicobacter*-derived Cas1 (C5ZYI4) and Cas2 (C5ZYI5) gene sequences were obtained from UniProt (on 07/02/2015), and BLASTN was used to detect sequence similarity between them and the genomes collected. A match was defined as any hit with expect values less than 0.05 and alignment lengths that covered more than 85% of the query protein length.

The complete genome coding sequences of 27 firmicutes were obtained as GenBank files from NCBI (Table 15). Of these genomes, 12 belong to *S. pyogenes* subspecies with CRISPR-Cas systems present (obtained on 01/01/2016). The remaining genomes, 8 of which belong to *Lactococcus* species (obtained on 01/01/2016), 2 to *Pediococcus* species (obtained on 01/01/2016), and 5 to *S. pyogenes* subspecies (obtained on 01/01/2016), did not code for CRISPR-Cas systems. The CRISPR database (http://crispr.u-psud.fr/crispr/ accessed on 12/28/2015) was used to identify groups of closely related firmicutes species believed to be absent of any CRISPR-Cas systems. *S. pyogenes*-derived Cas1 (J7M1S9) and Cas2 (J7M8B7) gene sequences were obtained from UniProt (on 01/02/2016), and BLASTN was used to detect sequence similarity between them and the genomes collected. A match was defined as any hit with expect values less than 0.05 and alignment lengths that covered more than 85% of the query protein length.

The complete genome coding sequences of 20 gammaproteobacteria were obtained as GenBank files from NCBI (Table 19). Of these genomes, 8 belong to *Shewanella* species with CRISPR-Cas systems present (obtained on 02/17/2016). The remaining genomes, 2 of which belong to *Idiomarina* species (obtained on 02/17/2016), 3 to *Marinobacter* species (obtained on 02/17/2016), and 7 to *Shewanella* species (obtained on 02/17/2016), did not code for CRISPR-Cas systems. The CRISPR database (http://crispr.u-psud.fr/crispr/ accessed on 02/09/2015) was used to identify groups of closely related gammaproteobacteria species believed to be absent of any CRISPR-Cas systems. *Shewanella*-derived Cas1 (E6XMP7) and Cas2 (A9KZ90) gene sequences were obtained from UniProt (on 02/19/2016), and BLASTN was used to detect sequence similarity between them and the genomes collected. A match was defined as any hit with expect values less than 0.05 and alignment lengths that covered more than 85% of the query protein length.

Gene sequences were manually deleted if the annotation provided by NCBI attributed their origin to be from a mobile or viral element, including genes associated with prophages, retroviruses, and transposases. An all-vs-all BLASTP with expect values less than $1 \times 10^{-20}$ and alignment lengths covering at least 85% of the query protein length were used to identify gene similarity between taxa. In order to organize this data into gene families, homologues were classified as the best reciprocal hits reported under these conditions, and were allocated to the same gene family. All potential paralogues were assumed to be the result of gene duplications and were clustered in the same family as well.

To build a phylogeny for these sequences, the nucleotide sequences of fifty genes found in all taxa were individually concatenated and aligned using MAFFT [26]. The individual alignments for each species were concatenated into a NEXUS file and provided to MrBayes [27]. A phylogenetic tree was created using a general time reversible substitution model with gamma distributed rate variation after 500,000 generations and a 25% burn-in. The resulting phylogeny was then rooted by out-group using Figtree [28].

From the all-vs-all BLASTP, genes with no identified homologues were searched against the nonredundant database (NR) using a one way BLASTP. Hits were retained if expect values were less than 0.05 and alignment lengths covered more than 85% of the query protein length. If a gene tested against the NR database hit a multispecies gene or hit to another taxa, it was automatically valid. However, if the gene being tested hit itself, the hit was disregarded. Gene families retained from the all-vs-all BLASTP were combined with unique genes retained from the search against the NR database to create a gene family matrix, with a zero in the matrix indicating absence of the gene family within the species, and a one indicating gene family presence.

Rates of LGT can be deduced from gene indel rates, while providing a maximum likelihood estimate of how much coding genetic material has been lost from congeneric species with active CRISPRs [14]. To accomplish this, the phylogenetic tree and gene family matrix were provided to the R package indelmiss [14]. Model 4 was fit to the data, and was used to calculate unique insertion and deletion rates among the species under consideration. Missing data proportions were only fit for taxa with CRISPR-Cas systems present.

A permutation test was computed to determine the distribution of the test statistic under the null hypothesis and measure its level of statistical significance. This required running 1000 replicates of random groups of 11 species through Model 4 of indelmiss. The missing values were averaged to yield a null hypothesis, which was then compared to the test statistic represented by the mean of the missing data proportions fit for the taxa with CRISPR-Cas systems present. To determine whether the difference between the null hypothesis and test statistic was statistically significant, the distribution of missing data parameters was ordered to find the percentile of data exceeded by the test statistic. The test statistic was deemed significantly different from the null hypothesis if it exceeded at least 95% of the missing data parameters obtained stochastically.

A simulation was conducted in order to test the ability of indelmiss to identify a difference between the gene insertion and deletion rates of congeneric species with and without CRISPRs. Gene family presence/absence data on 3839 genes for the 17 closely related *Helicobacter* species was simulated using R. To do this, a matrix of appropriate size was created and then altered according to set gene insertion and deletion rates. The gene deletion and insertion rates were tested separately by setting one of these rates to zero in each simulation. The rate being tested was set with each iteration to a random value between 0.5 and 1.5. The phylogeny including only *Helicobacter* species generated previously using MrBayes was used to simulate branch length heterogeneity. This was done using the geiger package in R, for which a scaling factor was selected and used to modify the branch lengths of the species with CRISPRs [29]. In estimating gene insertion and deletion rates, indelmiss creates a transition probability matrix using the product of both the branch lengths from the phylogeny and the gene family presence/absence data. By creating branch length heterogeneity, a difference in the gene insertion and deletion rates between the species with and without CRISPRs is simulated. To coincide with the initial hypothesis, the gene deletion rate of the species with CRISPRs was set to be higher than the species without CRISPRs, and the gene insertion rate to be lower. Indelmiss was run through 1000 iterations, and either the mu (deletion rate) or nu (insertion rate) was estimated independently for both the species with and without CRISPRs during each one. The mean of the 1000 mu or nu values was then taken for each group, and a Welch two-sample t-test was performed to obtain a measure of significance (p-value) between the two means.

Finally, Gene insertion and deletion rates were estimated uniquely for species with or without CRISPRs using markophylo [15]. To accomplish this, the phylogenetic tree and gene family matrix were provided to the R package, which fits maximum likelihood models onto the data to estimate gene insertion and deletion rates [15].

# 3 Results

## 3.1 Proteobacteria Analysis

A matrix containing gene family presence/absence data on 12,482 genes for 25 operational taxonomic units (OTUs) was used in this analysis (Table 10). The OTUs under consideration included 17 *Helicobacter* species, 4 *Edwardsiella* species, and 4 *Caulobacter* species (Figure 1). From these taxa, 9 of the *Helicobacter* species possessed a CRISPR-Cas system (Table 11). Indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system under the assumptions of Model 4. The gene insertion and deletion rates were estimated independently for each of the 3 major clades as highlighted on the phylogeny in Figure 1. Within each of the clades, the species were assumed to have the same gene insertion and deletion rates.

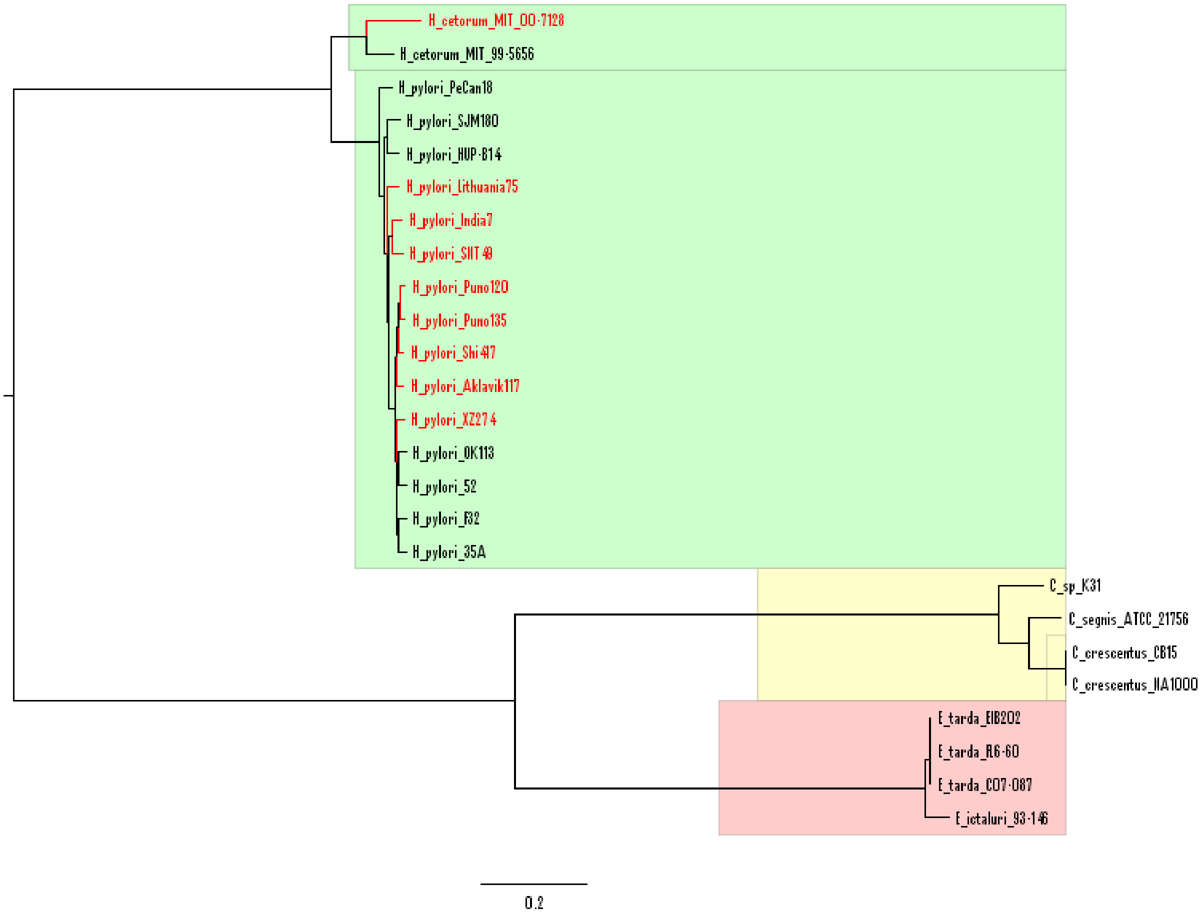
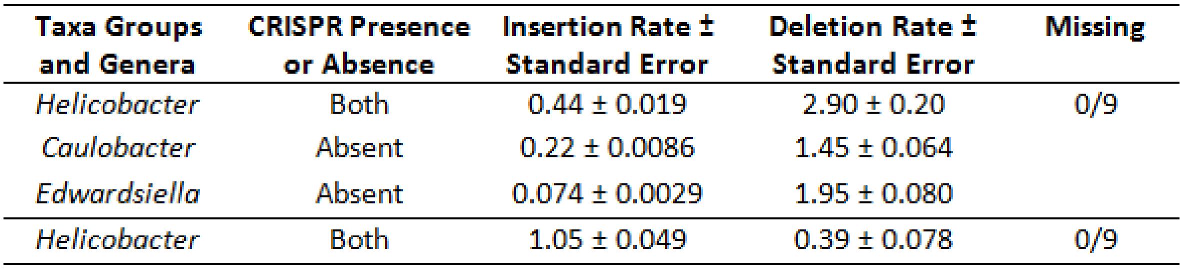


Figure 1: Gene tree showing branch lengths and topology for 3 closely related clades of *Helicobacter* (green), *Caulobacter* (yellow), *Edwardsiella* (pink) taxa. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

Table 1: Indel rate estimates made by indelmiss for the proteobacteria taxa with and without the out-group genera included in the analysis. The missing ratio numerator is indicative of the number of taxa with CRISPRs that had a high missing data proportion (greater than 0.10). The missing ratio denominator is indicative of the total number of taxa with CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error | Missing |
|---|---|---|---|---|
| *Helicobacter* | Both | 0.44 ± 0.019 | 2.90 ± 0.20 | 0/9 |
| *Caulobacter* | Absent | 0.22 ± 0.0086 | 1.45 ± 0.064 | |
| *Edwardsiella* | Absent | 0.074 ± 0.0029 | 1.95 ± 0.080 | |
| *Helicobacter* | Both | 1.05 ± 0.049 | 0.39 ± 0.078 | 0/9 |

The gene insertion and deletion rate estimates made by indelmiss have been summarized in Table 1. The *Helicobacter* clade had the highest estimated gene deletion rate of 2.898392 (se = 0.1996856), and the highest estimated gene insertion rate of 0.4438285 (se = 0.01919645). The *Edwardsiella* clade had an estimated gene deletion rate of 1.962281 (se = 0.06411055), and the lowest estimated gene insertion rate of 0.07428793 (se = 0.008572789). The *Caulobacter* clade had the lowest estimated gene deletion rate of 1.454248 (se = 0.07972343), and an estimated gene insertion rate of 0.2180159 (se = 0.002894523). All of the missing data proportions were estimated to be between 0 and 0.045, with a median value of 0.01692078 (se = 0.006550289).

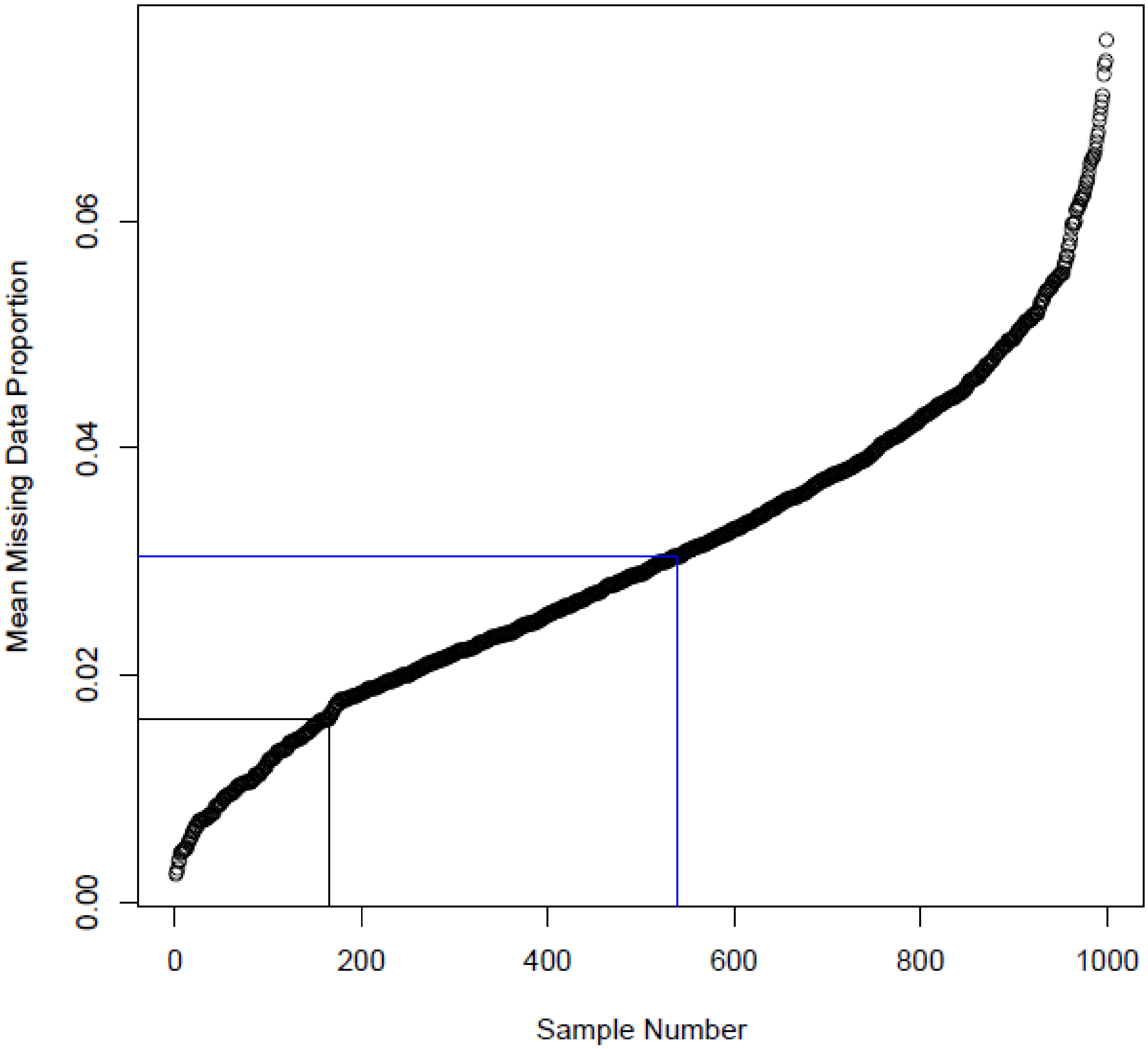
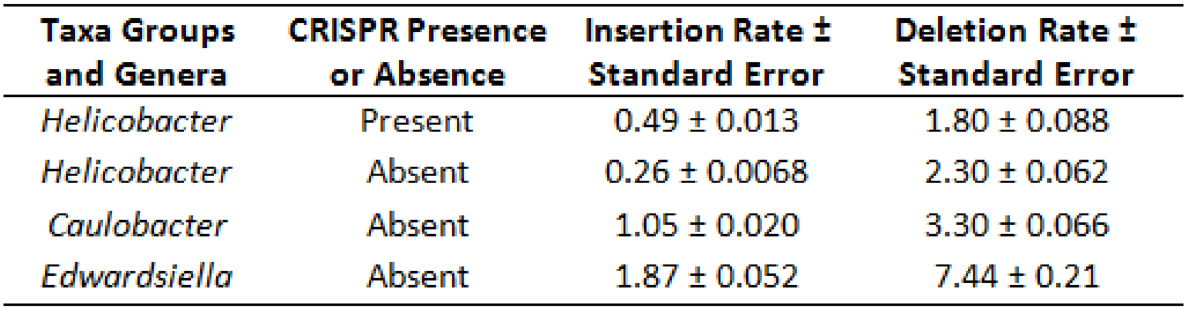


Figure 2: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 9 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

To assess the statistical significance of the estimated missing data proportions, a permutation test was conducted (Figure 2). The mean missing data proportion representative of the null hypothesis was found to be 0.03041044, while for the species possessing a CRISPR-Cas system it was 0.01615196. The difference between these two values was not deemed significant, as the mean missing data proportion for the species with CRISPRs was lower than only 83.5% of the 1000 random missing data proportions estimated.

Table 2: Indel rate estimates made by markophylo for the proteobacteria taxa with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Helicobacter* | Present | 0.49 ± 0.013 | 1.80 ± 0.088 |
| *Helicobacter* | Absent | 0.26 ± 0.0068 | 2.30 ± 0.062 |
| *Caulobacter* | Absent | 1.05 ± 0.020 | 3.30 ± 0.066 |
| *Edwardsiella* | Absent | 1.87 ± 0.052 | 7.44 ± 0.21 |

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each clade using markophylo. These gene insertion and deletion rate estimates made by markophylo have been summarized in Table 2. For the *Helicobacter* taxa with CRISPRs, the gene deletion rate was estimated to be 1.796414 (se = 0.08772410), while the gene insertion rate was estimated to be 0.493101 (se = 0.01251881). For the *Helicobacter* taxa without CRISPRs, the gene deletion rate was estimated to be higher at 2.3034123 (se = 0.062065468), while the gene insertion rate was estimated to be lower at 0.2626205 (se = 0.006805832). For the *Edwardsiella* taxa, the gene deletion and insertion rates were estimated to be highest at 7.435624 (se = 0.21034116) and 1.872357 (se = 0.05224986), respectively. For the *Caulobacter* taxa, the gene deletion and insertion rates were estimated to be 3.301165 (se = 0.06623118) and 1.052155 (se = 0.02024776) respectively. Overall, the indel rate estimates of the *Helicobacter* taxa with CRISPRs present are lower than the out-group genera without CRISPRs. In addition, the insertion rate estimate of the *Helicobacter* taxa with CRISPRs is higher than the congeneric species without CRISPRs, while the deletion rate estimate is lower. The difference between each of the insertion and deletion rate estimates is significant, due to the lack of overlap between the values and their standard errors.

There were concerns that the branch lengths across the phylogeny were too long between the 3 clades for indelmiss and markophylo to reliably function within their parameters. These concerns stemmed from the high indel rate estimates found for the *Idiomarina* taxa in the gammaproteobacteria group. To increase the accuracy of the programs, the data was reconsidered without the out-groups to include only the *Helicobacter* clade.

The matrix containing gene family presence/absence data on 3839 genes for the 17 closely-related *Helicobacter* species was used for the analysis (Table 12). As done previously, indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system under the assumptions of Model 4. Homogeneous indel rates were assumed for all branches of the phylogeny (Figure 3). The gene insertion and deletion rate estimates made by indelmiss have been summarized in Table 1. In contrast to the previous measurements, the gene insertion rate increased to an estimated 1.054827 (se = 0.0490973), while the gene deletion rate decreased to an estimated 0.3897539 (se = 0.07974148). The missing data proportions for all 9 of the species were estimated to be between 0 and 0.03, with a median value of 0.01916726 (se = 0.005084841).

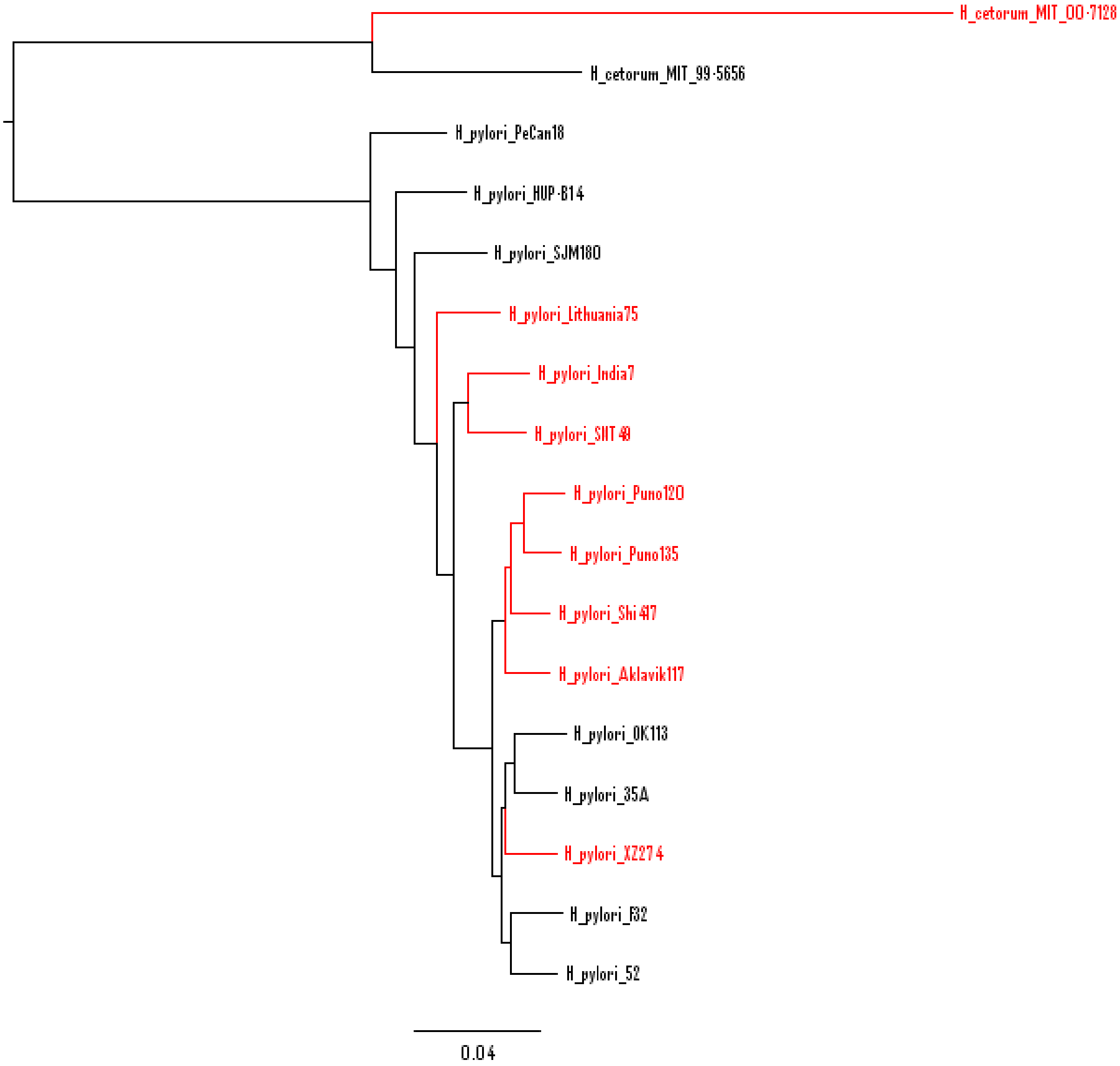


Figure 3: Gene tree showing branch lengths and topology for closely related species of *Helicobacter*. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

The statistical significance of the estimated missing data proportions was analysed using a permutation test (Figure 4). The mean missing data proportion representative of the null hypothesis was found to be 0.0156913, while for the species possessing a CRISPR-Cas system it was 0.008632798. The difference between these two values was not deemed significant, as the mean missing data proportion for the species with CRISPRs was lower than only 68.3% of the 1000 random missing data proportions estimated.

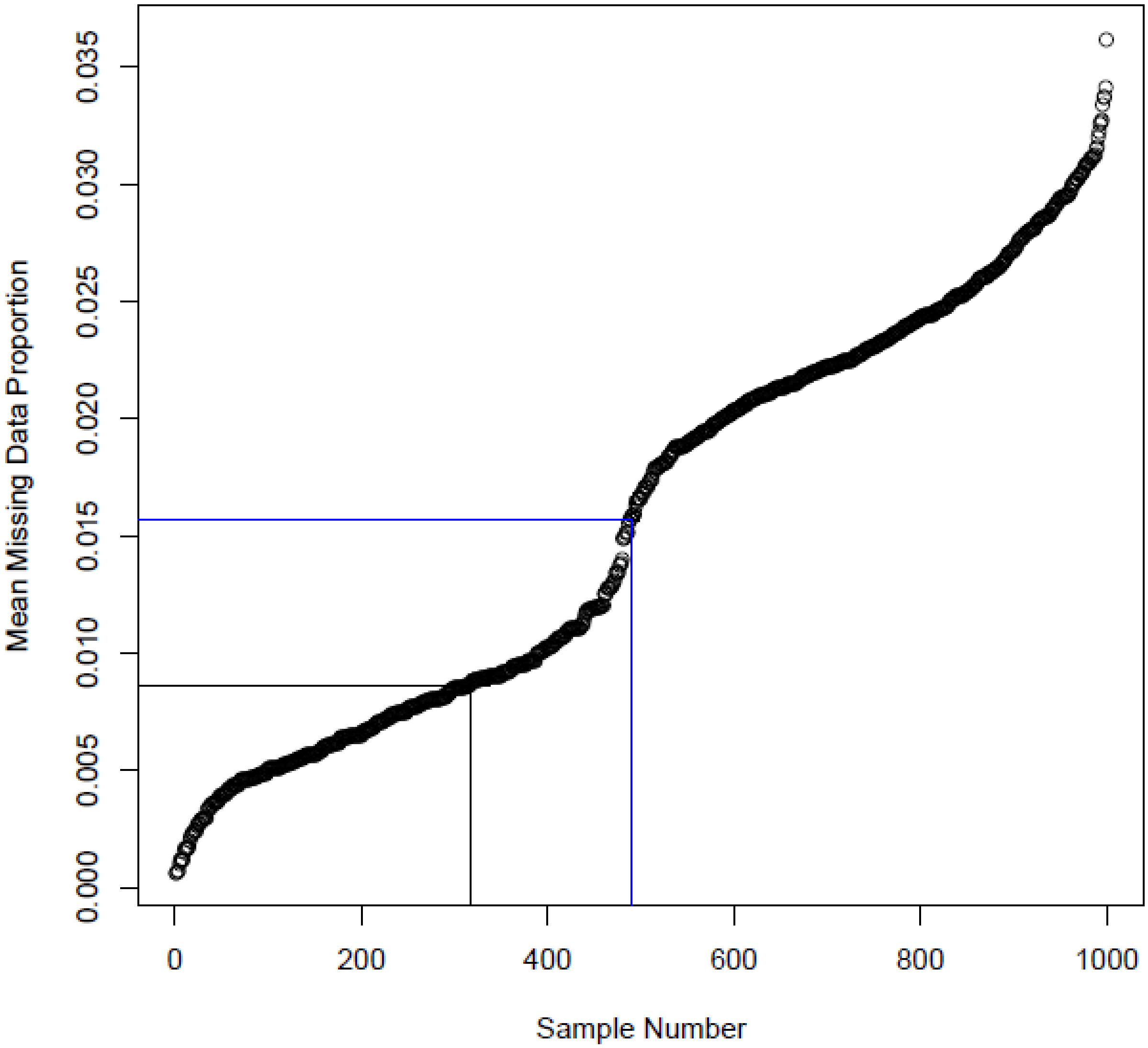
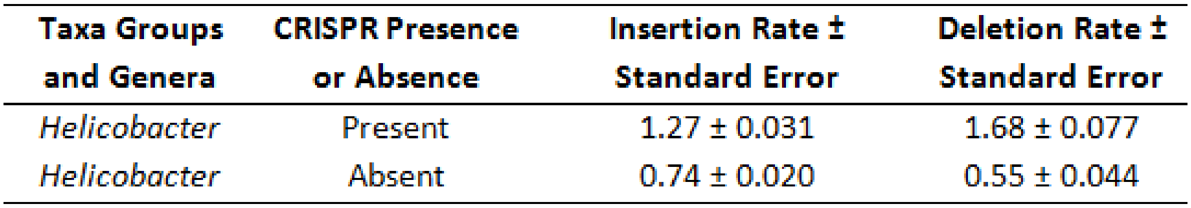


Figure 4: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 9 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each group using markophylo. These gene insertion and deletion rate estimates made by markophylo have been summarized in Table 3. For the taxa with CRISPRs, the gene insertion rate was estimated to be 1.682666 (se = 0.07683400), while the deletion rate was estimated to be 1.266831 (se = 0.03057321). For the taxa without CRISPRs, the gene insertion and deletion rates were estimated to be lower at 0.5498385 (se = 0.04389421) and 0.04389421 (se = 0.02000439) respectively. Overall, the indel rate estimates of the *Helicobacter* taxa with CRISPRs present are higher than the congeneric species without CRISPRs. The difference between these two rate estimates is significant between the two groups, due to the lack of overlap between the values and their standard errors.

Table 3: Indel rate estimates made by markophylo for the *Helicobacter* species with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Helicobacter* | Present | 1.27 ± 0.031 | 1.68 ± 0.077 |
| *Helicobacter* | Absent | 0.74 ± 0.020 | 0.55 ± 0.044 |

### 3.2 Firmicutes Analysis

Gene family presence/absence data on 7192 genes for 27 OTUs was used in this analysis(Table 14). The OTUs under consideration included 17 *Streptococcus pyogenes* subspecies, 8 *Lactococcus* species, and 2 *Pediococcus* species (Figure 5). From these taxa, 12 of the *S. pyogenes* subspecies possessed a CRISPR-Cas system (Table 15). Under the assumptions of Model 4, indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system. Each of the 3 major clades, as highlighted on the phylogeny in Figure 5, had unique gene insertion and deletion rates estimated. Within each of these clades, the species were assumed to have the same gene insertion and deletion rates.

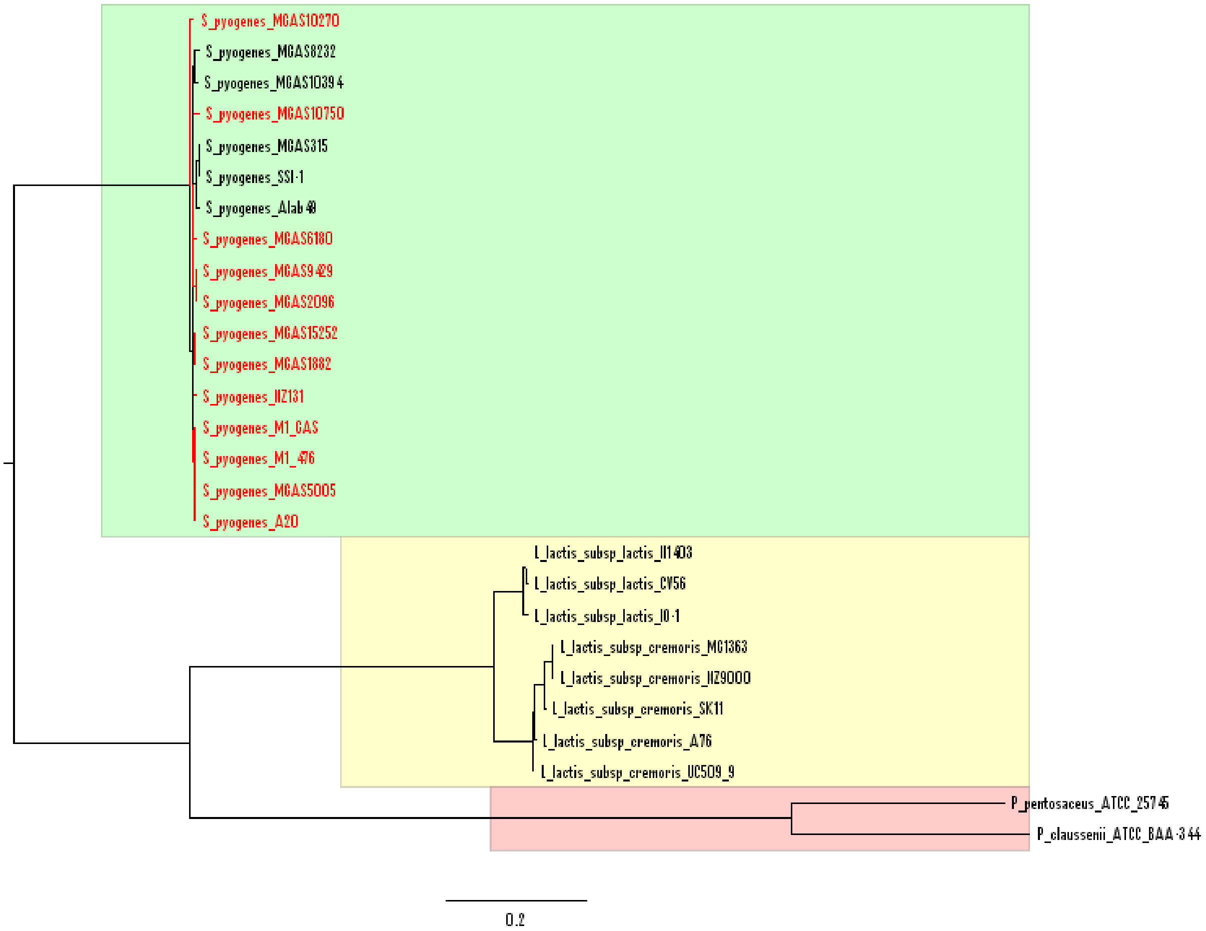
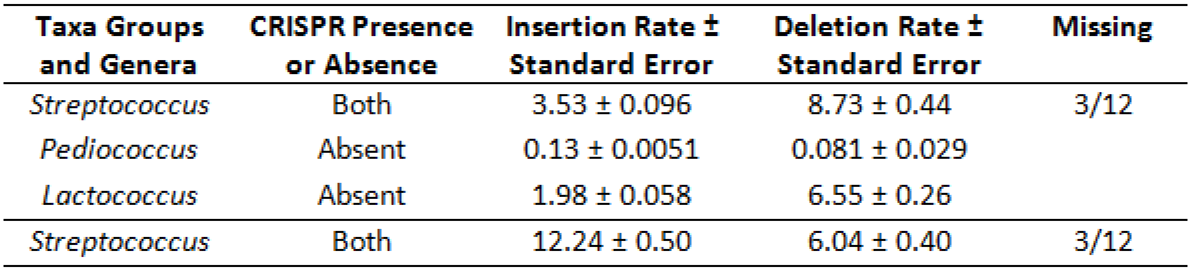


Figure 5: Gene tree showing branch lengths and topology for 3 closely related clades of *Streptococcus* (green), *Lactococcus* (yellow), *Pediococcus* (pink) taxa. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

Table 4: Indel rate estimates made by indelmiss for the firmicutes taxa with and without the out-group genera included in the analysis. The missing ratio numerator is indicative of the number of taxa with CRISPRs that had a high missing data proportion (greater than 0.10). The missing ratio denominator is indicative of the total number of taxa with CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error | Missing |
|---|---|---|---|---|
| *Streptococcus* | Both | 3.53 ± 0.096 | 8.73 ± 0.44 | 3/12 |
| *Pediococcus* | Absent | 0.13 ± 0.0051 | 0.081 ± 0.029 | |
| *Lactococcus* | Absent | 1.98 ± 0.058 | 6.55 ± 0.26 | |
| *Streptococcus* | Both | 12.24 ± 0.50 | 6.04 ± 0.40 | 3/12 |

The gene insertion and deletion rate estimates made by indelmiss have been summarized in Table 4. The *Pediococcus* clade had the lowest estimated gene insertion and deletion rates of 0.1271587 (se = 0.005054188) and 0.08054812 (se = 0.02870433), respectively. The *Lactococcus* clade had an estimated gene insertion rate of 1.984484 (se = 0.05775128), and an estimated deletion rate of 6.549941 (se = 0.2637572). The *Streptococcus* clade had the highest estimated gene insertion and deletion rates of 3.533837 (se = 0.09592228) and 8.731377 (se = 0.4422417), respectively. 9 of the missing data proportions were estimated to be between 0 and 0.05, while 3 missing data proportions were between 0.10 and 0.16. *S. pyogenes MGAS9429* had an estimated missing data proportion of 0.1555685 (se = 0.008985615), *S. pyogenes M1 476* of 0.1220621 (se = 0.009025957), and *S. pyogenes MGAS2096* of 0.1010776 (se = 0.007791975).

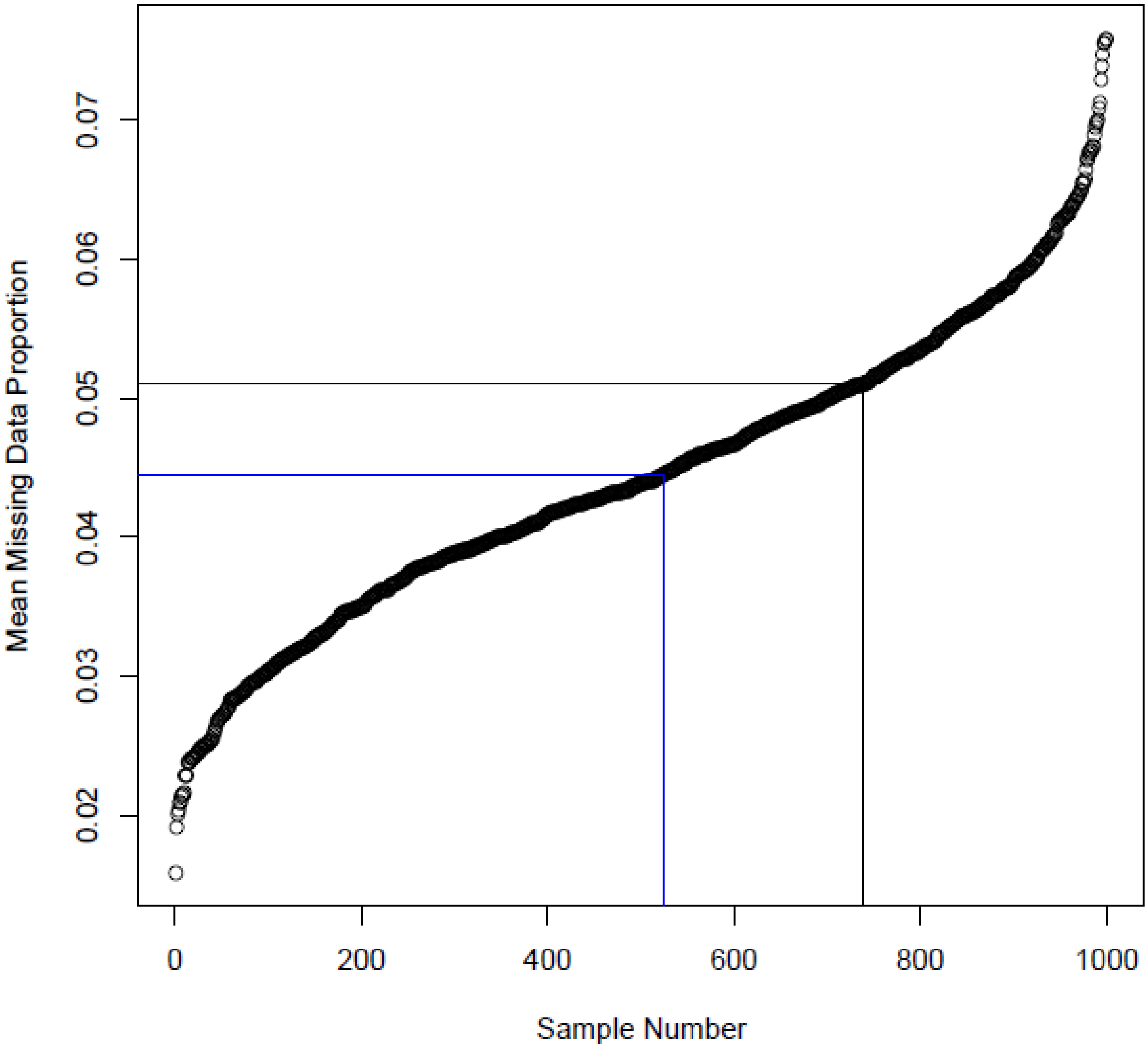
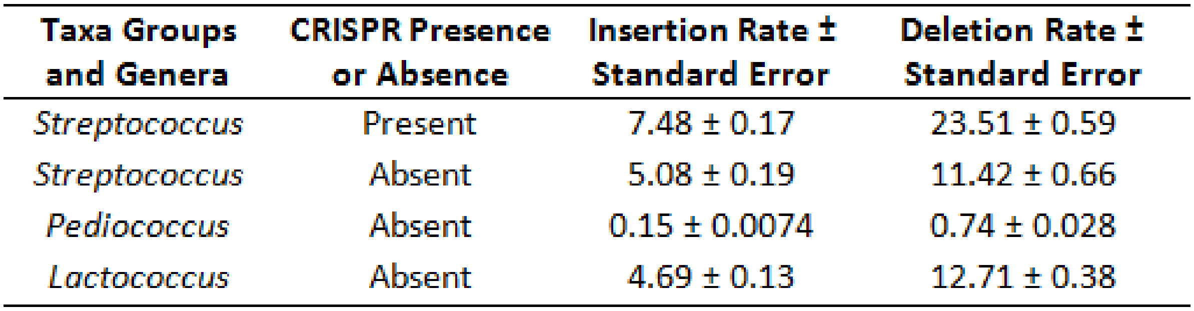


Figure 6: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 12 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

The statistical significance of the estimated missing value proportions was analysed using a permutation test (Figure 6). The mean missing value proportion representative of the null hypothesis was found to be 0.04442048, while for the species possessing a CRISPR-Cas system it was 0.05101146. The mean missing data proportion for the species with CRISPRs was higher than only 73.8% of the 1000 random missing value proportions estimated, and so the difference between it and the null hypothesis was not deemed significant.

Table 5: Indel rate estimates made by markophylo for the firmicutes taxa with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Streptococcus* | Present | 7.48 ± 0.17 | 23.51 ± 0.59 |
| *Streptococcus* | Absent | 5.08 ± 0.19 | 11.42 ± 0.66 |
| *Pediococcus* | Absent | 0.15 ± 0.0074 | 0.74 ± 0.028 |
| *Lactococcus* | Absent | 4.69 ± 0.13 | 12.71 ± 0.38 |

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each clade using markophylo. These gene insertion and deletion rate estimates made by markophylo have been summarized in Table 5. The *S. pyogenes* taxa with CRISPRs had the highest estimated gene deletion and insertion rates of 23.511697 (se = 0.5867292) and 7.484862 (se = 0.1735709), respectively. For the *S. pyogenes* taxa without CRISPRs, the gene deletion rate was estimated to be 11.422443 (se = 0.6593649), while the gene insertion rate was estimated to be 5.083424 (se = 0.1938592). The *Pediococcus* taxa had the lowest estimated gene deletion and insertion rates of 0.7396203 (se = 0.027824567) and 0.1467953 (se = 0.007432489), respectively. Similar to the *S. pyogenes* taxa without CRISPRs, the *Lactococcus* taxa had estimated gene deletion and insertion rates of 12.711737 (se = 0.3788496) and 4.691671 (se = 0.1319604) respectively. Overall, the indel rate estimates of the *Streptococcus* taxa with CRISPRs present are higher than both the out-group genera and the congeneric species without CRISPRs. The difference between each of the insertion and deletion rate estimates is significant, due to the lack of overlap between the values and their standard errors.

There were concerns that the branch lengths associated with the 3 clades of the phylogeny were too long for indelmiss and markophylo to reliably function. These concerns stemmed from the high indel rate estimates found for the *Idiomarina* taxa in the gammaproteobacteria group. The data was reconsidered without the out-groups to include only the *Streptococcus* clade so as to increase the accuracy of the programs.

Gene family presence/absence data on 6079 genes for the 17 closely-related *S. pyogenes subspecies* was used for the analysis (Table 16). Indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system under the assumptions of Model 4. Homogeneous indel rates were assumed for all branches of the phylogeny (Figure 7). The gene insertion and deletion rate estimates made by indelmiss have been summarized in

Table 4. In contrast to the previous assessment, the gene insertion rate for the *S. pyogenes* clade increased to an estimated 6.036089 (se = 0.4027212), while the gene deletion rate increased to an estimated 12.24041 (se = 0.5005551). 9 of the missing data proportions were estimated to be between 0 and 0.07, while 3 missing values were between 0.10 and 0.16. *S. pyogenes MGAS9429* had an estimated missing data proportion of 0.1553967 (se = 0.008949496), *S. pyogenes M1 476* of 0.1144709 (se = 0.008756957), and *S. pyogenes MGAS2096* of 0.1005255 (se = 0.007745599). a subset of the 17 OTUs used.

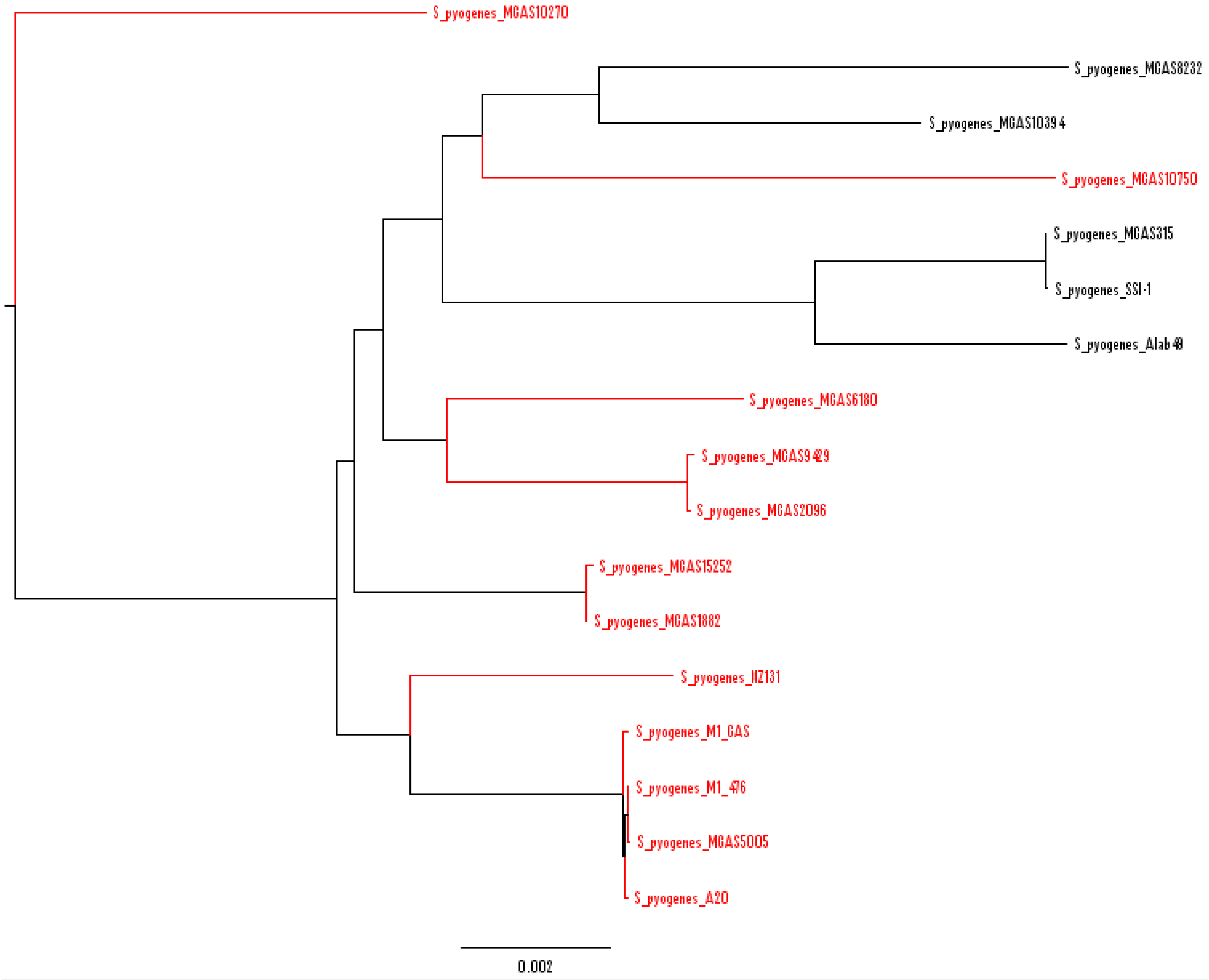


Figure 7: Gene tree showing branch lengths and topology for closely related subspecies of *S. pyogenes*. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

The statistical significance of the estimated missing data proportions was analysed using a permutation test (Figure 8). The mean missing data proportion representative of the null hypothesis was found to be 0.04441278, while for the species possessing a CRISPR-Cas system it was 0.04754556. The mean missing data proportion for the species with CRISPRs was higher than only 61% of the 1000 random missing value proportions estimated, and so the difference between it and the null hypothesis was not deemed significant.

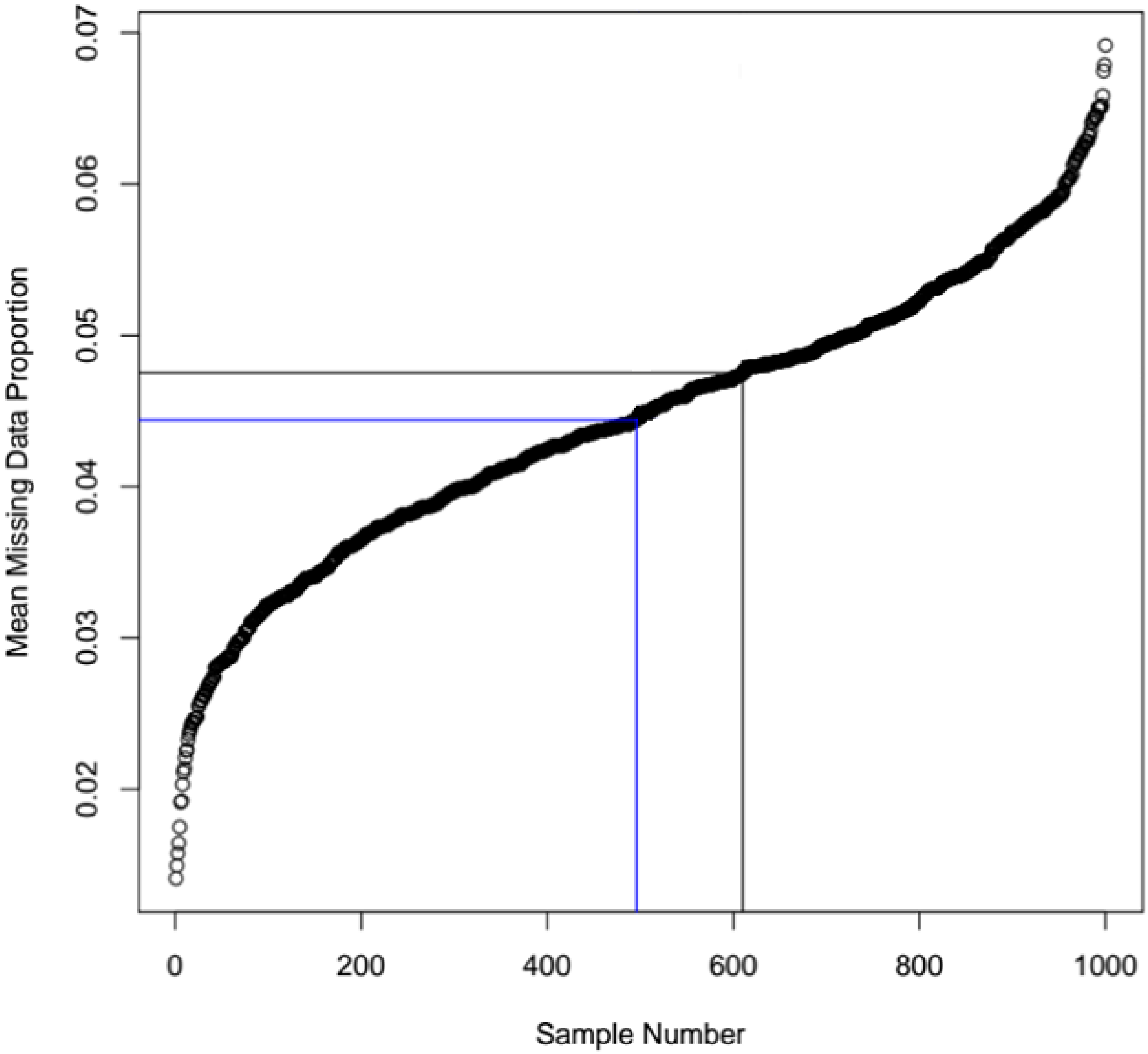
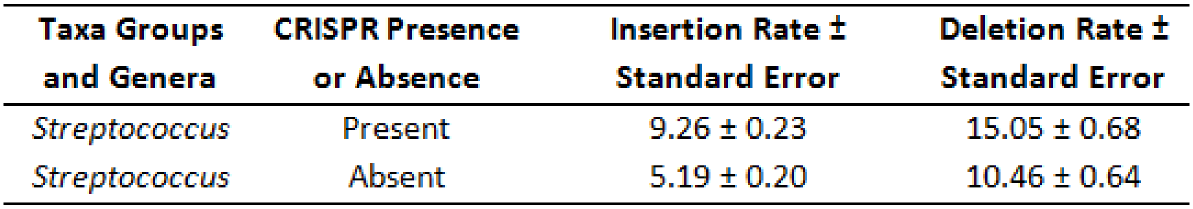


Figure 8: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 12 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each group using markophylo. These gene insertion and deletion rate estimates made by markophylo have been summarized in Table 6. For the taxa without CRISPRs, the gene insertion rate was estimated to be 10.459402 (se = 0.6419734), while the deletion rate was estimated to be 5.192652 (se = 0.1960772). For the taxa with CRISPRs, the gene insertion and deletion rates were estimated to be higher at 15.050502 (se = 0.6812003) and 9.259561 (se = 0.2288604) respectively. Overall, the indel rate estimates of the *Streptococcus* taxa with CRISPRs present are higher than the congeneric species without CRISPRs. The difference between these two rate estimates is significant between the two groups due to the large lack of overlap between the values and their standard errors.

Table 6: Indel rate estimates made by markophylo for the *Streptococcus* species with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Streptococcus* | Present | 9.26 ± 0.23 | 15.05 ± 0.68 |
| *Streptococcus* | Absent | 5.19 ± 0.20 | 10.46 ± 0.64 |

Gammaproteobacteria Analysis

Gene family presence/absence data on 13,702 genes for 20 OTUs was used in this analysis (Table 18). The OTUs under consideration included 15 *Shewanella* species, 3 *Marinobacter* species, and 2 *Idiomarina* species (Figure 9). From these taxa, 8 of the *Shewanella* species possessed a CRISPR-Cas system (Table 19). Under the assumptions of Model 4, indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system. Each of the 3 major clades, as highlighted on the phylogeny in Figure 9, had unique gene insertion and deletion rates estimated. Within each of these clades, the species were assumed to have the same gene insertion and deletion rates.

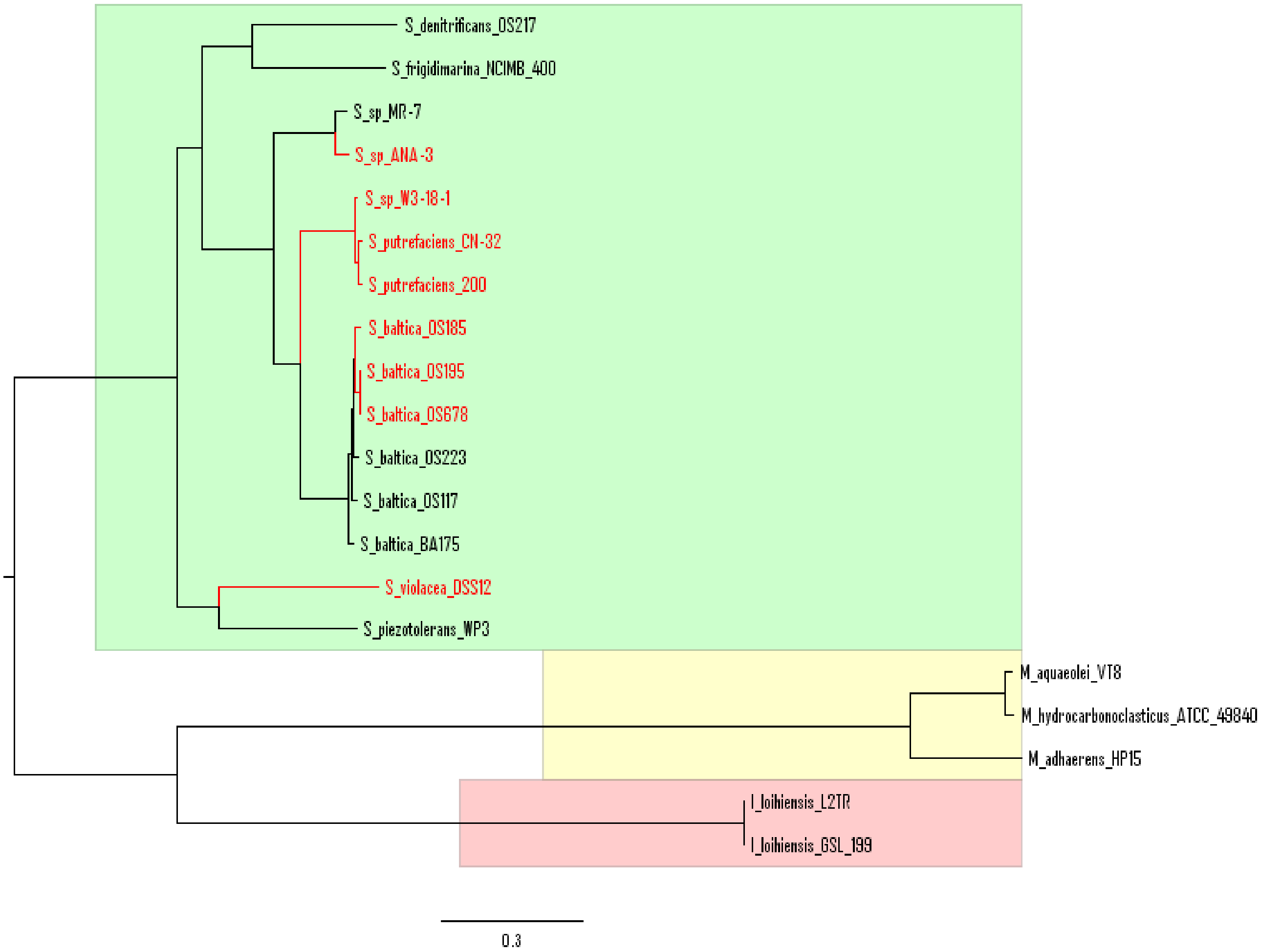


Figure 9: Gene tree showing branch lengths and topology for 3 closely related clades of *Shewanella* (green), *Marinobacter* (yellow), *Idiomarina* (pink) taxa. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

Table 7: Indel rate estimates made by indelmiss for the gammaproteobacteria taxa with and without the out-group genera included in the analysis. The missing ratio numerator is indicative of the number of taxa with CRISPRs that had a high missing data proportion (greater than 0.10). The missing ratio denominator is indicative of the total number of taxa with CRISPRs present.

| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error | Missing |
|---|---|---|---|---|
| *Shewanella* | Both | 0.30 ± 0.0073 | 1.38 ± 0.046 | 3/8 |
| *Idiomarina* | Absent | 100.00 ± 2.15 | 100.00 ± 5.15 | |
| *Marinobacter* | Absent | 0.32 ± 0.0046 | 0.75 ± 0.016 | |
| *Shewanella* | Both | 0.24 ± 0.0065 | 0.38 ± 0.012 | 0/8 |

The gene insertion and deletion rate estimates made by indelmiss have been summarized in Table 7. The *Marinobacter* clade had the lowest estimated gene insertion and deletion rates of 0.3248971 (se = 0.004622675) and 0.7492346 (se = 0.01620109), respectively. The *Idiomarina* clade had the highest estimated gene insertion and deletion rates of 100.0000 (se = 2.146956) and 100.0000 (se = 5.148462), respectively. The *Shewanella* clade had an estimated gene insertion rate of 0.3005813 (se = 0.007291252) and an estimated gene deletion rate of 1.378048 (se = 0.04553837). 5 of the missing data proportions were estimated to be between 0 and 0.03, while 3 missing data proportions were between 0.25 and 0.30.

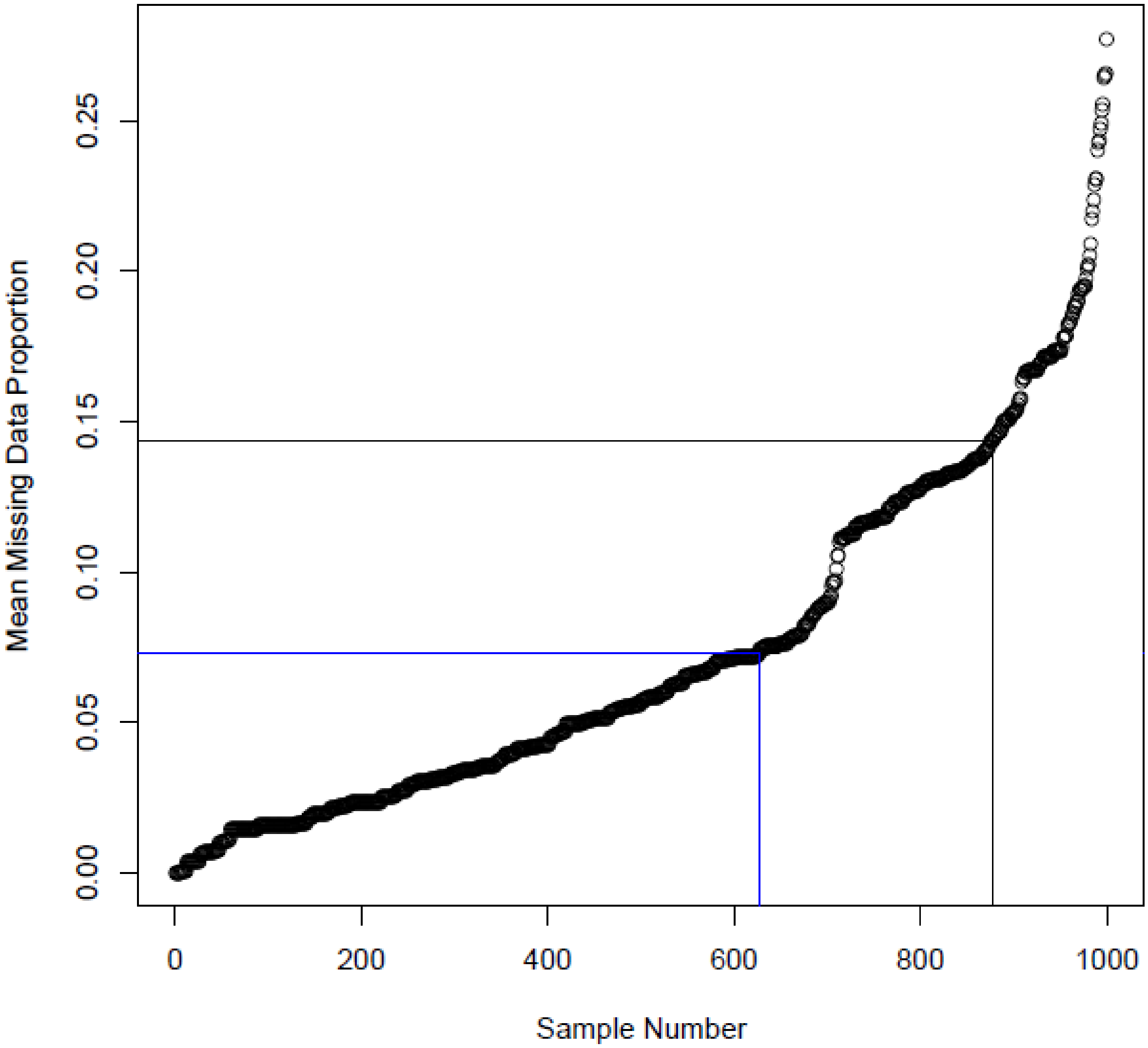
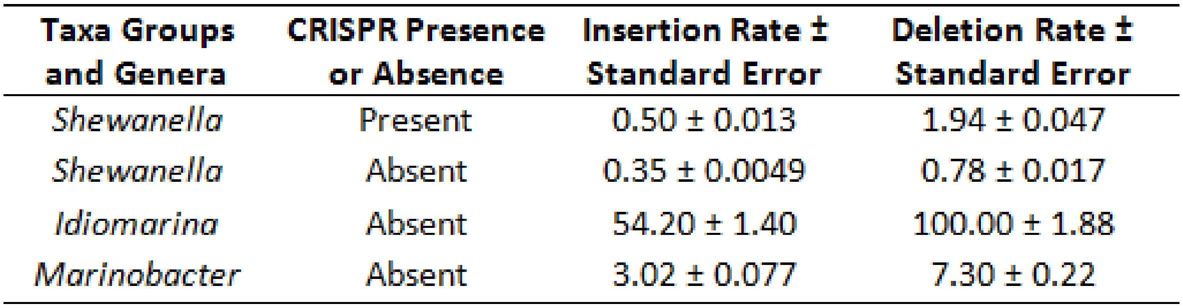


Figure 10: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 8 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

The statistical significance of the estimated missing data proportions was analyzed using a permutation test (Figure 10). The mean missing data proportion representative of the null hypothesis was found to be 0.0728491, while for the species possessing a

CRISPR-Cas system it was 0.1434643. The mean missing data proportion for the species with CRISPRs was higher than only 87.8% of the 1000 random missing data proportions estimated, and so the difference between it and the null hypothesis was not deemed significant.

Table 8: Indel rate estimates made by markophylo for the gammaproteobacteria taxa with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Shewanella* | Present | 0.50 ± 0.013 | 1.94 ± 0.047 |
| *Shewanella* | Absent | 0.35 ± 0.0049 | 0.78 ± 0.017 |
| *Idiomarina* | Absent | 54.20 ± 1.40 | 100.00 ± 1.88 |
| *Marinobacter* | Absent | 3.02 ± 0.077 | 7.30 ± 0.22 |

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each clade using markophylo. These gene insertion and deletion rate estimates made by markophylo have been summarized in Table 8. The *Shewanella* taxa with CRISPRs had an estimated gene deletion rate of 1.9357049 (se = 0.04679549) and an estimated gene insertion rate of 0.5049471 (se = 0.01277100). For the *Shewanella* taxa without CRISPRs, the gene deletion and insertion rates were estimated to be lower at 0.7837352 (se = 0.016572800) and 0.3492713 (se = 0.004940908), respectively. The *Idiomarina* taxa had the highest estimated gene deletion and insertion rates at 100.00000 (se = 1.882145) and 54.20031 (se = 1.402581), respectively. Gene insertion and deletion rate estimates of 100 are set at the upper limits of the program, and indicate program inaccuracy due to issues with the data. For the *Marinobacter* taxa, the gene deletion rate was estimated to be 7.300083 (se = 0.22158142), while the gene insertion rate was estimated to be 3.021449 (se = 0.07744549). Overall, the indel rate estimates of the *Shewanella* taxa with CRISPRs present are lower than the out-group genera without CRISPRs. In addition, the indel rate estimates of the *Shewanella* taxa with CRISPRs are higher than the congeneric species without CRISPRs. The difference between each of the gene insertion and deletion rate estimates is significant, due to the lack of overlap between the values and their standard errors.

Due to the high indel rate estimates found for the *Idiomarina* taxa, there were concerns that the branch lengths associated with the 3 clades of the phylogeny were too long for indelmiss and markophylo to reliably function. The data was reconsidered without the out-groups to include only the *Shewanella* clade so as to increase the accuracy of both programs.

Gene family presence/absence data on 10,006 genes for the 15 closely-related *Shewanella* species was used for the analysis (Table 20). Indelmiss was used to fit missing data proportions for all taxa with a CRISPR-Cas system under the assumptions of Model 4. Homogeneous indel rates were assumed for all branches of the phylogeny (Figure 11).

The gene insertion and deletion rate estimates made by indelmiss have been summarized in Table 7. In contrast to the previous assessment, the gene deletion and insertion rates for the *Shewanella* clade decreased to an estimated 0.3822619 (se = 0.01168795) and 0.2391315 (se = 0.006510731), respectively. All 8 of the missing data proportions were estimated to be between 0 and 0.05, with a median missing data proportion of 0.02175813 (se = 0.00325863). This decrease in the high missing data proportion estimates for 3 of the taxa found when the out-groups were included in the analysis demonstrates that the accuracy of the programs are increased when the branch lengths across the phylogeny are shortened.

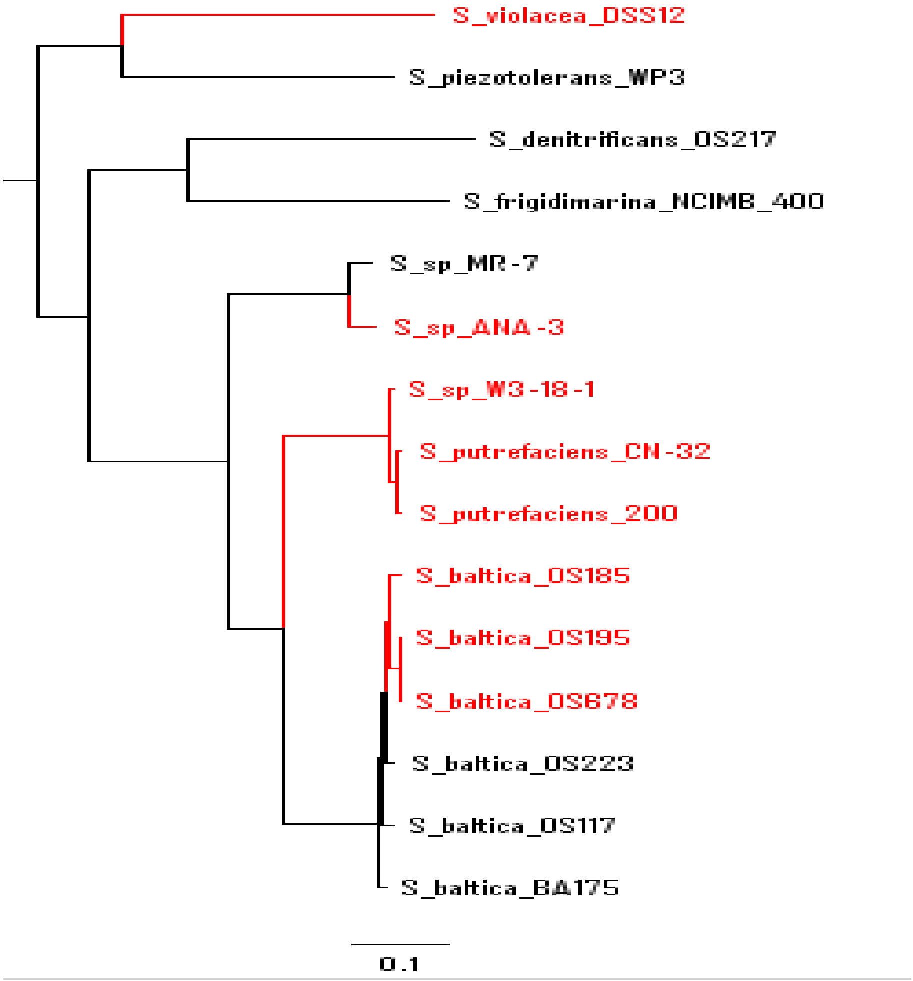


Figure 11: Gene tree showing branch lengths and topology for closely related species of *Shewanella*. Red branches denote the presence of a CRISPR-Cas system in the taxa at the tip, while black branches denote the absence of a CRISPR-Cas system in the taxa at the tip.

The statistical significance of the estimated missing data proportions was analyzed using a permutation test (Figure 12). The mean missing data proportion representative of the null hypothesis was found to be 0.02200131, while for the species possessing a

CRISPR-Cas system it was 0.02959815. The mean missing data proportion for the species with CRISPRs was higher than only 83.8% of the 1000 random missing data proportions estimated, and so the difference between it and the null hypothesis was not deemed significant.

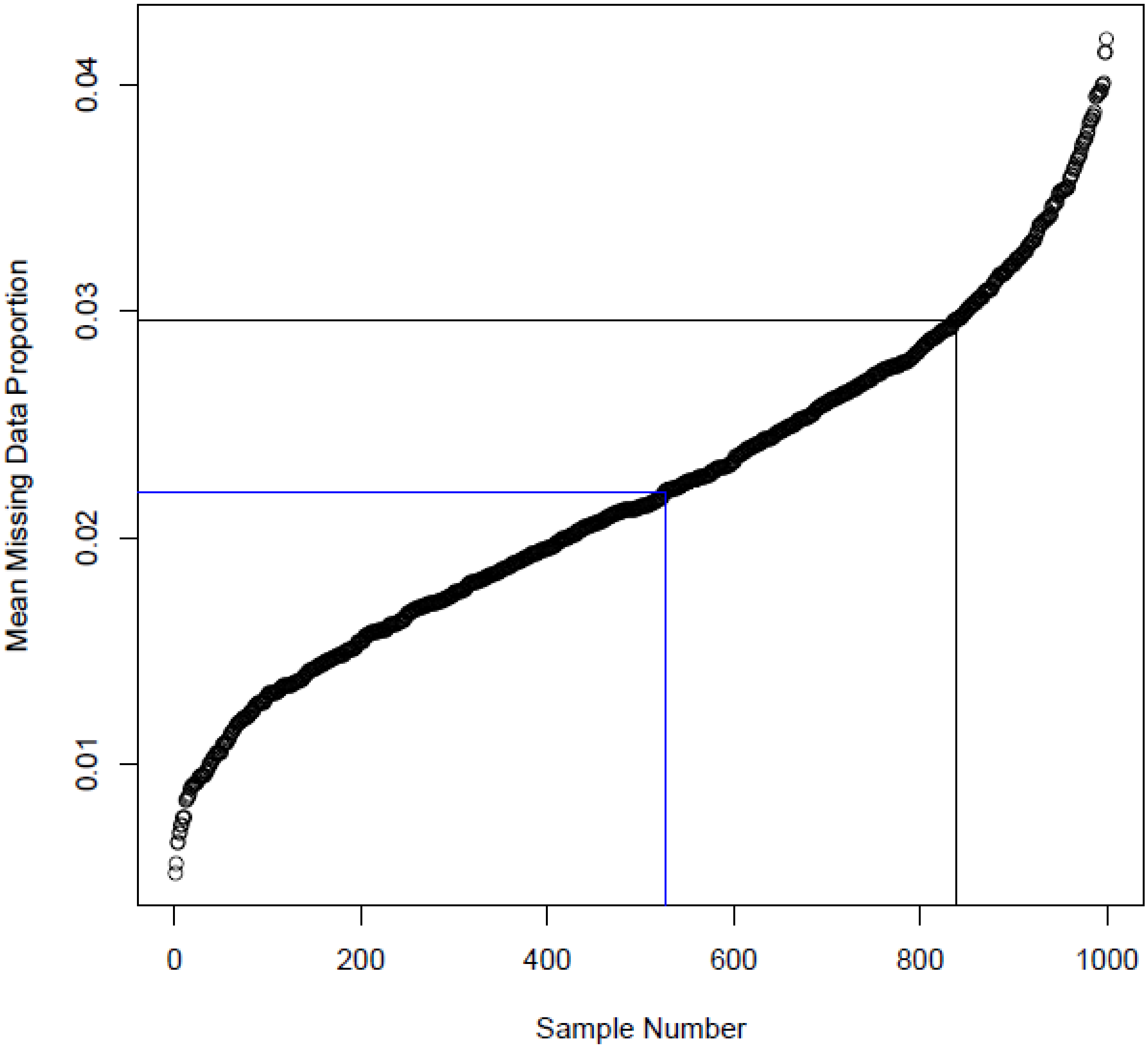
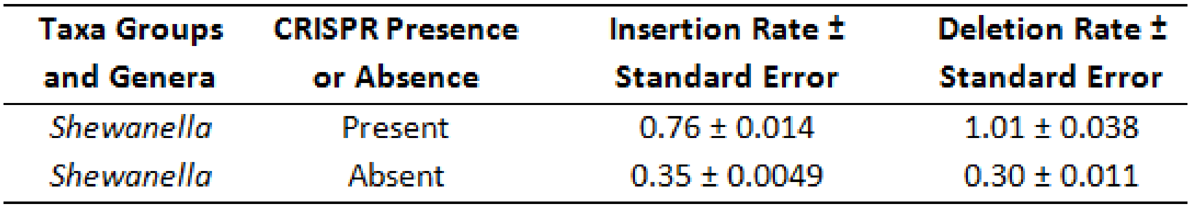


Figure 12: Plot of 1000 samples of missing data proportions (black circles) for random subsets of 8 taxa with the null hypothesis (blue line) and test statistic (black line) denoted.

To determine whether the taxa with CRISPRs were experiencing higher rates of LGT compared to those without CRISPRs, the gene insertion and deletion rates were estimated uniquely for each group using markophylo. The gene insertion and deletion rate estimates made by markophylo have been summarized in Table 9. For the taxa with CRISPRs, the gene deletion rate was estimated to be 1.0116474 (se = 0.03754635), while the insertion rate was estimated to be 0.7602147 (se = 0.01374589). For the taxa without CRISPRs, the gene deletion and insertion rates were estimated to be lower at 0.2962197 (se = 0.01128673) and 0.3537919 (se = 0.00489081), respectively. Overall, the indel rate estimates of the *Shewanella* taxa with CRISPRs present are higher than the congeneric species without CRISPRs. The difference between these two rate estimates is significant between the two groups due to the large lack of overlap between the values and their standard errors.

Table 9: Indel rate estimates made by markophylo for the *Shewanella* species with and without CRISPRs present.



| Taxa Groups and Genera | CRISPR Presence or Absence | Insertion Rate ± Standard Error | Deletion Rate ± Standard Error |
|---|---|---|---|
| *Shewanella* | Present | 0.76 ± 0.014 | 1.01 ± 0.038 |
| *Shewanella* | Absent | 0.35 ± 0.0049 | 0.30 ± 0.011 |

### 3.3 Indelmiss Simulation

This simulation was conducted in order to test the ability of indelmiss to detect a difference between the gene insertion and deletion rates of congeneric species with and without CRISPRs. This was done due to the lack of statistical power inherent in the permutation tests previously shown. Three different scaling factors were tested for both the gene deletion and insertion rate simulations, and their associated p-values have been plotted in Figure 13 and Figure 14 respectively.

The three scaling factors used to create differential gene deletion rates between the species with and without CRISPRs were 1.05, 1.1, and 1.3. Each of these values represents the product by which the branch lengths associated with the species with CRISPRs were scaled. A p-value was calculated for each of these scaling factors to measure the significance with which indelmiss is capable of distinguishing between the gene deletion rates of the two groups. When the scaling factor was set to 1.3, the p-value was lowest at $2.2 \times 10^{-16}$, representing a large degree of significance between the gene deletion rate estimates of the two groups. When the scaling factor was set to 1.1, the p-value was 0.002061, indicative of a lesser degree of significance between the gene deletion rate estimates of the two groups. When the scaling factor was set to 1.05, the p-value became 0.4609, representing a lack of significance between the gene deletion rate estimates of the two groups. An exponential regression was fit to the data, and was used to predict the scaling factor at which the p-value becomes 0.05. The model predicted that a scaling factor of 1.06011 would produce a p-value of 0.05, denoting that any deviations in the gene deletion rate between the CRISPR and non-CRISPR species groups below 6.011% are not reflected in the estimates of indelmiss. This demonstrates that any difference observed in the deletion rate estimates and their standard errors for species with and without CRISPRs is statistically significant, and that this difference is greater than a 6.011% deviation between the groups.

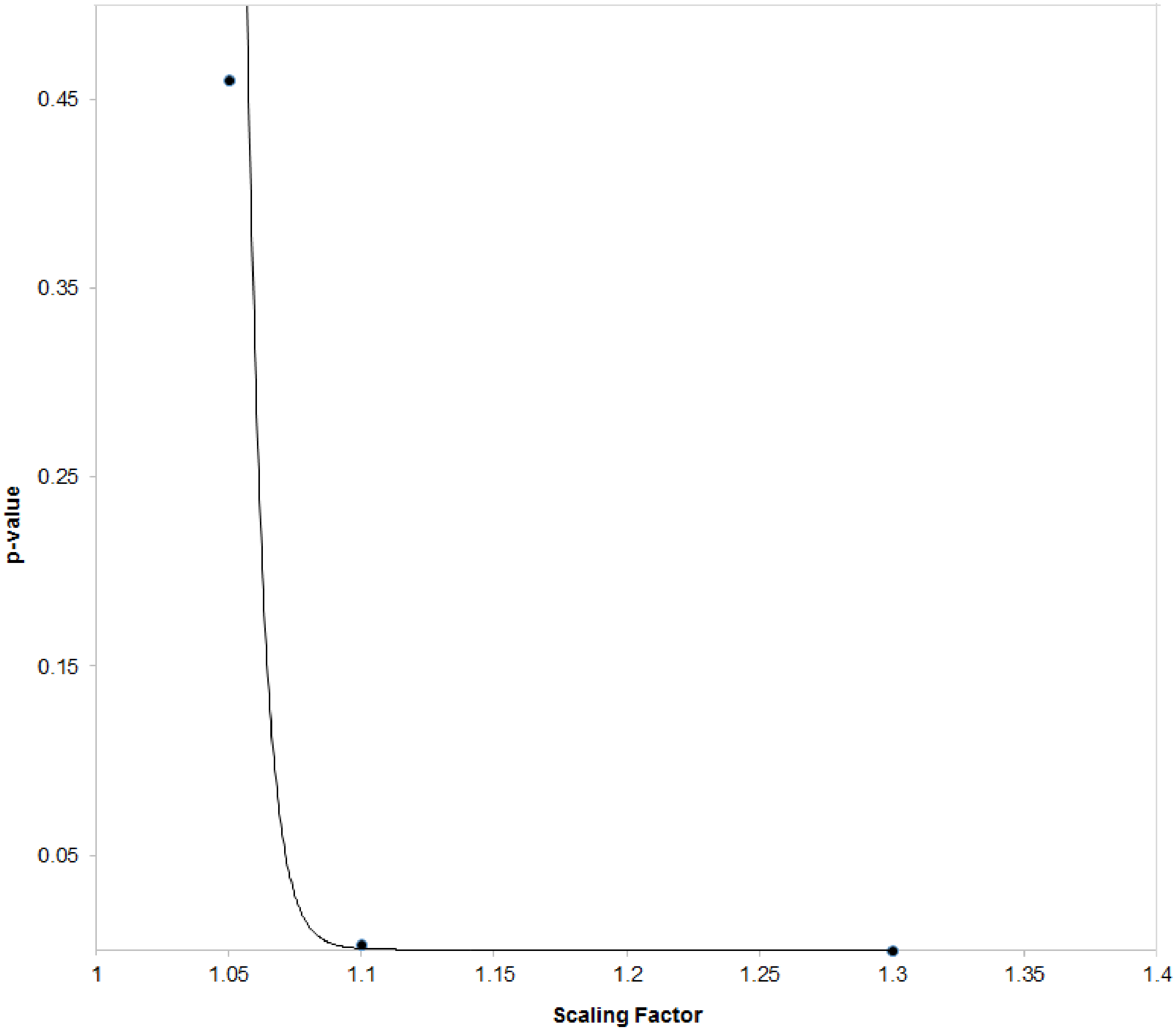


Figure 13: Plot of the gene deletion rate scaling factor used in the simulation and the associated p-value denoting the significance with which indelmiss estimates the inequality between the mu values of CRISPR and non-CRISPR species. An exponential regression has been fit to the points with an $R^2$ of 0.99477.

The three scaling factors used to create differential gene insertion rates between the species with and without CRISPRs were 0.75, 0.90, and 0.95. Each of these values represents the product by which the branch lengths of the species with CRISPRs were scaled. A p-value was calculated for each of these scaling factors to measure the significance with which indelmiss is capable of distinguishing between the gene insertion rates of the two groups. When the scaling factor was set to 0.75, the p-value was lowest at $2.2 \times 10^{-16}$, representing a large degree of significance between the gene insertion rate estimates of the two groups. When the scaling factor was set to 0.90, the p-value was $3.066 \times 10^{-5}$, indicative of a lesser degree of significance between the gene insertion rate estimates of the two groups. When the scaling factor was set to 0.95, the p-value became 0.3904, representing a lack of significance in the gene insertion estimates between the two groups.

An exponential regression was fit to the data, and was used to predict the scaling factor at which the p-value becomes 0.05. The model predicted that a scaling factor of 0.92157 would produce a p-value of 0.05, denoting that any deviations of the gene insertion rates between the CRISPR and non-CRISPR species groups below 7.843% are not reflected in the estimates of indelmiss. This demonstrates that any difference observed in the insertion rate estimates and their standard errors for species with and without CRISPRs is statistically significant, and that this difference is greater than a 7.843% deviation between the groups.

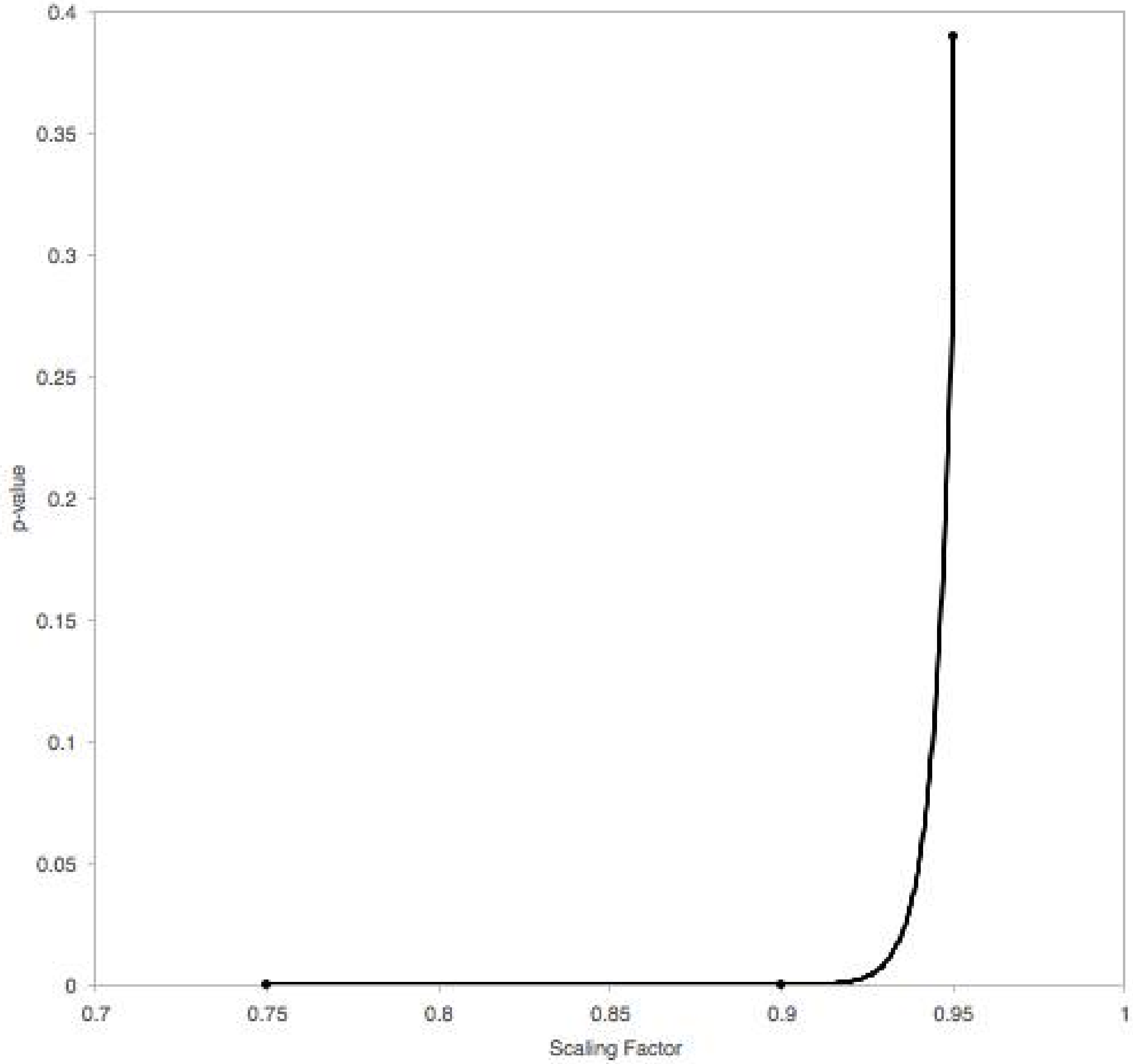


Figure 14: Plot of the gene insertion rate scaling factor used in the simulation and the associated p-value denoting the significance with which indelmiss estimates the inequality between the nu values of CRISPR and non-CRISPR species. An exponential regression has been fit to the points with an $R^2$ of 0.99958.

# 4 Discussion

It was originally hypothesized that lower rates of LGT would be observed in species with CRISPR-Cas systems compared to closely related species without them. There are three major findings relevant to this prediction that will be discussed in turn. Firstly, it was found that the hypothesis was only supported in data comparing proteobacteria species, and could not be replicated in groups of firmicutes. Subsequently, when comparing indel rates of species with CRISPR-Cas systems to closely related congeneric species without them, the opposite trend was consistently observed. Finally, the hypothesis was only supported when comparing rates of LGT between species with CRISPR-Cas systems and closely related genera that entirely lack them.

## 4.1 Proteobacteria and Firmicutes Analysis

The results support the hypothesis that lower rates of LGT are observed in species with CRISPR-Cas systems compared to closely related genera without them. However, this finding was only supported when comparing LGT rates between species of proteobacteria, and could not be replicated in firmicutes. Significantly lower gene insertion and deletion rates were observed in *Helicobacte*r species possessing a CRISPR-Cas system compared to both *Caulobacter* and *Edwardsiella* species without one. Similarly, significantly lower gene indel rates were observed in *Shewanella* species with a CRISPR-Cas system compared to both *Idiomarina* and *Marinobacter* species without one. The opposite trend was observed in the firmicutes species considered, for which subspecies of *S. pyogenes* in possession of a CRISPR-Cas system had significantly higher indel rates compared with out-group species of both the *Lactococcus* and *Pediococcus* genera.

This inherent difference in the effect of CRISPR-Cas system presence on LGT rates may be due to inherent ecological and organismal differences between the two phyla, and warrants further investigation. One possibility is that firmicutes species are often used in industry for food and beverage production, while the proteobacteria used here are often pathogenic and reside within organisms. Since many of these organisms adopt homeostatic regulation of their internal conditions, the proteobacteria may be presented with less variation in environmental conditions [16]. With greater niche specialization, novel genes may pose less of a potential benefit to proteobacteria species and may not be inserted or maintained as efficiently.

## 4.2 LGT Rates among Congeneric Species

The gene insertion and deletion rates of species in possession of a CRISPR-Cas system were consistently higher compared to closely-related congeneric species. This trend is the opposite of what was predicted based on the mechanistic action and purpose of CRISPR-Cas systems in preventing insertion of exogenous DNA. This outcome was consistently supported independent of whether the species under consideration were firmicutes or proteobacteria, despite their immense ecological differences and great evolutionary divergence. As evidence, significantly higher gene insertion and deletion rates were measured when comparing *Helicobacter* species with CRISPRs to those without. This result was replicated when comparing both *Shewanella* and *Streptococcus* species with CRISPRs to congeneric species without them. In addition, this effect was consistent irrespective of whether or not the out-group species were included in the analysis, and so branch lengths did not cause a difference in the gene indel rate estimates of markophylo.

Despite the significant increases found in LGT rates, an associated change in the missing data proportion of these species was not observed. Similar levels of statistical insignificance for the missing data proportions were found irrespective of whether out-group species were included in the analysis or not. However, branch lengths did seem to play a role in the accuracy of missing data proportion estimates by indelmiss. For instance, three *Shewanella* estimates exceeded 0.25 when out-groups were included in the analysis, which then decreased to less than 0.05 once the branch lengths were shortened. However, the high missing data proportions found for three *S. pyogenes* subspecies were consistent irrespective of whether or not out-group species were included in the analysis. This means that high missing data proportion estimates by indelmiss are not necessarily due to inaccuracy, but should be verified through shortening of the branch lengths on the phylogeny. No evidence of a deletion bias in the relevant *S. pyogenes* subspecies could be found, though further investigations are warranted to explain the high missing data proportions observed.

The lack of consistently high missing data proportion estimates for the species with CRISPRs suggests that the presence of a CRISPR-Cas system is not causing a change in the amount of genetic information present in the organism. In addition, the missing data proportions do not correlate with the gene insertion and deletion rate estimates of markophylo. For instance, both *Helicobacter* and *Shewanella* showed lower indel rates in species with CRISPRs compared to closely related genera. In addition, both of these genera also had higher indel rates in species with CRISPRs compared to congeneric species without them. Furthermore, both *Helicobacter* and *Shewanella* species with CRISPRs were found by markophylo to have deletion biases, while congeneric species without CRISPRs instead showed an insertion bias. Despite this, *Helicobacter* species with CRISPRs were estimated to have lower missing data proportions while *Shewanella* species with CRISPRs were estimated to have higher missing data proportions compared to the stochastic mean. In addition, the inherent genome sizes of the species being compared do not seem to be causing the effect either. For instance, while *Caulobacter* and *Edwardsiella* genomes average at approximately 4 Mb, *Helicobacter* genomes tend to average at around 2 Mb. Based strictly on genome size, the observed result that *Helicobacter* has an excess of genetic information contradicts what is expected. Based on this data, there seems to be some extraneous variable that is causing the difference in the direction of missing data proportion estimates. It is also possible that the lack of statistical power inherent in the permutation test is not showing a true interpretation of the significance of the missing data proportions.

A mechanism must be proposed to explain the higher gene insertion and deletion rates observed in species with a CRISPR-Cas system compared to closely-related congeneric species without them. This result contradicts the expectation that CRISPR-Cas systems function to reduce gene insertion by eliminating exogenous DNA for which a spacer exists. According to the trade-off hypothesis, two well-supported assumptions can be made about an organism's environment based on the presence or absence of a CRISPR-Cas system. In order for a CRISPR-Cas system to be maintained, the environment must possess either a high density of infectious phage elements that threaten the bacteria's genomic integrity, or a lack of exogenous DNA that can increase the fitness of the organism. In other words, if a bacterial species has selected against and therefore lacks a CRISPR-Cas system, it is likely that the environment lacks a high phage density and contains high quality exogenous DNA that can increase the fitness of the organism if laterally transferred. Since environments with a high phage density rely on the presence of a thriving bacterial population, it is more likely that such environments would have more exogenous DNA present [12]. If so, it is more likely that the CRISPR-Cas system inherent in these bacteria would reach its maximum efficiency in blocking the insertion of such DNA. Since the exogenous DNA is unlikely to infer a fitness advantage due to the sustained selection for the CRISPR system, the organism must adopt a high deletion bias to rid of the unnecessary information. In contrast, species that lack a CRISPR-Cas system have low phage presence in the environment, meaning that the bacterial population is at a lower density [12]. If so, then there is unlikely to be a high presence of exogenous DNA in the environment for insertion to take place. In addition, the available DNA is more likely to infer a fitness advantage since the CRISPR system has been selected against, and so it is less likely to be deleted. Both of these scenarios account for a higher indel rate to be present in bacteria with

CRISPR-Cas systems present than without, as was observed in the results. The *in-vitro* experiments that explore CRISPR-Cas dynamics limit themselves to considering only one or two exogenous genes [13]. This system of experimentation is not applicable to real environments in which these organisms may be presented with a high degree of exogenous DNA. Further investigations should attempt to mimic conditions in the field so as to gain a more applicable knowledge of CRISPR dynamics.

### 4.3 LGT Rates among Closely Related Genera

The initial hypothesis was only supported when comparing rates of LGT between species with CRISPR-Cas systems present and closely related genera that entirely lack them. The results demonstrated that *Helicobacter* species that possessed a CRISPR-Cas system had lower rates of gene insertion and deletion than species of both *Edwardsiella* and *Caulobacter.* Similarly, *Shewanella* species that possessed a CRISPR-Cas system had lower indel rates than species of both *Marinobacter* and *Idiomarina.* Based on this information, it can be inferred that these species also undergo lower rates of LGT. This result is still explained by the proposed mechanism in section 4.2 which accounts for the higher observed LGT rates in species with CRISPR-Cas systems present compared to congeneric species without them. Genera that have no CRISPR-Cas systems available to be selected for do not show the dynamics of CRISPR loss and gain. This being said, the environmental assumptions of the trade-off hypothesis no longer apply. In such a case, species may be exposed to a variety of different conditions with a range of phage densities and exogenous gene qualities. Throughout this environmental variation, the ability of the organism to engage in LGT is not restricted by the selection forces for and against a CRISPR-Cas system. This lack of restriction may therefore translate to the expected result of having higher rates of gene insertion and deletion.

In addition, a consideration of the accuracy of the programs used in the analysis is essential to discussing the validity of the results. Out-groups with only two species demonstrated strange results with biases for either very low or very high gene indel rate estimates by indelmiss. For instance, the gene deletion rate of *Idiomarina* species was estimated to be 100 (se $=$ 1.882145) by markophylo. This unusually high estimate may also be due to the long branch lengths associated with including the out-group genera on the phylogeny, which reduces the accuracy of the programs. However, when further genomes have been assembled and annotated for the relevant genera, the analysis should be replicated with a larger sample size.

In the article by Gophna et al. (2015), the authors conclude that on evolutionary timescales, the inhibitory effect of CRISPRs on LGT is not supported by evidence. Our findings contradict this conclusion and demonstrate that by using methods that estimate gene indel rates with high statistical power, the inhibitory effect of CRISPRs on LGT can be observed over short evolutionary timescales. In further investigations of the effects of CRISPR-Cas system presence on LGT, we suggest the use of methodologies that limit timescales to prevent an overwhelming amount of variation in the data, which can mask significant findings.

## References

[1] H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304, 2000.

[2] M.A. Ragan and R.G. Beiko. Lateral genetic transfer: open issues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:2241–2251, 2009.

[3] J. Davison. Genetic exchange between bacteria in the environment. *Plasmid*, 42:73–91, 1999.

[4] U. Gophna, D.M. Kristensen, Y.I. Wolf, O. Popa, C. Drevet, and E.V. Koonin. No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J*, 2015.

[5] H. Philippe and C.J. Douady. Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology*, 6:498–505, 2003.

[6] J. Bondy-Denomy and A.R. Davidson. To acquire or resist: the complex biological effects of CRISPR-Cas systems. *Trends Microbiol*, 22:218–225, 2014.

[7] A. Pawluk, J. Bondy-Denomy, V.H. Cheung, K.L. Maxwell, and A.R. Davidson. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR-Cas system of *Pseudomonas aeruginosa*i. *MBio*, 5:e00896, 2014.

[8] E.V. Koonin and K.S. Makarova. CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol*, 10:679–686, 2013.

[9] B. Shen, W. Zhang, J. Zhang, J. Zhou, J. Wang, L. Chen, L. Wang, A. Hodgkins, V. Iyer, X. Huang, et al. Efficient genome modification by crispr-cas9 nickase with minimal off-target effects. *Nature Methods*, 11:399–402, 2014.

[10] W. Jiang, I. Maniv, F. Arain, Y. Wang, B.R. Levin, and L.A. Marraffini. Dealing with the evolutionary downside of CRISPR immunity: bacteria and beneficial plasmids. *PLoS Genet*, 9:e1003844, 2013.

[11] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, and P. Horvath. Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315:1709–1712, 2007.

[12] E.V. Koonin and Y.I. Wolf. Evolution of the CRISPR-Cas adaptive immunity systems in prokaryotes: models and observations on virus-host coevolution. *Mol Biosyst*, 11:20–27, 2015.

[13] D. Bikard, A. Hatoum-Aslan, D. Mucida, and L.A. Marraffini. Crispr interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host and Microbe*, 12:177–186, 2012.

[14] U.J. Dang, A.M. Devault, H.N. Poinar, C. Pepperell, T. Mortimer, and G.B. Golding. Gene insertion deletion analysis while accounting for possible missing data. *in manuscript*, pages 1–28, 2015.

[15] U.J. Dang and G.B. Golding. markophylo: Markov chain analysis on phylogenetic trees. *Bioinformatics*, page btv541, 2015.

[16] S. Suerbaum and C. Josenhans. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nature Reviews Microbiology*, 5:441–452, 2007.

[17] M. Yang, Y. Lv, J. Xiao, H. Wu, H. Zheng, Q. Liu, Y. Zhang, and Q. Wang. *Edwardsiella* comparative phylogenomics reveal the new intra/inter-species taxonomic relationships, virulence evolution and niche adaptation mechanisms. *PloS One*, 7:e36987, 2012.

[18] E.A. O'Neill, R.H. Hynes, and R.A. Bender. Recombination deficient mutant of *Caulobacter crescentus*. *Molecular and General Genetics MGG*, 198:275–278, 1985.

[19] K. Todar. *Streptococcus pyogenes* and streptococcal disease. *Todarâs Online Textbook of Bacteriology, University of Wisconsin-Madison Department of Bacteriology*, 2008.

[20] K.H. Schleifer, J. Kraus, C. Dvorak, R. Kilpper-Bälz, M.D. Collins, and W. Fischer. Transfer of *Streptococcus lactis* and related streptococci to the genus lactococcus gen. nov. *Systematic and Applied Microbiology*, 6:183–195, 1985.

[21] T.M. Cogan, M. Barbosa, E. Beuvier, B. Bianchi-Salvadori, P.S. Cocconcelli, I. Fernandes, J. Gomez, R. Gomez, G. Kalantzopoulos, A. Ledda, et al. Characterization of the lactic acid bacteria in artisanal dairy products. *Journal of Dairy Research*, 64:409–421, 1997.

[22] H.P. Fleming, R.F. McFeeters, and M.A. Daeschel. The lactobacilli, pediococci, and leuconostocs: vegetable products. *Bacterial Starter Cultures for Foods*, pages 97–118, 1985.

[23] K.K. Sharma and U. Kalawat. Emerging infections: *Shewanella* - a series of five cases. *Journal of Laboratory Physicians*, 2:61, 2010.

[24] M.J. Gauthier, B. Lafay, R. Christen, L. Fernandez, M. Acquaviva, P. Bonin, and J.C. Bertrand. *Marinobacter hydrocarbonoclasticus* gen. nov., sp. nov., a new, extremely halotolerant, hydrocarbon-degrading marine bacterium. *International Journal of Systematic and Evolutionary Microbiology*, 42:568–576, 1992.

[25] S.P. Donachie, S. Hou, T.S. Gregory, A. Malahoff, and M. Alam. *Idiomarina loihiensis* sp. nov., a halophilic $\gamma$-proteobacterium from the lō âihi submarine volcano, hawai âi. *International Journal of Systematic and Evolutionary Microbiology*, 53:1873–1879, 2003.

[26] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30:3059–3066, 2002.

[27] F. Ronquist and J.P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574, 2003.

[28] A. Rambaut and A. Drummond. Figtree: Tree figure drawing tool, version 1.2. 2, 2008.

[29] L. Harmon, J. Weir, C. Brock, R. Glor, W. Challenger, and G. Hunt. Geiger: analysis of evolutionary diversification. r package version 1.3-1. *See http://CRAN.R-project.org/package=geiger*, 2009.