Is Sharing Caring?

Elucidating the Effects of the Presence of CRISPR-Cas Systems on Rates of Horizontal Gene Transfer Using Network Analysis

MolBiol 4C12 Thesis

Siddharth Reed*[1] and G. Brian Golding[1]
[1]Department of Biology, McMaster University, Hamilton, Canada

April 2, 2019

Abstract

Horiznotal Gene Transfer (HGT) is a mechanism by which organisms (mainly prokaryotes) can share genetic material outside of inheritance. HGT has proven to have significant effects on bacterial genome evolution, allowing for increased genetic diversity and niche adaptation. CRISPR associated (CRISPR-Cas) enzymes are an adaptive immune system in prokaryotes that has garnered much research attention due to it's application as a gene editing tool. While much of the focus on CRISPR-Cas systems has been related to this application, CRISPR-Cas has been shown to have a highly complex interaction with HGT. Effort has mostly been focused on how CRISPR-Cas systems affect the mechanisms of HGT and little is currently known about the effects of CRISPR-Cas on HGT rate, both for individuals and on a population level. Proposed here is a network-theoretic approach to further the understanding of the effects of the presence of CRISPR-Cas systems on HGT within a population. Network theory has already been applied to better model evolution and relatedness among bacteria, accounting for HGT which traditional phylogenetic methods ignore. This network-theoretic approach allows for study of CRISPR-Cas effects on individual bacteria as well as population level effects on HGT. Understanding the effects of CRISPR-Cas on HGT may help develop strategies to curb spreading antibiotic resistance, understanding bacterial evolution and extend the functionality of CRISPR-Cas gene editing systems.

*To whom correspondence should be addressed; reeds4@mcmaster.ca

Contents

1 CRISPR-Cas Systems

1.1 What Are They?

CRISPR-Cas systems are sets of nucleotide motifs (spacers) interspaced with nucleotide repeats (CRISPRs) and CRISPR-associated (Cas) proteins (usually adjacent to the CRISPR motifs) that have an adaptive immune function in many bacteria and archaea [1]. Each nucleotide motif is indicative of some DNA sequence that was taken up previously by the host and serves as a marker for the Cas proteins to degrade any foreign DNA matching this motif [1]. If a bacterium which possesses a CRISPR-Cas system is infected with a phage and survives, a motif representative of that phage can be integrated into the CRISPR repeats so that when the, bacterium is reinfected with the same phage strain it will be detected and degraded by a Cas protein before genomic integration can occur. CRISPR is an adaptive immune system, as the bacterium acquire resistance after an unsuccessful infection through spacer integration.

Although CRISPR-Cas is primarily considered an immune system, non-viral spacers representative of bacterial Mobile Genetic Element (MGE)s have been found to compose the majority of (detectable) CRISPR spacers [2]. In fact, many spacers have no detectable match to a viral sequence, being termed CRISPR "dark matter", indicating that knowledge about the acquisition of spacers and their effects on bacterial gene dynamics leaves much to be desired [2].

1.2 Diversity, Ubiquity And Detection

As of 2017, over 45% of bacterial genomes analyzed ($n = 6782$) appear to contain CRISPR motifs [3]. Moreover, CRISPR motifs show significant diversity between organisms, since they represent a chronological history of viral infection or MGE "infection" for that specific organism [1]. Cas proteins themselves show significant diversification, segregating into entirely different CRISPR-Cas systems [4]. There still exist many bacterial strains, and even entire genera with no *known* CRISPR-Cas systems, although they may have simply not been discovered yet [5, 6].

Between $11\% - 28\%$ of sequenced genomes have either only CRISPR repeats *or* Cas loci, but not both [5]. There also exist repetitive motifs that may superficially resemble CRISPRs, but have low spacer diversity and no Cas genes [5]. False detection of CRISPR systems is significantly increased by only examining repeat-spacer structural patterns, other parameters such as spacer dissimilarity and genomic context should be considered to reduce false positives [5]. Especially as sequencing efforts continue, better mechanistic understandings of CRISPR systems develop and CRISPR systems themselves propagate and transfer between bacteria they will continue to become more relevant and diverse [1]. Furthermore, the diversification of CRISPR-Cas systems is driven further by HGT acting on CRISPR and Cas components independently, adding another level of complexity to the propagation of CRISPR-Cas systems [1].

1.3 Applications In Biotechnology

While CRISPRs are interesting systems to study from a microbiological perspective, much of the current research interest (and funding) is motivated by applications of CRISPR to gene editing. The CRISPR-Cas9 system has been adapted into a simple, efficient tool for gene editing in both prokary-

otes and eukaryotes [1]. The Cas-9 protein induces a double-strand break to a region homologous to a guide RNA, which can be synthesized by a researcher. The break will then be re-annealed by DNA repair enzymes, often introducing errors (insertions,deletions,etc.) into the sequence, disrupting gene function [1]. A gene can also be inserted at the break point via homology directed repair by including DNA sequence with flanking arms homologous to the break region [1].

2 Horizontal Gene Transfer

HGT can be defined as the exchange of genetic information across lineages [7]. The word horizontal is in contrast to what can be referred to as vertical inheritance, between parents and offspring [8]. HGT is often a source of genetic variation, allowing organisms to respond to selective pressures much more quickly by copying an evolved function from another organism, rather than having to evolve new functions in genes themselves [8, 9].

2.1 Mechanisms

Transformation The uptake of free floating exogenous DNA by a bacterium and the incorporation of it into bacterium's genome [7]. Many factors can influence the competency (capability of transformation) of bacteria naturally, such as DNA damage, selective pressures, cell density and multiple methods have been found to induce competency for experimental purposes (cloning) [10].

Conjugation The sharing of genetic material through cell-to-cell bridges, usually carried on either a self-transmissible or non-self-transmissible plasmid [11].

Transduction The transfer of genes between bacteria through a bacteriophage [12]. When a donor cell infected by a phages is lysed, the lysed bacterial DNA fragments can accidentally be taken up into the phage head [12]. When the phage infects a new bacterium the lysed donor fragments are released into the recipient cell, where they can recombine into the genome [12]. The above method can transfer random fragments of DNA between bacterial cells, but there are more sequence specific methods of transfer through lysogenic phage [12]. Lysogenic phage incorporate themselves into specific regions of a bacterial genomes [12]. When they excise themselves they can accidentally incorporate bacterial DNA flanking the incorporated phage DNA and bring it with them to the next phage target [12].

It should be noted that *successful* HGT requires that a gene be maintained, either by genomic integration or plasmid replication. Frequently, putatively transferred genes are either lost quickly after transfer or evolve with little functional constraint, due to minimal selective pressure maintaining them [13].

2.2 Rate Influencing Factors

The rate of HGT in bacteria is constantly in flux, in part due to the amount of DNA available for transfer [14]. If there are low levels of exogenous DNA, low population density or low phage density, reduced HGT will be observed as less DNA available for transfer [7]. But just like mutation rates, HGT rates are thought to evolve in response to environmental factors or selective pressure [15,

16]. For strains of bacteria found in hospitals, the potential benefit of receiving antibiotic resistance genes via HGT may far outweigh any potential danger or metabolic cost, inducing an increase in a bacterium's uptake of foreign DNA. [17] There are clear metabolic costs for HGT, as host machinery to allow competency and conjugation are not trivial to synthesize [18]. Further, conjugation and transformation are not discriminatory processes, so DNA encoding for toxic products, having sub-optimal codon distribution or incompatible GC content may be taken up, but cannot be successfully incorporated or consistently expressed [18]. Conjugated plasmids may also be incompatible with a host due to the replication machinery required by the plasmid [19]. In fact it has been suggested that genes recently acquired via HGT are often quickly lost, having been lost for conferring no advantage or for conferring a specific advantage, that was lost with the removal of the maintaining selective pressure [13]. Ultimately HGT rates are influenced by a variety of factors related to fitness costs/benefits and mechanistic barriers associated with the genes being gained.

2.3 Pan-genomes

As sequencing costs have decreased, re-sequencing of strains and sequencing of many similar strains has grown drastically. The comparison of multiple genomes from strains of the same species yielding interesting results: many genes are not found in most of the strains sequenced [20]. This has lead to the concept of a pan-genome, the sum total of all unique genes among a set of strains [21]. A pan-genome has two parts: a core genome consisting of genes common to all strains in a species and an accessory genome, consisting of unique genes present in any of the strains [21]. In *Escherichia coli* (*E. coli*), as the total number of strains sequenced increases, the total number of unique genes increases logarithmically, meaning more unique genes are being identified with every new strain sequenced [22]. These accessory genes are prime candidates for HGT because they only appear in certain strains and may provide some niche-specific adaptation, such as antibiotic resistance [21]. The accessory genome can be considered a genetic toolbox that strains have access to through HGT, although this access is also limited by barriers to HGT (i.e. distance, genetic incompatibility etc.). Pan-genomes can further be categorized as open, if they appear to be expanding, adding more genes from more distant OTUs or closed, with the total number of unique genes plateauing as more strains are sequenced [21]. *E. coli* is an example of a open pan-genome, as the more sequences are obtained the larger the set of unique genes in all *E. coli* sequences becomes, with no clear asymptote visible [21].

2.4 Applications

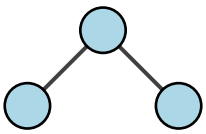
While the vast majority of HGT has been observed to occur between prokaryotes, cases have been identified between prokaryotes and eukaryotes. One particular case is the beetle, *Hypothenemus hampei*, gaining a gene from a bacterial strain colonizing its midgut becoming able to colonize coffee beans [23]. This pest beetle has become a huge issue for coffee farmers, estimated to result in over $20 million USD loses to rural farming families internationally [23].

Another much more important reason for studying HGT is that antibiotic resistance genes have been shown to transfer frequently [24]. The transfer of antibiotic resistance genes is so prevalent that the term resistome has be coined to refer to the set of resistance genes that an organism can acquire via HGT [24]. Understanding the dynamics of HGT and ways to limit or inhibit it specifically may prove integral to resolving the issue of the decreasing range of antibiotic effectiveness [24].

3 Network Theory

Network theory is an extension of graph theory, a branch of mathematics concerning the properties of "graphs". Graphs in this context refers to a set of nodes and a set of edges between those nodes, with edges typically representing some kind of relationship between those nodes [25]. Network theory focuses on modeling interactions using graphs and applying tools built to analyze graphs to gain an understanding of networks and how they function.

Consider a social networking site like Facebook, are people more likely to be friends with people who have a similar number of friends? These relationships can be modelled using a network, with nodes represent users and an edge between nodes represent whether two users are Facebook friends. To answer the above question, the assortativity of the network can be calculated. Assortativity is a measure of the network's nodes preference to form edges where more similar nodes [25]. Similarity here refers to the difference in the number of edges connected to each node, *i.e.* the number of friends each user has. Therefore if Bob and Alice have similar numbers of friends, they themselves are more likely to be friends with each other than people with different numbers of friends than them in a high assortativity network. If a network has a large assortativity value, similar nodes connect to each other more often than different ones [25]. Thus our question can be reformed through the lens of network theory as "Does a network constructed from Facebook user data have high assortativity?"



While the above example is a simple network, this model can be further extended through:

- Adding directions to edges (from one node to another)
- Adding weights to edges (often representing data about node interactions)
- Adding attributes to nodes themselves (binary,discrete or continuous)

An example of a directed, weighted biological network with low assortativity is a gene expression network (nodes:genes, edges:transcription level correlations) with few transcription factors which each modulates expression of multiple unrelated genes. For this project, nodes will represent OTUs and edges will represent genetic exchange between those OTUs, whereas in normal phylogenies, edges only represent taxonomic relationships. Despite the complexity of HGT, network theory allows a flexible theoretic framework to analyze these interactions that are normally ignored by traditional phylogenetic methods.

Phylogenomic Networks

HGT is an important factor in understanding evolution in prokaryotes. Since HGT has been found to be frequent throughout the prokaryotic tree of life, this has lead many to re-evaluate the concept of a "tree of life", which by definition ignores these horizontal interactions [26].

Prokaryotic Net Of Life

In graph theory a tree is defined as a graph where there is only one path between every pair of nodes. In phylogenetics this implies there is only one path for genetic material to transfer between organisms, that path being vertical inheritance. As HGT demonstrates, this tree model is clearly an incomplete representation of genetic relationships between OTUs. Genetic material can be transferred outside of reproduction, allowing for multiple paths by which a single gene can be found in two OTUs (either inheritance, transfer or some combination of the two) [7]. This prompted the idea of a prokaryotic network of life (as opposed to a tree), with edges indicating both vertical and horizontal transfers of genetic material [26]. Edges can now connect OTUs to closely or distantly related OTUs, and even extinct ancestral OTUs.

Detection

While understanding that HGT is important and networks provide a useful theoretic framework to study it, constructing such networks is not trivial. Many different strategies have been developed to detect potential HGT events given a phylogenetic tree, with some able to detect both recipients and donors [8]. There are two primary sets of methods for detecting HGT.

Parametric These methods rely on investigating the sequence composition (GC%, codon bias, etc.) in genes and when they deviate from the genomic average. Average GC content has been found to vary significantly between some organisms, even by up to 30% in closely related organisms [8]. The same is true for codon bias, where codons variants are observed with different frequency in different bacteria, dependent on the expression levels of the tRNAs in those respective organisms [8, 27]. For example, if *E. coli* contains more copies of a tRNA with the anti-codon TTA (Leucine) than CTC, genes will more likely encode the TTA codon to increase transcription efficiency [27]. If more TTA codons than CTC codons are observed in a gene in *Staphylococcus aureus* (*S. aureus*), assuming *S. aureus* has no leucin codon bias, one may be able to infer that the codon-biased gene was transferred horizontally [8]. Other metrics to consider are GC%, k-mer frequency or the presence of other features around the candidate gene, such as transposases or flanking sequences [8].

Phylogenetic These methods rely on recognizing discordance between gene trees and species trees. If a gene tree is found to have a significantly different topology from a species tree, this difference may be the result of an HGT event [28]. One can also compare the substructures of a gene trees and species trees (created by removing a set of edges leaving a set of sub-trees) to see if the tree substructures disagree [8]. Another strategy involves pruning (removing an edge to get 2 distinct trees) an internal branch and reattaching the subtrees at a different location. If the re-grafted tree has a better fit to the reference tree than the original, this may be indicative of an HGT event between the original node and the node the subtree was re-grafted to [8].

While HGTs can lead to these discordances, there are other series of evolutionary events than can produce the same results [28]. Events that may lead to false diagnosis of HGT are: incomplete lineage sorting, gene duplication followed by loss in one of the descendant lineages or homologous recombination [8, 28]. Strategies to account for these events, as well as account uncertainty in the tress themselves exist, but there still exist other sources that remain unaccounted for [28].

It should also be noted that many of these methods require heuristic solutions, as they are computationally expensive, and sometimes even entirely intractable, which creates further uncertainty in the results obtained [8]. As an example, finding the minimum edit path between 2 trees (as in the re-grafting method) is NP-Hard, but the solution space can be limited by not considering pruning branches between consistent nodes [8, 29].

Generally phylogenetic methods are preferred for multiple reasons:

- Can make use of multiple genomes at once [8]
- Require explicit evolutionary models, which come with their own framework for hypothesis testing and model selection [8].
- HGT events identified by parametric methods are often found by phylogenetic methods as well [8].
- In recent years, the requirements of computing power and multiple well sequenced genomes for phylogenetic methods have become easier and easier to meet [8].

While detecting HGT events with high degrees of certainty is still difficult, much progress has been made in recent years,especially using phylogenetic methods [8].

4 Do CRISPR Systems Affect Horizontal Gene Transfer?

Yes.

4.1 Interference Mechanisms

Since CRISPRs have been shown to be capable of interfering with conjugation (conjugative plasmid specific spacers) and transduction (phage immunity), it has been hypothesized that lower rates of HGT will be observed in strains with CRISPR-Cas systems [30]. CRISPR-Cas systems have also been found to interfere with transformation-mediated HGT, by degrading foreign DNA taken up by a cell [31].

4.2 Complexities And Costs Of CRISPR-Cas Systems

As noted above, CRISPR-Cas systems have been shown to interfere with plasmid conjugation in *S. aureus* by integrating a spacer targeting the *nickase* sequence, necessary for conjugation in *S. aureus* [30]. Since antibiotic resistance genes are often transferred on plasmids, this can incur a significant cost, especially in environments with large amounts of antibiotics (ex: hospitals, trees etc.) [17]. CRISPR-Cas systems incur a metabolic cost, as Cas proteins, guide RNAs, spacer acquisition proteins must all be expressed to maintain immunity [1]. Despite primarily being an immune system, the way CRISPR-Cas functions (degrading foreign DNA matching spacers motifs, resisting phage infection) can have off-target effects on HGT [32]. While resisting lytic phage infection clearly provides some fitness benefit, CRISPR-Cas has also been shown to resist prophage incorporation [32]. Prophages can serve as vectors for HGT, but they can also provide super-infection immunity,

and even reduce competitor bacterial populations through infection [32, 33]. It has also been shown that spacer sequences representative of a bacterium's own chromosomal DNA can be incorporated in to CRISPR array, leading to an auto-immune response where Cas proteins target native host DNA [34]. As CRISPR-Cas systems persist, anti-CRISPR mechanisms have evolved in certain phages, making them immune to CRISPR-Cas, denoted anti-CRISPRs [32]. This has a two-fold effect, as it can increase the susceptibility of the host to infection, reducing the fitness benefit of CRISPR-Cas, but it can also allow for more transduction-mediated HGT [32].

4.3 Potential Strategies For Reducing CRISPR-HGT Trade-off Costs

Due to the myriad of fitness costs associated with consistently expressing CRISPR-Cas systems, bacteria have appeared to develop strategies to mitigate these costs. While CRISPR-Cas systems can confer a fitness advantage by providing immunity to phage infection, the fitness cost associated is complex, especially as CRISPR-Cas systems themselves can be transferred horizontally, either on a plasmid or even through transduction [35]. It has been posited that CRISPR-Cas systems need only be present in a few members of a population at once and transferred between members to maintain phage immunity while reducing the cost of constantly maintaining CRISPR-Cas systems [32]. It has been found that the presence of a CRISPR system does not necessarily imply activity of the system, creating new mechanism(s) by which the fitness cost of CRISPR-Cas systems can be reduced [32]. The presence of CRISPR-Cas systems have also been shown to actually enhance HGT via transduction at the population level by reducing total phage abundance [33]. The presence of CRISPR-Cas systems in Firmicutes have been shown to be associated with increased levels of gene insertion and deletion compared to closely related outgroups, further demonstrating the complexity of this relationship [36]. The effects of CRISPR-Cas systems on rates of HGT are highly complex, owning in no small part to the broad range of CRISPR effects, how CRISPR activity can be modulated and the transfer of CRISPR systems themselves within a population [32]. Taking a systematic approach may help elucidate the dynamics between CRISPR system presence and HGT rate.

5 Hypothesis & Objectives

5.1 Hypothesis

The null hypothesis is that bacterial strains/genera with known CRISPR systems will show no significant differences in network statistics to those strains/genera without known CRISPR systems. Using sequenced genomes, the goal of this project is to construct phylogenetic networks for all strains within sets of genera with and without CRISPR-Cas systems.

5.2 Objectives

Ultimately the goal of this project is to examine the relationship of HGT rates and the presence of CRISPR-Cas systems, using a network theoretic approach. The following sets of comparisons will contribute to the understanding of this relationship:

Within Network Comparisons For genera with strains containing CRISPR and Non-CRISPR species, comparing the network dynamics of those sets of nodes across genera will elucidate if CRISPR-Cas systems affect the HGT rates or the association patterns of individual OTUs.

Gene Indel Rates Vs. Network Statistics Comparing insertion and deletion rates independantly can help further specify what mechanisms may be responsible for trends observed in network statistics. If a mixed network is found to be density connected, but also shows a deletion bias, this may imply that most of the genes being transferred may not confer a fitness advantage.

6 Methods

6.1 Summary

The goal of the project is to create a phylogenetic network from a set of protein fasta files and the corresponding nucleotide sequences. In this case all full genomes for a given bacterial genus, for analysis of HGT. The workflow is as follows for a single genus:

1. Download fasta files
2. Filter mobile genetic elements from genomes
3. Cluster all genes into families using Diamond (% identity $> 80\%$)
4. Construct a presence/absence matrix of gene families for organisms
5. Estimate gene family indel rates separately for the CRISPR and non-CRISPR containing genomes using the R package markophylo
6. Construct a species tree using all 16S rRNA genes that have 1 copy for each member of the genus
   (a) Align each 16S gene with mafft using defaul settings
   (b) Concatenate all alignments together as a nexus file
   (c) Build the tree using Mr Bayes (10000 generations, 25% burn in)
7. Construct the gene trees (≤ 1500)
   (a) Only consider families with a gene belonging in at least 40% of the genomes analyzed (ex: a family with 6 genes in 6 of 15 genomes)
   (b) Align each family using mafft with default settings
   (c) Build a tree for each alignment using Mr Bayes (10000 generations, 25% burn in)
8. Create 1000 subsets of 50 gene trees through bootstrap sampling
9. For each subset, use the program HiDe to infer a phylogenetic network from the species tree and the 50 gene trees.

10. Annotate each network with CRISPR data scraped from the CRISPR-one database.

11. Using the gene indel rates estimated and the annotated networks examine if there are any trends or effects on the dynamics of HGT between organisms with and without CRISPR-Cas systems.

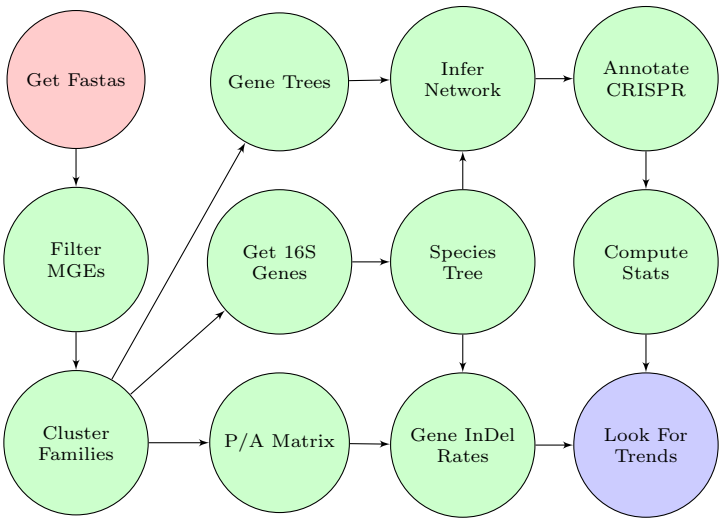


Figure 1: Diagram of the worfkflow for a single genus

6.2 Data Collection

Protein and nucleotide fastas of all CDS sequences were downloaded from NCBI RefSeq. CRISPR annotations of Cas and Cfp proteins from the CRISPRone tool from Zhang and Ye will be used to assess the presence of CRISPR systems [5].

6.3 Gene Presence/Absence Matrix

In order to use the program markophylo to estimate indel rates, a Presence/Absence (P/A) matrix of gene families and organisms and a species tree are required. First any genes classified as MGEs (from NCBI annotations) are removed. Next genes are grouped into families by reciprocal BLAST hits and single link clustering. The remaining unclassified genes are compared to the NCBI non-redundant database with BLAST to check if they are genes, and if they are then they are considered their own family with one member. The P/A matrix is constructed as follows, for each OTU a binary vector is created, where each entry represents a gene family and a 1 indicates that that OTU contains at least 1 gene in that family. This is repeated for all OTUs, creating a $G \times O$ binary matrix, where G is the total number of gene families and O is the number OTUs.

There are many ways to construct a species tree, but for this project the tree will be constructed with 16S rRNA genes, using Bayesian methods, as implemented in the program MrBayes.

6.4 Makophylo Rate Estimations

Given a species tree and a gene family P/A matrix for the OTUs of the species tree the R package *markophylo* can provide gene insertion and gene deletion rate estimates [37]. The presence or absence of gene families are considered 2 discrete states, for which a (2×2) transition rate matrix (of a Continuous-time Markov chain with finite state space (CTMC-FFS) model) can be estimated using maximum likelihood techniques. The values in this estimated transition matrix are the insertion rate (transition probability of gene absence $\rightarrow$ presence) and deletion rate (transition probability of gene presence $\rightarrow$ absence) [37].

6.5 Network Construction

Quartet decomposition is method by which HGT events can be identified using a set of gene trees and a species tree. Given a tree T a quartet is a subtree contain 4 of the leaf nodes in T, meaning that for a tree with N leaf nodes (or OTUs) there are $\binom{N}{4}$ unique quartets in that tree. A quartet Q is considered consistent with a tree if $Q = T|Le(Q)$ where $T|Le(Q)$ is the tree obtained by suppressing all degree-two nodes in $T[X]$ and $T[X]$ is the minimal subtree of T with all nodes in X, which is a leaf set of T [38]. To calculate the weight of an edge for the network, given a species tree S and a set of gene trees G [38]:

1. Pick a horizontal edge $H = ((u, v), (v, u))$ from S
2. Pick a gene tree G_i in G
3. Decompose G_i into it's set of quartets ϕ_i
4. Remove all quartets consistent with S or previously explained from ϕ_i
5. Set $RS((u, v), \phi_i)$ to be the number of quartets in ϕ_i that support the edge (u, v)
6. Set $NS((u, v), \phi_i)$ to be $RS((u, v), \phi_i)$ divided by λ, which is the total number of quartets in S that are consistent with the edge (u, v).
7. The score for the edge H for tree G_i is $max\{NS((u, v), \phi_i), NS((v, u), \phi_i)\}$
8. The total score for the edge H is the sum of scores for each tree G_i
9. This total score calculation is repeated for each horizontal edge H_i in S, resulting in a list of edges, which is a complete description of the network.

This is further explained in the original work, [38].

6.6 Network Statistics

All networks will be comprised of nodes representing OTUs and weighted edges represent the estimated amount of HGT events between the two incident nodes. As multiple sets of networks can be computed for a single set of genera (using different sets of gene trees), bootstrap support for edges and confidence intervals on edge weights can also be calculated. Given a network, with a set of nodes

$V = \{V_0 \dots V_i\}$ of cardinality N and a set of weighted edges (an unordered 2-tuple and weight) $T = \{((V_1, V_2), W_{1,2}) \dots ((V_i, V_j), W_{i,j})\}$ with cardinality E descriptive statistics can be computed as follows [39]:

- **Average Node Degree**: $\frac{1}{|N_u|} \sum_{uv}^{N_u} w_{uv}$ where N_u is the set of nodes incidenent to u
- **Average Edge Weight**: $\frac{1}{N_c} \sum_i w_i$, The average edge weight for all nodes with CRISPR or without CRISPR
- **Node Clustering Coefficient**:$\frac{1}{k_u(k_u-1)} \sum_{vw}^{T(u)} (\hat{w}_{uw}\hat{w}_{vw}\hat{w}_{uv})^{\frac{1}{3}}$ where $T(u)$ is the set of traingles containing u [40]
- **Node Assortativity**: $A = \frac{Tr(M)-||M^2||}{1-||M^2||}$ Where M is the mixing matrix of a given attribute and $||M||$ is the sum of all elements of M. $A \in [-1, 1]$.[41]
- **Network Modularity**: $Q = \frac{1}{2m} \sum_{uv}^{W} [W_{uv} - \frac{k_u k_v}{2m}] \delta(u, v)$ where m is the total weight of alledges, k_u is the degree of u and $\delta(u, v)$ is 1 if u and v both have or do not have CRISPR systems and 0 otherwise. $Q \in [-1, 1]$ [42]

Each statistic was computed, either separately for the CRISPR and Non-CRISPR nodes or for the entire network for each of the 1000 bootstraps replicates. Each replicate was produced with 50 gene trees sampled randomly from all gene trees produced for that genus.

7 Results

Note: The phrase indel refers to gene insertion/deletion events. It is impossible to tell between two OTUs if a gene was deleted from one or inserted in the other. Thus such discrepancies are referred to as indels, inferred to be the result of HGT between the two OTUs.

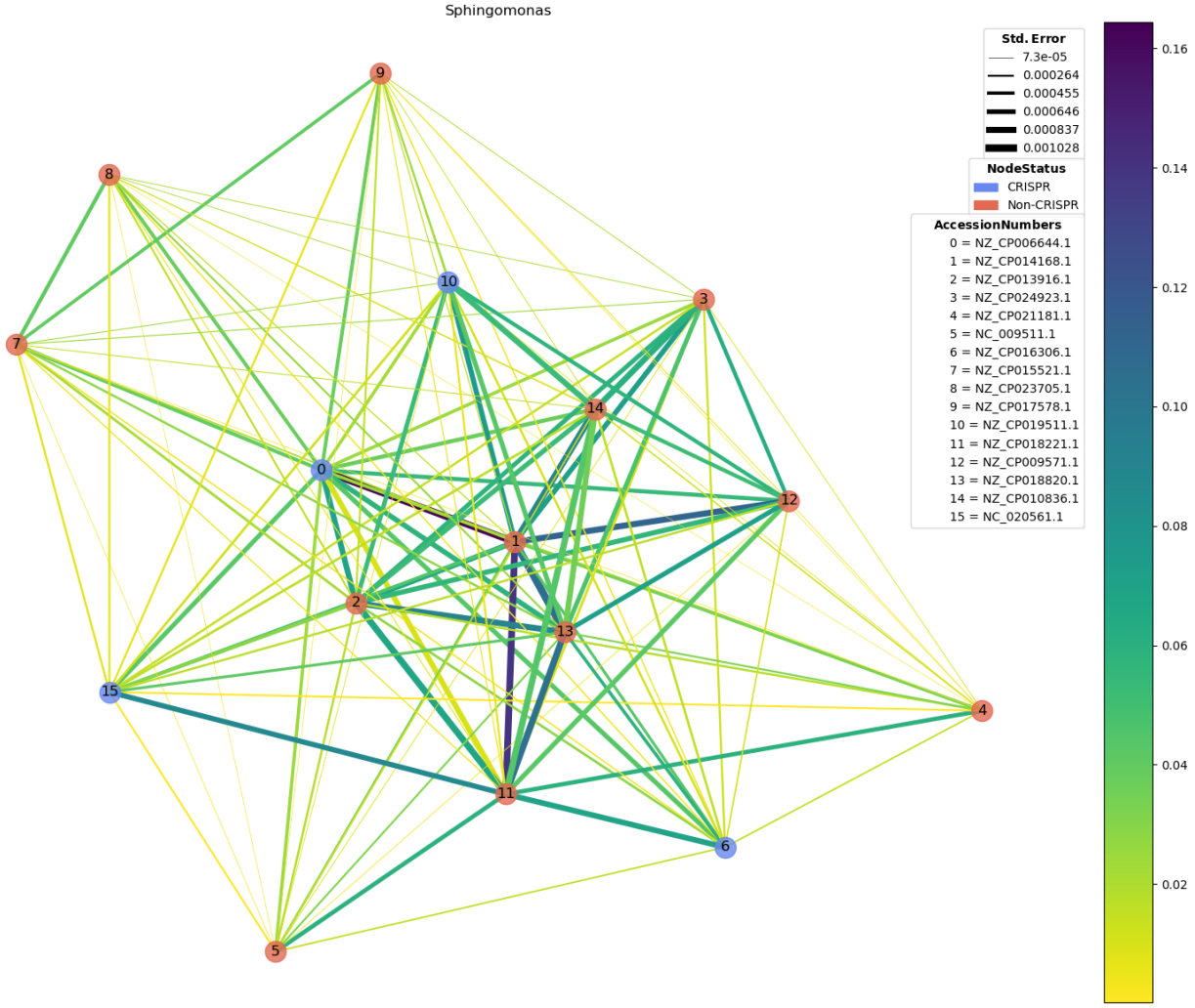


Figure 2: Example of a HGT network produced by HiDe. This network is a "consensus" over 1000 bootstap replicate networks, produced from sampling gene trees. Each bootstrap replicate was produced from 50 randomly sampled gene trees froma a total of 376 individual gene trees. Blue nodes were cassified as having a CRISPR system, red nodes were not. Color represents the fraction of genes examined that were transfered along that edge. Width represents the standard error of the edge value over the 1000 bootstraps. (Note the maximum value for an edge is 1.00, meaning all examined genes were transfered along that edge)

In figure 2 it appears that several nodes have weak connections with most other nodes but strong connections with a few nodes. Further both CRISPR and non-CRISPR nodes both show distributions of strong and weak connections with other CRISPR and non-CRISPR nodes both. Also the standard error of each edge appears proportional to it's weight. This is likely due to the sampling, as if more genes were transferred along an edge, the more likely some of those genes were left out of any individual bootstrap sample, as the size of each bootstrap sample was $\frac{50}{376}$ of the total number of gene trees.

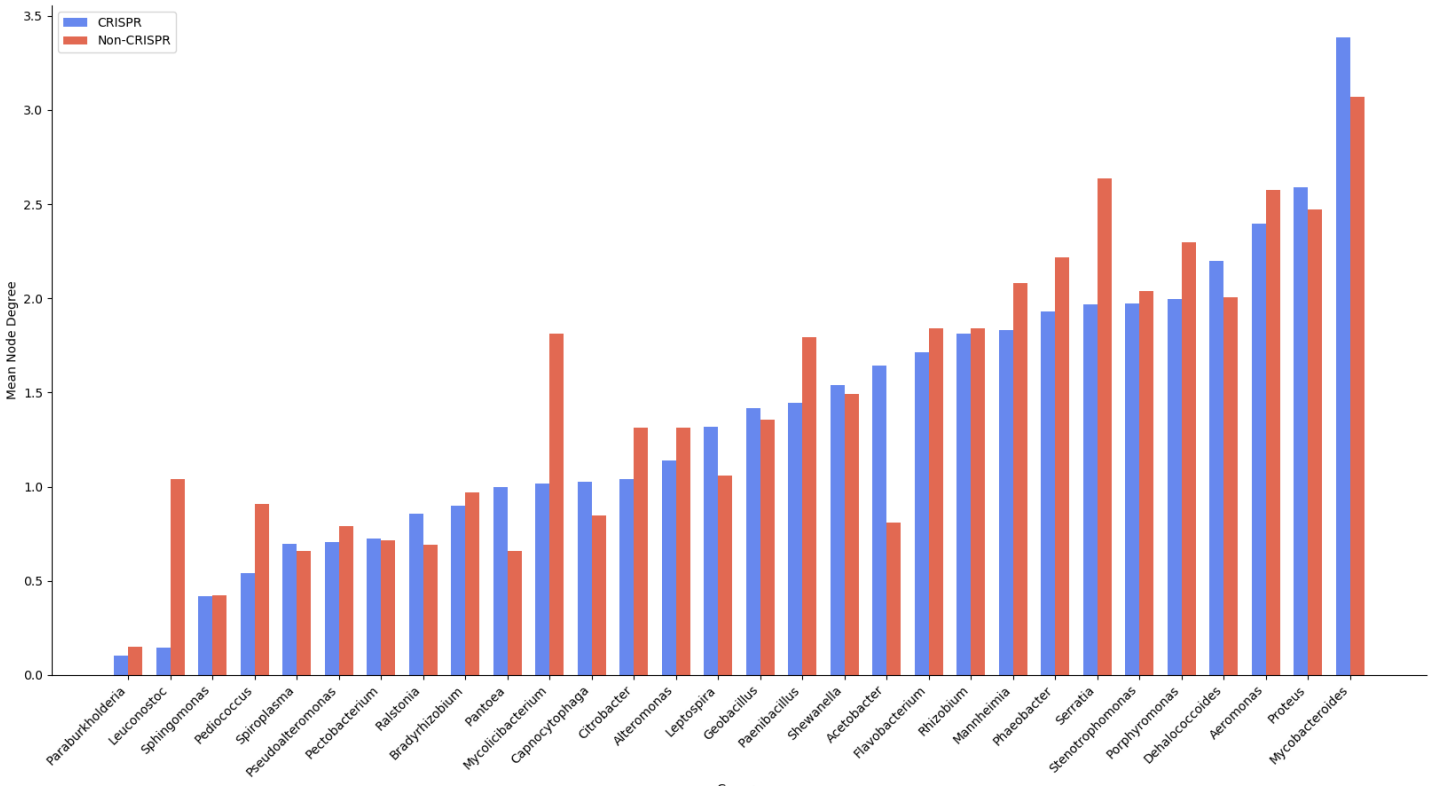


Figure 3: Mean node degree for either all CRISPR or Non-CRISPR nodes across all 1000 bootstrap replicates for each genus. There were 30 genera used.

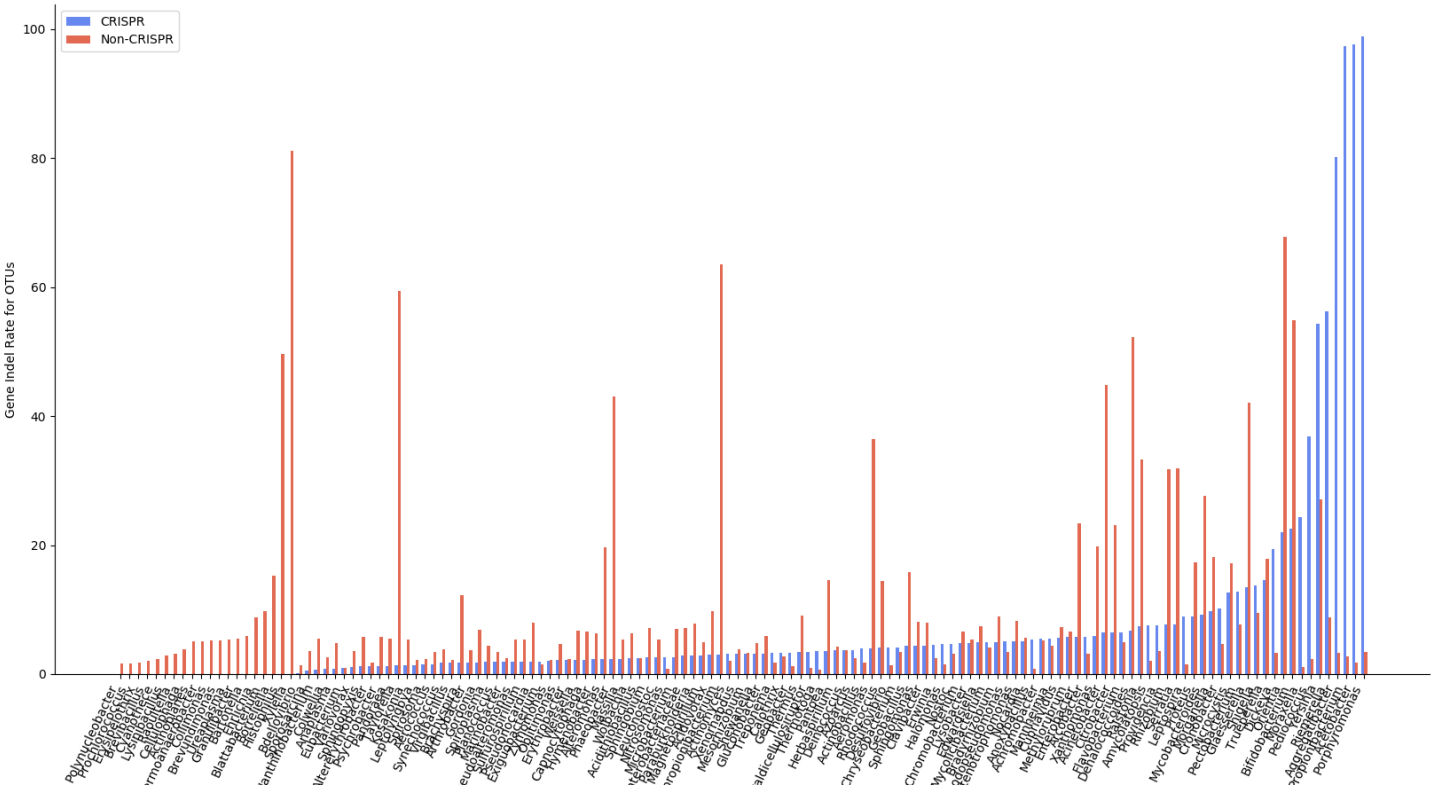


Figure 4: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus. Rate is indel events per base pair substitution. There were 140 genera used.

The mean node degree is much more similar between the CRISPR and non-CRISPR than the indel rate estimates (figures 3,4). Both show significant variability for the non-CRISPR nodes across genera, but the indel rates estimates for the CRISPR nodes are much less variable and generally smaller by comparison. Despite this there are clear exceptions where the indel rate is estimated to be much larger for CRISPR OTUs than non-CRISPR OTUs, specifically of Rhizobium, Acetobacter, Pediococcus and Moraxella.

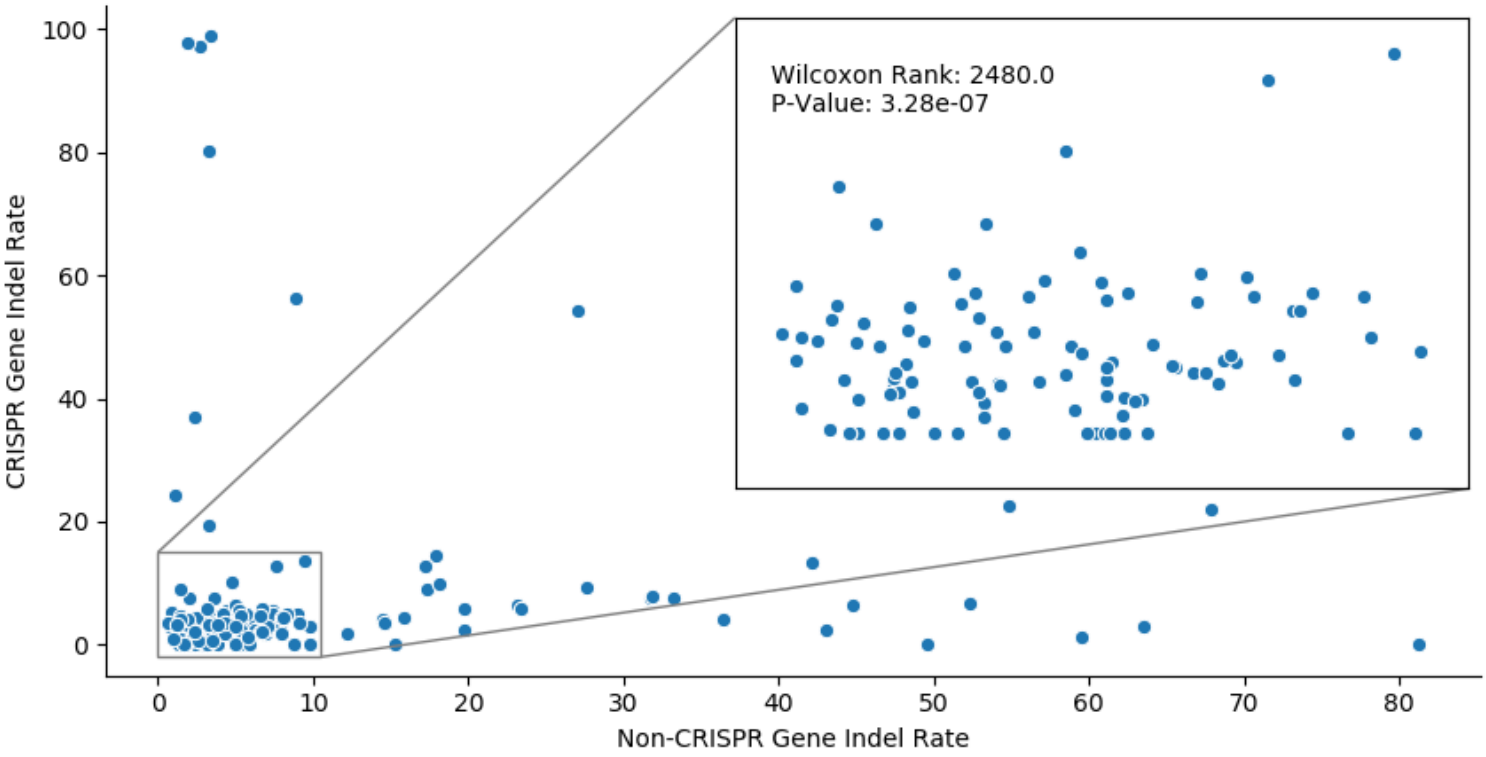


Figure 5: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus. Rate is indel events per base pair substitution. Each point represents a genus. There were 140 genera used.

Figure 5 further demonstrates these points, that CRISPR indel rates are smaller and less varied that the and non-CRISPR. This difference is quantified by the Wilcoxon signed rank test statistic of $191, 0$ and a corresponding p-value 0.02596.

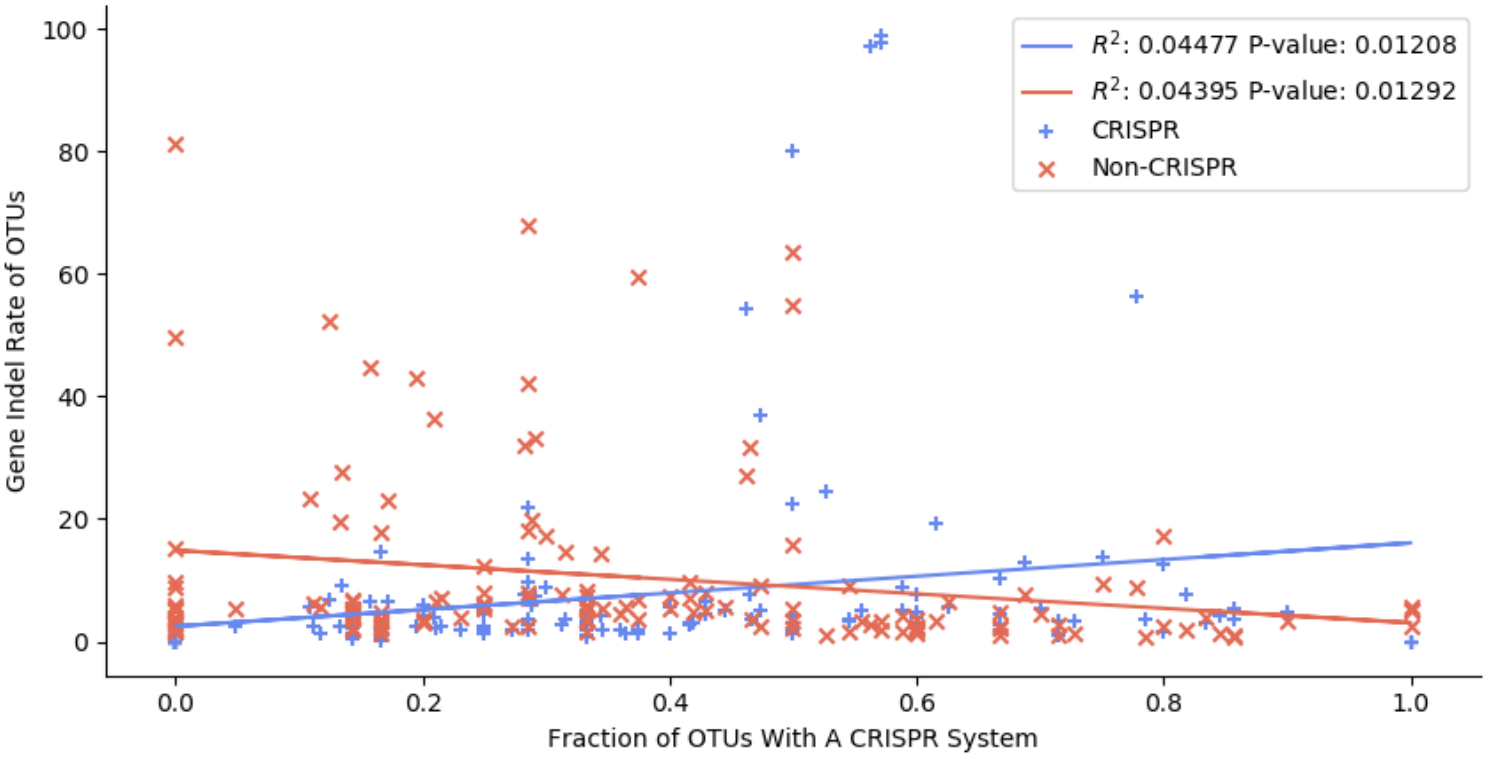


Figure 6: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus against the fraction of all OTUs in that genus that are annotated as having a CRISPR system. R^2 values are for linear regression lines fit to the CRISPR and non-CRISPR estimates. Rate is indel events per base pair substitution. Each point represents a genus. There were 140 genera used.

Figure 6 show that as the fraction of all OTUs in a genus with a CRISPR system increases, the gene indel rates appear to decrease for non-CRISPR OTUs remains mostly stagnant for CRISPR OTUs. The R^2 values in figure 6 are fairly small, implying a poor fit of a linear relationship to the data, but the p-values are significant, implying that some relationship does exist.

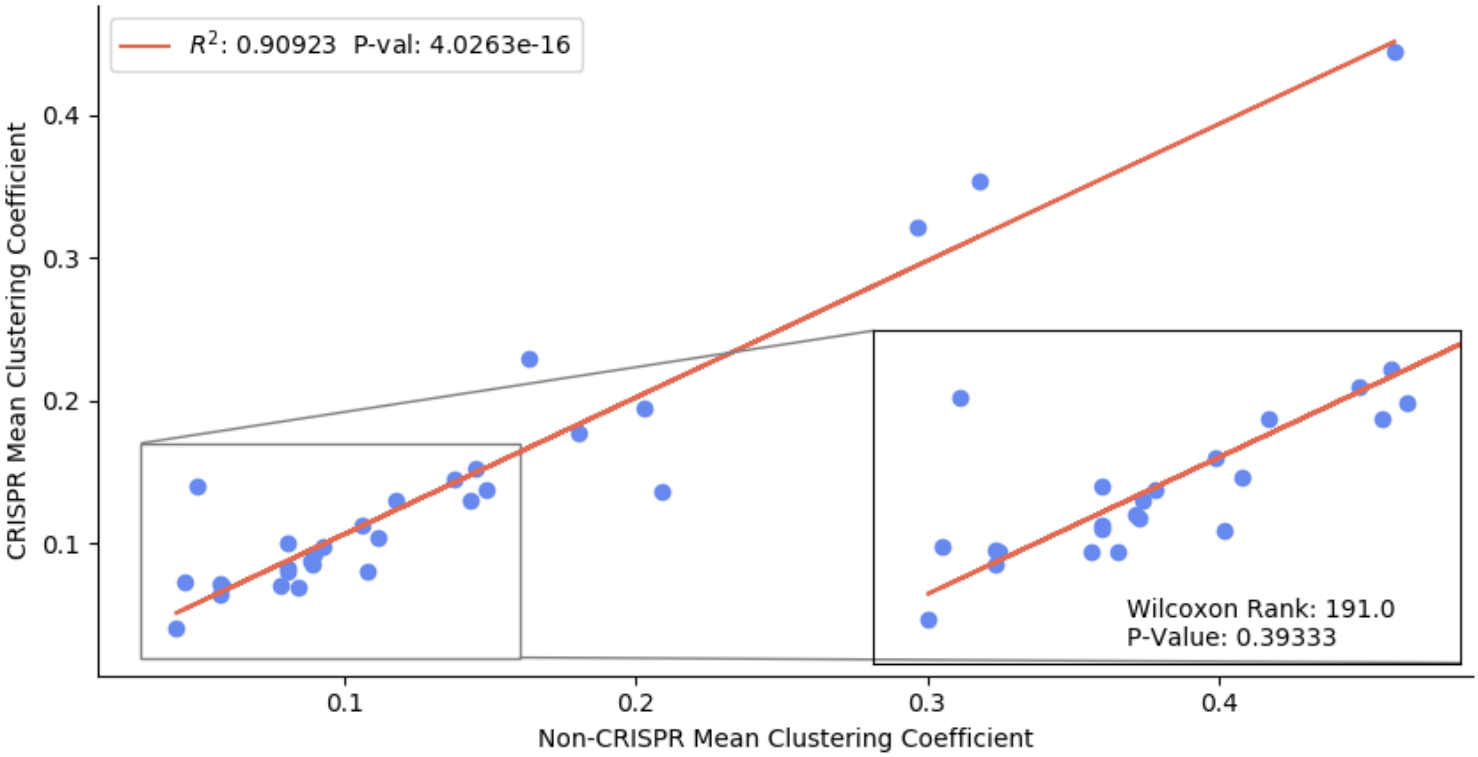


Figure 7: Mean over 1000 bootstraps of the clutering coefficients of the CRISPR OTUs for each genus against the non-CRISPR means over the 1000 bootstraps. R^2 value is for the linear regression fit to the CRISPR and non-CRISPR estimates. There were 140 genera used.

Figure 7 show the mean clustering coefficient over 1000 bootstraps for each genus for all CRISPR and non-CRISPR OTUs. There appears to be a clear linear relationship between the mean clustering coefficients of CRISPR and non-CRISPR OTUs. Clustering is also generally small in magnitude, with most of the data in the range of 0.0 to 0.2 (the maximum value is 1.0).

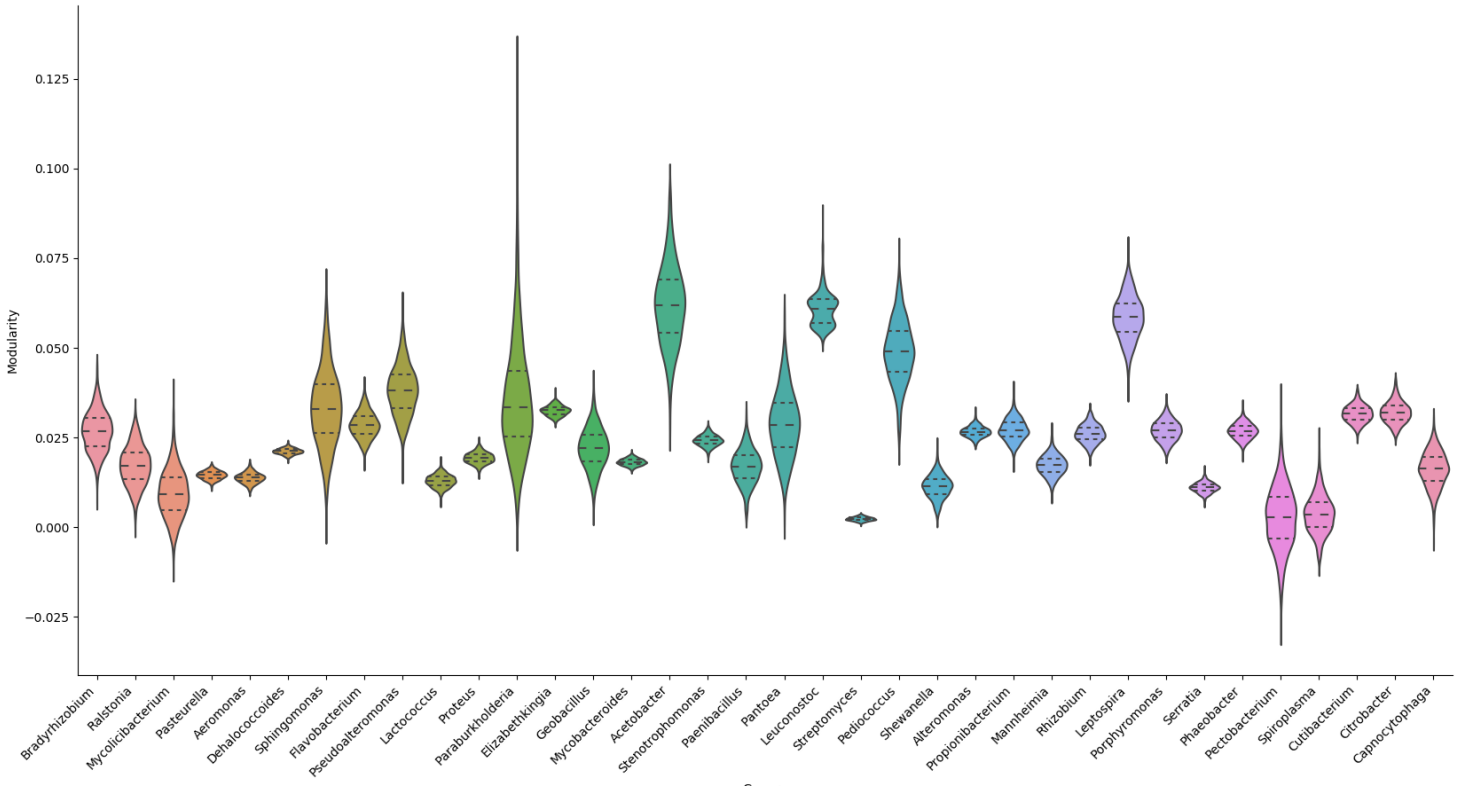


Figure 8: Distribution of network modularity over 1000 bootstrap repliactes for each genus. Plot is of the kernel density estimated from the observed values, with witdh proportional to the number of data points. Lines inside each distribution represent quartiles.

Figure 8 shows that the distribution of network modularity is centered near 0 for most networks, implying a lack of modularity between CRISPR and non-CRISPR OTUs. However the variability in the shape of each distribution, ranging from very narrow to very wide, with some being bimodal, should be noted.

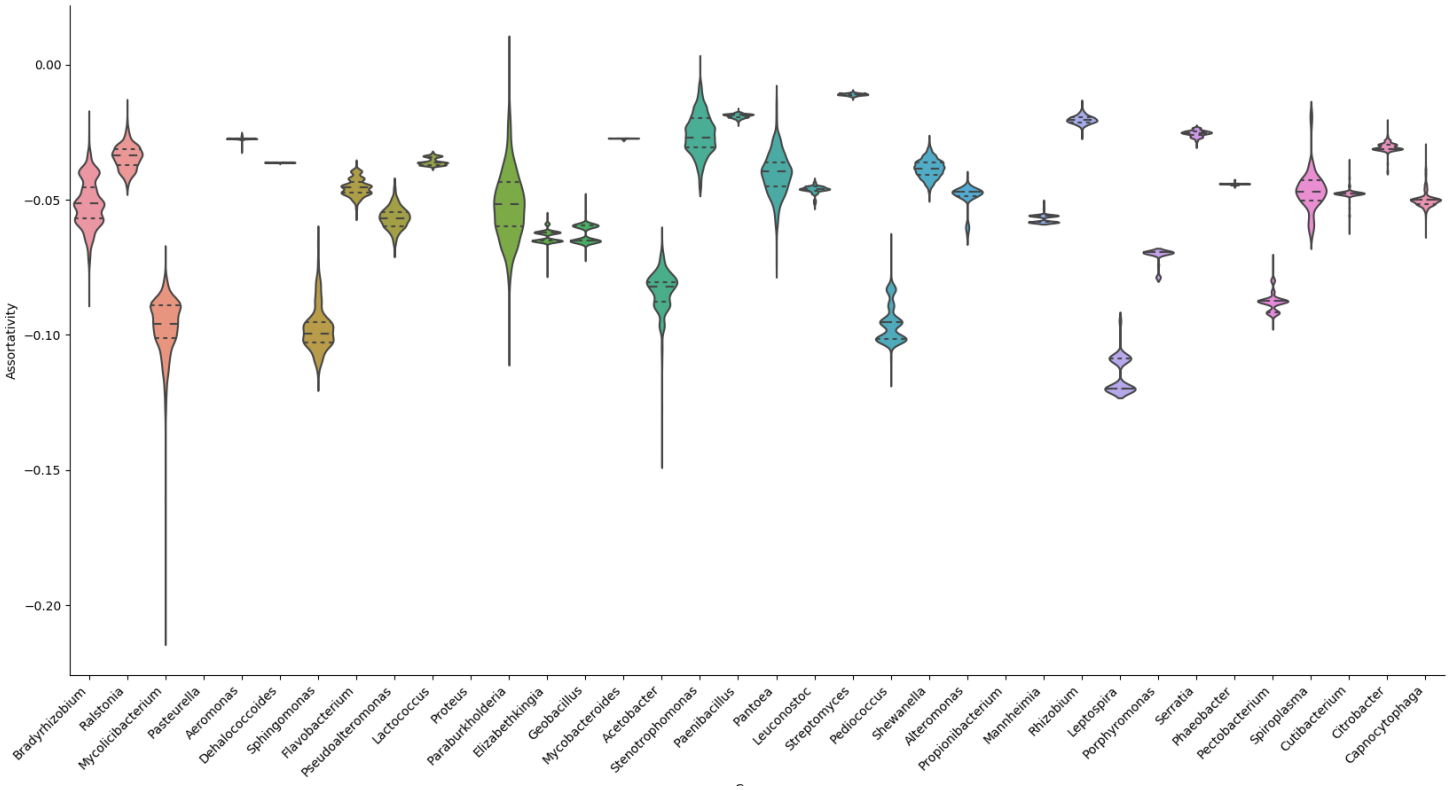


Figure 9: Distribution of network assortativity by CRISPR status (either CRISPR or non-CRISPR) over 1000 bootstrap repliactes for each genus. Plot is of the kernel density estimated from the observed values, with witdh proportional to the number of data points. Lines inside each distribution represent quartiles.

Figure 9 shows that the distribution of network assortativity is centered near -0.05 for most networks, implying a lack of assortativity between CRISPR and non-CRISPR OTUs. The variability in the shape of each distribution is much more pronounced than with modularity, with many distributions having several undulations or very sharp contrasts between different peaks or completely smooth and centered around the mean. This variability may be due to the variation in the fraction of OTUs with a CRISPR system. Some genera may only have one or two OTUs with a CRISPR system, thus limiting the range of values that assortativity can take on, due to how it is defined.

8 Discussion

8.1 Gene Indel Rates are Different for CRISPR and Non-CRISPR OTUs

For most genera, the gene indel rate for non-CRISPR genera is larger than for CRISPR genera. This is in-line with literature surrounding the mechanisms of HGT as CRISPR-Cas systems are meant to

stop the integration of foreign DNA into the bacterial genome. Despite this, the mean node degree of CRISPR and non-CRISPR OTUs is relatively similar across genera. One reason for this discrepancy may be that there are often more non-CRISPR OTUs than CRISPR OTUs, thus more genes are exchanged between all non-CRISPR OTUs, but each non-CRISPR OTU transfers genes at a similar rate to CRISPR OTUs. One possible explanation for certain genera having very high gene indel rates for CRISPR OTUs may be that it is an efficient way to acquire new spacers. CRISPR-Cas systems may enhance HGT to preemptively acquire new spacers from the environment in response to environmental phage density.

8.2 Phylogenomic Networks Have Low Assortativity

There seems to be significant HGT between CRISPR and non-CRISPR OTUs, with no clear clustering, assortativity or modularity among most of the networks examined. CRISPR-Cas systems do not appear to have a segregating effect on the network, but do appear to have a population level effect of decreasing the rate of HGT. As suggested by [33] CRISPR-Cas systems may have a population level effect on HGT rate, but it appears to be suppressive in this case, as opposed to theirs.

8.3 HGT Dynamics Vary Across Bacterial Genera

Despite some trends being observable, the one constant is that there is significant variability between genera with regards to HGT. HGT rates can be similar or significantly different between CRISPR and non-CRISPR OTUs, either can be larger, both can be similar and either large or small. While the means are similar, the shapes of the distributions of network assortativity and modularity are not homogeneous.

8.4 Limitations

Some factors that may have introduced bias or error into this analysis:

- **Ignored Singletons**: Genes that did not cluster into any families were ignored from future steps, but may have still represented horizontally transferred genes
- **Ignored Some Gene Families**: For time considerations, only 1500 gene trees were generated for each genus
- **Taxonomic Mistakes**: Inconsistencies in taxonomic labelling can result in ignored or misplaced OTUs.
- **Multifurcation Error**: Some species trees contained multifurcations, which were resolved randomly to generate a bifurcating tree. Estimating this error by examining variance over different resolutions is possible, but was not done here.

More insight may be gained by examining specific genera more in-depth, or considering new network based metric for understanding the dynamics of HGT.

9 Significance & Future Work

9.1 Significance

This work highlights the large degree of variability in HGT rate between bacterial genera. While there seems to be an association of CRISPR-Cas systems with decreased rates of HGT compared to OTUs without such systems, prominent exceptions exist. Further CRISPR or non-CRISPR OTUs do not appear to transfer genes preferentially to either CRISPR or non-CRISPR OTUs. Ultimately, this pipeline provides a fairly straightforward way to study trends in HGT for a set of bacterial OTUs. Clearly the dynamics of CRISPR and HGT warrent further investigation and should be studied within individual genera such as Streptomyces, which have been model systems for studying CRISPR-Cas previously.

9.2 Future Work

There are multiple ways to expand this analysis to answer other questions related to the transfer of genes. As HGT inference methods improve and it becomes possible to discern the direction of transfer with confidence, a whole new set of techniques become available for study. Some possible ways to extend this analysis are:

- **Inferring direction**: Directed networks have a host of available analytic tools undirected networks do not
- **Gene function analysis**: Considering the transfer dynamics of different functional classes of genes
- **Studying movement of CRISPR systems**: Studying how frequently CRISPR systems themselves are transferred from arrays, *Cas* genes
- **Intergenic comparisons**: Combine any set of fasta files from OTUs for analyzing transfer dynamics
- **Continuous CRISPR activity**: Labelling nodes by estimated CRISPR activity (array length, transciptomic data, etc.)
- **Considering bacterial ecology and environments**: Consider geographically close OTUs or differences between networks due to environmental factors

References

1. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117.** Special Issue: Regulatory RNAs, 119–128. ISSN: 0300-9084 (2015).

2. Shmakov, S. A. *et al.* The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* **8** (eds Gilmore, M. S., Sorek, R. & Barrangou, R.) (2017).

3. Grissa, I. and Drevet, C. and Couvin, D. *CRISPRdb* http://crispr.i2bc.paris-saclay.fr/. Online; accessed 22 October 2018. 2017.

4. Makarova, K. S. *et al.* Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.* **9,** 467–477 (2011).

5. Zhang, Q. & Ye, Y. Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics* **18,** 92. ISSN: 1471-2105 (Feb. 2017).

6. Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.* **1,** e60 (2005).

7. Zhaxybayeva, O. & Doolittle, W. F. Lateral gene transfer. *Current Biology* **21,** R242–R246. ISSN: 0960-9822 (2011).

8. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring Horizontal Gene Transfer. *PLoS Computational Biology* **11,** 1–16 (May 2015).

9. Marri, P. R., Hao, W. & Golding, G. B. The role of laterally transferred genes in adaptive evolution. *BMC Evol. Biol.* **7 Suppl 1,** S8 (2007).

10. Blokesch, M. Natural competence for transformation. *Current Biology* **26,** R1126–R1130. ISSN: 0960-9822 (2016).

11. Davison, J. Genetic exchange between bacteria in the environment. *Plasmid* **42,** 73–91 (1999).

12. Griffiths, A. J. F. *et al. An Introduction to Genetic Analysis* 7^{th} *Edition* (W.H. Freeman, 2000).

13. Hao, W. & Golding, G. B. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* **16,** 636–643 (2006).

14. Popa, O. & Dagan, T. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology* **14.** Antimicrobials/Genomics, 615–623. ISSN: 1369-5274 (2011).

15. Wielgoss, S. *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences* **110,** 222–227. ISSN: 0027-8424 (2013).

16. Mozhayskiy, V. & Tagkopoulos, I. Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution. *BMC Bioinformatics* **13,** S13. ISSN: 1471-2105 (June 2012).

17. Dzidic, S. & Bedeković, V. Horizontal gene transfer-emerging multidrug resistance in hospital bacteria. *Acta pharmacologica Sinica* **24,** 519–526 (2003).

18. Baltrus, D. A. Exploring the costs of horizontal gene transfer. *Trends in Ecology and Evolution* **28,** 489–495. ISSN: 0169-5347 (2013).

19. Novick, R. Plasmid Incompatibility. *Microbiol Rev* **51,** 381–95 (1987).

20. Rouli, L., Merhej, V., Fournier, P.-E. & Raoult, D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections* **7,** 72–85. ISSN: 2052-2975 (2015).

21. Guimaraes, L. C. *et al.* Inside the Pan-genome - Methods and Software Overview. *Curr. Genomics* **16,** 245–252 (2015).

22. Rasko, D. A. *et al.* The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. *Journal of Bacteriology* **190,** 6881–6893. ISSN: 0021-9193 (2008).

23. Acuña, R. *et al.* Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences.* ISSN: 0027-8424 (2012).

24. Von Wintersdorff, C. J. *et al.* Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. *Front Microbiol* **7,** 173 (2016).

25. Newman, M. E. J. Assortative Mixing in Networks. *Phys. Rev. Lett.* **89,** 208701 (20 Oct. 2002).

26. Kunin, V., Goldovsky, L., Darzentas, N. & Ouzounis, C. A. The net of life: reconstructing the microbial phylogenetic network. *Genome Res.* **15,** 954–959 (2005).

27. Kurland, C. Codon bias and gene expression. *FEBS Letters* **285,** 165–169.

28. Than, C., Ruths, D., Innan, H. & Nakhleh, L. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J. Comput. Biol.* **14,** 517–535 (2007).

29. Hickey, G., Dehne, F., Rau-Chaplin, A. & Blouin, C. SPR distance computation for unrooted trees. *Evol. Bioinform. Online* **4,** 17–27 (2008).

30. Marraffini, L. A. & Sontheimer, E. J. CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. *Science* **322,** 1843–1845. ISSN: 0036-8075 (2008).

31. Zhang, Y. *et al.* Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. *Molecular Cell* **50,** 488–503. ISSN: 1097-2765 (2013).

32. Bondy-Denomy, J. & Davidson, A. R. To Acquire Or Resist:The Complex Biological Effects Of CRISPR-Cas systems. *Trends Microbio.* **22,** 218–25 (2014).

33. Watson, B. N. J., Staals, R. H. J. & Fineran, P. C. CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *mBio* **9** (eds Bondy-Denomy, J. & Gilmore, M. S.) (2018).

34. Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genetics* **26,** 335–340. ISSN: 0168-9525 (2010).

35. Godde, J. S. & Bickerton, A. The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes. *Journal of Molecular Evolution* **62,** 718–729. ISSN: 1432-1432 (June 2006).

36. Zambelis, A., Dang, U. J. & Golding, G. B. *Effects of CRISPR-Cas System Presence On Lateral Gene Transfer Rates In Bacteria* (2015).

37. Dang, U. J. & Golding, G. B. markophylo: Markov chain analysis on phylogenetic trees. *Bioinformatics* **32,** 130–132 (2016).

38. Bansal, M. S., Banay, G., Harlow, T. J., Gogarten, J. P. & Shamir, R. Systematic inference of highways of horizontal gene transfer in prokaryotes. *Bioinformatics* **29,** 571–579 (2013).

39. Newman, M. The Structure and Function of Complex Networks. *SIAM Review* **45,** 167–256 (2003).

40. Onnela, J. P., Saramaki, J., Kertesz, J. & Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **71,** 065103 (2005).

41. Newman, M. E. Assortative mixing in networks. *Phys. Rev. Lett.* **89,** 208701 (2002).

42. Newman, M. E. Analysis of weighted networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **70,** 056131 (2004).