# Is Sharing Caring?

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

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

February 21, 2020

**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 advanced niche adaptation. CRISPR associated (CRISPR-Cas) is an adaptive immune system in prokaryotes that has garnered a lot of research attention recently, largely due to it's applications in gene editing. Due to the nature of how it works, using guide RNA to cut DNA, CRISPR-Cas systems have been thought to affect rates of HGT. Effort has mostly been focused on how CRISPR-Cas systems affect the mechanisms of HGT and thus little is known about its effects on HGT rates. This work uses is a network-theoretic approach to better characterize the effects of the presence of CRISPR-Cas systems on HGT rates within bacterial populations. This approaches makes use of phylogenetic methods for estimating HGT rates, improving on estimates using methods based on genome composition used previously. Understanding the effects of CRISPR-Cas on HGT may help uncover potential targets for curbing the spread antibiotic resistance genes.

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

# Contents

# 1 Background

## 1.1 What is CRISPR-Cas?

CRISPR-Cas systems are sets of nucleotide motifs (spacers) interspaced with nucleotide repeats (CRISPRs) and CRISPR-associated (Cas) proteins (with sequences usually adjacent to the CRISPR motifs) that have an adaptive immune function in many bacteria and archaea [1]. Each nucleotide motif is the result of a DNA molecule that was previously taken up by the host, serving as a marker for the Cas proteins to degrade any DNA matching the motif [1]. If a bacterium possessing a CRISPR-Cas system is infected with a phage and survives, a motif representative of that phage can be integrated as a spacer. If the bacterium is reinfected with the same phage strain the spacers will guide Cas proteins to the invading phage DNA and degrade it, hence adaptive immunity. Although CRISPR-Cas appears to have evolved to degrade viral DNA system, the majority of CRISPR spacers have been found to match bacterial Mobile Genetic Element (MGE)s with no known viral match [2].

As of 2017, over 45% of bacterial genomes analyzed ($n = 6782$) appear to contain CRISPR motifs [3]. Moreover, CRISPR motifs show significant diversity between individual bacteriums as they represent a chronological history of spacer acquisition (usually via viral infection or MGE “infection”) for that specific bacterium[1]. 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 [4, 5]. 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].

## 1.2 Horizontal Gene Transfer

### 1.2.1 Mechanisms

HGT can be defined as the exchange of genetic information across lineages [6], as opposed to vertical gene transfer between parents and offspring [7]. It is a source of genetic variation, allowing organisms to adapt quickly by copying a gene with a specific function, rather than evolving it themselves [7, 8]. There are 3 main mechanisms of HGT

**Transformation** Free floating DNA is taken up by a bacterium and incorporated into the genome [6]. DNA is not always incorporated successfully even if it taken up.

**Conjugation** The sharing of genetic material through cell-to-cell bridges, the genes for which are usually carried on a plasmid [9].

**Transduction** Genes or DNA fragments can be transferred through either lytic or lysogenic bacteriophages [10]. Bacterial DNA can be accidentally packaged into the lysogenic phage head during cell lysis and integrate into the next infected host. [10]. Lysogenic phages can take up bacterial DNA flanking the viral sequence and bring it with them to the next host[10].

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 or diverge quickly due to minimal selective pressure maintaining them [11].

### 1.2.2 Rate Influencing Factors

The rate of HGT in bacteria is constantly in flux[12]. The more exogenous DNA, higher population density or higher phage density means more DNA is available for transfer [6]. Just like mutation rates, HGT rates are also thought to evolve in response to environmental factors or selective pressure [13, 14]. The metaboliccost or the possibility of receiving toxic or incompatible genes can dis incentivize a cell to produce the machinery required for HGT [15]. But for bacteria in hospitals, the potential benefit of receiving antibiotic resistance genes can outweigh any potential danger or metabolic cost, inducing increased bacterial competence [16] It has been suggested that genes acquired via HGT are often quickly lost since they often confer no advantage to a cell's current selective pressures [11]. Ultimately HGT rates are influenced by a variety of factors related balancing the potential fitness costs and benefits.

## 1.3 Phylogenomic Networks

HGT is an important factor in understanding evolution in prokaryotes. 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. The existence of HGT demonstrates that the tree model is clearly an incomplete representation of genetic relationships between bacterial OTUs. Genetic material can be transferred outside of reproduction, creating multiple paths through which a gene can exist in two different OTUs [6]. The frequency of HGT among prokaryotes has lead many to re-evaluate the concept of a "prokaryotic tree of life", which ignores these horizontal interactions [17]. 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 [17]. Edges can now connect closely or distantly related OTUs if HGT has occurred between them.

## 1.4 Detection

While understanding that HGT is important to bacterial evolution and networks provide a useful theoretic framework to study it, constructing such networks is not trivial. Phylogenetic methods are often best at inferring HGT events. They 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 [18]. Generally phylogenetic methods are preferred for multiple reasons:

- Can make use of multiple genomes at once [7]
- Require explicit evolutionary models, which come with their own framework for hypothesis testing and model selection [7].
- HGT events identified by parametric methods are often found by phylogenetic methods as well [7].
- In recent years, the requirements of computing power and multiple well sequenced genomes for phylogenetic methods have become easier and easier to meet [7].

While detecting HGT events with high degrees of certainty is still difficult, much progress has been made in recent years,especially using phylogenetic methods [7]. 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 [7, 18].

## 1.5 How Does CRISPR Affect HGT?

### 1.5.1 Interference Mechanisms

CRISPR-Cas systems have also been found to interfere with transformation-mediated HGT, by degrading foreign DNA taken up by a cell [19]. They have been shown to interfere with conjugation by targeting genes on conjugative plasmid [20]. They have also been shown to interfere with transduction by creating immunity to phage infection [20]. Thus it been hypothesized that lower rates of HGT will be observed in bacterial strains with CRISPR-Cas systems than without [20].

### 1.5.2 Complexities And Costs Of CRISPR-Cas Systems

Since antibiotic resistance genes are often transferred on plasmids maintaining a CRISPR-Cas systems can present a large opportunity cost, especially in environments like hospitals or trees [16]. CRISPR-Cas systems incur a metabolic cost, as Cas proteins, guide RNAs and spacer acquisition proteins must all be expressed consistently[1]. Much like the human immune system, CRISPR-Cas systems can have off-target effects, sometimes affecting HGT [21]. While resisting lytic phage infection clearly provides some fitness benefit, CRISPR-Cas has also been shown to help resist prophages which can provide super-infection immunity or reduce competitor populations [21, 22]. It has also been shown that spacers targeting a bacterium's own DNA can be acquired, leading to an auto-immune response [23]. As CRISPR-Cas systems prevent phage infection, mechanisms, denoted as anti-CRISPRs, have evolved in certain phages making them immune to CRISPR [21]. This has a two-fold effect, as it can increase the susceptibility of the host to infection reducing the fitness benefit of CRISPR, but it can also increase transduction-mediated HGT [21].

### 1.5.3 Balancing the Cost of CRISPR and HGT

Due to the myriad of fitness costs associated with consistently expressing CRISPR-Cas systems, bacteria have appeared to develop strategies to mitigate these costs. It has been posited that CRISPR-Cas systems need only be present in some proportion of a population as they can be horizontally transferred themselves between members[24]. This allows populations to maintain phage immunity while isolating the metabolic cost to only a few organisms [21]. It should also be noted that CRISPR-Cas genes are not necessarily constitutively transcribed, potentially allowing bacteria to tune their CRISPR-Cas to suit their selective pressures [21]. The presence of CRISPR-Cas systems have also been shown to actually enhance HGT at a population level via transduction by reducing total phage abundance [22]. A bioinformatic analyssis has shown increased levels of gene insertion and deletion as in Firmicutes with CRISPR-Cas systems compared to closely related outgroups without [25]. The effects of CRISPR-Cas systems on HGT rates are highly complex, owning in no small part to the broad range of these effects, how CRISPR activity can be modulated and the transfer of CRISPR systems within a population [21]. Taking a systematic approach may help elucidate the dynamics between CRISPR system presence and HGT rate.

# 2 Objectives

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. 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.

**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.

# 3 Methods

## 3.1 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 ($\leq 1500$)
   (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.

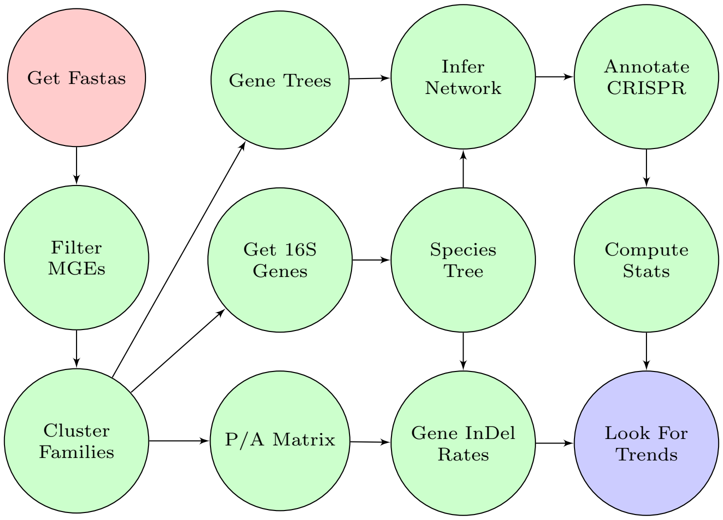

Figure 1: Diagram of the workfklow for a single genus

## 3.2 Data Collection

Protein and nucleotide fastas of all CDS sequences and whole genome fastas for all bacterial strains included 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 [4].

## 3.3 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.

## 3.4 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 [26]. The presence or absence of gene families are considered 2 discrete states, for which a $(2 \times 2)$ 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 → presence) and deletion rate (transition probability of gene presence → absence) [26].

## 3.5 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 $T$ [27]. To calculate the weight of an edge for the network, given a species tree $S$ and a set of gene trees $G$ [27]:

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 $\phi_i$
4. Remove all quartets from $\phi_i$ either consistent with $S$ or discordant with $S$ but accounted for previously by a quartet set from another tree $G_j \in G$
5. Set $RS((u, v), \phi_i)$ to be the number of quartets in $\phi_i$ that support the existenc of edge $(u, v)$
6. Set $NS((u, v), \phi_i) = \frac{RS((u,v),\phi_i)}{\lambda}$, where $\lambda$ is the total number of quartets in $S$ that are consistent with the existence of edge $(u, v)$.
7. The score for the edge $H$ for tree $G_i$ is $max\{NS((u, v), \phi_i), NS((v, u), \phi_i)\} \in [0, 1]$
8. The total score for the edge $H$ is the sum of scores for each tree $G_i \in G$

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, [27].

### 3.6 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 [28]:

- **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 $u$ [29]
- **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 $M$. $A \in [-1, 1]$.[30]
- **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. $Q \in [-1, 1]$ [31]

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.

# 4 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.

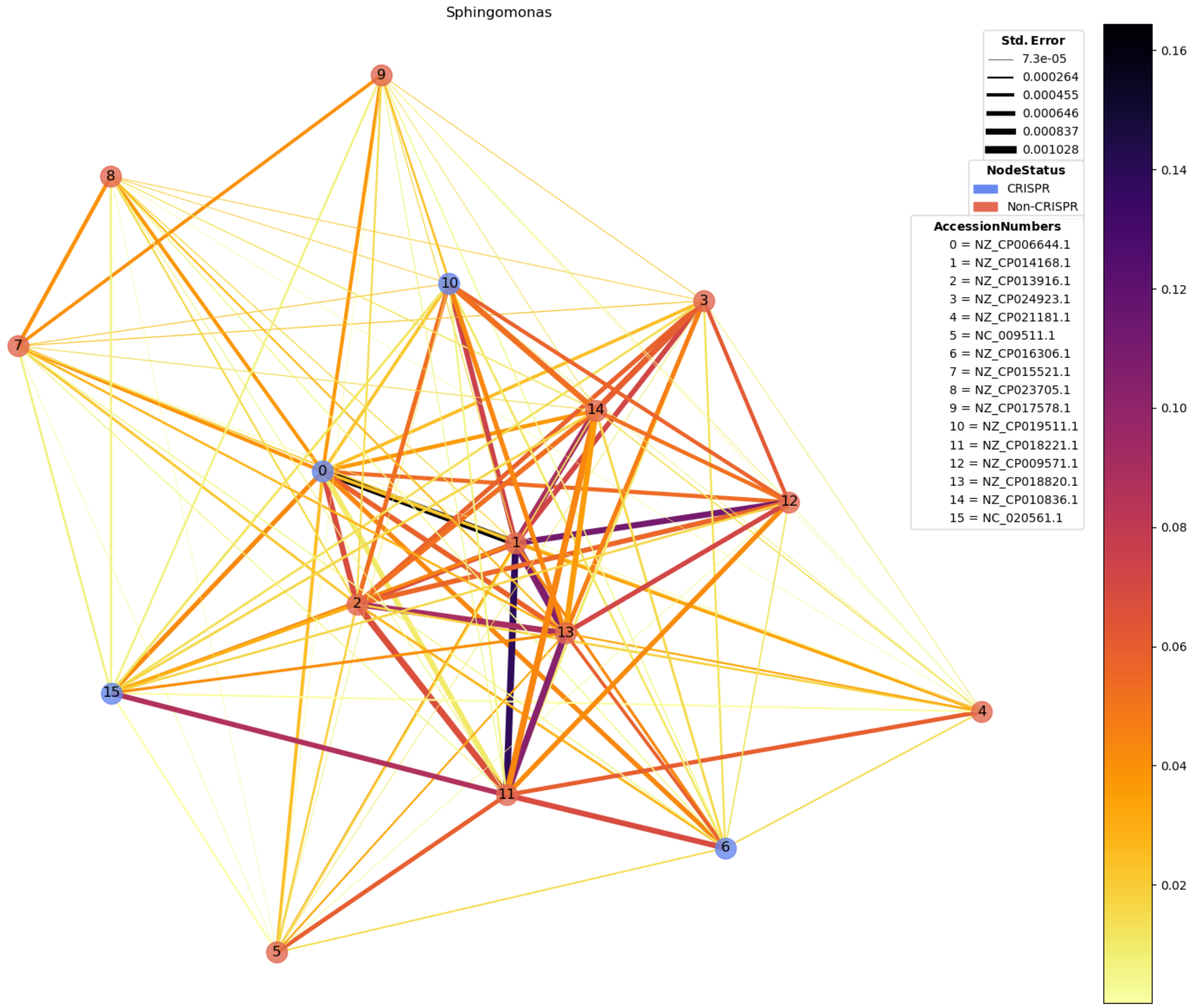

Figure 2: Example of a HGT network produced by HiDe. This network is a “consensus” over 1000 bootstap replicate networks, produced from sampling gene trees. Each bootstrap replicate was produced from 50 randomly sampled gene trees froma a total of 376 individual gene trees. Blue nodes were cassified 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.

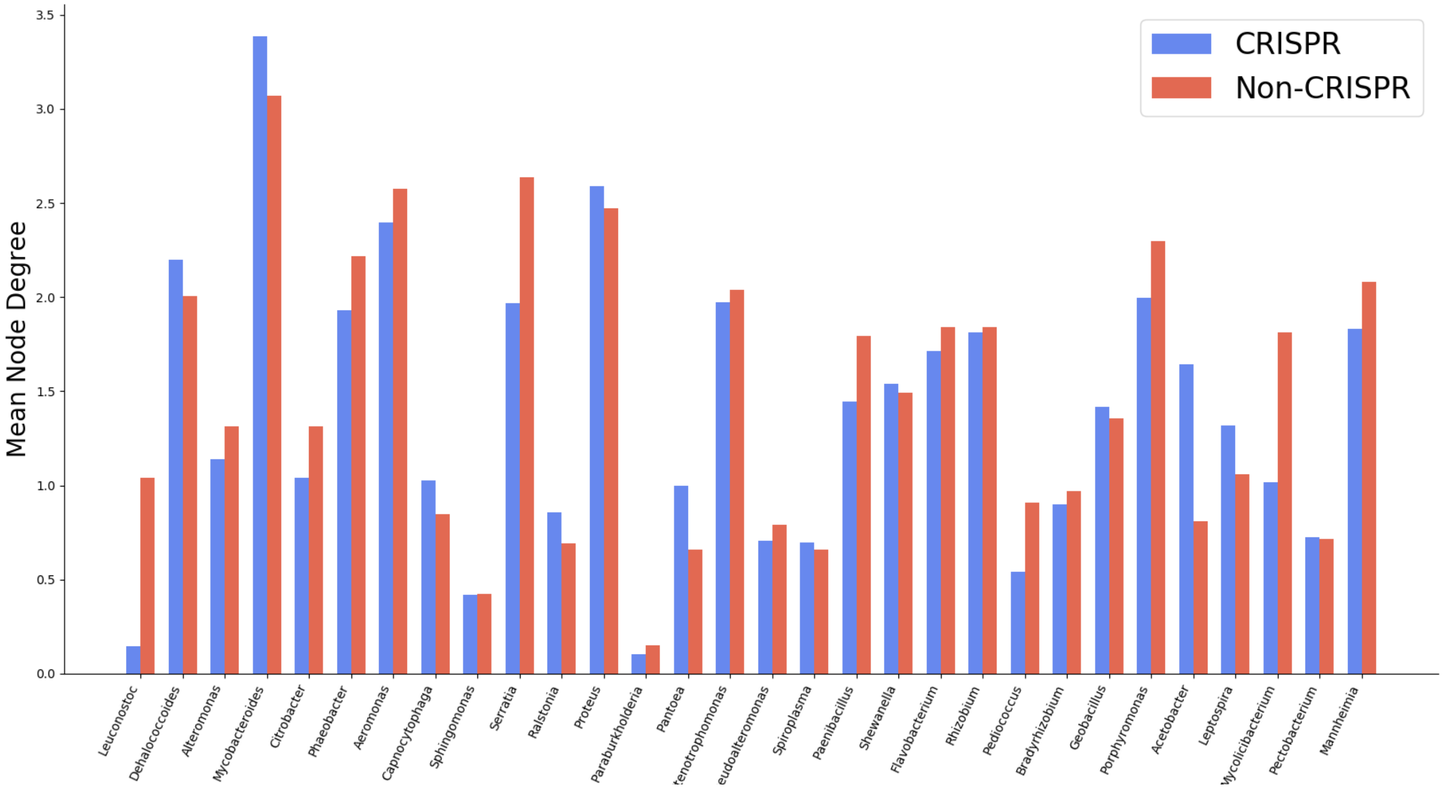

Figure 3: Mean node degree for either all CRISPR or Non-CRISPR nodes across all 1000 bootstrap replicates for each genus. There were 30 genera used.

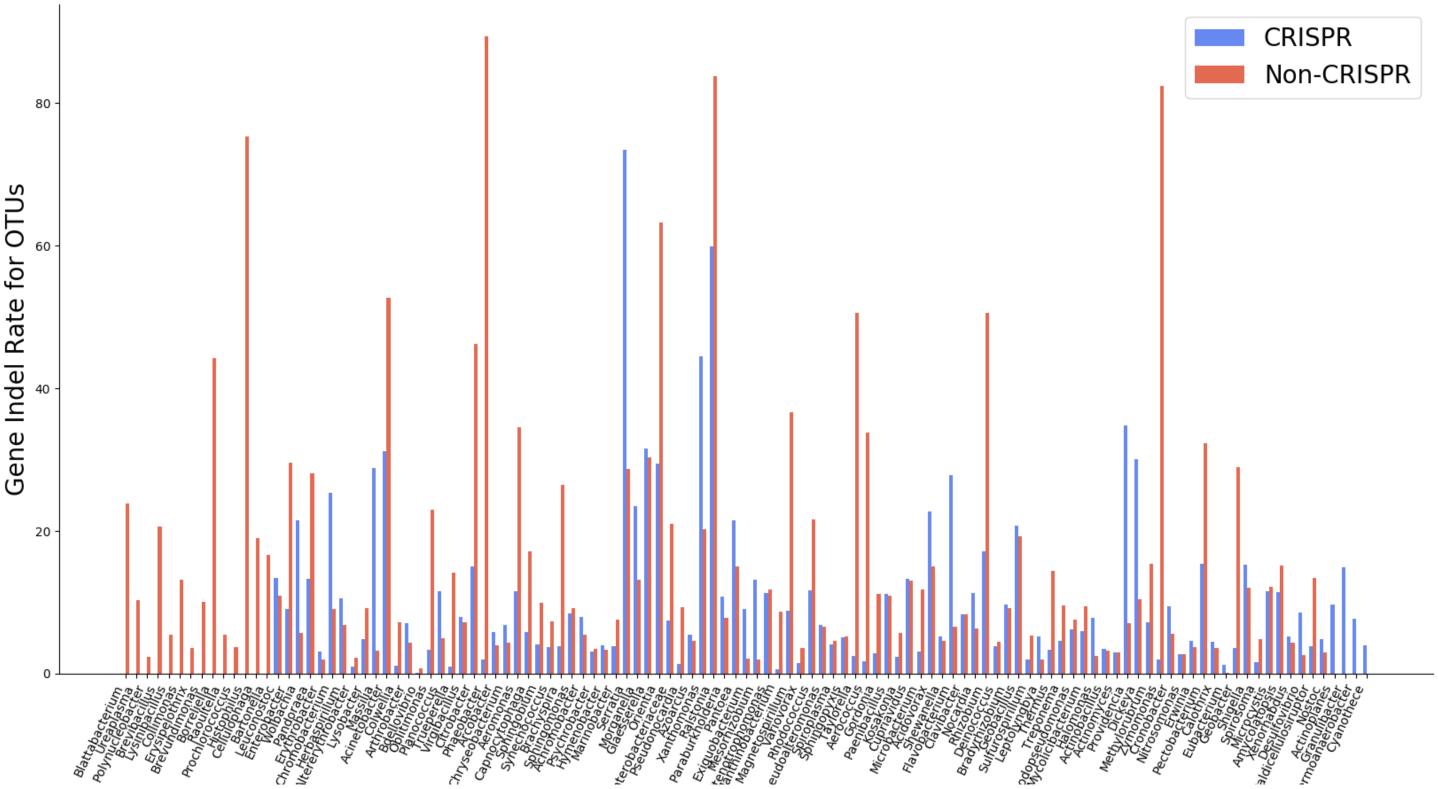

Figure 4: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus. Rate is indel events per base pair substitution. There were 140 genera used.

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.

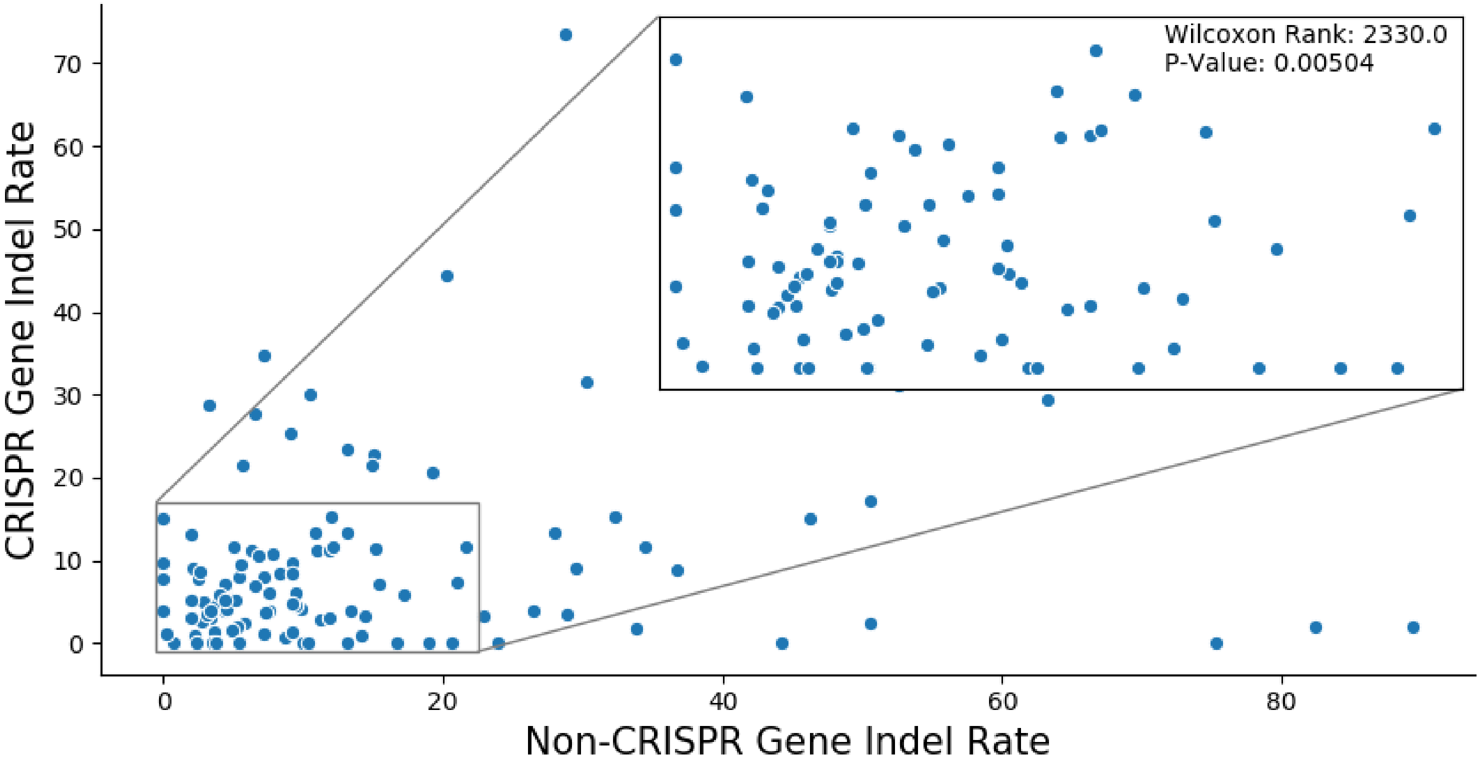

Figure 5: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus. Rate is indel events per base pair substitution. Each point represents a genus. There were 140 genera used.

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.

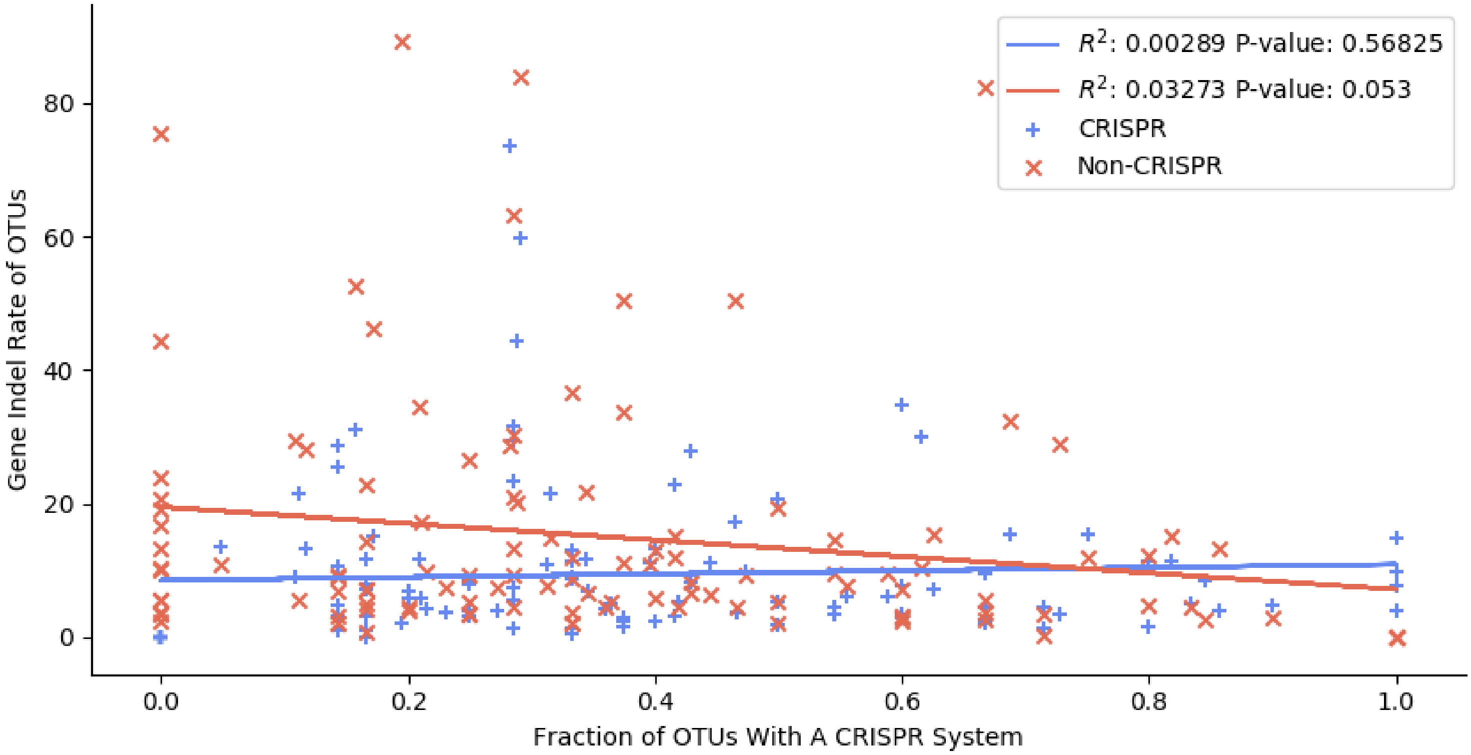

Figure 6: Markopholo estimate of gene indel rates for the partitions of CRISPR and Non-CRISPR OTUs for each genus against the fraction of all OTUs in that genus that are annotated as having a CRISPR system. $R^2$ values are for linear regression lines fit to the CRISPR and non-CRISPR estimates. Rate 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.

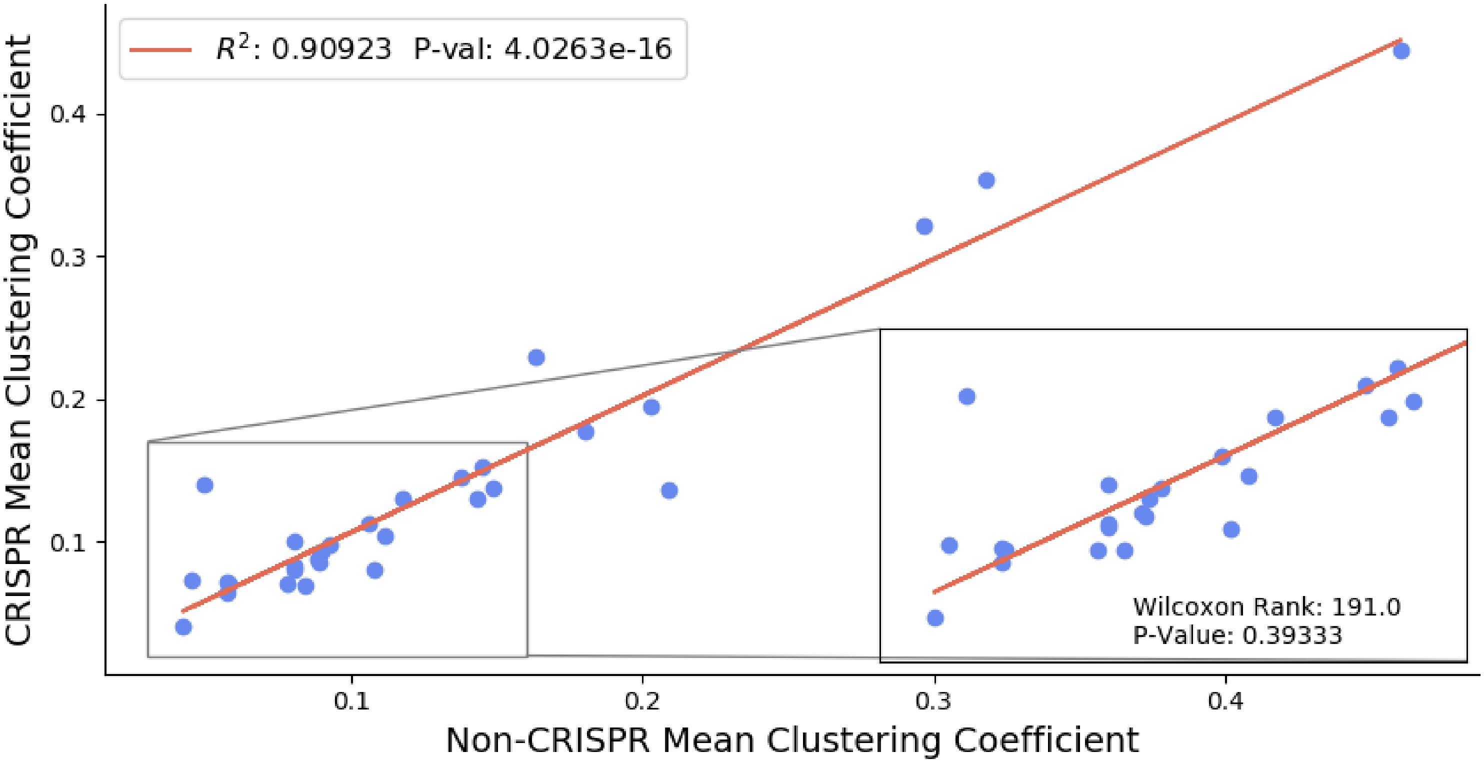

Figure 7: Mean over 1000 bootstraps of the clutering coefficients of the CRISPR OTUs for each genus against the non-CRISPR means over the 1000 bootstraps. $R^2$ value is for the linear regression fit to the CRISPR and non-CRISPR estimates. There were 140 genera used.

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).

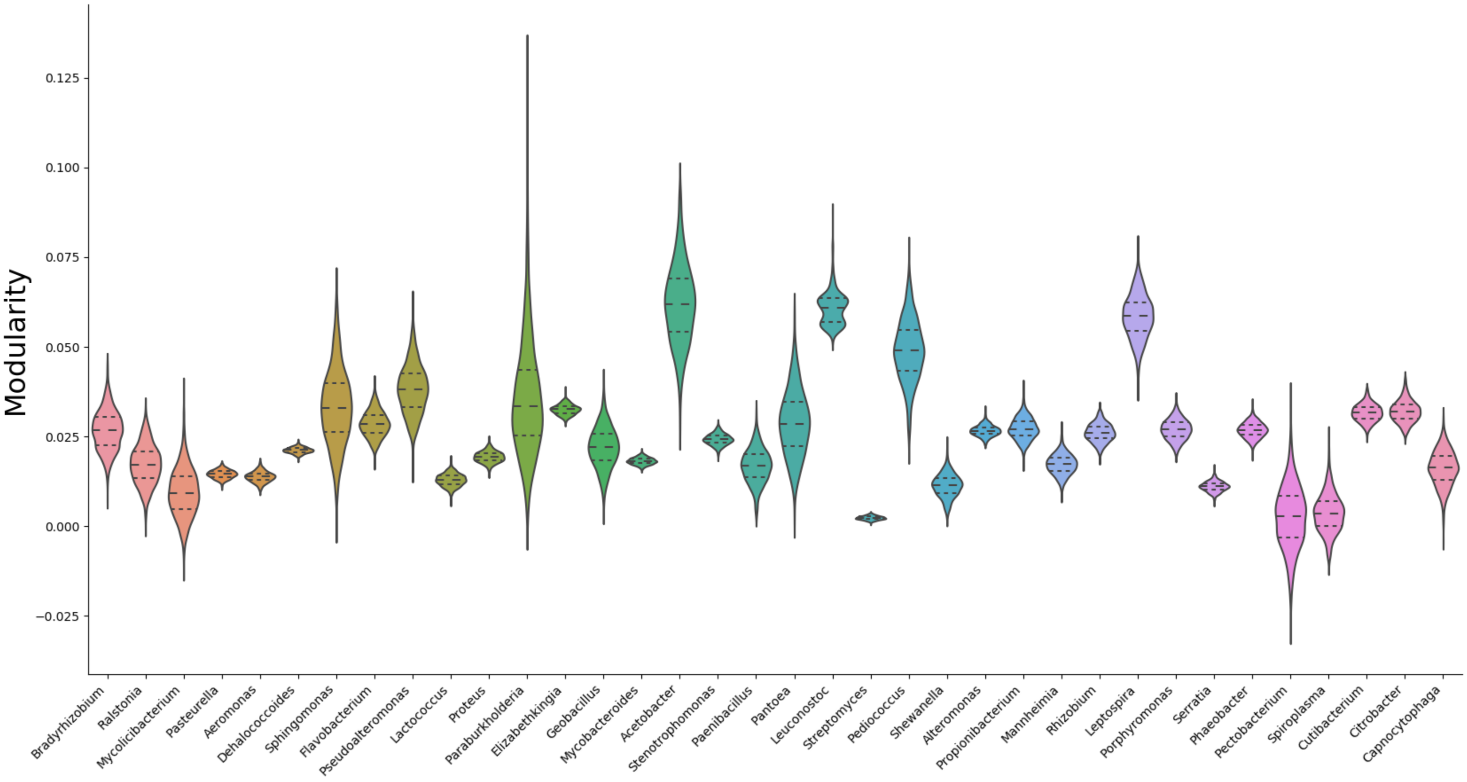

Figure 8: Distribution of network modularity over 1000 bootstrap repliactes for each genus. Plot is of the kernel density estimated from the observed values, with witdh proportional to the number of data points. Lines inside each distribution represent quartiles.

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.

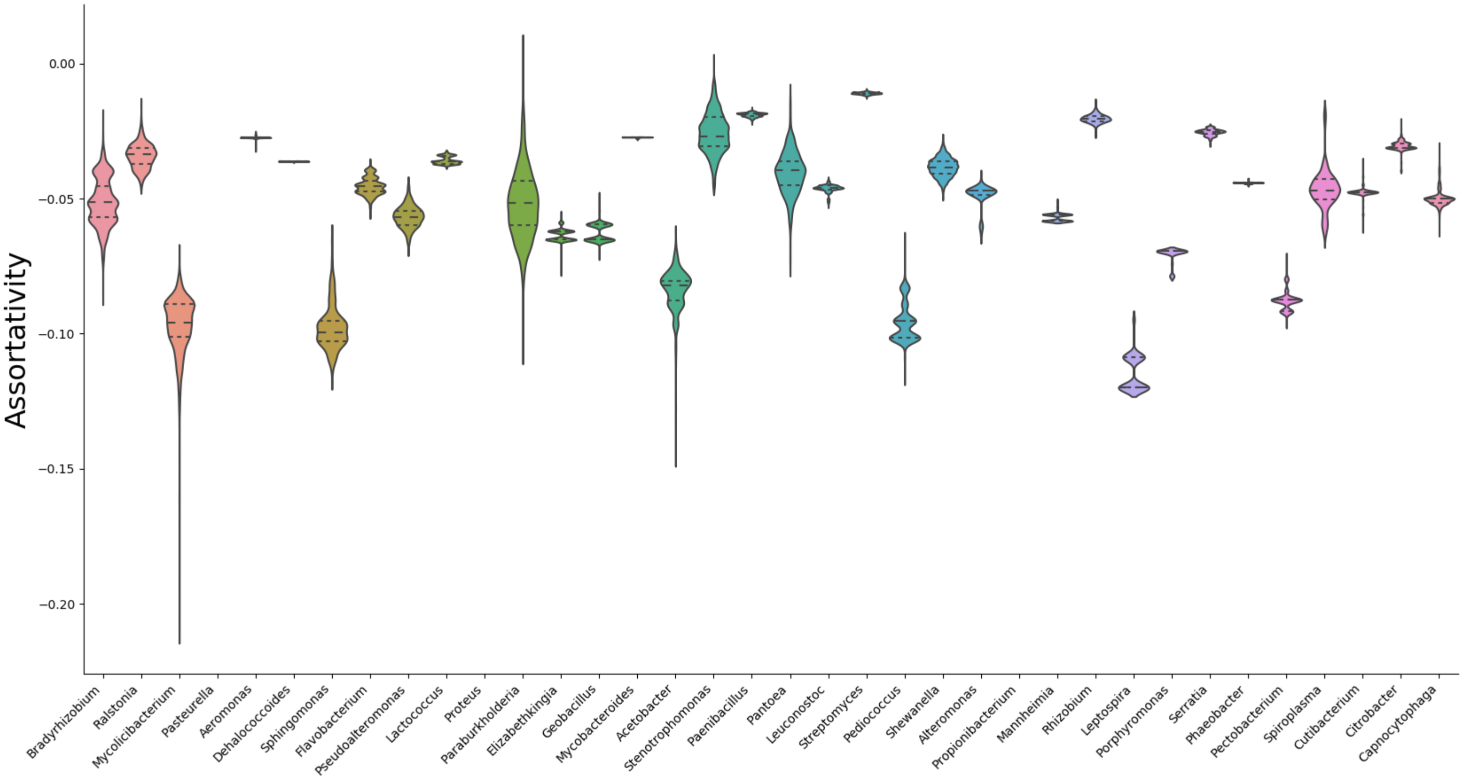

Figure 9: Distribution of network assortativity by CRISPR status (either CRISPR or non-CRISPR) over 1000 bootstrap repliactes for each genus. Plot is of the kernel density estimated from the observed values, with witdh proportional to the number of data points. Lines inside each distribution represent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

# 5 Discussion

## 5.1 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.

### 5.2 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 [22] 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.

### 5.3 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.

## 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. 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).

5. 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).

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

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

8. 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).

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

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

11. 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).

12. 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).

13. 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).

14. 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).

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

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

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

18. 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).

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

20. 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).

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

22. 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).

23. 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).

24. 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).

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

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

27. 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).

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

29. 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).

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

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