Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original Paper

Phylogenetics

Evolink: A Phylogenetic Approach for Rapid Identification of Genotype-Phenotype Associations in Large-scale Microbial Multi-Species Data

Yiyan Yang[1] and Xiaofang Jiang[1,*]

[1]National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The discovery of the genetic features that underly a phenotype is a fundamental task in microbial genomics. With the growing number of microbial genomes that are paired with phenotypic data, new challenges and opportunities are arising for genotype-phenotype inference. Phylogenetic approaches are frequently used to adjust for the population structure of microbes but scaling them to trees with thousands of leaves representing heterogeneous populations is highly challenging. This greatly hinders the identification of prevalent genetic features that contribute to phenotypes that are observed in a wide diversity of species.

Results: In this study, Evolink was developed as an approach to rapidly identify genotypes associated with phenotypes in large-scale multi-species microbial datasets. Compared to other similar tools, Evolink was consistently among the top-performing methods in terms of precision and sensitivity when applied to simulated and real-world flagella datasets. In addition, Evolink significantly outperformed all other approaches in terms of computation time. Application of Evolink on flagella and gram-staining datasets revealed findings that are consistent with known markers and supported by the literature. In conclusion, Evolink can rapidly detect phenotype-associated genotypes across multiple species, demonstrating its potential to be broadly utilized to identify gene families associated with traits of interest.

Availability and implementation: The source code, docker container and web server for Evolink are freely available at https://github.com/nlm-irp-jianglab/Evolink.

Contact: xiaofang.jiang@nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identifying relationships between genotypes and phenotypes is a crucial task in biology. One vital methodology in this field is genome-wide association studies (GWAS), which has led to tremendous advances in understanding complex traits and have uncovered hundreds of genetic variants in humans over the past several decades (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; O'Brien *et al.*; The Electronic Medical Records and Genomics (eMERGE) Consortium *et al.*, 2014). GWAS approaches have been successfully adopted in the field of microbiology and are known as microbial GWAS (mGWAS) (Power *et al.*, 2017; Falush, 2016). Unlike human genetic variations that are reassorted by meiosis, microbial genetic changes can reach high frequency on multiple genetic backgrounds, causing standard GWAS methods to predict many false associations. To address this issue, mGWAS methods were developed to control for the influence of microbial population structure (San *et al.*, 2020; Chen and Shapiro, 2015). Notably, the use of phylogeny-based solutions has been shown to be an effective approach (Farhat *et al.*, 2013; Sheppard *et al.*, 2013; Weimann *et al.*, 2016).

mGWAS methods are limited to analyzing a single species or a few sub-species (Collins and Didelot, 2018; Lees *et al.*, 2018, 2020, 2016; Earle *et al.*, 2016; Saund and Snitkin, 2020). However, phenotypic traits can be

shared across different species and even phyla, especially in cases of convergent evolution. Therefore, it is questionable whether mGWAS methods could be applied to distantly related species (Dunn and Munro, 2016) because they are not specifically aimed at accounting for phylogenetic relationships across multiple species. Compared with single nucleotide polymorphisms (SNPs), which are used in most mGWAS tools, gene gains and losses may contribute equally, if not more, to phenotypic evolution across broader taxonomic ranges. So far, only a few approaches (Kowalczyk *et al.*, 2019; Prudent *et al.*, 2016; Nagy *et al.*, 2014, 2020) have been proposed to perform phylogeny-aware multi-species comparative genomic analysis, but their primary focus has been eukaryotic species. In microbiology, the rapid progress in sequencing technologies has led to the accumulation of microbial species genomes, gene family pools (Huerta-Cepas *et al.*, 2019), and phylogenetic trees with tens of thousands of leaves representing heterogeneous populations (Parks *et al.*, 2022; Zhu *et al.*, 2019). This presents a highly challenging situation that has grown beyond the ability of most tools and highlights the need for a method that is as powerful as traditional mGWAS approaches but can be applied in comparative genomics analyses over large taxonomic scales. In this study, we proposed an efficient and easy-to-interpret index, the Evolink index, to measure genotype-phenotype associations across multiple microbial species while explicitly accounting for the phylogenetic relationships of those species. We also presented a tool named Evolink to facilitate the calculation of this index. Compared with alternative methods, Evolink demonstrated its robustness and promising performance on both simulated and empirical datasets, ranking among the highest-performing methods in F1 scores and having the shortest runtimes in various scenarios.

When applied to two real-world datasets, the genes identified by Evolink showed a high association with the phenotypes and were well supported by previous studies. Evolink facilitates expeditious genotype-phenotype association detection in multiple species, addressing the increasing need for efficient microbial data analysis. The source code and documentation are freely available at https://github.com/nlm-irp-jianglab/Evolink. We also provided a docker container (https://hub.docker.com/r/nlmirpjianglab/evolink) and a web portal (https://jianglabnlm.com/evolink).

2 Methods

2.1 Simulated datasets

In this study, we created four groups of simulated datasets to evaluate Evolink and other comparable methods (Table S1). The first group was designed to examine each method's robustness to the distribution of positive phenotypes across species and included five datasets where the phenotypes ranged from 10% to 90% in prevalence. The second group tested the stability of the methods with increasing levels of phenotype phylogenetic overdispersion (*i.e.,* decreasing levels of phenotype clustering), and included five datasets (details in Figure S1). The third group, comprised of six datasets, was used to evaluate runtime with a fixed number of 10,000 gene families but varying numbers of species (100 to 3,200). The fourth group, consisting of five datasets, was used to evaluate runtime with a fixed number of 1,000 species but varying numbers of gene families (10K to 160K).

To simulate gene presence/absence changes along the species phylogeny, we used a continuous-time Markov chain, an approach that has been widely adopted for simulating gene family evolution in prokaryotes (Cohen and Pupko, 2010; Cohen *et al.*, 2010, 2013; Zamani-Dahaj *et al.*, 2016; Collins and Didelot, 2018). We first devised four instantaneous transition rate matrices. The Q_{pos} and Q_{neg} matrices are used to simulate positively phenotype-associated and negatively phenotype-associated gene families, respectively. The Q_p matrix represents the instantaneous transition rate matrix used for phenotype evolution simulation.

$$Q_{pos} = \begin{array}{c} \\ G_0P_0 \\ G_0P_1 \\ G_1P_0 \\ G_1P_1 \end{array} \begin{array}{c} \begin{array}{cccc} G_0P_0 & G_0P_1 & G_1P_0 & G_1P_1 \end{array} \\ \begin{pmatrix} -2x & x & x & 0 \\ \alpha x & -2\alpha x & 0 & \alpha x \\ \alpha x & 0 & -2\alpha x & \alpha x \\ 0 & x & x & -2x \end{pmatrix} \end{array}$$

$$Q_{neg} = \begin{array}{c} \\ G_0P_0 \\ G_0P_1 \\ G_1P_0 \\ G_1P_1 \end{array} \begin{array}{c} \begin{array}{cccc} G_0P_0 & G_0P_1 & G_1P_0 & G_1P_1 \end{array} \\ \begin{pmatrix} -2\alpha x & \alpha x & \alpha x & 0 \\ x & -2x & 0 & x \\ x & 0 & -2x & x \\ 0 & \alpha x & \alpha x & -2\alpha x \end{pmatrix} \end{array}$$

$$Q_p = \begin{array}{c} \\ P_0 \\ P_1 \end{array} \begin{array}{c} \begin{array}{cc} P_0 & P_1 \end{array} \\ \begin{pmatrix} -(x+\alpha x) & x+\alpha x \\ x+\alpha x & -(x+\alpha x) \end{pmatrix} \end{array}$$

The x represents the total number of substitutions, which was calculated as 80 divided by the total length of species tree branches. The parameter α represents the level of association between the gene family and the phenotype and was set to 15.

For non-associated gene families that are independent of phenotype, a $Q_{non-associated}$ matrix is designed as follows:

$$Q_{non-associated} = \begin{array}{c} \\ G_0 \\ G_1 \end{array} \begin{array}{c} \begin{array}{cc} G_0 & G_1 \end{array} \\ \begin{pmatrix} -p_{gene} \cdot 2(x+\alpha x) & p_{gene} \cdot 2(x+\alpha x) \\ (1-p_{gene}) \cdot 2(x+\alpha x) & -(1-p_{gene}) \cdot 2(x+\alpha x) \end{pmatrix} \end{array}$$

$$p_{gene} = \frac{species\ number\ having\ the\ gene}{species\ number}, 0 < p_{gene} \leq 1$$

where p_{gene} represents a broad range of gene prevalence with the number of species having this gene being sampled from a beta-binomial distribution $Betabinom(size = species\ number, \alpha = 0.007, \beta = 0.75)$, to mimic the distribution of gene prevalence observed in the empirical data (Figure S2).

To simulate the data, we followed these steps:

1. The phylogenetic species trees were simulated with the "pbtree" function in the phytools R package (v1.0.3) (Revell, 2012). The tree branch lengths resembled those of the empirical species trees in terms of their distribution and descriptive statistics (Figure S3). The same species trees were used to analyze datasets with varying levels of phenotype prevalence and phenotype phylogenetic overdispersion, respectively.
2. The phenotype data for datasets with varying phenotype prevalence was generated by randomly assigning corresponding percentages of tree leaves with the presence of the phenotype. To simulate phenotype phylogenetic overdispersion, the species tree was cut into multiple clusters and the phenotype presence/absence was allocated to the interleaved clusters.
3. Given the assigned phenotype states on tree leaves and the Q_p matrix, the probabilities of phenotype presence/absence at each internal node were calculated through 10000 repetitions of stochastic character mapping using the "make.simmap" function in the phytools R package (v1.0.3) (Revell, 2012).
4. The positively and negatively phenotype-associated gene families were simulated along the tree branches based on Q_{pos} or Q_{neg}, respectively and adjusted by the pre-calculated phenotype presence/absence probabilities at each node of the tree. Each simulated dataset contained ten positively associated and ten negatively associated gene families. The non-associated genotypic states were simulated along the tree via the $Q_{non-associated}$ matrix. If a gene family to simulate has low prevalence, it may be absent in all species after simulation. In this case, the simulated data for that gene family would be discarded and rerun.

Descriptive statistics of species tree branch lengths and phenotype and gene gain/loss rates in the simulated data are provided in Table S1. The simulated datasets are available as Supplementary Data S1 to S4 and the scripts to generate them are provided at https://github.com/nlm-irp-jianglab/Evolink/tree/main/Evolink_paper.

2.2 Empirical dataset

Madin *et al.* provided a synthesis of bacterial and archaeal phenotypic trait datasets by unifying multiple microbial trait sources (https://github.com/bacteria-archaea-traits/bacteria-archaea-traits/releases/tag/v1.0.0) (Madin *et al.*, 2020). From the genomes that could be mapped to the 5,709 WoL (Reference Phylogeny for Microbes) bacterial species (Zhu *et al.*, 2019), a total of 8,172 genomes labeled with "flagella" or "no" in the motility category were extracted. The "flagella" label was a subcategory under "motility". Genomes labeled "yes" were excluded, while genomes labeled with "no" were treated as lacking both motility and flagella, and were used as negative samples, resulting in phenotypic assignments for 1,978 WoL species. Since including species with ambiguous phenotypes can potentially introduce noise and reduce the accuracy of the results, we excluded species that had strains with mixed labels, resulting in 1,948 unambiguously labeled species, including 284 species with flagella function and 1,664 without. A rooted phylogenetic species subtree for the above species was extracted from the WoL reference species tree. These species represent a broad taxonomic range, including 31 phyla, 57 classes and 264 families (Figure S4). Protein coding genes for the species-representative genomes were predicted using Prokka v1.14.6 (Seemann, 2014) with default settings and bacterial COG annotation was performed using eggNOG-mapper v2.1.6 (Cantalapiedra *et al.*, 2021). A total of 149,316 gene families were converted into a binary gene presence/absence matrix with rows corresponding to gene families and columns to species. The flagella ground truth, comprised of 24 structural genes that are hypothesized to have been conserved in the ancestral bacterial genome, is an ideal resource for investigating genotype-phenotype associations in bacteria (Liu and Ochman, 2007). These 24 genes were assigned to 21 gene families using eggNOG-mapper (with multiple genes sometimes assigned to a single family) and were used as the minimum reference set for flagella-associated gene families in the empirical dataset (Supplementary Data S5).

To further demonstrate the utility of our method using another real-world example, we used the gram-staining dataset from the study of Madin *et al.* (Madin *et al.*, 2020). The data was processed in the same way as the flagella dataset. A total of 30,493 genomes labeled with "Gram-negative" (labeled as phenotype presence) or "Gram-positive" (labeled as phenotype absence) in the gram-staining category were extracted and associated with 4,472 WoL species. After excluding species containing mixed gram strains, there were 4,104 unambiguously labeled species consisting of 2,503 Gram-negative species and 1,601 Gram-positive species (Supplementary Data S6).

To generate subsets from these large datasets, we utilized Treemmer v0.3 (Menardo *et al.*, 2018) to randomly sample 10% of the species from the original species tree. This approach allowed us to create flagella and gram-staining subset trees with minimal loss of phylogenetic diversity.

2.3 Benchmarking

The benchmarking was done using both simulated data and empirical data with flagella as a phenotype. Tetrachoric correlation, a measure of the correlation between two binary variables was used as the baseline (Divgi, 1979). Evolink was also compared with the other two types of phylogeny-aware methods. The multi-species methods considered were COMPARE (comparative phylogenomic analysis of trait evolution) (Nagy *et al.*, 2014), RERconverge (Kowalczyk *et al.*, 2019), ForwardGenomics (Prudent *et al.*, 2016), and Phylolm (Bradley *et al.*, 2018), however, only ForwardGenomics and Phylolm were used in the comparison due to the unavailability of open-source software or a webserver implementation for COMPARE and the computational cost of constructing gene trees for each gene in RERconverge. Only the "GLS (generalized least square)" method was used for ForwardGenomics as the "branch" method prepares ancestral reconstructions for all gene families, which is computationally costly. From the available mGWAS methods, Bugwas (Earle *et al.*, 2016), Pyseer (Python sequence element enrichment analysis) (Lees *et al.*, 2020, 2018, 2016), treeWAS (Collins and Didelot, 2018), and Hogwash (Saund and Snitkin, 2020) were chosen because they incorporated phylogenic information. Bugwas was modified to handle binary phenotypes by converting the genotype binary matrix into a biallelic SNP-like format. Hogwash was limited to using the synchronous test and phyC test due to the binary phenotypes used in the comparison. The parameters and descriptions of the methods tested in this study are provided in Table S2.

All methods were executed on a high-performance computing node with 32 CPUs and 240 GB of memory. A series of metrics including precision, recall, F1 score, balanced accuracy, the area under the precision-recall curve (PRAUC), false positive rate and runtime were utilized to evaluate the performances of these methods.

3 Results

3.1 Design of the Evolink index

The Evolink index is based on the calculation of four Faith's phylogenetic diversities for species with and without a gene family in the subtrees based on the presence and absence of the phenotype (G_1P_1, G_0P_1, G_1P_0 and G_0P_0 in Fig.1A). In this study we focus on the presence or absence of gene families, but these features could represent other genetic characteristics in a more general sense. Faith's phylogenetic diversity (PD) of a set of species is defined as the sum of the lengths of all those branches on the tree that span the members of the set (Faith, 1992; Faith and Richards, 2012). Due to Faith's PD's clear rationale and simplicity, other metrics derived from it, such as Functional diversity, RecPD, and Unifrac have been proposed and widely utilized (Petchey and Gaston, 2002; Bundalovic-Torma *et al.*, 2022; Lozupone and Knight, 2005; Lozupone *et al.*, 2006, 2011).

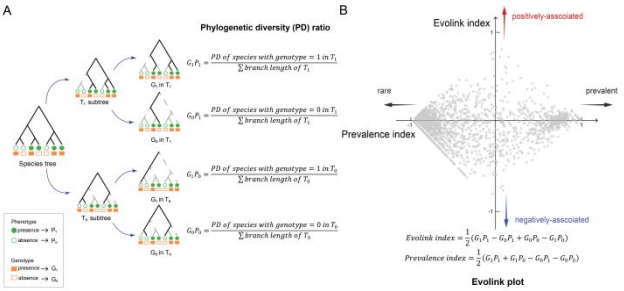


Fig. 1. Schematic diagrams for the Evolink index and the Evolink plot. (A) The schematic diagram for the Evolink index. The calculation of the Evolink index is based on four Faith's phylogenetic diversities (Faith's PD) for species with and without a gene family

in the phenotype-positive and negative subtrees. (B) The schematic diagram for the Evolink plot. Each point in the Evolink plot represents a gene family with its Evolink index on the y-axis and its Prevalence index on the x-axis. The Evolink index quantifies the association of a gene family with the phenotype, whereas the Prevalence index indicates the gene prevalence across species.

We define $\varphi(X,t)$ as a function to extract a subtree from tree t with a subset of species X and $\varphi(X,t) \subseteq t$. We also define $PD(X, t)$ as a function to calculate the Faith's phylogeny diversity for a set of species X on tree t, and $bl(t)$ as a function to get the sum of branch lengths of t. Two functions, $pheno(x)$ and $gene_i(x)$, to get the presence and absence binary status of the phenotype and any gene family i for a species x, are defined as follows (Fig.1A):

$$pheno(x) = \begin{cases} 1, & if\ species\ x\ has\ the\ phenotype \\ 0, & if\ species\ x\ doesn't\ have\ the\ phenotype \end{cases}$$

$$geno_i(x) = \begin{cases} 1, & if\ species\ x\ has\ gene\ family\ g_i \\ 0, & if\ species\ x\ doesn't\ have\ gene\ family\ g_i \end{cases}$$

Given a phylogenetic tree T with a set of species $S = \{sp_1, sp_2, ..., sp_i, ..., sp_N\}$ which contain a set of gene families $G = \{g_1, g_2, ..., g_i, ..., g_M\}$, the Evolink index for each gene family g_i is defined as follows:

$$Evolink\ index = \frac{1}{2}\left(G_1^i P_1 - G_0^i P_1 + G_0^i P_0 - G_1^i P_0\right)$$

where

$$G_x^i P_y = \frac{PD(\{sp_i : geno_i(sp_i) == x\}, T_y)}{bl(T_y)}$$

is defined as a phylogenetic diversity ratio, with $x, y \in \{0, 1\}$ and $T_y = \varphi\left(\{sp_i : pheno(sp_i) == y\}, T\right)$.

The higher the Evolink index, the more positively associated the gene family is with the phenotype, while a smaller index value indicates a more negative association with the phenotype. The Evolink index ranges from -1 to 1, with a maximum or minimum value indicating that a gene family has the exact same or opposite presence/absence pattern as the phenotype, respectively.

Likewise, a Prevalence index can also be defined as:

$$Prevalence\ index = \frac{1}{2}\left(G_1^i P_1 + G_1^i P_0 - G_0^i P_1 - G_0^i P_0\right)$$

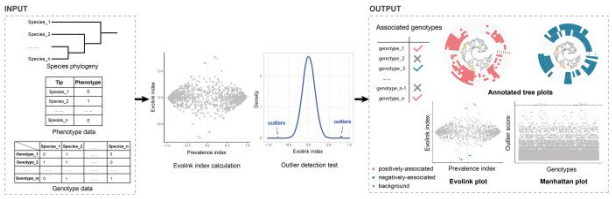
The higher the Prevalence index is, the more prevalent the gene family is across the species. It also ranges from -1 to 1 and is highly correlated with the gene family prevalence in species (Figure S5).

3.2 Implementation of Evolink

Evolink requires three inputs: a rooted phylogenetic tree with leaves representing species, a gene family presence/absence matrix and a phenotype presence/absence list mapped to species (presence=1 and absence=0) (Figure 2). The gene family matrix can be a binary presence/absence matrix or a numeric matrix (representing gene copies, k-mer counts, etc.). To ensure proper processing of numeric matrix inputs, users must inform Evolink of this using the "-c/--copy-number" option to convert the matrix to binary format. The Evolink and Prevalence indices are computed by extracting phenotype-based subtrees from the species tree and combining the four Faith's PD values. Phenotype-associated gene family candidates can then be identified with various available methods in the Evolink such as the Generalized Extreme Studentized Deviate (GESD) test (Rosner, 1983), Isolation Forest (Liu *et al.*, 2012), by using z-score or using a custom-defined Evolink index threshold to detect outliers. The GESD test is a statistical test designed for detecting outliers in a univariate dataset. In this test, the p-value is used to determine whether an observation in the dataset deviates significantly from expected values and is therefore considered an outlier. In the context of Evolink, the p-value measures the probability that the Evolink index value of a gene family is an outlier. Isolation Forest, a machine learning anomaly detection algorithm, detects gene families whose absolute Evolink index values significantly deviate from the background. The threshold for identifying gene families with significant associations is determined by the upper bound of maximal differences of the sorted outlier scores. Isolation Forest is used by default in the Evolink. By providing multiple options for outlier detection, Evolink aims to empower users to conduct extensive exploration of their data.

To facilitate the visualization of results from Evolink, we designed the Evolink plot, a type of scatter graph based on the Evolink and Prevalence indices, with each point representing a single gene family (Fig.1B). The Evolink index naturally ranks widespread and rare gene families lower than those who are not. In the Evolink plot, the most positively associated gene families (with higher Evolink indices on the top) and negatively associated gene families (with lower Evolink indices on the bottom) tend to have moderate Prevalence index values. Apart from the Evolink plot, a zipped input file for users to visualize the tree with iTOL v5.0 (Letunic and Bork, 2021), a species tree with leaves annotated with the presence/absence of phenotypes and the top associated gene families, and a Manhattan plot are generated (Fig.2).

Fig. 2. The workflow of Evolink. With a species tree, a binary phenotypic list, and a binary gene family matrix as input, the Evolink index for each gene family is calculated to



investigate its association with the phenotype. An outlier detection approach (by default isolation forest) is applied to identify significant gene families that are associated with the phenotype. These results can be visualized using tree plots, Evolink plots, and Manhattan plots.

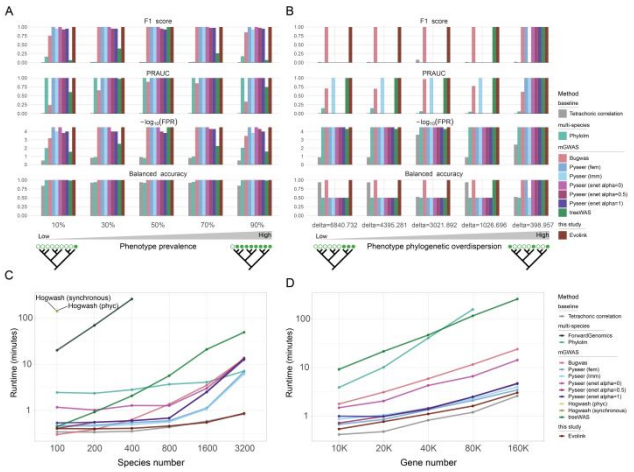
3.3 Evaluation of Evolink Performance on Simulated Datasets

We compared Evolink with twelve other methods on four types of simulated data to assess its performance under various conditions (Table S2). These compared methods included tetrachoric correlation, phylogeny-aware multi-species methods (ForwardGenomes, Phylolm) and repurposed mGWAS methods (Bugwas, Pyseer using different linear models, treeWAS and Hogwash using different tests).

First, we tested how the ratio of positive and negative phenotypes would impact the performance of the methods being evaluated. In many cases, phenotypic information is unavailable for a substantial number of strains and the distribution of phenotypes is often highly variable and uneven (Madin *et al.*, 2020; Nayfach *et al.*, 2021; Mukherjee *et al.*, 2021). We used five datasets with different phenotype prevalence values across species to mimic these situations and benchmarked the performance of each tool. Most methods perform well except for the tetrachoric correlation when the phenotype is evenly distributed among species, but their accuracy decreases when the phenotype prevalence deviates from 50%, particularly in terms of precision (Fig.3A). Among these methods, the performance of treeWAS is the most susceptible to phenotype prevalence. Unlike most methods, Phylolm, predicted many more

associated genes than the ground truth gene set, resulting in much lower precision and a higher false positive rate (Table S3).

Fig. 3. Comparing Evolink with alternative methods on simulated datasets. (A-B)



Comparison of F1 scores, the areas under the precision-recall curve (PRAUC), false positive rates (FPR), and balanced accuracies of methods on datasets with a range of (A) phenotype prevalence and (B) phenotype phylogenetic overdispersion. Note that because treeWAS uses three different strategies (namely terminal score, simultaneous score, and subsequence score), the PRAUCs for treeWAS were the maximum PRAUC of the three strategies on each dataset. (C-D) Comparison of runtimes among methods on datasets with a variety of (C) species numbers and (D) gene family numbers. Note that only methods with runtime less than 5 hours (300 minutes) on the datasets are shown. Pyseer (fem): Pyseer using the fixed effects model; Pyseer (lmm): Pyseer using linear mixed model; Pyseer (enet): Pyseer using the elastic net model. Please refer to Table S2 for parameters and descriptions of the tested methods.

Next, these methods were tested on datasets with different levels of phenotype phylogenetic overdispersion. A phenotype phylogenetic overdispersion value measures the spread of a phenotype across different species on a phylogenetic tree. It provides information about how evenly or unevenly the phenotype is distributed across the tree. All the tested methods, except for Pyseer using the linear mixed model and Phylolm, were able to accurately predict the ground truth on the dataset with the highest phenotype phylogenetic overdispersion. However, most of the methods compromised, in varying degrees, in F1 scores when the phenotype was more clustered since they tended to predict fewer significant genes. When evaluating the F1 scores on the dataset with the lowest phenotype phylogenetic overdispersion, only Bugwas and Evolink accurately identified the ground truth (Fig.3B, Table S3). In summary, Evolink's performance was stable and remained the top method on a broad range of phenotype prevalence and phenotype phylogenetic overdispersion values, indicating its robustness to changing phenotype distributions (Fig.3A-B).

Lastly, to evaluate runtime performance, two groups of datasets were used, with varying numbers of species and gene families, respectively. The first group had a fixed gene family number (N=10K) and varying number of species, and all methods except Hogwash and ForwardGenomics finished within 25 minutes. Evolink performed well and exhibited a notable advantage in speed when analyzing larger species sizes (Fig.3C). The second group had a fixed species number (N=1000) and a varying number of gene families, with Evolink showing comparable speed to the fastest mGWAS methods, producing results in under 5 minutes for all the datasets (Fig.3D). Both Hogwash and ForwardGenomics failed to produce results on the smallest dataset. The results of testing with these simulated datasets indicate that Evolink's performance is highly robust and that it is the fastest among all the tested methods except for the baseline method (Figure S6, Table S3).

It should be noted that despite using the coefficient with the highest F1 score as the threshold to identify significant genes, the baseline tetrachoric correlation method was outperformed by phylogeny-aware methods on the simulated data, indicating the importance of phylogeny information in identification of microbial genotype-phenotype associations (Fig.3). Overall, Evolink demonstrated superior performance on simulated data while maintaining an efficient runtime. Nevertheless, these simulation results only provide a partial evaluation of the method's performance, and thus an evaluation based on empirical data was further performed.

3.4 Evaluation of Evolink Performance on an Empirical Dataset with Flagella as a Phenotype

We further compared Evolink with eight other approaches on an empirical dataset using flagella as the phenotype of interest containing 1,948 species and 149,316 gene families (Fig.4). The approaches involved in the comparison include Phylolm, Bugwas, Pyseer using different linear models and treeWAS. We selected flagella as a phenotype because it plays a crucial role in several essential bacterial functions, such as bacterial motility, adhesion, biofilm formation, and host invasion (Kirov, 2003; Haiko and Westerlund-Wikström, 2013).

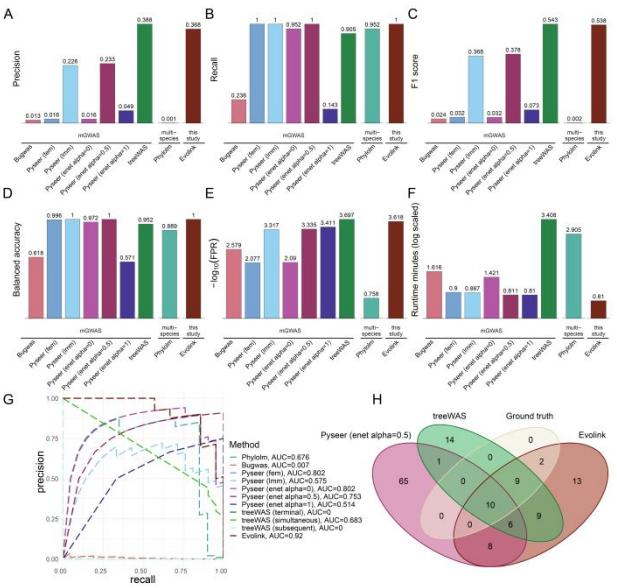


Fig. 4. Comparing Evolink with alternative methods on the flagella dataset. The (A) precisions, (B) recalls, (C) F1 scores, (D) false positive rates (FPR), (E) balanced accuracies, and (F) runtimes of the selected nine methods on the flagella dataset. Note that because treeWAS uses three different strategies (namely terminal score, simultaneous score, and subsequence score), the PRAUCs for treeWAS were the maximum PRAUC of the three strategies on each dataset. (G) The precision-recall curves and the areas under the curves for all the methods. (H) The Venn diagram showing the intersection of the ground truth and the flagella-associated genes identified by the top three methods ranked by F1 scores. Pyseer (fem): Pyseer using the fixed effects model; Pyseer (lmm): Pyseer using linear mixed model; Pyseer (enet): Pyseer using the elastic net model. Please refer to Table S2 for parameters and descriptions of the tested methods.

None of the tested methods achieved a precision of 1.0. One possible explanation for this is that the ground truth gene set used consists of a core set of genes that are exclusively and universally associated with flagella. This means that accessory or taxonomically restricted genes may be excluded, despite potentially being strongly associated with flagella. Nevertheless, Evolink ranked second with a high precision of 0.368, only slightly lower than treeWAS's precision of 0.388 (Fig.4A, Table S4). The top four methods based on F1 scores were Pyseer using the linear mixed model, Pyseer using the elastic net model (alpha=0.5), treeWAS and Evolink with F1 scores of 0.368, 0.378, 0.543 and 0.538, respectively (Fig.4C, E). Although treeWAS had a higher F1 score than other methods, it only predicted 19 out of 21 ground truth genes, resulting in a lower recall of 0.905 (Fig.4B). Evolink was the quickest method with a runtime of 4 minutes, while treeWAS was the most time-intensive, taking over 42 hours to complete (Fig.4F). Furthermore, the areas under the precision-recall curve (PRAUC) of all methods were calculated to compare their rankings of flagella-associated gene families. Although Pyseer using the linear mixed model, Pyseer using the elastic net model (alpha=0.5), and treeWAS achieved similar F1 scores to Evolink, they had lower PRAUC scores than Evolink, which achieved a PRAUC of 0.92 (Fig.4G). To make evaluating ForwardGenomics and Hogwash feasible, we utilized a subset of the data. Even though the runtime of the subset data was still longer than that of other methods on the full dataset, ForwardGenomics demonstrated fair performance with an F1 score of 0.678 and a PRAUC of 0.754 (Table S4). Despite their time-consuming nature, the performances of treeWAS and ForwardGenomics on the flagella (sub)dataset suggest their potential application for analyzing small datasets. To summarize, we have shown that Evolink is a highly competitive approach when compared with alternative methods in the empirical dataset, achieving high F1 scores while maintaining the shortest runtime.

Evolink predicted 57 gene families that were positively associated with the flagella phenotype, and no negatively associated gene families (Fig.5A, Table S5). Ten of these were universally shared with the ground truth and the other top two methods based on F1 score, while 13 were uniquely found by Evolink (Figure 4H). The presence/absence of the top five genes ranked by the Evolink index was mapped to the species tree, demonstrating a strong association with the presence and absence of the flagellar function at leaves (Fig.5B). The 36 gene families predicted by Evolink yet not in the ground truth were either directly related to flagella such as *FliJ*, *FliK*, and *FliO*, or involved in chemotaxis, type III secretion system, and ATP-dependent protease. Among the 13 genes unique to Evolink, most were related to flagella, chemotaxis, and signal transduction. However, two sRNA-binding regulator proteins (COG1551, CsrA/RsmA; COG1923, Hfq), a disulfide bond formation protein (COG1495, DsbB), and an uncharacterized conserved protein (COG1671, UPF0178 family) were also identified (Table S5). It has been widely reported that CsrA/RsmA proteins can directly regulate the flagella expression and the RNA-binding protein Hfq acts as a cofactor during the process (Mika and Hengge, 2013; Wei *et al.*, 2001; Timmermans and Van Melderen, 2010). An early study reported that mutations in *DsbB* can disrupt flagellar assembly in *Escherichia coli* (Dailey and Berg, 1993). To further measure if these gene families are correlated with flagella, we assigned each gene family an average of STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database association scores (Szklarczyk *et al.*, 2021) between itself and the ground truth genes. Forty-five of the predicted genes reached scores greater than 700, indicating their strong association to known flagella markers (Table S5).

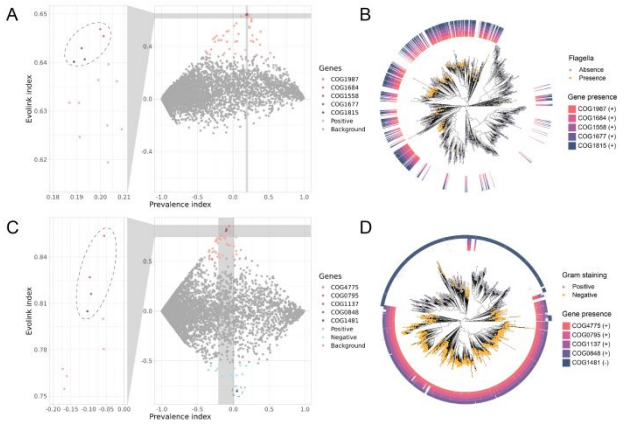


Fig. 5. The flagella-associated and gram staining-associated gene families identified with Evolink. (A) Evolink plot for flagella data. The left panel shows the top five positively flagella-associated genes (COG1987, COG1684, COG1558, COG1677, and COG1815) are highlighted in a zoom-in view. The right panel shows the overview of the Evolink plot including other positively associated (light red) and non-associated (grey) genes. (B) The presence and absence of flagellar function and the top five genes were mapped to the species tree (n=1,948). (C) Evolink plot for gram staining data. The left panel shows the top four Gram-negative-associated genes (COG4775, COG0795, COG1137, and COG0848) and the top one Gram-positive-associated gene (COG1481) in a zoom-in view. The right panel shows the overview of the Evolink plot including other positively associated (light red), other negatively associated (light blue) and non-associated (grey) genes. (D) The states of gram staining and the top five associated gene presence and absence were mapped to the species tree (n=4,104). Note that if a species is Gram-negative, its phenotype is positive (or labeled as "1"). "+"/ "-" represents positively/negatively associated genes, respectively. Trees are displayed by the ggtree R package (v2.4.2) (Yu, 2020).

3.5 Application of Evolink to a Real-World Dataset with Gram Staining as a Phenotype

To demonstrate the practical applications of Evolink, we applied it to a real-world gram staining dataset including 4,104 species and 191,099 gene families. This dataset, which comprises bacteria classified into two categories based on cell membrane structure as Gram-positive (monoderm) with a single cell membrane and Gram-negative (diderm) with an outer membrane containing lipopolysaccharides (Sutcliffe, 2010), serves as an example of Evolink's ability to effectively analyze real-world multi-species comparative genomics data. Compared with the flagella dataset, this dataset has a larger number of species and a less biased phenotype prevalence (Fig.5D). Although there is no standard ground truth for genes to distinguish Gram-negative and Gram-positive bacteria, the LPS synthesis genes (*lpxABCD*) have been widely recognized as being unique to Gram-negative bacteria, making them useful markers for this group of bacteria and a good test set for Evolink (Opiyo *et al.*, 2010; Taib *et al.*, 2020).

Among the methods applied to this gram-staining dataset, only Evolink, Pyseer using the elastic net model (alpha=0), and Pyseer using the fixed effects model correctly detected all four *lpxABCD* genes (Table S6). Evolink identified 56 significant genes in total, with 41 being associated with Gram-negative bacteria and 15 with Gram-positive bacteria (Fig.5C, Table S7). We further used the average STRING database association scores between a set of genes and *lpxABCD* (referred to as the *lpxABCD* association score) to evaluate the methods' performance. Evolink ranked first with a *lpxABCD* association score of 445.57, while the scores of other methods were below 300 (Table S6). In addition, a comparison of the gene families shared by the above three approaches and the *lpxABCD* gene set showed that the 30 genes uniquely identified by Evolink have a *lpxABCD* association score of 304.48, while 26 genes uniquely found by Pyseer

using the fixed effects model have a score of 292.72, the 5 genes uniquely found by Pyseer using the elastic net model (alpha=0) have a score of 172.90 and the 90 genes uniquely shared by both Pyseer-based models have a score of 134.45 (Figure S7). These collectively suggest that Evolink performed better than the other methods on the gram staining dataset. Moreover, Evolink had the fastest performance compared to other methods on this dataset (Table S6).

4 Discussion

The existing methods for multi-species comparative genomic analysis are primarily geared towards analyzing a limited number of genomes from closely related species and are not equipped to handle datasets containing tens of thousands of species. In response to this, we propose the use of the Evolink index as a tool for measuring genotype-phenotype associations based on phylogeny. The Evolink index provides an efficient way to calculate these associations and is easy to interpret. The constrained values of the index facilitate comparative analyses and the integration of the Evolink index into other large-scale genomics investigations. The utilization of the Evolink index is expected to contribute to the advancement of multi-species comparative genomic analysis and provide valuable insights into the evolution of microbial populations.

Despite its strengths, the Evolink index has several limitations that must be taken into consideration when using it for analysis. First, Evolink is not the optimal solution for detecting horizontally transferred genes such as mobile genetic elements. If a gene is rapidly exchanged between species, the phylogenetic relationships that Evolink leverages are not helpful in detecting the association. Phylogeny-unaware methods such as correlation analysis may be sufficient for identifying the associated genotypes in these cases. Second, Evolink was designed to be used with large datasets. For smaller datasets, other methods like ForwardGenomics and treeWAS could be more accurate, although more time-consuming. We also observed that several mGWAS methods, including Bugwas and Pyseer, exhibited satisfactory performance and speed in certain scenarios, but their suitability for a particular study would depend on the specific characteristics of the dataset and research question being investigated.

Future improvements are possible for Evolink. Currently, Evolink supports only binary phenotype and gene family inputs, but future versions could include the option to convert categorical data into unique binary representations and automatically convert continuous phenotypic/genotypic inputs into binary based on cutoffs determined by the software or provided by the users. Additionally, it may incorporate conversion from genes to higher-level biological ontologies to better address convergent evolution and to provide more flexibility in how genetic features are represented. Overall, Evolink promises to be a useful tool for the further analysis of microbial genomic data and the continued expansion of Evolink will provide new ways to study the genetic basis of traits across biological fields.

Acknowledgements

We gratefully acknowledge the help of Keith Dufault-Thompson on the revision of the manuscript and his constructive suggestions throughout this project. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Funding

Y.Y and X.J. are supported by the Intramural Research Program of the NIH, National Library of Medicine.

Conflict of Interest: none declared.

References

Bradley,P.H. *et al.* (2018) Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLoS Comput Biol*, **14**, e1006242.

Bundalovic-Torma,C. *et al.* (2022) RecPD: A Recombination-aware measure of phylogenetic diversity. *PLoS Comput Biol*, **18**, e1009899.

Cantalapiedra,C.P. *et al.* (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, **38**, 5825–5829.

Chen,P.E. and Shapiro,B.J. (2015) The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, **25**, 17–24.

Cohen,O. *et al.* (2013) CoPAP: Coevolution of Presence–Absence Patterns. *Nucleic Acids Research*, **41**, W232–W237.

Cohen,O. *et al.* (2010) GLOOME: gain loss mapping engine. *Bioinformatics*, **26**, 2914–2915.

Cohen,O. and Pupko,T. (2010) Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution*, **27**, 703–713.

Collins,C. and Didelot,X. (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput Biol*, **14**, e1005958.

Dailey,F.E. and Berg,H.C. (1993) Mutants in disulfide bond formation that disrupt flagellar assembly in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 1043–1047.

Divgi,D.R. (1979) Calculation of the tetrachoric correlation coefficient. *Psychometrika*, **44**, 169–172.

Dunn,C.W. and Munro,C. (2016) Comparative genomics and the diversity of life. *Zool Scr*, **45**, 5–13.

Earle,S.G. *et al.* (2016) Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat Microbiol*, **1**, 16041.

Faith,D. and Richards,Z. (2012) Climate Change Impacts on the Tree of Life: Changes in Phylogenetic Diversity Illustrated for Acropora Corals. *Biology*, **1**, 906–932.

Faith,D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

Falush,D. (2016) Bacterial genomics: Microbial GWAS coming of age. *Nat Microbiol*, **1**, 16059.

Farhat,M.R. *et al.* (2013) Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet*, **45**, 1183–1189.

Haiko,J. and Westerlund-Wikström,B. (2013) The Role of the Bacterial Flagellum in Adhesion and Virulence. *Biology*, **2**, 1242–1267.

Huerta-Cepas,J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, **47**, D309–D314.

Kirov,S.M. (2003) Bacteria that express lateral flagella enable dissection of the multifunctional roles of flagella in pathogenesis. *FEMS Microbiology Letters*, **224**, 151–159.

Kowalczyk,A. *et al.* (2019) RERconverge: an R package for associating evolutionary rates with convergent traits. *Bioinformatics*, **35**, 4815–4817.

Lees,J.A. *et al.* (2020) Improved Prediction of Bacterial Genotype-Phenotype Associations Using Interpretable Pangenome-Spanning Regressions. *mBio*, **11**, e01344-20.

Lees,J.A. *et al.* (2018) pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, **34**, 4310–4312.

Lees,J.A. *et al.* (2016) Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*, **7**, 12797.

Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, **49**, W293–W296.

Liu,F.T. *et al.* (2012) Isolation-Based Anomaly Detection. *ACM Trans. Knowl. Discov. Data*, **6**, 1–39.

Liu,R. and Ochman,H. (2007) Stepwise formation of the bacterial flagellar system. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 7116–7121.

Lozupone,C. *et al.* (2006) UniFrac – An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.

Lozupone,C. *et al.* (2011) UniFrac: an effective distance metric for microbial community comparison. *ISME J*, **5**, 169–172.

Lozupone,C. and Knight,R. (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Appl Environ Microbiol*, **71**, 8228–8235.

Madin,J.S. *et al.* (2020) A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data*, **7**, 170.

Menardo,F. *et al.* (2018) Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics*, **19**, 164.

Mika,F. and Hengge,R. (2013) Small Regulatory RNAs in the Control of Motility and Biofilm Formation in E. coli and Salmonella. *IJMS*, **14**, 4560–4579.

Mukherjee,S. *et al.* (2021) Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research*, **49**, D723–D733.

Nagy,L.G. *et al.* (2014) Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun*, **5**, 4471.

Nagy,L.G. *et al.* (2020) Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Research*, **48**, 2209–2219.

Nayfach,S. *et al.* (2021) A genomic catalog of Earth's microbiomes. *Nat Biotechnol*, **39**, 499–509.

O'Brien,P.A. *et al.* Host-Microbe Coevolution: Applying Evidence from Model Systems to Complex Marine Invertebrate Holobionts. *mBio*, **10**, e02241-18.

Opiyo,S.O. *et al.* (2010) Evolution of the Kdo2-lipid A biosynthesis in bacteria. *BMC Evol Biol*, **10**, 362.

Parks,D.H. *et al.* (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, **50**, D785–D794.

Petchey,O.L. and Gaston,K.J. (2002) Functional diversity (FD), species richness and community composition. *Ecol Letters*, **5**, 402–411.

Power,R.A. *et al.* (2017) Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet*, **18**, 41–50.

Prudent,X. *et al.* (2016) Controlling for Phylogenetic Relatedness and Evolutionary Rates Improves the Discovery of Associations Between Species' Phenotypic and Genomic Differences. *Mol Biol Evol*, **33**, 2135–2150.

Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things): *phytools: R package*. *Methods in Ecology and Evolution*, **3**, 217–223.

Rosner,B. (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, **25**, 165–172.

San,J.E. *et al.* (2020) Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front. Microbiol.*, **10**, 3119.

Saund,K. and Snitkin,E.S. (2020) Hogwash: three methods for genome-wide association studies in bacteria. *Microbial Genomics*, **6**.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Sheppard,S.K. *et al.* (2013) Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol Ecol*, **22**, 1051–1064.

Sutcliffe,I.C. (2010) A phylum level perspective on bacterial cell envelope architecture. *Trends in Microbiology*, **18**, 464–470.

Szklarczyk,D. *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, **49**, D605–D612.

Taib,N. *et al.* (2020) Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nat Ecol Evol*, **4**, 1661–1672.

The Electronic Medical Records and Genomics (eMERGE) Consortium *et al.* (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, **46**, 1173–1186.

Timmermans,J. and Van Melderen,L. (2010) Post-transcriptional global regulation by CsrA in bacteria. *Cell. Mol. Life Sci.*, **67**, 2897–2908.

Wei,B.L. *et al.* (2001) Positive regulation of motility and flhDC expression by the RNA-binding protein CsrA of Escherichia coli: Effects of CsrA on flhDC expression. *Molecular Microbiology*, **40**, 245–256.

Weimann,A. *et al.* (2016) From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. *mSystems*, **1**, e00101-16.

Yu,G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, **69**.

Zamani-Dahaj,S.A. *et al.* (2016) Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Mol Biol Evol*, **33**, 1843–1857.

Zhu,Q. *et al.* (2019) Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun*, **10**, 5477.