Title

CORGIAS: identifying correlated gene pairs by considering evolutionary history in a large-scale prokaryotic genome dataset

Authors

Yuki Nishimura[*1], Kimiho Omae[1,2], Kento Tominaga[1,3], Wataru Iwasaki[1]

*To whom correspondence should be addressed.

Affiliations

1. Department of Integrated Biosciences, Graduate School of Frontier Sciences, the University of Tokyo, Chiba, Japan
2. Geobiology and Astrobiology Laboratory, RIKEN Pioneering Research Institute, Saitama, Japan,
3. Department of Natural Environmental Studies, Graduate School of Frontier Sciences, the University of Tokyo, Chiba, Japan

Abstract

Motivation: Phylogenetic profiling is a method for identifying functionally related genes using co- or anti-occurrence of gene distributions. Although some methods developed in previous studies are available for large-scale data, they do not consider evolutionary times or gene gain/loss are not counted properly. Moreover, the results can be altered just by rotating the phylogenetic tree.

Results: We proposed two new phylogenetic methods, Ancestral State Adjustment (ASA) and Simultaneous EVolution test (SEV), which consider the ancestral state of gene presence/absence to evaluate the evolution of ortholog groups (OGs) more properly and guarantee the robustness of the results. In evaluations using three distinct prokaryotic datasets, ASA and SEV showed better performance than methods previously proposed for large-scale data. We compared the functionally related genes detected by each method and found that SEV and its predecessor can identify slowly evolving genes, such as housekeeping genes. In contrast, ASA and its predecessors can detect functionally related genes that tend to be gained or lost in a fixed-order, indicating a strong evolutionary constraint that provides clues for functional prediction. Using matrix multiplication, we showed that SEV is scalable in the latest genome databases.

Availability and implementation: Our methods were implemented in a Python package called CORGIAS (CORelated Genes Identifier by Considering Ancestral State) and are freely available at https://github.com/ynishimuraLv/corgias.git.

Contact: ynishimura{at}edu.k.u-tokyo.ac.jp

1. Introduction

Advancements in sequencing technology have allowed access to a number of genomes of uncultured prokaryotes, including those forming large lineages that occupy the ignorable parts of prokaryotic diversity (Brown *et al.* 2015, Hug *et al.* 2016). These ‘microbial dark matter’ harbor many ortholog groups (OGs) showing no similarity in amino acid sequences to those functionally annotated in previous studies (Nayfach *et al.* 2021, Almeida *et al.* 2021, Baltoumas *et al.* 2023, Río *et al.* 2024). As functionally unknown OGs in microbial dark matter are estimated to constitute up to 50% of their genomes (Castelle *et al.* 2018), it is important to reveal their function to understand their ecology and evolution and to apply genetic resources to industry, agriculture, and medicine.

Phylogenetic profiling is a method to find the co-evolved OGs that are expected to be functionally related from a matrix indicating the presence/absence or copy number of genes in OGs, so-called profile, in many genomes (Pellegrini *et al.* 1999). Once a set of correlated OGs has been identified, they are expected to be functionally related. Thus, the function of unannotated OGs can be inferred from the annotations of other OGs in the correlated gene set (Tabach *et al.* 2013, Dey *et al.* 2015, Moi *et al.* 2020, Dembech *et al.* 2023, Li *et al.* 2023, Fukunaga *et al.* 2024).

In addition to positively correlated OGs, negatively correlated OGs can help understand the function of OGs (Kumagai *et al.* 2018, Morett *et al.* 2003). Recently, we proposed an approach that uses negatively correlated OG pairs for functional prediction as 'contrapositive genetics'. As a proof of concept, we showed that the function of unannotated OGs frequently found in the genome lacking those for replicative helicase loaders work as helicase loaders (Tominaga *et al.* 2024).

One of the challenges in phylogenetic profiling is distinguishing co- and anti-occurring OGs that are functionally related from those just inherited from a common ancestor and have no functional relation (Fig. 1A). Several methods have been developed to overcome this problem. These methods can be classified into two types: model-based and heuristic. A model-based approach explains the OG distribution patterns based on a statistical evolutionary model. For example, in Pagel's method (Pagel 1994), likelihoods of the paired OG distribution in the phylogenetic tree under dependent (co-evolved) and independent models are evaluated. Then, a likelihood ratio test is performed to assess whether the paired OG is co-evolved or not. Although Pagel's method and other model-based approaches are designed to evaluate paired OG (Cohen *et al.* 2013, Li *et al.* 2014), the evolCCM can model interactions including more than three OGs (Liu *et al.* 2022). However, model-based approaches are computationally expensive and difficult to apply to large prokaryotic datasets that can include thousands or tens of thousands of species and OGs these days (Parks *et al.* 2018, Rinke *et al.* 2021, Mendler *et al.* 2019).

In contrast, heuristic methods compare profiles of paired OGs, trying to reduce the phylogenetic effect. For example, in Run-length encoding (RLE, Fig. 1B), a pair of profiles is firstly ordered along genomes in a species tree, and consecutive chunks of profiles sharing the presence/absence state ("runs") are compressed and counted as one (Cokus *et al.* 2007). Although RLE is so fast as to be used to find co-evolved pairs in approximately thirty thousand bacterial genomes (Tremblay *et al.*, 2021), it has two problems: tree rotation makes different results from the same data (Fig. 1B), and genomes in the same run are not necessarily evolutionarily close. These problems can be avoided by clade-wise adjustment (CWA; Fig. 1B), which compresses the

state-sharing profiles only when they form a monophyly (Mering *et al.* 2003). Still, it cannot compress the profiles sharing the states derived from a common ancestor and form a paraphyletic or polyphyletic clade. Ruano-Rubio et al. utilized ancestral state reconstruction to select genomes that are compressed more rationally (Ruano-Rubio *et al.* 2009). However, these 'weighted methods' ignore the evolutionary time the genomes have undergone as they compress the profiles sharing presence/absence state into just one. A corrected number in profiles should be determined considering the evolutionary time and how many genomes are included.

In contrast to weighted methods, Dembech et al. proposed a novel heuristic approach, cotransitions, in which profiles are firstly ordered along a phylogenetic tree as well as RLE but focus on the number of presence/absence changes in rows of profiles (Dembech *et al.* 2023). Then, a statistical test using the numbers of genomes and changes is conducted to detect a correlation in the paired OG (Fig. 1D). Although cotransitions is highly scalable, the results can be changed by tree rotation as well as RLE. In fact, the authors analyzed the same dataset in three different orientations and reported that only 22% of the detected co-evolved pairs overlapped among the analyses (Dembech *et al.* 2023). In addition, gain/loss events cannot be counted correctly (Fig. 1D).

Here, we propose two new heuristic approaches that utilize the reconstructed ancestral states of the profile to evaluate the evolution of OGs properly and obtain robust results against tree rotation. One is the Ancestral State Adjustment (ASA), a variant of the weighted methods that calculates the corrected number of genomes by considering ancestral states and evolutionary time. The second is a Simultaneous EVolution test (SEV) modified from cotransitions. The two proposed methods showed

better or comparable performances to the previous heuristic methods in evaluation using three distinct prokaryotic datasets. SEV also showed high scalability that can be applied to large-scale datasets. Comparison of the co-evolved OG pairs detected by each method revealed that SEV and cotransitions can find pairs that have been gained/lost infrequently and those that are unevenly distributed among the genomes. In contrast, pairs detected only by weighted methods tend to have a specific order of gene gain/loss. These characteristics can be useful for estimating the functions of co-evolved OGs identified by phylogenetic profiling. ASA and SEV are freely available in a Python package called CORelated Genes Identifier by Considering Ancestral State (CORGIAS), with the reimplementation of RLE, CWA, and cotransitions at https://github.com/ynishimuraLv/corgias.git.

2. Materials and Methods

2.1 Ancestral state adjustment (ASA)

We introduced an ASA that corrects the number of paired OG profiles sharing presence/absence states from a common ancestor, considering the evolutionary time. The ancestral states of the profiles at each node in a species tree should be given, but ambiguous states are allowed. The methods inferring the ancestral state are arbitrary; however, in this study, we adopted a maximum-likelihood method using pastML (Ishikawa *et al.* 2019). In the first step of ASA, a tree is divided into subtrees based on the estimated ancestral states of the paired OGs (Fig. 1C, left panel). Each subtree has the same profile of a paired OG from the root to the leaves. This means that all descendants inherited the profile that emerged at the root of the subtree, and each subtree is subject to adjustment, corresponding to 'run' in RLE. The root of the subtree

is identified by tracing back from the leaves (extant species) to their ancestral nodes until the ancestral profile is changed or uncertain. Then, the corrected number of profiles in the respective subtrees is calculated by the following equation:

$$M \times \frac{\sum_{n \in N} b_n}{\sum_{n \in N} b_n \times l_n}$$

where M is the number of genomes in the subtree; N is the set of all leaves and nodes in the subtree; b_n is the branch length between node *n* and its ancestral node; and l_n is the number of leaves derived from node n, respectively. The second term is equivalent to the sum of the branch lengths in the subtree divided by the sum of the branch lengths in the star tree, converted from the subtree. Therefore, the corrected number of genomes decreases as the number of leaves increases, and the branch lengths representing the evolutionary time are longer. The corrected numbers of profiles are summed by the presence/absence state of the paired OG and subjected to Fisher's exact test to determine the significance of their correlation. When the sum of the corrected number of genomes is not an integer, it is rounded up before performing Fisher's exact test.

2.2 Simultaneous EVolution test (SEV)

SEV is very similar to cotransitions (Dembech *et al*. 2023) but focuses on presence/absence changes at the branch, not in rows of profiles, as is counted in cotransitions (Fig. 1D). A concordance change, in which both OGs are simultaneously gained or lost between the parent and its child nodes, is a sign of co-evolution. In contrast, a discordance change in which one OG is gained while the other is lost at the same branch indicates contrapositive evolution, in which two OGs have evolved to avoid each other. SEV is also similar to the "Simultaneous Score" in treeWAS (Collins and Didelot 2018), but the statistical test was inspired by cotransitions. As with ASA,

SEV requires a phylogenetic tree and reconstructed ancestral presence/absence states of the profiles. However, the parallel gain or loss of an OG at the two branches derived from their common parent node is prohibited because of the constraint in the statistical test (see below). The easiest method satisfying this constraint is the maximum parsimony method. A one-tailed Fisher's exact test from a 2 × 2 contingency table below can determine whether an OG pair is correlated.

$\lvert k \rvert$	$t_1 - \lvert k \rvert$
$t_2 - \lvert k \rvert$	$n - t_1 - t_2 - \lvert k \rvert$

where k is the difference in the number of concordance and discordance changes, t_1 and t_2 are the numbers of presence/absence changes in each OG, and n is the number of internal nodes of the tree. Under the constraint mentioned above, n is not the number of branches, because the parallel OG gain/loss at a node is never inferred by the maximum parsimony method. When the ancestral state is uncertain at the nodes, the change is regarded as not having occurred at the branches extending from the nodes. When the obtained p-value is significant and k is positive, the pair OG can be regarded as co-evolved, whereas a negative k indicates contrapositive evolution. k for all the paired OG combinations in the dataset can be calculated by multiplying a matrix comprising the vector representing each OG gain/loss at the branch, thus providing scalability to the SEV. In the matrix, 1 and -1 indicate the gain and loss of an OG, respectively, while 0 indicates no change. When the ancestral state of the node is uncertain, 0 is placed at the corresponding position in the vector. Matrix manipulation is also available in cotransitions; however, the authors' implementation does not adapt to it. Thus, we reimplemented the cotransitions to utilize matrix manipulation for performance

evaluation.

2.3 Preparing the prokaryotic datasets for performance evaluation

The proteome data inferred from representative genomes belonging to three distinct lineages, Order Pseudomonadales (gram-negative), Order Mycobacteriales (gram-positive), and Domain Archaea, were downloaded from GTDB Release 214 (Parks *et al.* 2021). The qualities of the downloaded proteomes were assessed using checkM2 (Chklovski *et al.* 2023). Proteomes with less than 90% completeness or > 10% contamination were excluded from the analysis. The remaining 2,181, 1,627, and 1,922 proteomes of Pseudomonadales, Mycobacteriales, and Archaea, respectively, were used for further analysis.

All proteins in the datasets were submitted to COGclassifier (https://github.com/moshi4/COGclassifier), which assigns proteins to Clusters of Orthologous Genes (COGs (Galperin *et al.* 2020)). The results were used to prepare profiles for each dataset. As it is difficult to identify co-evolved OGs contained in almost all or few genomes by phylogenetic profiling, only COGs found in 1-99% of the genomes were retained, resulting in 3,233, 2,815, and 3,308 COGs in Pseudomonadales, Mycobacteriales, and Archaea, respectively. A phylogenetic tree of archaea with 1,922 species was prepared from the ar53_r214.tree provided by GTDB by subtracting the genomes of the discarded proteomes. Phylogenetic trees of Pseudomonadales and Mycobacteriales were prepared as well, but from the bac120_r214.tree. Ancestral state reconstructions were performed on each COG profile and tree using pastML (Ishikawa *et al.* 2019) by MPPA (the maximum likelihood method recommended by the developers) for ASA and ACCTRAN for SEV, because they always showed the best

area under the precision-recall curve (PRAUC) in all datasets (Supplementary Table S1).

The COG pair scores were derived from the STRING database v12 (Szklarczyk *et al.* 2022). As STRING scores were calculated from gene neighborhood conservation, gene fusion, co-expression, protein interaction experiments, other databases, text mining, and occurrence patterns, the scores were recalculated without occurrence to make the score independent of co-occurrence. COG pairs with a recalculated score larger than 0.9 were regarded as positive; all others were considered negative. The number of COG pairs scored in STRING and the positive pairs in the datasets are summarized in Supplementary Table S2.

2.4 Preparing the simulation datasets for evaluation of runtime performance

Two groups of datasets were simulated to compare the runtimes of each method. One group contained six datasets with 1,000 genomes and varying numbers of OGs. To this end, a phylogenetic tree was generated with the "pbtree" function in phytools (Revell 2012) and used for simulating 1,000, 2,000, 4,000, 8,000, 16,000, and 32,000 OG pairs by evolCCM (Liu *et al.* 2022). The evolCCM uses five parameters to generate pairs of OGs. One of these parameters is the coefficient of interaction between the two OGs. To generate co-evolved OG pairs, the coefficients were sampled from a uniform distribution ranging from 0.2 to 0.75 while setting the coefficient to zero for non-co-evolved OGs. Co-evolved and non-co-evolved pairs were generated at a ratio of 1:99. Two parameters representing the intrinsic rate of each OG were randomly sampled from a uniform distribution from -0.5 to 1. The last two parameters indicated half the difference between the gain and loss rates of each OG and were sampled from a uniform

distribution from -0.5 to 0.3. Profiles with 2,000-64,000 OGs were prepared by concatenating the simulated OG pairs within the datasets. The second group contains six datasets with 4,000 OGs and various numbers of genomes (500, 1,000, 2,000, 4,000, 8,000, and 16,000). Phylogenetic trees and profiles were prepared in the same way as in the first group, but with different numbers of OG pairs and genomes.

The runtimes for calculating the number of genomes (in the naïve method), the corrected number of genomes (in RLE, CWA, and ASA), or k, t1, t2, and n (in cotransitions and SEV) were measured from a phylogenetic tree, a profile, and the inferred ancestral state of OGs in each dataset of the two groups. As the naïve method, cotransitions, and SEV included multiplications of matrices, calculations were performed both with and without a GPU. All runtime measurements were executed on a computer with 56 CPUs of 2.60 GHz, 512 GB of RAM, and a single NVIDIA RTX A6000. Analyses that lasted longer than one day were terminated.

3. Results

3.1 Performances in the Prokaryotic Datasets

Three distinct prokaryotic datasets containing 1,600-2,200 genomes were prepared: Order Pseudomonadales (gram-negative bacteria), Order Mycobacteriales (gram-positive bacteria), and Domain Archaea. All combinations of COG pairs in the datasets were evaluated using naïve, RLE, CWA, ASA, cotransitions, and SEV. As the datasets are unbalanced, with a few positive (co-evolved) pairs compared to negative (non-co-evolved) pairs (Supplementary Table S2), the performances were evaluated based on the PRAUC. We also compared the number of detected positive pairs at a true positive rate (TPR) of 0.7.

The cotransitions and SEV (hereafter, transition methods) showed the best PRAUC in all three datasets (Fig. 2). At TPR = 0.7, SEV detected the most positive pairs in all datasets. SEV identified more than 88% of the pairs detected by all methods (Table 1). In addition, the number of positive pairs detected only by SEV was 3.6~6.5 times larger than those detected only by cotransitions (Supplementary Table S3). Considering the above performance and robustness against tree rotation, SEV can be regarded as a revision of cotransitions.

The PRAUC and the number of detected positive pairs of all weighted methods were far from those of the transition methods (Fig. 2 and Supplementary Table S1). Even if the positive pairs detected by the weighted methods accumulated, each transition method still detected more positive pairs (Table 1). The number of pairs detected by either of the weighted methods, but not by either of the transition methods (i.e., pairs only detected by the weighted methods), was less than 10% of all detected pairs (Supplementary Table S3). In contrast, 20-40% of the pairs detected by either of the transition methods were not found by any weighted method (Supplementary Table S3).

Among the weighted methods, ASA always showed the best PRAUC, although the differences from the second-best method (CWA) were small in the Pseudomonadales and Archaea datasets (Fig. 2). At TPR = 0.7, CWA detected more positive pairs than any other weighted method in the Pseudomonadales dataset (Table 1). However, the superiority of CWA is tentative, and ASA often won at the other TPR and datasets (Fig. 2 and Table 1). In addition, CWA results were unstable among the datasets. For example, the ratio of pairs detected by CWA to those detected by either method in the Mycobacteriales dataset was less than half of that in the Pseudomonadale dataset. Similar to CWA, RLE decreased the number of detected positive pairs and was inferior

even to the naïve method in the Mycobacteriales dataset (Table 1), suggesting that the performance of CWA and RLE depend heavily on the datasets. In contrast, ASA showed stable performance in all three datasets, demonstrating the importance of considering the ancestral state and branch lengths. Considering these results, ASA can be regarded as a revision of the previously proposed weighted methods.

3.2 Transition methods can detect slowly evolving and unevenly distributed OG pairs

To clarify the characteristics and advantages of transition methods over weighted methods, we calculated the coefficients between all positive pairs detected by either the weighted or transition methods using evolCCM. The coefficient of evolCCM determines the transition rate between the presence/absence states of the paired OG; the larger the coefficient, the faster the transition rate from the current state to other states.

Comparing the coefficient distribution between the pairs detected only by the weighted and transition methods, those of the transition methods had peaks at low values (1-1.5) in all datasets (Fig. 3A), suggesting that the transition methods can detect slowly evolving OG pairs. In other words, transition methods can detect more conserved OG pairs than weighted methods. For example, many pairs of COGs encoding ribosomal proteins and translation factors (classified as COG category J) were detected only by the transition methods and had coefficients of approximately 1 in all three datasets (Supplementary Fig. S1). As ribosomes are indispensable for cellular life (Ramulu *et al.* 2014), it is reasonable that their gains and losses rarely occur, resulting in low coefficients. Owing to the high conservation of ribosomal proteins, almost all genomes have COGs detected by the transition methods; the ribosomal protein COGs

consisting of the positive pairs only detected by transition methods are possessed by 96, 84, and 91% of Pseudomonadales, Mycobacteriales, and Archaea genomes in the dataset, respectively. These high possession rates lower the corrected number of profiles, rendering them undetectable by weighted methods.

In addition to the peak around 1, a peak ranging from 2-3 was observed in the coefficient distribution in the Mycobacteriales dataset (Fig. 3A). This peak mainly comprised gene pairs of flagellar COGs belonging to categories N and/or U (Supplementary Fig. S1). We noticed that the differences in the number of presence/absence state changes across the phylogenetic tree (i.e., t_1 and t_2 in SEV) were greater for flagellar COG pairs detected only by the transition methods than the other pairs, except for two that were only detected by ASA (Fig. 3B and Supplementary Figs. S2-S4). The asymmetry in t_1 and t_2 led to an uneven distribution of the nine COGs found only in the pairs detected by the transition methods and their co-evolved partners (Supplementary Fig S5 and Fig. 3C). Therefore, it is difficult to identify them using weighted methods. The reason why the nine genes were independently gained or lost from the other flagellar proteins is unclear, but they can have another function or work with other flagellar proteins that have not yet been defined in the COG. It is also possible that some of the homologous genes of the nine COGs have evolved to dissimilar amino acid sequences because of different evolutionary pressures, preventing them from being classified into the same COGs.

3.3 Weighted methods can identify the gain/loss order of the gene pairs

Although the transition methods found more positive pairs than any other weighted methods, there are pairs detected only by weighted methods and not by transition

methods. Considering that transition methods require simultaneous gains and losses of OGs, they can miss pairs evolving through the intermediate state, where only one OG is present. In contrast, weighted methods do not matter if the OGs have evolved simultaneously. To investigate this, we examined the changes that led to the current states of the pair profiles. Specifically, we divided the phylogenetic tree into subtrees based on the reconstructed ancestral state of the pair profiles, as in the first step of ASA, and then checked the state change between the root of each subtree and its ancestral node. As expected, the transitions from/to the intermediate state were more frequently observed at the root of the subtrees in pairs detected only by weighted pairs than in those detected only by transition methods (Supplementary Fig. S6).

We further investigated which intermediate state is preferred in each pair. If one of the intermediate states appears frequently across the phylogenetic tree and the other does not, there should be a strong evolutionary constraint on the order of OG gain/loss. Intriguingly, most of state changes in the pair detected only by the weighted method in the Pseudomonadales dataset were biased to one direction (Fig. 4). The same trends were observed in the other two datasets, but they are not as clear as in the Pseudomonadales dataset (Fig. 4), partially because a smaller number of pairs were only detected by the weighted methods.

Among the three datasets, the COGs for Type I restriction-modification (RM) systems were commonly found in the lists of co-evolved pairs with fixed-order evolution (Supplementary Tables S4-S6). The genes for methyltransferase (MTase; COG0286) and specificity (S) subunit (COG0732) were obtained before the acquisition of restriction endonucleases (REases; COG4096), or alternatively, REases may have been lost earlier than MTase and the S subunit genes. In prokaryotes, Type I

restriction-modification (RM) system works as an immunity against foreign DNA, such as phages and plasmids (Loenen *et al.* 2014). REase digests foreign, non-methylated DNA, but does not digest host DNA methylated by MTase. The S subunit is a characteristic of Type I RM system and determines the specificity of REase and MTase. As REase without MTase and the S subunit may have caused host DNA digestion, it is reasonable to avoid the state with REase but without MTase and the S subunit during evolution.

Similarly, a state with antitoxin alone is permitted for the toxin-antitoxin (TA) system in Pseudomonadales and Mycobacteriales. The toxin is neutralized by forming a complex with the antitoxin when a bacterial cell is not infected by the phage. Once a bacterial cell is infected, the toxin is activated, leading to cell death to prevent phage production (Sberro *et al.* 2013). Therefore, the state with a toxin alone is harmful to the cell, and it is reasonable that the TA system has evolved through the other state.

The examples mentioned above are only part of the co-evolved pairs with a fixed intermediate state, and their biological interpretation remains to be resolved in future studies.

3.4 Runtime performance in simulated datasets

The runtimes for each method were evaluated using two groups of simulated datasets. One group contained 1,000 genomes with varying numbers of genes ranging from 2,000 to 64,000. As CWA and ASA have to parse all nodes in the tree for all combinations of gene pairs, they require a longer time than the other methods, and their runtime increases with the square of the number of OGs (Fig. 5A). The RLE runtime also increased in the same order, but the time required for the one pair comparison was much

shorter (Fig. 5A). Compared to these three methods, naïve, cotransitions, and SEV performed very quickly, as they can leverage matrix multiplication for all-vs-all gene comparisons (Fig. 5A). Calculation with GPU further accelerates these methods; a dataset with 64,000 genes (~20 billion pair comparisons) was calculated within 20 min.

The other groups of simulated data included varying numbers of genomes, from 500 to 1,6000 and 4,000 genes. Again, ASA was the most time-consuming method, indicating that it is difficult to apply to a large-scale dataset (Fig. 5B). The runtime increased less linearly with the number of genomes, except for ASA and naïve without GPU.

AS the runtime of each method consistently increased with the number of genomes and genes, we inferred the runtime for multiple regression to predict the runtime of arbitrary numbers of genomes and genes. The regressions fitted the data well, with $R^2 > 0.95$ (Supplementary Table S7). Transition methods with GPU were predicted to conduct comparisons of all gene pairs in the datasets as large as the latest OrthoDB (v12) and GTDB (Release 220) within an hour (Supplementary Table S7), suggesting their high scalability.

4. Discussion

In this study, we developed two novel methods for phylogenetic profiling: ASA and SEV. Considering the ancestral state and gene gain/loss events were handled more properly than in the previous method, their results are robust to tree rotation. In addition, ASA showed better recall than similar methods in the validation datasets. SEV also detected more co-evolved pairs at a practical true positive rate than its predecessor, cotransitions, although the difference was small. Furthermore, SEV and cotransitions

implemented with matrix manipulation showed high scalability, which enables us to investigate the latest large-scale databases. ASA, SEV, and the other phylogenetic profiling methods compared in this study are freely available in a Python package CORGIAS (CORelated Genes Identifier by Considering Ancestral State; https://github.com/ynishimuraLv/corgias.git).

Although the weighted methods, including ASA, were outperformed by the transition methods in the number of detected positive pairs, they are able to capture the co-evolved pairs that evolved through the intermediate states. Unexpectedly, in the Pseudomonadales dataset, most of the pairs detected only by the weighted methods have evolved through only one of the two possible intermediate states. The reason this trend is remarkable in the Pseudomonadales dataset alone is unclear, but the largest number of genomes among the three datasets can affect it because the trend gets weaker as the number of genomes decreases. The accuracy and granularity of ortholog clustering (e.g., COG classification is regarded as coarser than the KEGG ortholog) and other possible factors should be investigated in future studies. To further investigate the characteristics of OG pairs detected only by weighted methods, ASA must be accelerated, as it is the slowest method in this study. From the perspective of evolutionary biology, the interpretation of the order of gene gain/loss is another interesting topic (Konno and Iwasaki 2023).

Both ASA and SEV depend on how the ancestral states of the profile are reconstructed. The ancestral state reconstruction methods that we adopted were quick but simple and ignored the intrinsic complexity of evolution. For example, the gain/loss rate heterogeneity among genes, lineages, (Fukunaga and Iwasaki 2021, 2022b; Yamanouchi *et al.* 2024) and horizontal gene transfers (Dagan and Martin 2007) can

mislead inferences under the simple evolutionary model. Although ancestral state reconstruction methods considering complex evolutionary backgrounds are computationally expensive and contain many uncertainties (Williams *et al.* 2024), it is worth investigating whether they can improve ASA and SEV.

The phylogenetic profiling methods developed in this study can be applied to any binary dataset and are not limited to gene absence/presence data. They can be useful in finding genotype-phenotype associations when gene profiles and binary trait data are used, as in a trait-based approach (Collins and Didelot 2018; Yang and Jiang 2023). Improvements to allow for the handling of categorical and quantitative data are a possible direction for future studies. As the methods in this study can detect only paired relationships, detecting the relationship among more than three genes (Liu *et al.*, 2022) and/or considering the global metric (Kim and Price 2011, Fukunaga and Iwasaki 2022a) is another direction. Machine learning approach may be a promising approach in these directions (Konno and Iwasaki 2023, Beavan *et al.* 2024).

In conclusion, CORGIAS is an accurate and scalable phylogenetic profiling method. Our method is powerful for narrowing down OGs with a target function from a number of unannotated OGs encoded in microbial dark matter. Applying the candidate OGs identified by CORGIAS to other bioinformatics methods, such as those using synteny (Shmakov *et al.* 2019, Botas *et al.* 2022), co-expression (Dam *et al.* 2017, Chang *et al.* 2023), and protein structure (Jumper *et al.* 2021, Abramson *et al.* 2024), further provides clues for functional prediction.

Figures and Tables

Figure 1. Schematic diagram of phylogenetic profiling. (A) Phylogenetic profiles of a non-co-evolved (top) and a co-evolved gene pair (below). Whether a pair of genes is co-evolved or not is determined by the statistical significance of the co-occurrence. In the naïve phylogenetic profiling methods, both co-evolved and non-co-evolved pairs are tested with the same contingency table (bottom). Thus, they are statistically indistinguishable. (B) Application examples of RLE and CWA to the non-co-evolved pair in (A). Profiles surrounded by black and orange dotted squares are compressed to one profile in RLE and CWA, respectively. In the middle, the leftmost two branches and their profiles of the top figure are rotated. The resultant contingency tables are shown at the bottom. (C) The schematic diagram of ASA. A phylogenetic tree is divided into subtrees based on the reconstructed ancestral states of a paired-gene profile (left panel). The tree branches are colored by the result of ancestral state reconstruction. A number of genomes in a subtree with more than two leaves (surrounded by a dotted circle) are submitted to correction, according to evolutionary time (right panel). The corrected numbers of genomes are shown below the dotted circles, where b_n represents the branch length of node n. In the method proposed by Ruano-Rubio et al., each subtree is counted as one. (D) The schematic diagram of cotransitions and SEV. The curved arrows indicate presence/absence changes counted in cotransitions, while triangles indicate those counted in SEV. Concordance and discordance changes are shown in red and blue, respectively, while gray represents changes only in either of the genes. The significance of the correlation of a pair of genes is tested by using the contingency table in the top right panel. The notation in the contingency table is shown in the middle right panel (Although Dembech et al. designated n as the number of genomes, the number of

genomes minus one is correct because transitions can be placed between genomes). The number of concordances, disconcordances changes, t_1, t_2 in cotransitions, and SEV are shown in the bottom right panel. Note that the concordance gene gain occurred only once (red triangle) but was counted as two transitions in cotransitions.

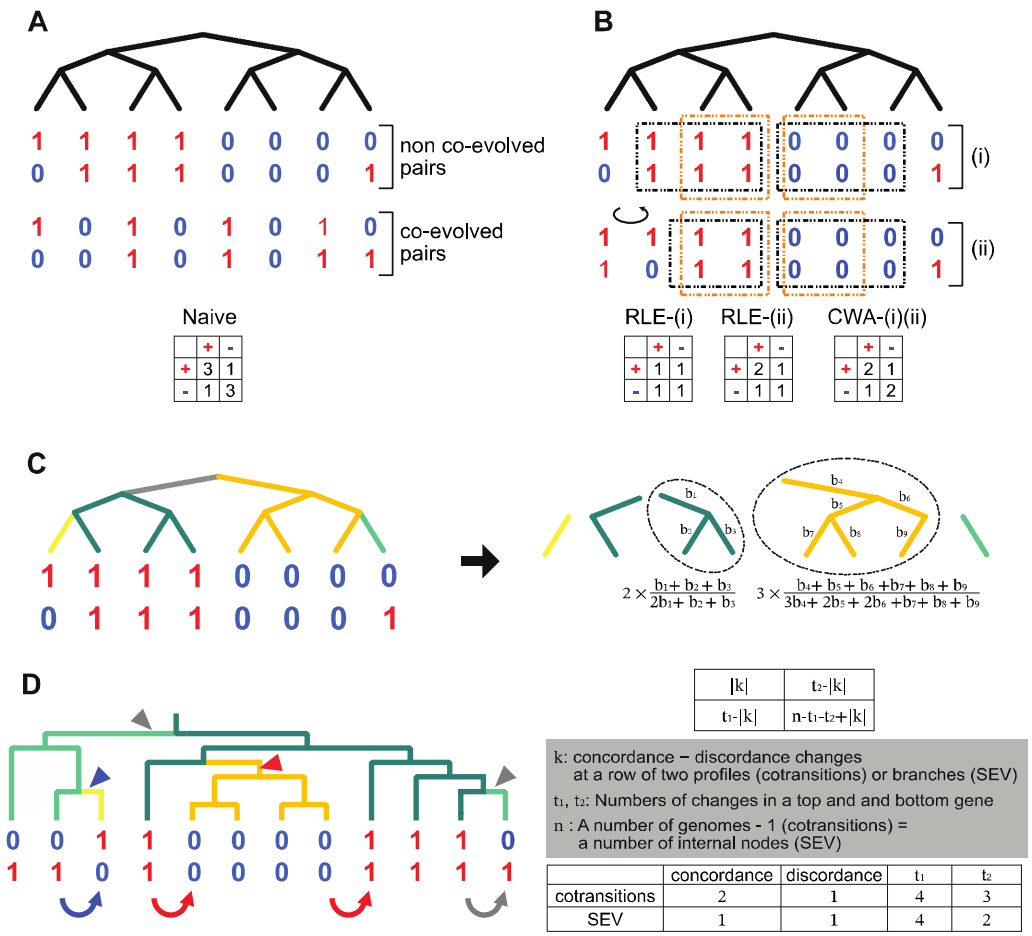


Figure 2. Performance comparison of phylogenetic profiling methods using PR curves. PRAUCs are shown in the parentheses.

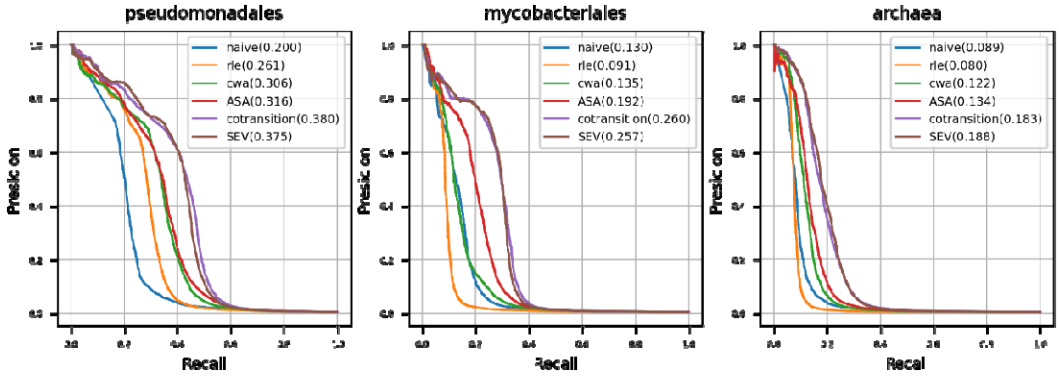


Fig 3. Characteristic of transition methods. (A) The evolCCM coefficient distribution of the co-evolved pairs only detected by transition methods (blue) and weighted methods (orange). (B) The scatter plot of the coefficient (x-axis) and the differences between t_1 and t_2 in co-evolved pairs of COGs belonging to category N (y-axis). t_1 and t_2 are the number of presence/absence state changes in each COG of the pairs across the phylogenetic tree. Blue dots indicate the co-evolved pairs only detected by transition methods, while green ones indicate pairs detected by both transition and weighted methods. Note that no pairs were detected only by the weighted methods, except for two with large differences between t_1 and t_2. The plots for the other categories in all datasets are shown in Supplementary Figs. S3-S5. (C) Phylogenetic distribution of co-evolved genes belonging to COG category N in Mycobacteriales. The top panel shows the species tree of Mycobacteriales used in this study. The middle and bottom heatmap shows the presence/absence of the genes only detected by transition methods and those detected by both weighted and transition methods.

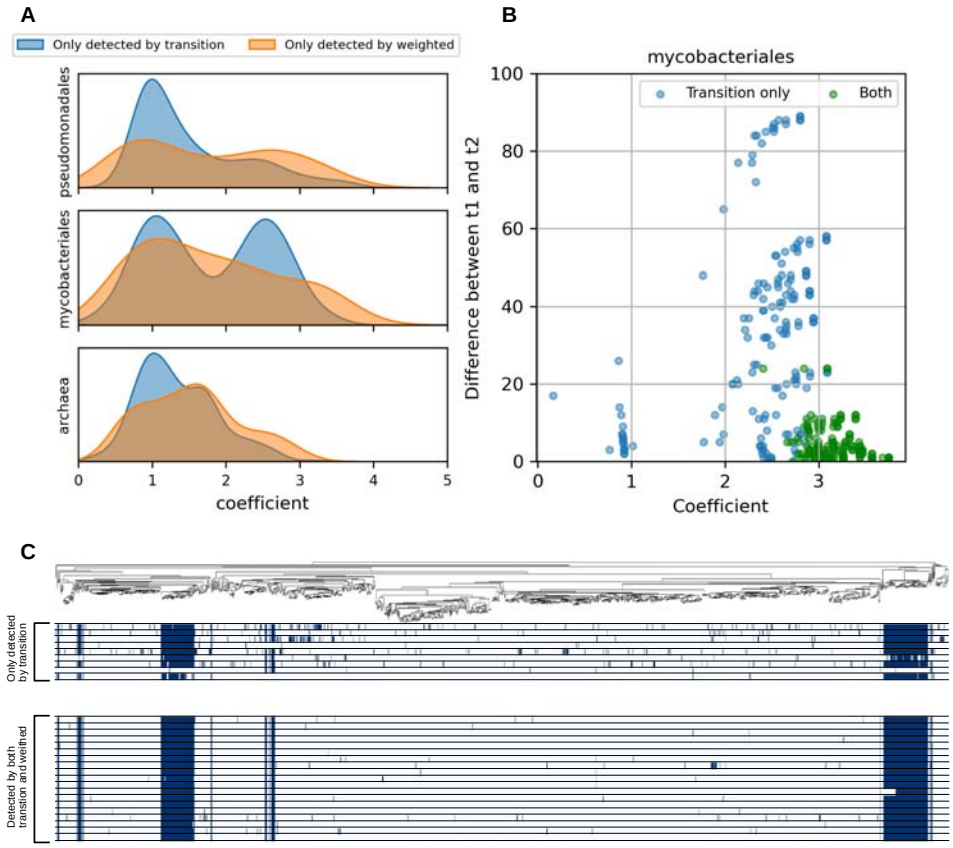

A
Only detected by transition
Only detected by weighted
pseudomonadales
mycobacteriales
archaea
0
1
2
3
4
5
coefficient
B
mycobacteriales
Transition only
Both
Difference between t1 and t2
100
80
60
40
20
0
0
1
2
3
Coefficient
C
Only detected by transition
Detected by both transition and weighted

Fig 4. Characteristic of the weighted methods. Comparison of the numbers of changes from/to the intermediate state of the co-evolved gene pairs only detected by the transition method (blue) and weighted methods (orange). The number of changes was counted from the ancestral state reconstruction result of the co-evolved gene pairs. The numbers of the changes from/to the intermediate state are indicated on the x-axis, and the biases to the changes to one of two intermediate states are indicated on the y-axis. The transition bias was calculated by dividing the larger number of changes from/to the intermediate state of pair genes by their sum.

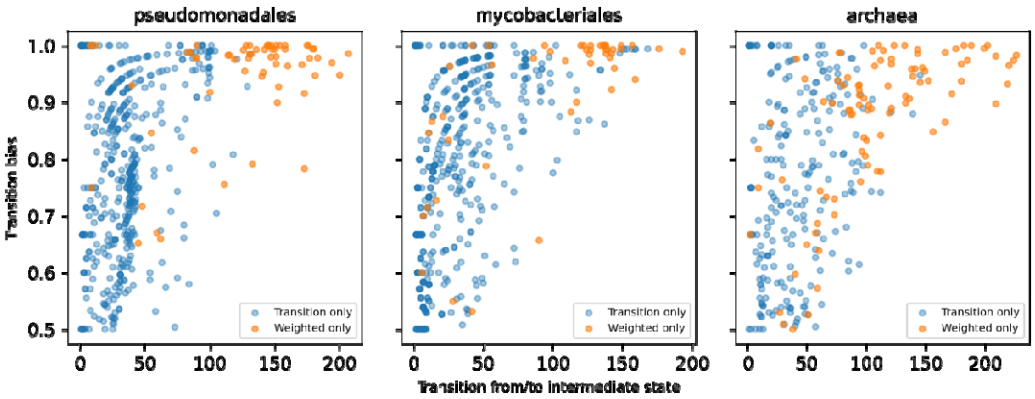


Fig 5. Runtimes comparison among phylogenetic profiling methods with (A) 1,000 genomes and a variety number of genes and (B) 4,000 genes and a variety number of genomes. The runtimes of naïve, RLE, CWA, ASA, cotransition, and SEV are indicated by the black, yellow, green, blue, sky blue, and red circles, respectively, and the lines indicate the regression curve of the runtimes. The runtimes with the GPU are connected by the dotted lines.

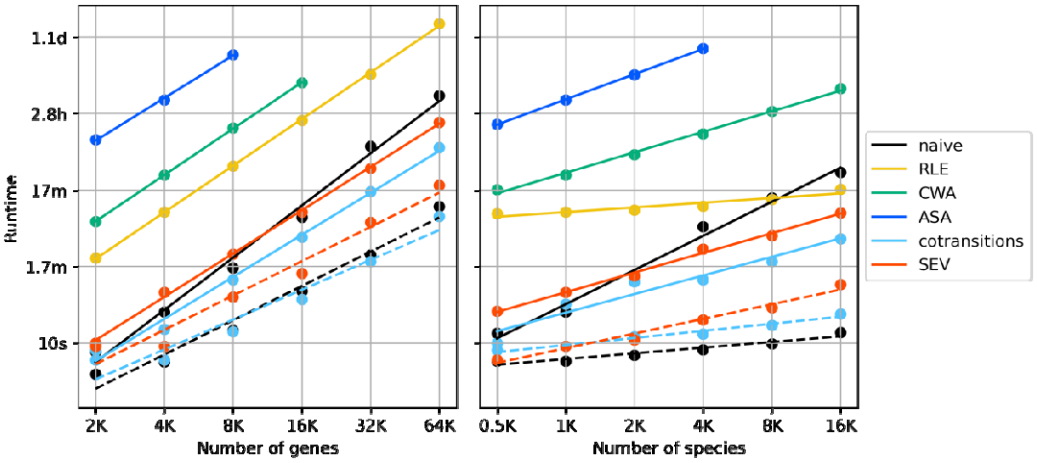

1.1d
2.8h
17m
1.7m
10s
Runtime
2K
4K
8K
16K
32K
64K
Number of genes
0.5K
1K
2K
4K
8K
16K
Number of species
naive
RLE
CWA
ASA
cotransitions
SEV

Table 1. The number of detected positive pairs at TPR = 0.70 by phylogenetic methods investigated in this study. The totals for the weighted (naïve, RLE, CWA, and ASA), transition (cotransitions and SEV), and all methods are also shown. The coverage of the total of all methods is indicated in parentheses.

	Pseudomonadales	Mycobacteriales	Archaea
naïve	990 (45.4%)	321 (29.1%)	366 (38.6%)
RLE	1349 (46.1%)	259 (23.5%)	350 (36.9%)
CWA	1636 (75.0%)	365 (33.2%)	527 (55.6%)
ASA	1497 (68.6%)	607 (55.1%)	605 (63.8%)
weighted	1710 (78.6%)	671 (60.5%)	724 (75.5%)
cotransitions	1966 (90.1%)	994 (90.3%)	805 (84.9%)
SEV	2073 (95.0%)	1034 (93.9%)	838 (88.4%)
transition	2111 (97.1%)	1058 (95.3%)	867 (90.4%)
Total	2173	1110	959

Data availability

The implementation of CORGIAS is freely available at https://github.com/ynishimuraLv/corgias.

Acknowledgment

The authors thank Dr. Wataru Iwasaki's lab members for the fruitful discussion.

Fundings

This work was supported by the Japan Society for the Promotion of Science (KAKENHI, Grant Number 22H04925) and Japan Science and Technology Agency (CREST JPMJCR19S2 and GteX Program JPMJGX23B2).

References

Abramson J *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493–500.

Almeida A *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol* 2021;**39**:105–114.

Baltoumas FA *et al.* NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res* 2023;**52**:D502–D512.

Beavan AJS *et al.* Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. *Proc Natl Acad Sci* 2024;**121**:e2304934120.

Botas J *et al.* GeCoViz: genomic context visualisation of prokaryotic genes from a functional and evolutionary perspective. *Nucleic Acids Res* 2022:**50**:W352–W357.

Brown CT *et al.* Brown et al-2015-Nature - supplement.pdf. *Nature* 2015;**523**:208–211.

Castelle CJ *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* 2018;**16**:629–645.

Chang L-Y *et al.* Gene set correlation enrichment analysis for interpreting and annotating gene expression profiles. *Nucleic Acids Res* 2023;**52**:e17–e17.

Chklovski A *et al.* CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–1212.

Cohen O *et al.* CoPAP: Coevolution of Presence–Absence Patterns. *Nucleic Acids Res* 2013;**41**:W232–W237.

Cokus S *et al.* An improved method for identifying functionally linked proteins using phylogenetic profiles. *Bmc Bioinformatics* 2007;**8**, S7–S7.

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

Dagan T and Martin W. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci* 2007;**104**:870–875.

Dam S van *et al.* Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* 2017;**19**:575–592.

Dembech E. *et al.* Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc Natl Acad Sci* 2023;**120**: e2218329120.

Dey G. *et al.* Systematic Discovery of Human Gene Function and Principles of Modular Organization through Phylogenetic Profiling. *Cell Rep* 2015;**10**:993–1006.

Fukunaga T. *et al.* Phylogenetic Profiling Analysis of the Phycobilisome Revealed a Novel State-Transition Regulator Gene in Synechocystis sp. PCC 6803. *Plant Cell Physiol* 2024;**65**:1450–1460.

Fukunaga T and Iwasaki W. Inverse Potts model improves accuracy of phylogenetic profiling. *Bioinformatics* 2022a;**38**:1794–1800.

Fukunaga T and Iwasaki W. Mirage 2.0: fast and memory-efficient reconstruction of gene-content evolution considering heterogeneous evolutionary patterns among gene families. *Bioinformatics* 2022b;**38**:4039–4041.

Fukunaga T and Iwasaki W. Mirage: estimation of ancestral gene-copy numbers by considering different evolutionary patterns among gene families. *Bioinform Adv* 2021;**1**:vbab014.

Galperin MY *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* 2020;**49**:D274–D281.

Hug LA *et al.* A new view of the tree of life. *Nat Microbiol* 2016;**1**:16048.

Ishikawa SA *et al.* A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Mol Biol Evol* 2019;**36**:2069–2085.

Jumper J *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–589.

Kim P-J and Price ND. Genetic Co-Occurrence Network across Sequenced Microbes. *PLoS Comput Biol* 2021;**7**:e1002340.

Konno N and Iwasaki W. Machine learning enables prediction of metabolic system evolution in bacteria. *Sci Adv* 2023;**9**:eadc9130.

Kumagai Y *et al.* Solar-panel and parasol strategies shape the proteorhodopsin distribution pattern in marine Flavobacteriia. *Isme J* 2018;**12**:1329–1343.

Li Y *et al.* Expansion of Biological Pathways Based on Evolutionary Inference. *Cell* 2014;**158**:213–225.

Li Y *et al.* PPNet: Identifying Functional Association Networks by Phylogenetic Profiling of Prokaryotic Genomes. *Microbiol Spectr* 2023;**11**:e03871–22.

Liu C *et al.* The Community Coevolution Model with Application to the Study of Evolutionary Relationships between Genes Based on Phylogenetic Profiles. *Syst Biol* 2022;**72**:559–574.

Loenen WAM *et al.* Type I restriction enzymes and their relatives. *Nucleic Acids Res* 2014;**42**:20–44.

Mendler K *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* 2019;**47**:4442–4448.

Mering C von *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**:258–261.

Moi D *et al.* Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLoS Comput Biol* 2020;**16**:e1007553.

Morett E *et al.* Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat Biotechnol* 2003;**21**:790–795.

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

Pagel M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc R Soc Lond Ser B: Biol Sci* 1994;**255**:37–45.

Parks DH *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004.

Parks DH *et al.* (2021) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2021;**50**:D785–D794.

Pellegrini M *et al.* Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci* 1999;**96**:4285–4288.

Ramulu HG *et al.* Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol Phylogenetics Evol* 2014;**75**:103–117.

Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol Evol* 2012;**3**:217–223.

Rinke C *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat Microbiol* 2021;**6**:946–959.

Río ÁR del *et al.* Functional and evolutionary significance of unknown genes from uncultivated taxa. *Nature* 2024;**626**:377–384.

Ruano-Rubio V *et al.* Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinform* 2009;**10**:383–383.

Sberro H *et al.* Discovery of Functional Toxin/Antitoxin Systems in Bacteria by Shotgun Cloning. *Mol Cell* 2013;**50**:136–148.

Shmakov SA *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nat Protoc* 2019;**14**:3013–3031.

Szklarczyk D *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2022;**51**:D638–D646.

Tabach Y *et al*. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* 2013;**493**:694–698.

Tominaga K *et al*. Frequent nonhomologous replacement of replicative helicase loaders by viruses in Vibrionaceae. *Proc Natl Acad Sci United States Am* 2024;**121**:e2317954121.

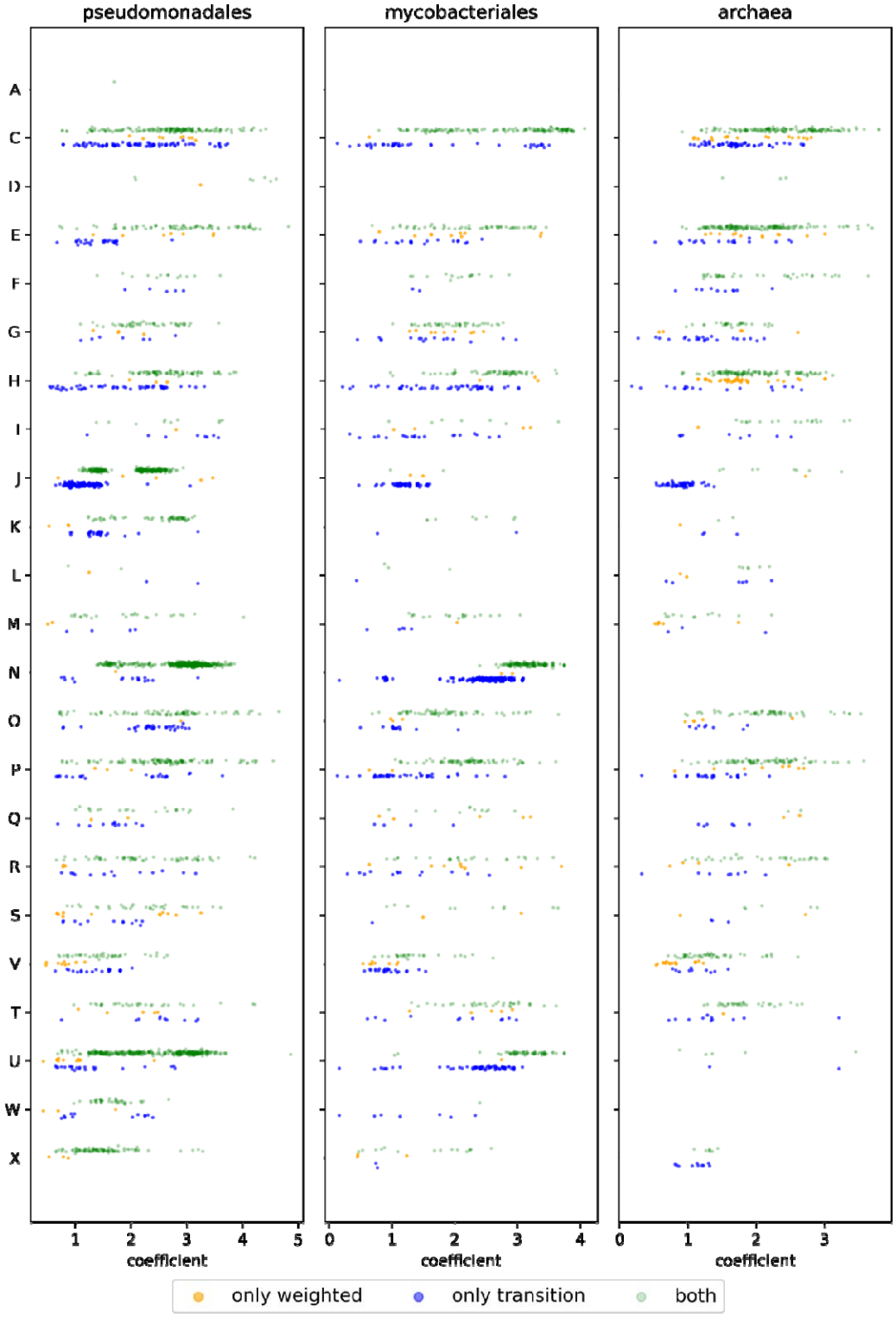
Tremblay BJM *et al.* PhyloCorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. *Bioinform Oxf Engl* 2021;**37**:17–22.

Williams TA *et al*. Phylogenetic reconciliation: making the most of genomes to understand microbial ecology and evolution. *ISME J* 2024;**18**:wrae129.

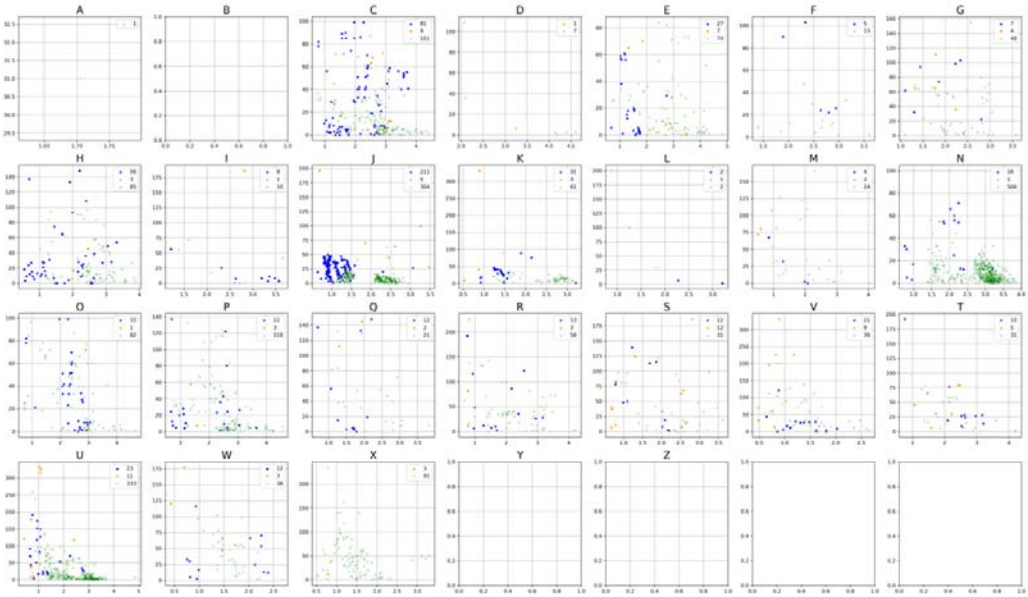
Yang Y and Jiang X. Evolink: a phylogenetic approach for rapid identification of genotype–phenotype associations in large-scale microbial multispecies data. *Bioinformatics* 2023;**39**:btad215.

Supplementary Figures and Tables

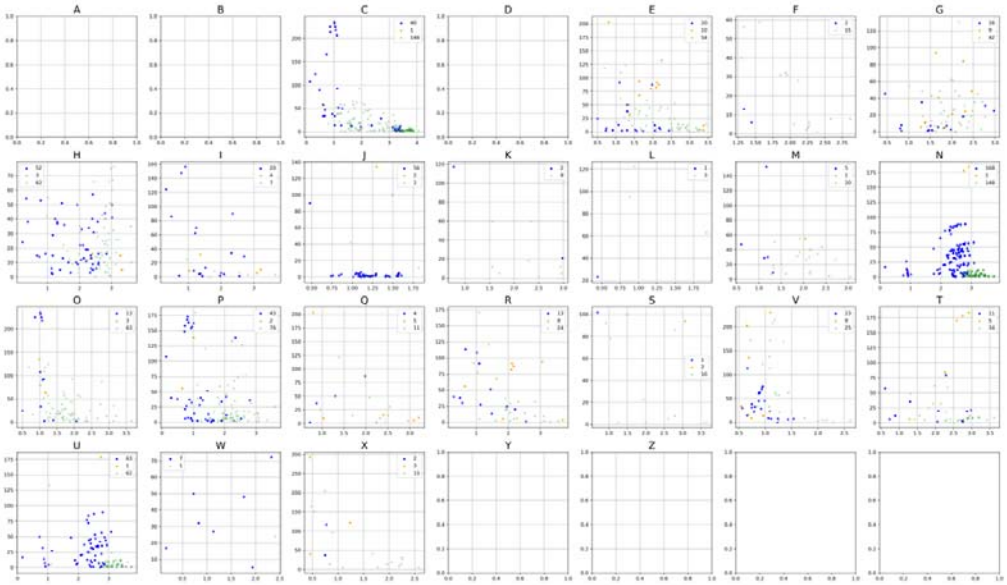
Supplementary Figure S1. The distribution of coefficients of the detected positive pairs only detected by transition methods (green), weighted methods (orange), and both (blue). The pairs are classified by the categories of COGs that are contained. As a COG has more than one COG category and a pair contains COGs belonging to different COG categories, a pair can appear in more than one category. Alphabets in the y-axis represent COG categories as follows. A: RNA processing and modification, C: Energy production and conversion, D: Cell cycle control, cell division, chromosome partitioning, E: Amino acid transport and metabolism, F: Nucleotide transport and metabolism, G: Carbohydrate transport and metabolism, H: Coenzyme transport and metabolism, I: Lipid transport and metabolism, J: Translation, ribosomal structure, and biogenesis, K: Transcription, L: Replication, recombination, and repair, M: Cell wall/membrane/envelope/ biogenesis, N: Cell motility, O: Posttranslational modification, protein turnover, and chaperones, P: Inorganic ion transport and metabolism, Q: Secondary metabolites biogenesis, transport, and metabolism, R: General function prediction only, S: Function unknown, T: Signal transduction mechanism, U: Intercellular trafficking, secretion, and vesicular transport, V: Defense mechanisms, W: Extracellular structure, X: Mobilome, prophages, and transposons.


pseudomonadales
mycobacteriales
archaea
A
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
V
T
U
W
X
1 2 3 4 5
0 1 2 3 4
0 1 2 3
coefficient
coefficient
coefficient
only weighted
only transition
both

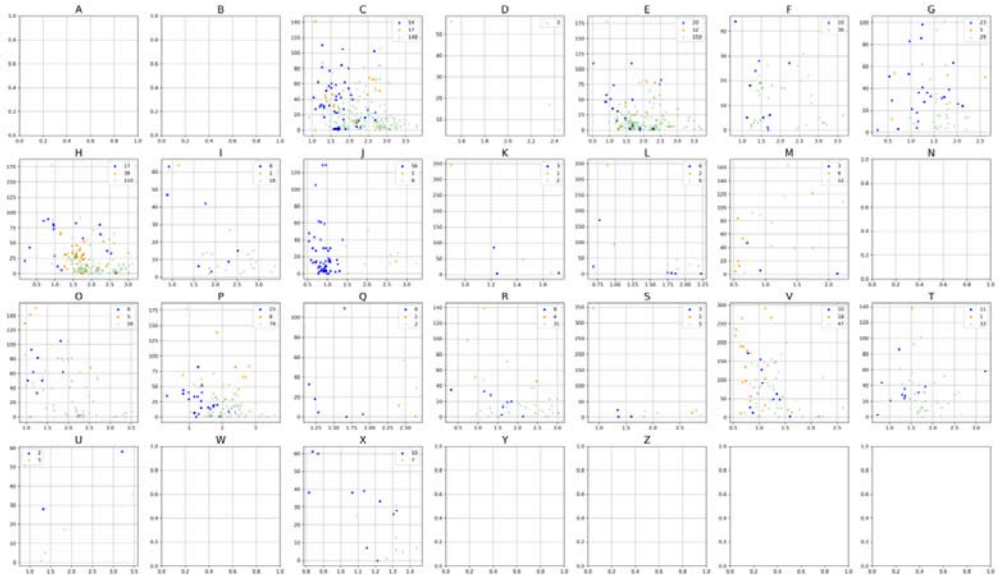
Supplementary Figure S2. The scatter plot of the coefficient (x-axis) and the differences between t_1 and t_2 in co-evolved pairs in the Pseudomonadales dataset. The details are the same as Fig. 3B. The COG categories represented by alphabets are described in the legend of Supplementary Figure S1.



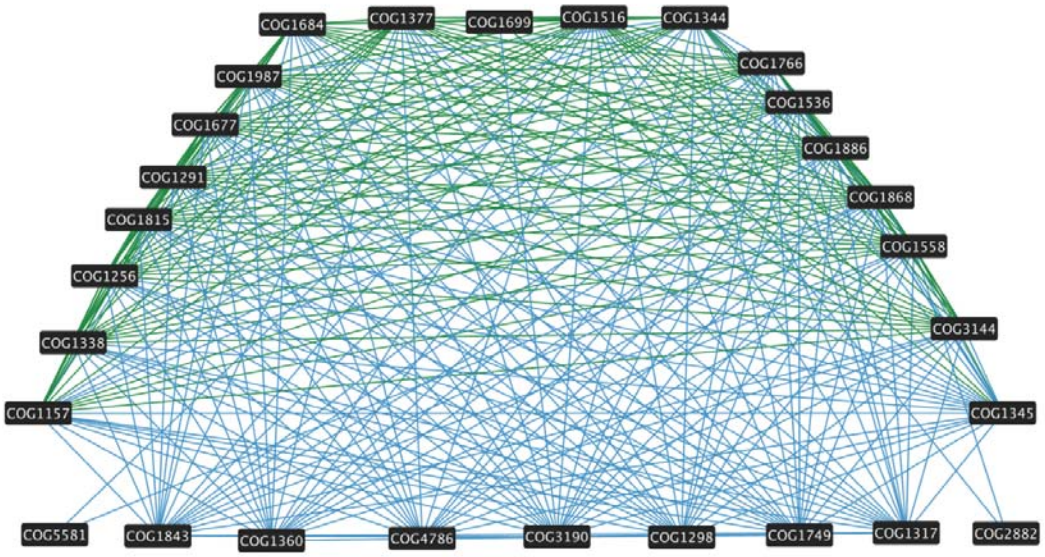
Supplementary Figure S3. The scatter plot of the coefficient (x-axis) and the differences between t_1 and t_2 in co-evolved pairs in the Mycobacteriales dataset. The details are the same as Fig. 3B. The COG categories represented by alphabets are described in the legend of Supplementary Figure S1.



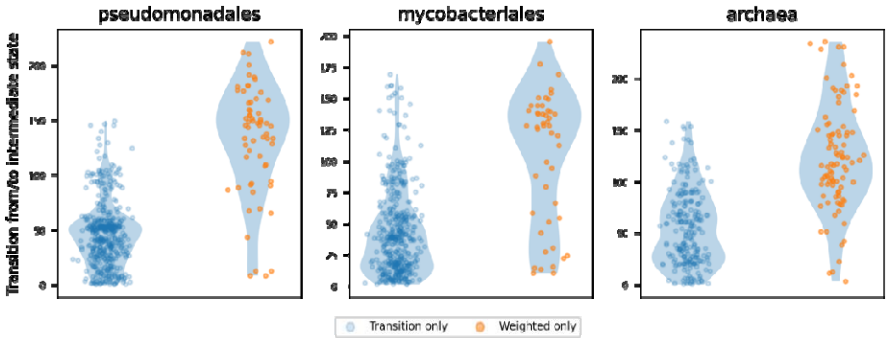
Supplementary Figure S4. The scatter plot of the coefficient (x-axis) and the differences between t_1 and t_2 in co-evolved pairs in the Archaea dataset. The details are the same as Fig. 3B. The COG categories represented by alphabets are described in the legend of Supplementary Figure S1.



Supplementary Fig S5. A network of co-evolved COG pairs detected by phylogenetic profiling in the Mycobacteriales dataset. The COGs belonging to category N (cell motility) are connected by the edges when they are detected as co-evolved pairs. The pairs only detected by transition methods are colored in blue and otherwise in green. Only the maximum connected component is shown. The nine COGs at the bottom only have blue edges, indicating that they were not in the pairs detected by weighted methods.



Supplementary Figure S6. Comparison of the number of transitions from/to intermediate state between the pairs only detected by transition methods (blue) and weighted methods (orange).



Supplementary Table S1. PRAUC of ASA and SEV using various methods for ancestral state reconstruction. “MPPA” is the maximum likelihood method, and the other are the maximum parsimony methods. The best AUC in the dataset and the phylogenetic profiling methods (ASA or SEV) are shown in bold.

Phylogenetic profiling	Dataset	Ancestral state reconstruction	PRACU
ASA	Pseudomonadasles	MPPA	**0.316**
		DOWNPASS	0.257
		ACTRAN	0.296
		DELTRAN	0.290
	Mycobacteriales	MPPA	**0.192**
		DOWNPASS	0.098
		ACTRAN	0.158
		DELTRAN	0.146
	Archaea	MPPA	**0.134**
		DOWNPASS	0.132
		ACTRAN	0.133
		DELTRAN	0.118
SEV	Pseudomonadasles	DOWNPASS	0.371
		ACTRAN	**0.375**
		DELTRAN	0.370
	Mycobacteriales	DOWNPASS	0.251
		ACTRAN	**0.257**
		DELTRAN	0.251
	Archaea	DOWNPASS	0.187
		ACTRAN	**0.188**
		DELTRAN	0.185

Supplementary Table S2. The summary of the datasets used in the evaluation.

Dataset	Genomes	COGs	Scored pairs	Positive pairs
Pseudomonadales	2,182	3,234	1,069,619	5,870
Mycobacteriales	1,677	2,815	846,124	4,040
Archaea	1,921	3,308	1,297,337	6,003

Supplementary Table S3. The number of positive pairs only identified by each phylogenetic profiling method. The numbers of positive pairs detected by either weighted methods but not the transition methods, and vice versa, are also shown. The ratios of the number of pairs detected by each method are shown in parentheses.

Dataset	Pseudomonadales	Mycobacteriales	Archaea
naïve	5 (0.23%)	10 (0.90%)	17 (0.18%)
RLE	6 (0.28%)	9 (0.81%)	0 (0%)
CWA	11 (0.51%)	5 (0.45%)	10 (1.0%)
ASA	4 (0.18%)	12 (1.1%)	24 (1.5%)
weighted methods	62 (2.9%)	52 (4.7%)	92 (9.6%)
cotransitions	19 (0.87%)	13 (1.1%)	15 (1.6%)
SEV	144 (6.6%)	57 (5.1%)	54 (5.6%)
transition methods	463 (21.3%)	439 (39.5%)	235 (24.5%)

Supplementary Table S4. COG pairs only detected by weighted methods in the Pseudomonadales dataset at TPR = 0.7.

Supplementary Table S5. COG pairs only detected by weighted methods in the Mycobacteriales dataset at TPR = 0.7.

Supplementary Table S6. COG pairs only detected by weighted methods in the Archaea dataset at TPR = 0.7.

Supplementary Table S7. Runtime regression of the phylogenetic profiling methods used in this study. Runtimes (t) were regressed on the following: $t = \alpha_1 \log(\text{number of species}) + \alpha_2 \log(\text{number of genes}) + \beta$. The predicted runtimes of the latest GTDB (Release 220, 107,000 species and 6,000 OGs) and OrthoDB (v12, 28,000 species, and 61,000 OGs) are also shown. The numbers of OGs were obtained from previous studies (Tremblay *et al.* 2021, Dembech *et al.* 2023).

	α_1	α_2	β	R^2	OrthoDB	GTDB
naïve	1.52	2.23	-25.6	0.990	31 d	54 h
naïve w/ GPU	0.219	1.52	-12.2	0.979	5.8 m	10 m
RLE	0.205	2.02	-12.0	0.999	19 h	51 h
CWA	0.905	2.00	-15.4	0.997	63 d	22 d
ASA	1.09	1.85	-13.2	0.999	2.9 y	195 d
cotransitions	0.847	1.79	-17.6	0.982	10.7 h	3.8 h
cotransitions w/ GPU	0.336	1.27	-10.7	0.951	8.3 m	8.4 m
SEV	0.884	1.851	-17.7	0.998	27 h	9.0 h
SEV w/ GPU	0.521	1.61	-14.5	0.954	48 m	38 m