Heinrich Heine
University
Düsseldorf

Inference and Evolutionary Impact of Horizontal Gene Transfer in Bacteria: Quality Over Quantity on the Road to New Environments

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Swastik Mishra

geboren in
Bhubaneswar, India

Düsseldorf, August 2025

aus der Arbeitsgruppe für Computational Cell Biology
am Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Martin J. Lercher

2. Prof. William F. Martin

Tag der mündlichen Prüfung: 06.11.2025

"Haar ke jeetne wale ko baazigar kehte hain"
(A gambler is the one who wins by losing)
- Baazigar (1993)

Abstract

Regarded as a cornerstone of prokaryotic innovation, the process of horizontal gene transfer (HGT) drives genetic exchange between organisms and, in doing so, is widely believed to significantly enhance the adaptability and evolutionary success of prokaryotes. HGT is associated with bacterial colonization of diverse environments, from antibiotic-resistant pathogens in clinical settings to extremophiles in harsh ecological niches. However, large-scale empirical studies of HGT's role in environmental adaptation remain limited, constrained by significant methodological challenges in accurately detecting and validating transfer events. This cumulative thesis addresses both methodological issues and the evolutionary significance of horizontal gene transfer.

Manuscript 1 systematically evaluates the performance of diverse HGT inference methods. By testing whether HGT events inferred from real genomic data follow expected patterns, such as neighboring genes being co-transferred on the same DNA fragment, we identify best practices for recovering biologically realistic HGT events. Contrary to prevailing assumptions, our results demonstrate that implicit phylogenetic approaches using gene presence-absence matrices outperform gene tree-based and parametric sequence composition-based methods, yielding fewer false positives and thus more meaningful biological inferences.

Based on these methodological insights, Manuscript 2 applies the best-performing HGT inference technique to probe the relationship between HGT and environmental transitions at scale. By integrating gene and ecosystem presence-absence data across a wide array of bacterial lineages, we reveal that bacteria colonizing new ecosystems generally possess smaller genomes and undergo lower rates of HGT – with both observations contradicting *a priori* expectations. The reduced HGT frequency is entirely explained by genome size, with no independent effect of environmental change, challenging prevailing ideas that genomic flexibility via increased HGT rates is essential.

In Manuscript 3, we investigate the evolutionary fate of transferred genes, uncovering a two-phase process: rapid loss of most long-distance HGTs shortly after transfer, followed by persistence in the remaining cases. Remarkably, a minority of bacterial genomes harbor the majority of inter-phylum transfers. The retention of these genes is gene function-dependent, illustrating the swift and selective nature of bacterial genome streamlining following HGT.

The final chapter, Chapter 5, synthesizes the main findings and explores their broader implications for prokaryotic genome evolution. It discusses how the methodological and conceptual advances developed in this thesis can inform future research, and proposes new directions for advancing our understanding of horizontal gene transfer and genome dynamics in prokaryotes.

In sum, this dissertation demonstrates that the evolutionary success of bacteria hinges on the discerning retention of functionally relevant genes and the efficient purging of superfluous ones – embodying a "quality over quantity" principle at the genomic level. This same philosophy of "quality over quantity" applies to the inference of horizontal gene transfer: methods grounded in limited but high-quality information yield more precise and meaningful evolutionary insights than approaches that merely maximize the information used for inferences. By developing and applying robust strategies for both HGT detection and interpretation, this work provides essential conceptual and methodological advances for the broader study of prokaryotic genome evolution.

Acknowledgements

They say a PhD is a solitary pursuit, but anyone who has watched me scramble through code, paperwork, existential crises, and caffeine knows that that's not true. Allow me to indulge in the least formal, least scientific, and most heartfelt acknowledgements section I can manage.

To my supervisor, Martin: Thank you for enduring my rants, repeated confusion, and poorly timed attempts at creativity. Of all the things I have learned from you – how to do science, how to persevere, and how to be patient – perhaps most importantly, I have discovered that it is possible to scale both scientific mountains and actual bouldering walls with equal grace.

I began my work full of optimism but with little knowledge of the field. Thank you to Mayo, Tin, and Peter for shepherding me through the hazardous early stages, listening to my ideas, and helping me find the right direction with the right resources. Thanks to Chilperic and Sajjad for pushing me (and, at times, pulling me) to keep myself physically fit. The office would not have been the banter-filled haven it was without you two, along with Jeremie, Lea, and Antonio. Who knew bad puns and occasional cakes would become so important? To everyone in the Computational Cell Biology group, thank you for making it feel comfortable. Some in the group have stayed, while others drifted away, but each left a lasting impression and possibly some memes, on my brain.

As I peek out beyond our group, I realize the entire university and the next, had my back. When Covid struck and I was stranded in India, dazed, confused, and blaming the cosmos for my luck, Sigrun at JUNO personally took on dragons for me. By this, I mean she conquered bureaucratic horrors, delivered my visa, and probably increased my life expectancy. Without Sigrun and the JUNO team, neither this thesis nor my sanity would have survived the onslaught of paperwork over the years. A special shoutout to Wolf Frommer and Oliver Ebenhöh's groups: your kindness and encouragement made it worthwhile. To everyone in the CRC 1310: your seminars and discussions (scientific, philosophical, or just plain odd) kept the inspiration flowing.

Finally, I need to leave the academic world: it takes a *Dorf* to raise a doctoral student. Thanks to Gaby and Vishak, for making Dusseldorf be less of a Drizzledorf and more the cozy D-dorf I grew to love. To the book club, the improv troupe, and my TTRPG crew – especially you, Ghi – I am glad I met you and took a leap of faith. All thanks to my chaotic crew for keeping me sane.

My deepest gratitude to my not-so-near-yet-ever-dear ones: Danish, Vishrut, Reema, Upasana, and of course my family; you kept my Signal/Telegram notifications alive, even when in-person hugs were sparse. Thank you for the support. You deserved a thesis in your honor, but for now, this section will have to do.

To round it off: it's only fair to recognize the unsung heroes of open knowledge: those internet outlaws and open-source sages. So thank you to Sci-Hub and LibGen, which I never use, obviously! Here's to the many online anarchists who have written blogs, open source code, and tutorials, without any requirement for payments. My heartfelt gratitude to Lo-Fi hip-hop playlist curators and video game composers; your beats provide the official soundtrack to this thesis. So long, and thanks for all the hits.

Contents

Chapter 1

Introduction

The dynamic nature of prokaryotic evolution

Prokaryotic evolution is shaped by a variety of mechanisms, with HGT believed to play a significant role in driving adaptation to new environments. HGT allows for the exchange of genetic material between organisms, bypassing vertical inheritance. Recent estimates suggest that HGT is highly prevalent in prokaryotes, with some studies inferring at least one HGT event in nearly 75% of gene families across diverse prokaryotic taxa (Kloesges et al., 2011). While HGT is a common phenomenon in prokaryotes, it is relatively rare in eukaryotes, where gene duplication is the primary mechanism for generating genetic diversity (Doolittle, 1999; Martin, 2017); Conversely, duplications are rarer than HGT in prokaryotes (Treangen and Rocha, 2011; Tria and Martin, 2021).

HGT occurs through several mechanisms, including transformation (uptake of free DNA), transduction (phage-mediated transfer), and conjugation (cell-to-cell transfer), and can involve anything from small DNA segments to multi-gene cassettes (Arnold et al., 2022; Dilthey and Lercher, 2015; Pang and Lercher, 2017, 2019). This prevalent and mechanistically diverse process has far-reaching consequences for bacterial genomes: it can obscure phylogenetic relationships among genomes (Garud et al., 2019) and enable rapid adaptation to new niches even with just a few recombination events (Frazão et al., 2019).

In recent years, genomic studies have increasingly focused on how HGT and selection interact to shape bacterial evolution. Researchers now seek not only to detect HGT, but also to identify which transferred genes are adaptive in particular environments (Bradley et al., 2018; Dmitrijeva et al., 2024; Smillie et al., 2011) or to assess whether pangenomes are shaped primarily by selection (Andreani et al., 2017; Vos and Eyre-Walker, 2017). Uncovering such signatures of adaptive HGT is key to understanding the evolutionary pressures shaping prokaryotic populations and their genomic diversity.

In this thesis, I address both the methodological challenges of HGT inference and the conceptual questions surrounding the role of HGT in prokaryotic evolution. The findings reveal the importance of quality over quantity in the contribution of HGT to bacterial adaptation. Methodologically, simpler methods using less – but more reliable – data appear to be more reliable than more involved approaches (Chapter 2). Environmental transitions are not characterized by a flood of new genes acquired via HGT, suggesting that a small number of additional genes is sufficient for adaptation (Chapter 3). More generally, only a functionally biased minority of the genes acquired via HGT are retained in the recipient genomes (Chapter 4).

Methods and challenges in inferring horizontal gene transfer

Central to the study of HGT in is its identification from genomic data. Outside of controlled experimental settings, HGT can be detected through methods that fall into two main categories: parametric and phylogenetic approaches (Ravenhall et al., 2015). Parametric methods search for genes with unusual sequence composition, such as GC content or codon usage, compared to the host genome. These approaches can miss events between closely related taxa and are less effective for older transfers, as sequences gradually adapt to the host background.

Phylogenetic methods use evolutionary relationships to infer transfer events. Explicit approaches compare gene trees with species trees to detect discordant patterns, while implicit approaches analyze patterns of gene presence or absence across a species phylogeny. Explicit methods can infer both donor and recipient but are sensitive to gene tree reconstruction errors. Implicit methods are faster and avoid gene tree errors but typically identify only the recipient lineage. All phylogenetic methods assume an underlying species tree, which may not be well-defined in the presence of frequent transfer events. For a comprehensive review of HGT inference methods, see Ravenhall et al. (2015).

Given the importance of HGT in prokaryotic evolution, robust methods for its inference are essential. Inference methods are, however, difficult to validate. HGT signals are often obscured by subsequent gene loss and sequence amelioration, making reliable validation possible only for very recent events, typically observed in laboratory experiments. Such experiments, however, cannot capture the breadth of HGT seen over longer timescales, including transfers between distant lineages (Cordero and Hogeweg, 2009; Sheinman et al., 2021). While data resampling can measure robustness, it does not address the suitability of model assumptions (Felsenstein, 2003).

As a result, most HGT inference methods rely on simulated data for validation. Simulations offer a controlled setting and a known ground truth, allowing systematic assessment of method performance and sensitivity. Nevertheless, simulation outcomes are shaped by their underlying assumptions, which may not fully reflect biological realities and can bias results (Feyerabend, 1993). For example, simulations might treat genes as indivisible units, overlook intragenic changes or population dynamics, or ignore unsampled taxa. Furthermore, inference methods may perform well on simulations designed with similar assumptions, limiting their ability to generalize to real data (Kapust et al., 2018). Few methods have been tested and compared using shared simulation frameworks or empirical data.

In Chapter 2, we benchmark widely used HGT inference methods by evaluating how well they match expected patterns in empirical data. We focus on the observation that genes co-transferred in a single HGT event enter the recipient genome together on the same DNA fragment and tend to remain adjacent immediately after transfer (Dilthey and Lercher, 2015; Pang and Lercher, 2017). Although genomic rearrangements may eventually separate these genes, many co-transferred pairs should still be physically close in present-day genomes. If a method infers a higher proportion of neighboring gene pairs as having been acquired simultaneously (for example, on the same branch of the species tree), it is less likely to identify false-positive HGT events. This expectation underlies our comparative approach.

While it might seem that methods using explicit gene tree reconciliation methods utilize more information and therefore would outperform others, our results show that implicit phylogenetic methods actually recover more biologically meaningful HGT events and exhibit lower false positive rates than either explicit phylogenetic or parametric approaches. The reason gene tree-reconciliation methods perform poorly is likely due to their sensitivity to gene tree reconstruction errors (Than et al., 2008, 2007), which can lead to false positive HGT inferences. Inference methods therefore need to balance the use of available information with robustness to noise. This also means that for better HGT inferences, one important approach can be to improve gene tree reconstruction methods, which I touch upon in Chapter 5.

Conceptual questions regarding acquisition or loss of genetic material in bacterial evolution

Bacteria often adapt to new habitats such as the gut, skin, fungal partners, and marine plastics (Caruso, 2020; Lieberman, 2022; Richter et al., 2024; Zheng et al., 2020). These transitions impose new selection pressures and population bottlenecks that reshape genomes. Large genomes can ease

transitions by offering a wider functional "toolbox" (Maslov et al., 2009; Szappanos et al., 2016) and more homologous landing pads for incoming DNA via HGT (Taylor et al., 2024). Small genomes tend to reflect niche specialisation and reduced flexibility (Serra Moncadas et al., 2024). Meanwhile, HGT supplies ready-made genes or operons that accelerate adaptation (Arnold et al., 2022; Pál et al., 2005; Pang and Lercher, 2019), and environmental change may raise both the benefit and the opportunity for transfer (Dmitrijeva et al., 2024; Engelstädter and Moradigaravand, 2014). Laboratory studies confirm bursts of HGT under stress, mainly for antibiotic-resistance genes in limited bacterial species (Acar Kirit et al., 2020; Dadeh Amirfard et al., 2024; Goh et al., 2024; Woods et al., 2020), yet large-scale studies are non-existent, leaving genome-wide patterns poorly resolved. Furthermore, studies on sets of closely related species indicate that the retention of newly acquired genes is often transient; empirical studies indicate that the majority of horizontally transferred genes are rapidly lost following their acquisition (Hao and Golding, 2006; Lerat et al., 2005; Puigbò et al., 2014). However, it is unclear if this pattern of rapid initial loss holds generally across bacteria.

These observations raise fundamental questions about the interplay between genome architecture, environmental transitions, and evolutionary adaptation. How do population bottlenecks and shifting selection pressures alter the evolutionary trajectories of bacterial lineages during habitat transitions? Which specific genomic characteristics enable some bacteria to successfully colonise new environments while others fail? To what extent does HGT contribute to adaptation during these ecological shifts, and how do the competing dynamics of gene acquisition and gene loss together sculpt prokaryotic genome evolution over time?

We address these issues in Chapter 3 and Chapter 4. Chapter 3 investigates the relationship between HGT and environmental transitions at scale, applying the most reliable inference methods identified in Chapter 2. Contrary to expectations, bacteria colonising new habitats typically possess smaller genomes and exhibit lower HGT rates. Once genome size is controlled for, environmental transitions show no additional effect on HGT frequency. These findings suggest that genomic streamlining, rather than expanded gene repertoires, facilitates successful ecological transitions, prompting a fundamental reassessment of HGT's role in bacterial adaptation. In a complementary analysis, Chapter 4 examines the evolutionary fate of horizontally acquired genes, revealing a biphasic pattern of gene retention and loss. Inter-phylum HGT events are rare overall but concentrate in a subset of bacterial lineages, where they undergo rapid initial purging followed by much slower secondary loss. This selective retention process highlights how bacterial genomes maintain only those horizontally acquired functions that provide lasting adaptive value, further supporting the importance of genome streamlining in bacterial evolution.

Finally, in Chapter 5, I summarise the main findings of this dissertation and discuss their implications for our understanding of prokaryotic genome evolution. I highlight the issues of methodological validation and HGT inference and the questions raised by the counter-intuitive findings presented in this thesis as directions for future research.

Chapter 2

Horizontal Gene Transfer Inference: Gene presence-absence outperforms gene trees

This section consists of an article that has been published in Molecular Biology and Evolution, Volume 42, Issue 7, July 2025, msaf166, https://doi.org/10.1093/molbev/msaf166.
The article contains references to supplementary materials available online, which are included at the end of this chapter.
I designed the study, performed the analyses, and drafted the manuscript.

Molecular Biology and Evolution, 2025, **42**, 1–9
https://doi.org/10.1093/molbev/msaf166
Advance access publication 28 July 2025
Discoveries

MBE OXFORD

Horizontal Gene Transfer Inference: Gene Presence–Absence Outperforms Gene Trees

Swastik Mishra, Martin J. Lercher *

Institute for Computer Science and Department of Biology, Heinrich Heine University, Düsseldorf, Germany

***Corresponding author**: E-mail: martin.lercher@hhu.de.
Associate editor: Keith Crandall

Abstract

Horizontal gene transfer is a fundamental driver of prokaryotic evolution, facilitating the acquisition of novel traits and adaptation to new environments. Despite its importance, methods for inferring horizontal gene transfer are rarely systematically compared, leaving a gap in our understanding of their relative strengths and limitations. Validating horizontal gene transfer inference methods is challenging due to the absence of a genomic fossil record that could confirm historical transfer events. Without an empirical gold standard, new inference methods are typically validated using simulated data; however, these simulations may not accurately capture biological complexity and often embed the same assumptions used in the inference methods themselves. Here, we leverage the tendency of horizontal gene transfer events to involve multiple neighboring genes to assess the accuracy of diverse horizontal gene transfer inference methods. We show that methods analyzing gene family presence/absence patterns across species trees consistently outperform approaches based on gene tree-species tree reconciliation. Our findings challenge the prevailing assumption that explicit phylogenetic reconciliation methods are superior to simpler implicit methods. By providing a comprehensive benchmark, we offer practical recommendations for selecting appropriate methods and indicate avenues for future methodological advancements.

Keywords: horizontal gene transfer, HGT inference, gene presence–absence profiles, reconciliation models, implicit phylogenetic methods, explicit phylogenetic methods

Introduction

Horizontal gene transfer (HGT) is recognized as a major driving force in the evolution of prokaryotes. (Arnold et al. 2021) However, the inference of HGT events is challenging, since HGT is commonly followed by other evolutionary processes such as sequence amelioration and gene loss. In the best-case scenario, HGT events can be traced between closely related genomes in lab evolution experiments to validate HGT inferences. This limits any rigorous validation of HGT inference methods to only extremely recent HGT events. However, on longer timescales, transfers occur frequently even between distantly related organisms (Cordero and Hogeweg 2009; Sheinman et al. 2021), something that laboratory evolution experiments cannot replicate. Although one can test the robustness of the HGT inference methods, for example by data resampling, that does not validate the model choices made during the inference process (Felsenstein 2003).

As a result of these challenges, HGT inference methods are typically validated using simulated data. Simulated data offers key advantages over empirical data for method validation: it provides a known ground truth and allows direct assessment of robustness and sensitivity by varying parameter values or violating specific assumptions. However, simulations are based on a set of assumptions about the evolutionary process, potentially biasing their results in comparison to real data—even if these assumptions are accepted principles or "natural interpretations" of the community (Feyerabend 1993). Examples of potentially biasing assumptions are: considering genes as evolutionary units, ignoring intragenic rearrangements; neglecting population-level processes such as drift; assuming that sites evolve independently from each other and from structural constraints; and ignoring extinct or unsampled species. Using simulations to compare different HGT inference methods is complicated by the differences in assumptions made by each method: a particular method will perform well in the simulated world designed for it. The vast majority of methods, when published, have not been tested on a common reference set of simulated data, nor have they been compared to one another. Moreover, methods not tested in the real world may fail to capture real-world complexities (Kapust et al. 2018). It is thus valuable to validate and compare methods based on their performance on the same empirical data set.

In the following, we develop a comparative analysis of HGT inference methods using genomics data. Our goal is to provide an approximate ranking of HGT inference methods, delineating which methods infer results that are meaningful in a given biological context, as well as to have a nuanced view of how these methods compare to each other.

HGT inference methods can be broadly classified into two categories: *parametric* (or sequence composition) methods and *phylogenetic* (or tree-based) methods (Ravenhall et al. 2015).

Parametric methods operate under the assumption that the *sequence composition* of the transferred gene differs between the donor and recipient organisms. Sequence features of a genome, such as GC content, codon usage, and oligonucleotide frequencies, are often analyzed. However, they are limited in their effectiveness when the donor and recipient are closely

Received: February 11, 2025. **Revised:** May 28, 2025. **Accepted:** June 23, 2025

related and thus have similar sequence features, or when dealing with ancient HGT events where enough time has passed for the acquired nucleotide sequence to evolve to the sequence features of the host genome (sequence amelioration).

Phylogenetic methods, in contrast, analyze sets of sequences by leveraging phylogenetic trees. These methods can be categorized into explicit and implicit approaches.

Explicit phylogenetic methods compare the branching patterns (topologies) of gene trees, which show the history of individual genes, with species trees, which show how species are related based on vertical inheritance. When the topology of a gene tree deviates from that of the species tree, such discrepancies can not only indicate the occurrence of an HGT event, but also help infer the direction of transfer – identifying both the likely donor and recipient species. For example, imagine a gene tree where a gene from species A is very similar to genes from a distant clade that contains species B. This mismatch suggests that species A likely received the gene from the lineage of species B through HGT. In the simplest case of a single HGT event, the method reconciles the gene tree with the species tree by introducing a horizontal branch connecting the donor and recipient lineages. However, in bacterial evolution we often need to consider multiple events of duplication, transfer, or loss (DTL) to fully explain the observed differences between gene and species trees, and there may be several plausible reconciliation scenarios for any given pair of trees. The frequency with which a particular DTL event appears across these scenarios reflects the confidence of the method in that inferred event. For example, a transfer event that appears in 90% of the reconciliations indicates a high confidence in that event. Among explicit methods, tools such as RANGER-DTL (Bansal et al. 2018) and ALE (Szöllosi et al. 2012; Szöllősi et al. 2013) use a maximum likelihood (ML) framework to estimate the rates of duplication, transfer, and loss events that best explain the observed gene and species trees. These methods optimize DTL rate parameters to maximize the overall likelihood, and then sample reconciliation scenarios probabilistically according to these rates. In contrast, methods like AnGST (David and Alm 2011) operate within a maximum parsimony (MP) framework, inferring DTL events by minimizing a cost function for each type of event.

Implicit methods avoid such a direct comparison and hence do not require a gene tree. Instead, they aim to infer the gene gain and loss events that explain the observed gene distribution across genomes in a genome tree. BLAST-hit methods such as DLIGHT (Dessimoz et al. 2008) find very similar gene sequences in distantly related organisms. Methods based on phyletic patterns such as GLOOME (Cohen et al. 2010) and Count (Csűös 2010) use a statistical framework instead.

Since implicit phylogenetic methods do not require a gene tree, they are faster to execute and are not compromised by the potential inaccuracies of gene tree reconstruction. However, unlike explicit methods, they can only infer the *gain* or *loss* of a gene in a given recipient branch of the species tree, without providing information about the donor of the gene. Additionally, one can save time by not estimating a species tree beforehand and instead let for example GLOOME infer the species tree topology based on the phyletic pattern itself, even if this may result in a less accurate tree topology. Note that all phylogenetic inference methods assume the existence (and correctness) of a species tree describing purely vertical inheritance; however, such a species tree may not always be well defined, especially in cases of highly reticulate evolution.

For a more comprehensive review of HGT inference methods, we refer the reader to Ravenhall et al. (2015).

Although phylogenetic methods offer valuable insights, they are not without limitations. Unrecognized paralogy resulting from duplication followed by loss can be misinterpreted as HGT. Furthermore, tree reconstruction errors, particularly in gene trees, can lead to false HGT inferences (Than et al. 2007). However, with increases in computing power, studies on HGT have increasingly used phylogenetic methods, not least because they are not compromised by sequence amelioration and can thus infer older HGT events than parametric methods.

Parametric methods were developed first and have been benchmarked elsewhere (Becq et al. 2010), with oligonucleotide signature-based methods found to outperform alternative approaches. We thus use an oligonucleotide signature-based method, Wn (Tsirigos and Rigoutsos 2005), as a representative parametric method for our benchmark, and focus primarily on the comparison of phylogenetic methods. Explicit phylogenetic methods included in our study are ALE, RANGER-DTL, and AnGST. For RANGER-DTL, in addition to the full version, we also include a heuristic "fast" version that samples a smaller space of optimal reconciliations. Implicit methods included in our study are GLOOME and Count, two phyletic pattern-based methods. GLOOME and Count were each run with both ML and MP settings. With GLOOME, each of those were also run with and without an input species tree. Unless stated otherwise, we refer to GLOOME (ML or MP) as versions with an input species tree.

Our selection of HGT inference methods for the benchmark aims to provide a comprehensive overview of different state-of-the-art methods. We preferentially include methods that are widely used by the community and have user-friendly, publicly available implementations. There are many other methods that have not been included in this study. These include methods that have no publicly available implementations of the algorithm (e.g. DLIGHT Dessimoz et al. 2008); methods that cannot handle modern large datasets with multiple genes of the same gene family in a taxon (especially subtree-prune-and-regraft based methods, e.g. RIATA-HGT (Nakhleh et al. 2005), SPRIT (Hill et al. 2010), EEEP (Beiko and Hamilton 2006)); methods that focus on the presence of HGT in the evolution of a gene but not on the inference of exact transfer events (e.g. JPrime (Sjöstrand et al. 2014); and additional methods discussed in Poptsova 2009).

Here, we benchmark the selected methods by evaluating how well they reproduce expected patterns in empirical data. Specifically, we expect that genes transferred together in a single HGT event often arrive on the same DNA fragment and initially remain neighbors in the recipient genome. Over time, genomic rearrangements may separate them, but a substantial fraction should still be found close together in extant genomes. Thus, if we find a higher percentage of neighboring gene pairs among the genes inferred by a given method to have been acquired around the same time (i.e. on the same branch of the species tree), this method likely captures true HGT events more reliably than alternative methods. This principle forms the core of our comparative analysis.

Although intuitively, the usage of explicit gene tree information may seem to be an advantage, our findings indicate that implicit phylogenetic methods infer more biologically meaningful HGT events than explicit phylogenetic methods or parametric methods, having a lower false positive rate.

Results

For our benchmark of HGT inference methods, we used the dataset of orthologous gene families in Gammaproteobacteria in the EggNOG database v6 (Hernández-Plaza et al. 2023). EggNOG terms these gene families non-supervised orthologous groups, or *NOGs*. We focused on the 359 Gammaproteobacterial taxa in EggNOG that are featured in the tree of life provided by ASTRAL (Zhu et al. 2019); we used the subtree with these taxa as the input genome tree for the phylogenetic methods. We then sampled a representative set of 1286 gene families (NOGs) that occur in at least 10 of these taxa. We obtained the corresponding gene trees from EggNOG, rooting them using Minimum Ancestor Deviation (Tria et al. 2017), which has been shown to be more accurate than other rooting methods (Wade et al. 2020). We used these gene families to infer HGT events using each of the tested methods.

For each method, one can restrict the set of HGT inferences to the most reliable ones, either by modifying an input parameter (the gain/loss penalty ratio for Count (MP) and GLOOME (MP)) or by removing inferences labeled by the inference method as less reliable (e.g. based on the frequency of gain events across reconciliation models by ALE). In this way, each method's number of HGT inferences can be varied, with lower numbers corresponding to higher "stringency", i.e. more reliable inferences. Because each method quantifies confidence differently, these thresholds are not directly comparable across methods. To enable fair comparisons, we instead use the numbers of inferred co-acquired gene pairs as common reference points at which we compare the methods.

For the explicit phylogenetic methods, we defined stringency as the mean number of transfers inferred per reconciliation model. In GLOOME (ML) and Count (ML), we used the minimum posterior probability of gene gain events as the stringency. For GLOOME (MP) and Count (MP), we use the gain/loss penalty ratio as the stringency parameter. In the parametric method Wn, we fixed the oligonucleotide window size to 8 and defined stringency as the ratio of absolute deviation to the median absolute deviation (MAD; see Methods for details).

Implicit and Explicit Methods Infer Very Different Sets of HGT Events

To assess similarities and differences in the sets of HGT events inferred by each method, we first clustered the methods based on the similarity between pairs of these sets. As a similarity measure, we used the Overlap Coefficient, which is the ratio of the size of the intersection to the size of the smaller set in the pair. Because for each method the set of infered HGT events can expand or contract massively by changing the stringency, we measured similarities as the maximum possible Overlap Coefficient across all possible pairs of thresholds.

As expected, the explicit phylogenetic methods (ALE, RANGER, AnGST) cluster together, as do the implicit phylogenetic methods (GLOOME, Count) (Fig. 1). The biggest differences in terms of inferred HGT events are seen between explicit and implicit phylogenetic methods. The parametric method Wn clusters with the implicit phylogenetic methods. Note that the maximum Overlap Coefficient between all the implicit methods is 1 (or very close to 1 when involving GLOOME ML and MP without a species tree), i.e. for each pair of methods, there are thresholds such that one set of HGT inferences is a subset of the other. The same is true for ALE and AnGST.

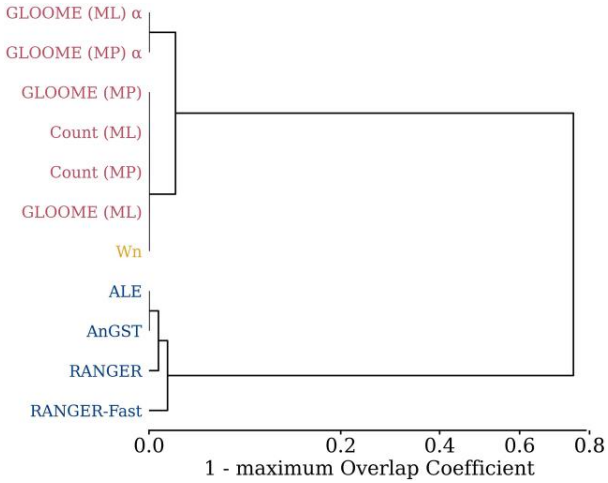


Fig. 1. Clustering analysis of the inferred sets of HGT events. Shown here is a UPGMA dendrogram of the tested methods based on their maximum Overlap Coefficient. Overlap Coefficient is the ratio of the size of the intersection to the size of the smaller set (of inferred HGT events) in the pair of methods. Maximum Overlap Coefficient is the maximum of this metric across all stringencies of any pair of methods. This is a similarity measure between 0 and 1, and 1 minus this value is used here as a distance measure. Red: implicit phylogenetic methods. Blue: explicit phylogenetic methods. Yellow: parametric method Wn. To enhance clarity, a power transformation with exponent 0.6 was applied to the *x*-axis, spreading apart internal nodes that would otherwise appear too close to zero. *α*: GLOOME without species tree.

Supplementary fig. S1, Supplementary Material online confirms the very low overlap between implicit and explicit phylogenetic methods shown in Fig. 1. Within these two groups, we generally see relatively low overlap coefficients at higher stringencies but strong overlaps at low stringencies. This indicates that high-confidence inferences tend to be different even across similar methods, while for low-confidence inferences, one set of inferences tends to be a subset of the other for any pair of methods from the same class (implicit or explicit).

Implicit Phylogenetic Methods Infer a Higher Percentage of Co-acquisitions as Neighbors

When a method infers that two genes from different families were acquired on the same terminal branch of the genome tree, we refer to them as co-acquisitions. If both genes are also inferred to originate from the same donor branch, we call them co-transfers. These terms do not imply that the genes were acquired in a single HGT event. However, some such pairs likely indeed stem from the transfer of a single DNA fragment containing multiple adjacent genes, which would initially remain neighbors in the recipient genome (Dilthey and Lercher 2015; Pang and Lercher 2017). While independent transfers could also result in neighboring genes by chance, this is expected to be rare, as quantified below.

We do not know in advance how often true co-acquired gene pairs are the result of single HGT events, and how often they should still be neighbors. However, if one method finds a much higher fraction of neighboring co-acquisitions than another, this is unlikely to be the result of chance—instead, it suggests a lower rate of false-positive inferences by the first method.

For each method, the number of inferred co-acquisitions varied based on the stringency used for each inference method.

As we decrease the stringency, the number of co-acquisitions inferred by ALE, Wn, Count, and GLOOME (ML) varies from about 100 to 5×10^5. In contrast, GLOOME (ML, without tree), and RANGER infer much higher numbers of co-acquisitions in general, varying from about 2×10^4 to 10^6 (supplementary fig. S2b, Supplementary Material online). For a fair comparison, the percentage of neighbors should be compared at comparable levels of stringency, i.e. at the same number of inferred co-acquisitions.

The percentage of co-acquisitions in which the co-acquired gene pairs are separated by zero or one intervening gene is notably high, but this percentage declines sharply as the number of intervening genes increases (supplementary fig. S5, Supplementary Material online). For the main text, we define two genes as neighbors if they are separated by at most one other gene. supplementary table S1, Supplementary Material online summarizes the results for other values of the maximum number of intervening genes.

At any given number of co-acquisitions—i.e. any given level of stringency—the implicit phylogenetic methods Count (ML or MP) and GLOOME (ML or MP) almost always infer a higher percentage of neighboring co-acquisitions than the explicit phylogenetic methods or Wn (Fig. 2, Table 1). For example, at 10^3 co-acquisitions, Count (MP) infers >8 times the number of neighboring co-acquisitions inferred by ALE. Count and GLOOME make the exact same inferences with the MP approach. Across stringencies, they outperformed all other methods, with GLOOME (ML) taking second place except at the most stringent thresholds. GLOOME and Count rely solely on the presence and absence of genes across the genome tree.

If one or both genes inferred to have been co-acquired by a given method are false inferences, or if they result from independent HGT events, there is no reason to expect them to be genomic neighbors. Indeed, the expected percentage of neighboring pairs among randomly located genes is very low—typically between 0.09% and 0.18%, with most values around 0.13% (supplementary fig. S2a, Supplementary Material online). Correcting for this background expectation does not alter the results.

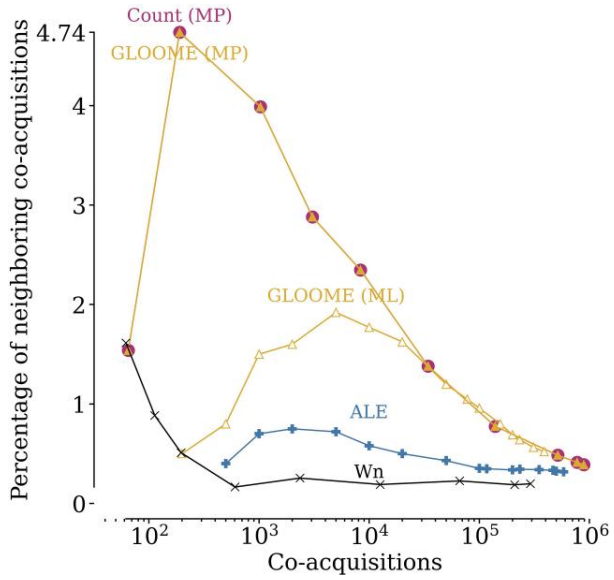


Fig. 2. Count (MP) and GLOOME (MP) outperform other methods at inferring neighboring co-acquisitions. The *y*-axis shows the percentage of neighboring co-acquisitions inferred. Higher stringencies of individual methods are expected to lead to fewer false positive predictions; to make results comparable across methods, the *x*-axis indicates the number of co-acquisitions. Additional methods are shown in supplementary fig. S2b, Supplementary Material online.

It appears likely that gene trees are more accurate when they are based on more sequence information, that is, for longer gene sequences. As expected, methods that utilize the gene tree information perform slightly better on longer genes (supplementary fig. S3, Supplementary Material online); however, the explicit phylogenetic methods still infer lower percentages of neighboring co-acquisitions compared to the well-performing implicit methods. In sum, the use of gene trees appears to lead to a higher fraction of false-positive HGT inferences.

Gene acquisitions inferred on short terminal branches of the species tree must be (relatively) recent, while gene acquisitions inferred on long terminal branches can be recent, but can also be much older. Thus, we expect that the events inferred for short branches are on average younger than those inferred for long branches. As shown in supplementary fig. S4, Supplementary Material online, method performance ranking varies little with branch length, suggesting that our results generalize across phylogenetic depths.

With GLOOME, one can also infer HGTs without providing a *species tree*, in which case the method will infer a tree topology that is most consistent with the phyletic patterns. Without an input species tree, GLOOME consistently performs worse, inferring lower percentages of neighboring co-acquisitions, than the corresponding case of ML or MP with an input species tree (Fig. 3)

We note here that given the small percentages of neighboring co-acquisitions, one must be cautious in interpreting the results. For example, in supplementary fig. S2b, Supplementary Material online, the observed percentage of neighboring co-acquisitions for several methods (e.g. Count (ML), Wn) increases from 0.5% to 1% when we go from about 200 co-acquisitions to 100 co-acquisitions. This means that the same, single neighboring co-acquisition is inferred by the respective method, but the total number of co-acquisitions

Table 1 Count (MP) infers a higher percentage of neighboring co-acquisitions

Method	Mean	std	Max	Stringency
Count (MP)	1.893	1.554	4.737	7.000
GLOOME (MP)	1.893	1.554	4.737	7.000
GLOOME (ML)	1.076	0.491	1.920	0.956
Count (ML)	0.896	0.523	2.600	1.000
GLOOME (MP) α	0.896	0.409	1.376	8.000
GLOOME (ML) α	0.499	0.119	0.711	1.000
Wn	0.470	0.489	1.613	13.000
ALE	0.452	0.158	0.750	0.770
AnGST	0.337	0.000	0.337	1.000
RANGER-Fast	0.334	0.001	0.336	1.000
RANGER	0.319	0.036	0.409	1.000

Shown here are the mean, standard deviation, and maximum of the percentage of neighboring co-acquisitions, as well as the stringency at which the maximum occurs. This table shows only $t = 1$, where t is the maximum number of intervening genes between two co-acquired genes to be considered neighbors. supplementary table S1, Supplementary Material online shows the same for other values of t. Methods using presence–absence outperform the others for all tested values of t, with Count (MP) leading among them. Methods such as ALE or Wn have lower variation across their stringencies and low mean percentages of neighboring co-acquisitions. α : GLOOME without species tree

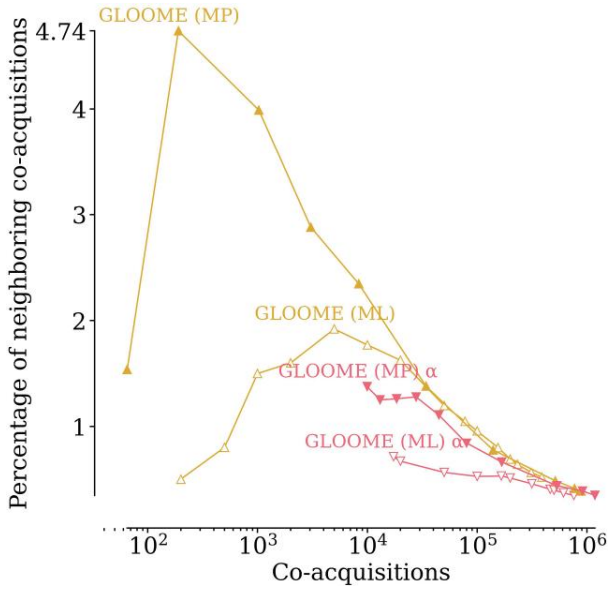


Fig. 3. GLOOME performs worse when run without a species tree. Shown here is the percentage of neighboring co-acquisitions inferred by GLOOME (ML) and GLOOME (MP) with and without a species tree (the latter is shown in red and marked with a α).

inferred is halved. Such effect of small numbers greatly affects the apparent performance of the methods. This type of effect also explains why the parametric method Wn, which generally performs worse than all other methods, appears to improve in performance at very high stringencies (Fig. 2).

The explicit phylogenetic methods in our analysis—ALE, RANGER, and AnGST—also provide information on the donors of the acquired genes. Genes acquired on the same branch of the genome tree could originate from the same donor or from different donors. We have no strong *a priori* expectation about the relative probability of these two types of events. However, it appears unlikely that a method would infer the same donor just by chance. Thus, if we see notable differences between methods in the fraction of inferred co-acquisitions that are also inferred to be co-transfers, then this indicates *a posteriori* that (i) some co-acquisitions are indeed co-transfers and that (ii) the method inferring a higher fraction of co-transfers has a lower false-positive rate. Thus, just like for neighboring co-acquisitions, we also expect that more reliable methods should infer more co-transfers among co-acquisitions; further, they should also infer more neighboring co-transfers among co-acquisitions.

Where results are available for both explicit phylogenetic methods, RANGER almost always infers higher percentages of co-transfers among co-acquired genes than ALE (Fig. 4a). At increasing stringencies for ALE – where no RANGER results are available—ALE infers higher percentages of co-transferred genes.

In terms of the inferred percentage of co-transferred neighbors among co-acquisitions, RANGER and ALE perform similarly (Fig. 4b). The results for ALE at higher stringency – where we have no data from RANGER—can again be attributed to the effects of small numbers. Given that the performance of the two methods in this test is similar, one may choose ALE over RANGER simply because it is faster to run. In our analyses, ALE finished running on the dataset using 100 threads in $\approx$ 18 min, whereas RANGER took $\approx$ 4.5 h. AnGST and RANGER-Fast, while faster, perform much worse: they consistently infer very similar percentages of co-transferred genes and co-transferred neighbors, which are substantially lower than those inferred by ALE and RANGER.

At relatively high stringencies ($<10^4$ co-acquisitions), ALE infers a relatively large proportion (>13%) of gene pairs as co-transfers (Fig. 4a). According to Fig. 4b, only about 0.2% of co-acquired gene pairs are co-transferred neighbors at this stringency. This indicates that only about 1.5% (i.e. 0.2%/13%) of co-transferred genes are neighbors. This low fraction is roughly consistent with Fig. 2, where about 0.5% of gene pairs inferred by ALE to be co-acquired are neighbors. The low fraction of neighbors indicates that the majority of what we call co-transfers (and, consequently, also co-acquisitions) are the result of independent HGT events. Importantly, that does not invalidate our basic assumption: that a higher fraction of inferred neighboring co-acquisitions (or co-transfers) indicates that the corresponding inference method is more reliable. Instead, it just emphasizes that a perfect inference method would not predict that 100% of co-acquisitions are neighbors.

Discussion

The present study is the first systematic comparison of phylogenetic methods for inferring HGT. Our analysis reveals several important insights that can guide the choice of method for HGT inference in a given biological study.

Implicit phylogenetic methods, which rely solely on phylogenetic profiles, more accurately capture the spatial clustering of horizontally transferred genes in recipient genomes compared to explicit phylogentic methods. The latter infer HGT events from topological conflicts between gene and species trees, making them vulnerable to gene tree reconstruction errors that can inflate false positive rates for acquisitions and co-acquisitions. The idea that gene tree noise undermines the reliability of explicit methods is supported by the slight performance increase of the explicit phylogenetic methods for long compared to short genes (supplementary fig. S3, Supplementary Material online). This aligns with earlier studies showing that statistical errors in gene tree reconstruction can cause topological incongruities with species trees, leading to spurious HGT inferences. Including low-confidence branches exacerbates this problem, increasing false positive rates (Than et al. 2007, 2008). In contrast, implicit methods bypass sequence-level reconstruction and instead rely on gene family assignments, which are considerably less error-prone.

It is surprising to find that the maximum parsimony approaches (of Count and GLOOME) perform on par or often better than their maximum likelihood counterparts, given that maximum likelihood is often considered a superior approach, e.g. in phylogeny reconstruction (Gadagkar and Kumar 2005). The difference between the two versions of Count is larger when the stringency is higher. The strong increase in the percentage of neighboring co-acquisitions indicates that the strategy of increasing Count (MP)'s stringency by increasing the gain/loss penalty ratio works well, leading to fewer, but more accurate HGTs. In contrast, increasing the stringency of maximum likelihood inferences by selecting events with increasingly higher inferred posterior probabilities appears to be less effective. Future work should explore alternative ways to vary the stringency of maximum likelihood inferences.

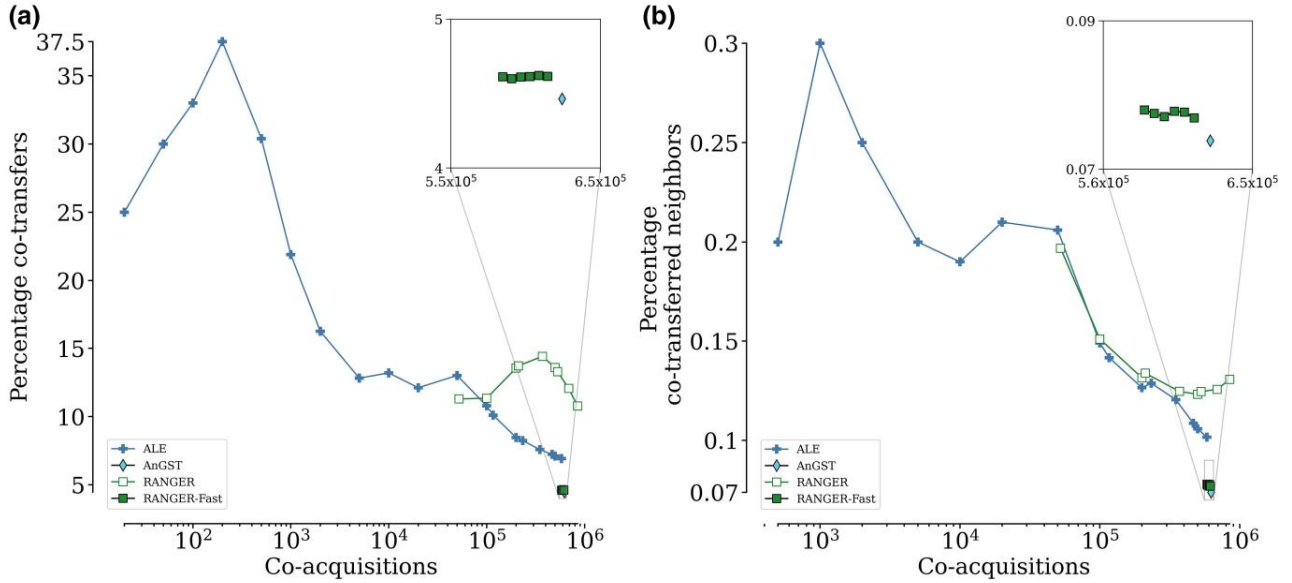


Fig. 4. RANGER performs better or similar to ALE at inferring co-transferred genes and co-transferred neighbors. At lower stringency, RANGER infers higher a) percentages of co-transfers and b) percentages of co-transferred neighbors than ALE. Although at higher stringency, ALE generally makes better inferences, it is either similar to RANGER or the performance can not be compared as no data from RANGER is available. AnGST and RANGER-Fast generally provide much worse inferences than ALE and RANGER. Note that only explicit phylogenetic methods are shown here, since other methods do not provide information about the source of transfer.

Among tree reconciliation methods, RANGER and ALE generally performed better or similar to the others at inferring co-transfers and co-transferred neighbors. Given that ALE is aware of extinct taxa (Szöllosi et al. 2012), it is surprising that it does not perform better than RANGER where results are available for both.

Our methodology is—at least in principle—biased in favor of Wn, the only sequence composition-based method included in our study. Wn is based on oligonucleotide frequencies, which may differ systematically across genomic regions. Accordingly, we may expect Wn to often flag neighboring genes as co-acquisitions, regardless of whether they were indeed the result of HGT. Despite this bias, Wn still performs worse than the phylogenetic methods tested.

Keeping these results in mind, we recommend the use of the implicit phylogenetic methods Count or GLOOME to identify horizontal gene acquisitions. In particular, the MP versions of Count and GLOOME outperform all other methods in identifying neighboring co-acquisitions, especially at higher stringencies (gain/loss penalty ratios 4 to 7; Fig. 2). If inferring the donor of HGT events is a requirement, explicit phylogenetic methods must be used. In this case, both ALE and RANGER are adequate choices, although ALE is computationally more efficient than RANGER. ALE can be made more stringent than RANGER by choosing a high stringency (>0.86 transfers per reconciliation model). At this level of stringency, its inferences are the most reliable among all tested scenarios for explicit methods, as indicated by the highest percentages of neighboring co-acquisitions or co-transfers (supplementary fig. S2c, Supplementary Material online). On that note, for any method used, we generally recommend using a minimum stringency (see Table 1) and focus on a subset of reliable HGT inferences.

Although Gammaproteobacteria are a well-studied and diverse bacterial clade, these findings might vary for other datasets. Specifically, results could differ when analyzing only recent gene transfers or examining shallower phylogenetic trees, such as those limited to *Escherichia coli* strains. However, this seems unlikely, given that the relative performance ranking of the methods remains mostly consistent across branch lengths (supplementary fig. S4, Supplementary Material online). Because many users will not spend time optimizing most of the parameters offered by the various programs, we ran each program only with their default settings. We note, however, that these parameters can dramatically change their performance.

Our conclusions rely mainly on the analysis of neighboring co-acquisitions and co-transfers. A limitation of this approach is the lack of an *a priori* expectation for how often gene pairs co-acquired on the same branch of the species tree indeed result from the same HGT event, and how often these are expected to still be neighbors in extant genomes. As a result, our method can provide relative rankings of inference methods, but cannot assess their reliability in absolute terms. Nonetheless, the substantial differences between methods in the fraction of neighboring co-acquisitions suggest that some of these co-acquisitions truly reflect single HGT events. If all methods had similar error rates, we would expect similar fractions of neighboring and co-transferred genes across methods – which is not the case. These differences therefore indicate varying reliability and provide a basis for improving HGT inference methods.

Methods

Genomic Data Selection and Cleaning

We used the Gammaproteobacterial dataset (NCBI taxonomic ID: 1236) from the EggNOG database v6 (Hernández-Plaza et al. 2023) for orthologous genes (non-supervised orthologous groups, or *NOGs*). The corresponding nucleotide sequence information (for Wn execution) was downloaded from the NCBI GenBank database (Sayers et al. 2022) using

the NCBI Datasets CLI program. The reference species tree downloaded from ASTRAL WoL (Zhu et al. 2019). Specifically, the `astral - branch length - cons - astral.cons.nwk` file was used. In brief, this means that the topology of the tree was inferred using ASTRAL and the branch lengths were re-estimated with maximum likelihood, using the 100 most conserved sites per gene. We retain only the taxa belonging to the maximum overlap of taxa between the set of NOGs, such that their sequence data could be retrieved and that these taxa could be reliably mapped onto the species tree. Furthermore, to ensure that gene families were large enough to infer meaningful HGT events, NOGs containing <10 genes or <30 taxa were removed.

NOGs were then selected around the average sized NOGs such that they together represented all the 359 taxa in the dataset, resulting in a set of 1,286 NOGs. The gene trees were taken from the EggNOG database and rooted using Minimum Ancestor Deviation (Tria et al. 2017).

Inferring HGT

All the programs were executed separately, using default parameters. For ALE, the *ALEml_undated* program was used, which does not require a dated species tree (Szöllősi et al. 2015). RANGER-DTL v2.0 was executed with and without the "fast" option. AnGST was executed with parameters as suggested in the manual.

The stringency of ALE, RANGER, and RANGER-Fast is defined as the mean number of transfers inferred per reconciliation model. At increasing stringency levels, we exclude HGT events that are inferred with lower frequencies across all of the output the reconciliation models. For Count, the "Asymmetric Wagner Parsimony" method was used as the maximum parsimony method. Unlike GLOOME where we also ran without a species tree for both ML and MP settings, Count could be run only with a species tree. For both of these implicit phylogenetic methods, the stringency is defined as the gain/loss penalty ratio given as input. Unlike the explicit phylogenetic methods where we vary the stringency on the output of a single run, with the implicit phylogenetic methods we run the program multiple times, each time with a different gain/loss penalty ratio. These phylogenetic HGT inference methods have publicly available implementations that we used with default settings. In contrast, we implemented the parametric method Wn ourselves, as described below, since there is no publicly available implementation.

The parametric method Wn was implemented based on the original publication (Tsirigos and Rigoutsos 2005), and as suggested, we used an oligonucleotide size of 8 nucleotides. Wn calculates a *typicality score* of how similar the gene sequence is to the genome sequence based on the oligonucleotide frequencies in both. These frequencies were efficiently calculated using KMC (Kokot et al. 2017). In the original paper for Wn, the authors used an automatic threshold detection method. This method rank orders the genes based on their typicality scores, smoothens this curve (requiring a rolling window average), takes a derivative of this curve, and then considers transferred genes as those that are above the point where the derivative becomes *approximately* constant. Although this method is termed *automatic*, it still depends on both the rolling window size used for smoothing and the user's notion of when exactly does the derivative become approximately constant (i.e. the double derivative is approximately zero). Instead of making the choice of when the double derivative is approximately zero, we used multiples of the *median absolute deviation* (MAD) of the double derivative around zero, which also helps to set the stringencies. These multiples represent the *stringency* parameter that goes from 4 times the MAD to 13 times the MAD. Note that it is generally recommended to use at least 2.5 times the MAD, and so we are already being highly strict even when Wn is run at the lowest stringency. The rolling window size was automatically determined by iteratively finding the longest Savitzky-Golay filter such that the root mean square error of the smoothed curve with respect to the original curve stops decreasing. The code for Wn is available at the code repository linked in the Data Availability section.

Co-acquisition and Co-transfer Analysis

To limit the effect of small numbers, chromosomes with <1,000 genes were excluded from calculations of co-acquisitions, co-transfers, and neighbors. At any given stringency of a method, if the number of co-acquisitions was < 20, that stringency was excluded from the analysis. The co-acquisitions discussed in this study are those where both genes have known positions in their respective chromosomes. In any gene family, genes from EggNOG that could not be retrieved from GenBank, or pair of genes co-acquired by the same taxon but not in the same chromosome/contig in the GenBank data, were excluded from this analysis.

We calculated the expected percentage of neighboring co-acquisitions under the null model that the two genes were acquired in independent HGT events or that one or both HGT inferences are false. We first calculate the expected number of neighboring co-acquisitions for each chromosome individually, based on the chromosome size and the number of inferred co-acquisitions for that chromosome (see below). The randomly expected number of neighboring co-acquisitions is the sum of these expected numbers across chromosomes. The expected percentage of neighboring co-acquisitions is this expected number, divided by the total number of co-acquisitions. Specifically, if t is the maximum number of intervening genes between the two genes in the pair (in our case $t = 1$), and n is the number of chromosomes among the inferred co-acquisitions for a method, the expected fraction of neighboring co-acquisitions is

$$\frac{2\sum_{i=1}^{n}(t+1)(c_i/g_i)}{\sum_{i=1}^{n}c_i}$$

where c_i and g_i are the number of co-acquisitions and the number of genes in chromosome i, respectively. This is because for each horizontally transferred gene in a chromosome, the probability of another gene being transferred as a neighbor is $2(t+1)/g_i$ (i.e. approximating for boundary genes, there are $t+1$ neighboring positions on either side of a gene out of g_i positions in the chromosome). The expected number of neighboring co-acquisitions is then this probability multiplied by the number of co-acquisitions for that chromosome. The expected *fraction* of neighboring co-acquisitions is the sum of the expected number of neighboring co-acquisitions (per chromosome) divided by the sum of co-acquisitions for all chromosomes.

The expected percentages differ across methods because they are calculated from the number of inferred HGTs for each given genome, considering the genome size. Since these numbers differ across methods, there is some variation across methods in the overall expected number of neighboring co-acquisitions.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Tal Dagan for discussions about the evaluation of HGT inference methods. We thank Carolina Fanalista for her assistance in setting up the project in its early stages.

Funding

This work was supported by funding to M.J.L. from the Volkswagenstiftung in the "Life?" initiative and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1310.

Data Availability

All of the code used in this study is available at the following Gitlab repository: https://gitlab.cs.uni-duesseldorf.de/general/ccb/hgt_inference_comparative_study The data used in this study can be retrieved from Zenodo: https://doi.org/10.5281/zenodo.15535009.

Conflict of interests statement

None declared.

References

Arnold BJ, Huang I-T, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*. 2021:20(4):206–218. ISSN 1740-1534. https://doi.org/10.1038/s41579-021-00650-4.

Bansal MS, Kellis M, Kordi M, Kundu S. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*. 2018:34(18):3214–3216. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/bty314.

Becq J, Churlaud C, Deschavanne P. A benchmark of parametric methods for horizontal transfers detection. *PLoS One*. 2010:5(4):e9989. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0009989.

Beiko RG, Hamilton N. Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol*. 2006:6(1):1–17. ISSN 1471-2148. https://doi.org/10.1186/1471-2148-6-15.

Cohen O, Ashkenazy H, Belinky F, Huchon D, Pupko T. GLOOME: gain loss mapping engine. *Bioinformatics*. 2010:26(22):2914–2915. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btq549.

Cordero OX, Hogeweg P. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci U S A*. 2009:106(51):21748–21753. https://doi.org/10.1073/pnas.0907584106.

Csűös M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*. 2010:26(15):1910–1912. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btq315.

David LA, Alm EJ. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*. 2011:469(7328):93–96. ISSN 1476-4687. https://doi.org/10.1038/nature09649.

Dessimoz C, Margadant D, Gonnet GH. DLIGHT – lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In: Vingron M, Wong L, editors. Research in computational molecular biology. Berlin, Heidelberg: Springer; 2008. p. 315–330. ISBN 978-3-540-78839-3. https://doi.org/10.1007/978-3-540-78839-3_27.

Dilthey A, Lercher MJ. Horizontally transferred genes cluster spatially and metabolically. *Biol Direct*. 2015:10(1):72. ISSN 1745-6150. https://doi.org/10.1186/s13062-015-0102-5.

Felsenstein J. Inferring phylogenies. Sunderland (MA): Sinauer Associates; 2003.

Feyerabend P. Against method. London: Verso Books; 1993.

Gadagkar SR, Kumar S. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol Biol Evol*. 2005:22(11):2139–2141. ISSN 0737-4038. https://doi.org/10.1093/molbev/msi212.

Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, Kirsch R, Rattei T, Letunic I, Jensen LJ, *et al*. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res*. 2023:51(D1):D389–D394. ISSN 0305-1048. https://doi.org/10.1093/nar/gkac1022.

Hill T, Nordström KJV, Thollesson M, Säfström TM, Vernersson AKE, Fredriksson R, Schiöth HB. SPRIT: identifying horizontal gene transfer in rooted phylogenetic trees. *BMC Evol Biol*. 2010:10(1):42. ISSN 1471-2148. https://doi.org/10.1186/1471-2148-10-42.

Kapust N, Nelson-Sathi S, Schönfeld B, Hazkani-Covo E, Bryant D, Lockhart PJ, Röttger M, Xavier JC, Martin WF. Failure to recover major events of gene flux in real biological data due to method misapplication. *Genome Biol Evol*. 2018:10(5):1198–1209. https://doi.org/10.1093/gbe/evy080.

Kokot M, Długosz M, Deorowicz S. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*. 2017:33(17):2759–2761. ISSN 1367-4803. https://doi.org/10.1093/bioinformatics/btx304.

Nakhleh L, Ruths D, Wang L-S. RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In: Wang L, editor. Computing and combinatorics, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2005. p. 84–93. ISBN 978-3-540-31806-4. https://doi.org/10.1007/11533719_11.

Pang TY, Lercher MJ. Supra-operonic clusters of functionally related genes (SOCs) are a source of horizontal gene co-transfers. *Sci Rep*. 2017:7(1):40294. ISSN 2045-2322. https://doi.org/10.1038/srep40294.

Poptsova M. Testing phylogenetic methods to identify horizontal gene transfer. In: Walker JM, Gogarten MB, Gogarten JP, Olendzenski LC, editors. Horizontal gene transfer. Vol. 532. Totowa, NJ: Humana Press; 2009. p. 227–240. ISBN 978-1-60327-852-2 978-1-60327-853-9. https://doi.org/10.1007/978-1-60327-853-9_13.

Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. *PLoS Comput Biol*. 2015:11(5):e1004095. ISSN 1553-734X. https://doi.org/10.1371/journal.pcbi.1004095.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022:50(D1):D20–D26. ISSN 1362-4962. https://doi.org/10.1093/nar/gkab1112.

Sheinman M, Arkhipova K, Arndt PF, Dutilh BE, Hermsen R, Massip F. Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *Elife*. 2021:10:e62719. ISSN 2050-084X. https://doi.org/10.7554/eLife.62719.

Sjöstrand J, Tofigh A, Daubin V, Arvestad L, Sennblad B, Lagergren J. A Bayesian method for analyzing lateral gene transfer. *Syst Biol*. 2014:63(3):409–420. ISSN 1063-5157. https://doi.org/10.1093/sysbio/syu007.

Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci U S A*. 2012:109(43):17513–17518. ISSN 1091-6490. https://doi.org/10.1073/pnas.1202997109.

Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos Trans R Soc Lond B Biol Sci*. 2015:370(1678):20140335. ISSN 1471-2970. https://doi.org/10.1098/rstb.2014.0335.

Szöllősi GJ, Tannier E, Lartillot N, Daubin V. Lateral gene transfer from the dead. *Syst Biol*. 2013:62(3):386–397. ISSN 1063-5157. https://doi.org/10.1093/sysbio/syt003.

Than C, Jin G, Nakhleh L. Integrating sequence and topology for efficient and accurate detection of horizontal gene transfer. In: Nelson CE, Vialette S, editors. Comparative genomics. Berlin, Heidelberg. Springer; 2008. p. 113–127. ISBN 978-3-540-87989-3. https://doi.org/10.1007/978-3-540-87989-3_9.

Than C, Ruths D, Innan H, Nakhleh L. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions.

J Comput Biol. 2007:14(4):517–535. https://doi.org/10.1089/cmb.2007.A010.

Tria FDK, Landan G, Dagan T. Phylogenetic rooting using minimal ancestor deviation. *Nat Ecol Evol*. 2017:1(1):1–7. ISSN 2397-334X. https://doi.org/10.1038/s41559-017-0193.

Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res*. 2005:33(3):922–933. ISSN 1362-4962. https://doi.org/10.1093/nar/gki187.

Wade T, Rangel LT, Kundu S, Fournier GP, Bansal MS. Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. *PLoS One*. 2020:15(5):e0232950. ISSN 1932-6203. https://doi.org/10.1371/journal.pone.0232950.

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, *et al*. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun*. 2019:10(1):5477. ISSN 2041-1723. https://doi.org/10.1038/s41467-019-13443-4.

Supplementary Information

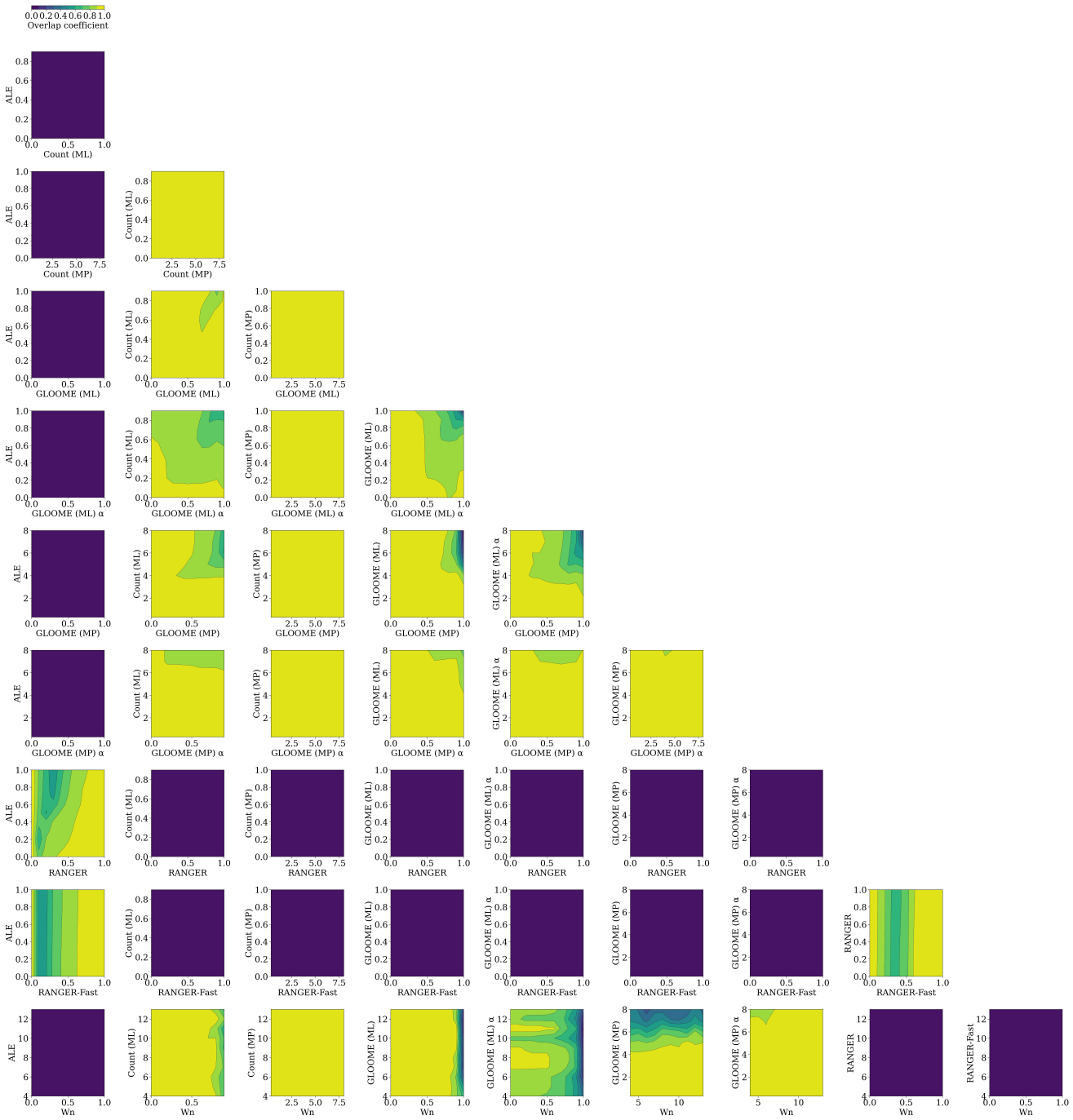


Fig. S1. Contour plots of Overlap Coefficient between pairs of inference methods. Figure shows the Overlap Coefficient between pairs of inference methods, for all gene families in our dataset. Each axis subplot is for one pair of methods, and each axis in each subplot shows stringency values for the corresponding method in the axis label. Colors indicate the Overlap Coefficient between the two methods, with warmer colors indicating higher overlap. The Overlap Coefficient is calculated as the size of the intersection of two sets divided by the size of the smaller set. Here, the two sets are sets of inferred HGT events. Implicit and explicit phylogenetic methods have very low overlap, while within these two categories pairs of methods generally have higher overlap at lower stringencies.
Alt text: Figure showing contour plots for each pair of methods. Method names are axis labels and the stringency values are axis ticks. The color of the contour plots indicates the overlap coefficient between the two methods. α : GLOOME without species tree.

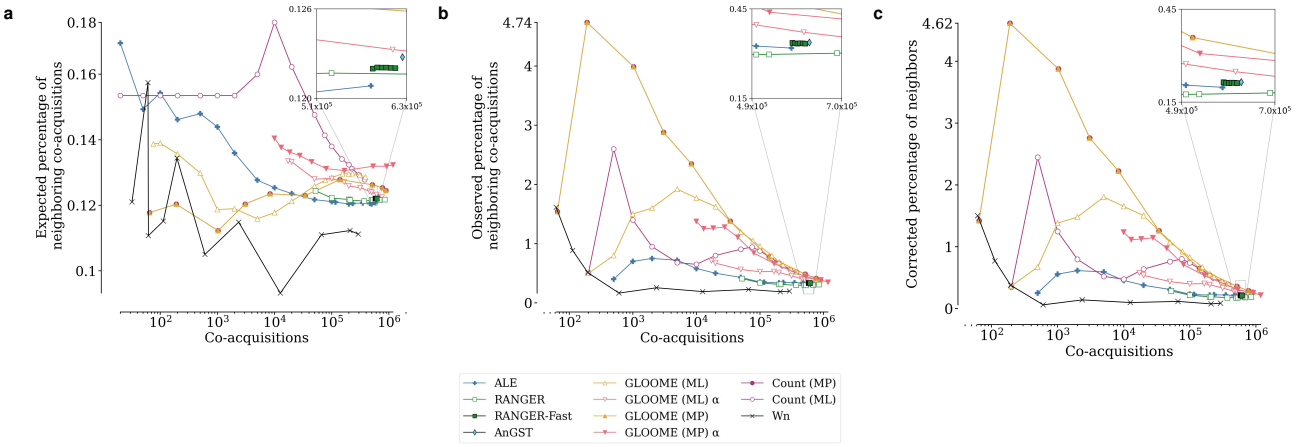


Fig. S2. (a) Expected, (b) observed, and (c) observed minus expected percentage of neighboring co-acquisitions, vs number of co-acquisitions inferred. The expected percentage of neighboring co-acquisitions changes with the number of co-acquisitions inferred at various thresholds, and varies between methods, since it depends on the set of chromosomes that those co-acquisitions are inferred on.
Alt text: Figure showing subplots for expected, observed, and observed minus expected percentage of neighboring co-acquisitions respectively. The x-axis shows the number of co-acquisitions inferred, and the y-axis shows the percentage of neighboring co-acquisitions. The legend shows the methods used for inference. α : GLOOME without species tree.

Method	Mean (t=0)	std (t=0)	Max (t=0)	Mean (t=2)	std (t=2)	Max (t=2)	Mean (t=3)	std (t=3)	Max (t=3)
Count (MP)	1.324	1.187	3.684	2.198	1.666	4.737	2.396	1.716	4.961
GLOOME (MP)	1.324	1.187	3.684	2.198	1.666	4.737	2.396	1.716	4.961
GLOOME (ML)	0.657	0.335	1.180	1.347	0.593	2.380	1.547	0.692	2.840
GLOOME (MP) α	0.544	0.255	0.814	1.192	0.549	1.888	1.410	0.640	2.250
Count (ML)	0.460	0.256	1.200	1.124	0.641	3.200	1.355	0.739	3.800
Wn	0.330	0.501	1.613	0.573	0.493	1.613	0.654	0.447	1.613
ALE	0.314	0.160	0.600	0.578	0.165	0.880	0.672	0.170	1.020
GLOOME (ML) α	0.289	0.065	0.405	0.672	0.171	0.995	0.817	0.212	1.232
AnGST	0.196	0.000	0.196	0.454	0.000	0.454	0.561	0.000	0.561
RANGER-Fast	0.194	0.001	0.195	0.444	0.001	0.446	0.546	0.001	0.547
RANGER	0.186	0.026	0.250	0.425	0.046	0.539	0.528	0.059	0.673

Table S1. Mean, standard deviation, and maximum of the percentage of neighboring co-acquisitions inferred by each method, for $t \neq 1$, where t is the maximum number of intervening genes between two co-acquired pair of genes to be considered neighbors.
α : GLOOME without species tree.

Short gene length		Long gene length	
Method	Mean	Method	Mean
Count (MP)	1.459	Count (MP)	1.556
GLOOME (MP)	1.459	GLOOME (MP)	1.556
GLOOME (ML)	1.183	GLOOME (MP) α	1.128
GLOOME (MP) α	1.002	GLOOME (ML)	1.100
Count (ML)	0.694	Count (ML)	0.900
GLOOME (ML) α	0.530	GLOOME (ML) α	0.646
ALE	0.324	ALE	0.607
RANGER	0.274	RANGER	0.449
Wn	0.225		

Table S2. Ranking of methods according to mean percentages of neighboring co-acquisitions inferred by each method, for gene families with mean gene length shorter or longer than the median value of 799.86 bp. Only co-acquisitions in the range of 10^3 to 10^5 were considered, to exclude the effect of small numbers at lower number of co-acquisitions and the effect of low confidence inferences at higher number of co-acquisitions. Note that the relative ranking of the methods remains similar (see also Supplementary Figure S3). α : GLOOME without species tree.

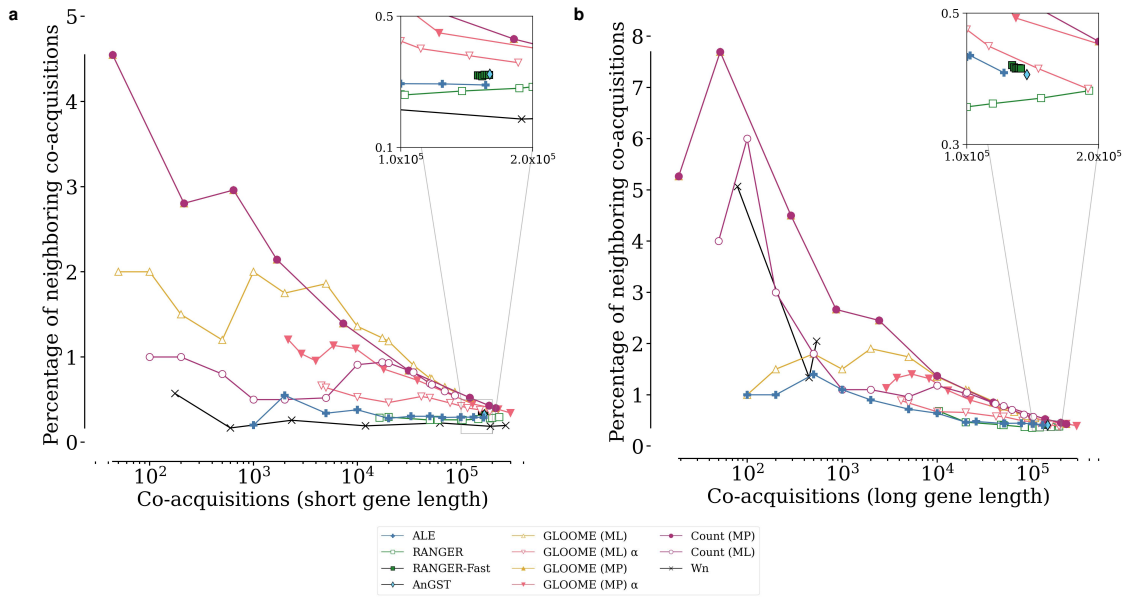


Fig. S3. Percentage of neighboring co-acquisitions vs number of co-acquisitions inferred, shown separately for gene families with mean gene length (a) shorter and (b) longer than the median value of 799.86 bp (i.e., the lower and upper halves of the gene length distribution in our dataset). The mean gene length is calculated as the average length of all genes in a gene family. Explicit phylogenetic methods perform slightly better on longer genes. For example, at 2000 co-acquisitions, the percentage of neighboring co-acquisitions inferred by ALE is approximately 0.5% for short genes and approximately 1% for long genes. However, the relative ranking of the methods remains similar (see also Supplementary Table S2) α : GLOOME without species tree.

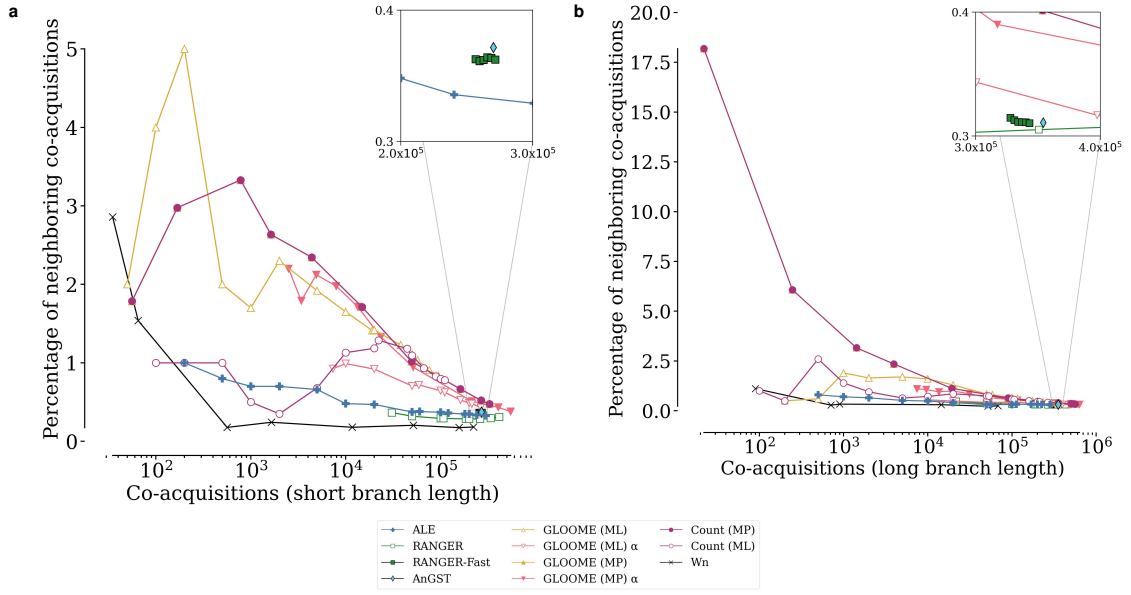


Fig. S4. Percentage of neighboring co-acquisitions vs number of co-acquisitions inferred, on species tree branches that are (a) shorter or (b) longer than the median branch length of 0.076 substitutions per site. The relative ranking of the methods remains similar (see also Supplementary Table S3). α : GLOOME without species tree.

Short branches		Long branches	
Method	Mean	Method	Mean
Count (MP)	1.927	Count (MP)	1.820
GLOOME (MP)	1.927	GLOOME (MP)	1.820
GLOOME (MP) α	1.726	GLOOME (ML)	1.305
GLOOME (ML)	1.411	GLOOME (MP) α	0.933
Count (ML)	0.906	Count (ML)	0.830
GLOOME (ML) α	0.818	GLOOME (ML) α	0.488
ALE	0.517	ALE	0.463
RANGER	0.326	RANGER	0.391
Wn	0.209	Wn	0.259

Table S3. Ranking of methods according to mean percentages of neighboring co-acquisitions inferred by each method, for species tree branches shorter or longer than the median branch length of 0.076 substitutions per site. Only co-acquisitions in the range of 10^3 to 10^5 were considered, to exclude the effect of small numbers at lower number of co-acquisitions and the effect of low confidence inferences at higher number of co-acquisitions. Note that the relative ranking of the methods remains similar (see also Supplementary Figure S4). α : GLOOME without species tree.

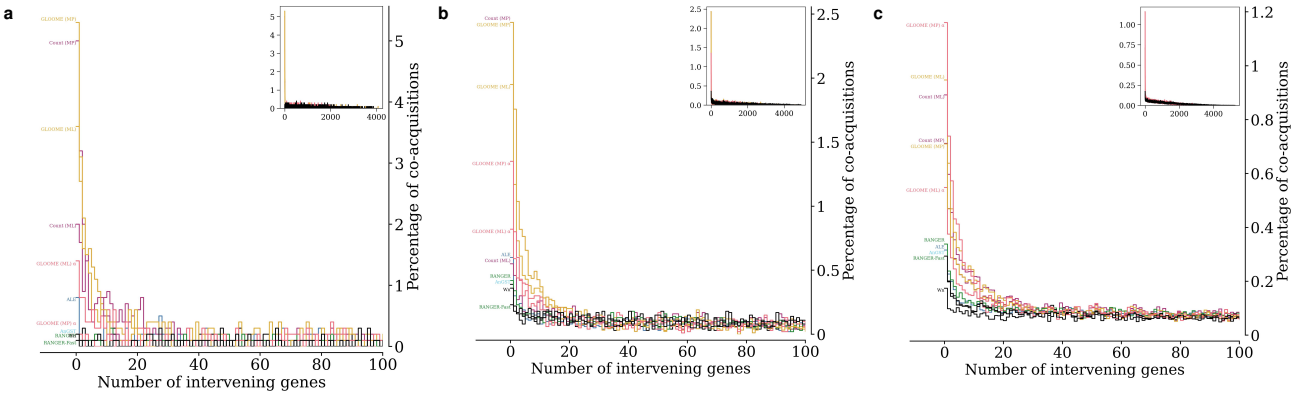


Fig. S5. Distribution of distances between co-acquired genes. Shown here are histograms of number of intervening genes between co-acquired genes at three different stringency levels such that the number of co-acquisitions is (a) 10^3, (b) 10^4, and (c) 10^5. The x-axis shows the number of intervening genes between co-acquired genes, and the y-axis shows the corresponding percentage of co-acquisitions. The inset shows the histogram across the full range of intervening genes across all co-acquisitions, while the main plot shows the histogram for the range of intervening genes between 0 and 100. α : GLOOME without species tree.

Chapter 3

Streamlined genomes, not horizontal gene transfer, mark bacterial transitions to new environments

At the time of submission of the thesis, this manuscript is under consideration at Nature Communications.
I designed the study, performed the analyses, and drafted the manuscript.

Streamlined genomes, not horizontal gene transfer, mark bacterial transitions to new environments

Swastik Mishra[1] **and Martin J. Lercher**[1] ✉

[1]Institute for Computer Science and Department of Biology, Heinrich Heine University, Düsseldorf, Germany

Bacterial colonization of unfamiliar environments constitutes a drastic evolutionary transition, temporarily altering patterns of natural selection. This process is often assumed to be associated with bursts of horizontal gene transfers (HGTs), a major driver of bacterial adaptation. Larger genomes are thought to facilitate such adaptations by providing broader functional repertoires and more integration sites for foreign DNA. Here, we systematically test these ideas across a broad bacterial phylogeny, linked to environmental transitions inferred from metagenomics data. Contrary to expectations, we find that bacteria entering new environments typically have smaller genomes and experience lower rates of HGT. The reduction in HGT is fully explained by genome size, with no residual effect of environmental transitions once size is controlled for. These findings suggest that successful bacterial colonizers rely less on genomic plasticity through HGT than previously assumed, highlighting gaps in our understanding of microbial evolutionary dynamics.

Horizontal Gene Transfer, Genome Size, Bacterial Evolution, Ecosystem Transitions, Colonization, Pathogenicity Gains, Purifying Selection

Correspondence: ***martin.lercher@hhu.de***

Introduction

Bacterial strains frequently adapt to new ecosystems, with prominent examples reported from gut microbiota (Zheng et al., 2020), skin microbiota (Lieberman, 2022), bacterial-fungal associations (Richter et al., 2024), and colonization of marine plastics (Caruso, 2020). Here, we explore the genomic correlates of such colonization events. Specifically, we test three expectations: (1) Because ecosystem transitions are often accompanied by population bottlenecks, we expect a temporary weakening of purifying selection at these events; moreover, as outlined below, there are reasons to hypothesize that colonization is (2) associated with large genomes and (3) accompanied by bursts of horizontal gene transfers (HGT).

Large genomes may facilitate transitions to new ecosystems by providing a broader "toolbox" of molecular functions that can be employed in new combinations to address novel challenges (Maslov et al., 2009, Szappanos et al., 2016). In contrast, bacteria with small genomes often exhibit niche specialization, which can limit their adaptability when environmental conditions shift beyond their optimized range (Serra Moncadas et al., 2024). In addition, larger genomes are more likely to contain homologous genes that can serve as landing pads for horizontal gene transfer (HGT), potentially enhancing access to foreign DNA (Taylor et al., 2024).

HGT plays a central role in bacterial adaptation by enabling the acquisition of genomic fragments – entire genes or operons – from other organisms (Pang and Lercher, 2019, Pál et al., 2005, Arnold et al., 2021). This mechanism enables the rapid acquisition of novel traits and may be particularly important during ecological transitions. Environmental change has therefore been hypothesized to increase HGT rates (Engelstädter and Moradigaravand, 2014), not only by creating new selective pressures but also by exposing bacteria to unfamiliar pangenomes that can be accessed via HGT (Dmitrijeva et al., 2024).

Several studies have shown that environmental change can indeed trigger significant spikes in HGT rates (Woods et al., 2020, Goh et al., 2024, Dadeh Amirfard et al., 2024, Acar Kirit et al., 2020). However, these findings primarily derive from laboratory experiments focused on specific environmental transitions and specific gene families, particularly antibiotic resistance genes; in contrast, HGT can affect virtually any bacterial gene (van Dijk et al., 2020, Coluzzi et al., 2023). Systematic analyses of such patterns across bacterial pangenomes in natural settings have been limited by the challenge of reconstructing environmental transitions along bacterial evolutionary histories.

In this study, we tested these three expectations using the GOLD database (Mukherjee et al., 2023), which provides metagenomic data for hundreds of bacterial genomes across diverse environments. From these data, we inferred transitions to new ecosystems and linked them to nucleotide substitutions, ancestral genome sizes, and HGT rates across a deep bacterial phylogeny. The data indicate that (1) purifying selection weakens during these transitions, as expected. Surprisingly, however, we found that (2) genomes involved in transitions to new ecosystems tend to be small, and (3) there is no evidence for increased HGT rates at such ecosystem transitions.

Results

Dataset construction. We used extant bacterial genomic data obtained from the EggNOG database v6 (Hernández-Plaza et al., 2023), which clusters genes into non-supervised orthologous groups (NOGs), referred to as gene families below. For the inference of transitions to new environments – which we also refer to as ecosystem gains – we used the GOLD database, which contains metagenomic data from a variety of ecosystems. We limited our analyses to bacterial genomes in GOLD with at least 3 'Ecosystem Type' la-

bels. To facilitate the inference of HGT events, we restricted our dataset further to those gene families that are present in at least 4 of these genomes. The final dataset comprises 8197 gene families across 159 taxa, represented in 80 diverse Ecosystem types encompassing both natural and human-made environments; examples include 'fermented vegetables', 'fetus', 'fish products', 'freshwater', and 'industrial wastewater'. We retrieved gene family trees and sequence alignments from EggNOG.

To infer ecosystem transitions and HGT events, we required a species tree representing the phylogenetic relationships between the genomes included. For its reconstruction, we applied the summary-based species tree method ASTRAL-Pro 2 (Zhang and Mirarab, 2022), using the 233 gene family trees of single-copy gene families present in at least 95% of all taxa as input. The branch lengths of this tree were estimated with IQ-TREE 2 (Minh et al., 2020) based on a concatenation of the multiple sequence alignments.

To map ecosystem gains on branches of the genome tree, we used the asymmetric Wagner parsimony algorithm implemented in Count (Csűös, 2010). While originally developed for inferring gene gains and losses from presence–absence patterns, this method is equally applicable to infer gains and losses of other discrete traits. We applied it to a presence–absence matrix of GOLD Ecosystem Type labels across genomes, allowing us to identify branches with ecosystem gains (EG) and those with no ecosystem gains (NEG).

Purifying selection weakens during transitions to new ecosystems, but only for a subset of gene families. Transitions to new ecosystems are likely often accompanied by population bottlenecks, leading to a global decrease in purifying selection. Due to changed interactions with the environment, patterns of natural selection are also likely to change in gene-specific ways: genes no longer required in the new ecosystem may experience relaxed selection, while genes involved in new or modified interactions may experience positive selection. These changes in natural selection are expected to affect nucleotide substitutions in protein-coding genes in the same direction, increasing $\omega = dN/dS$, the ratio of non-synonymous nucleotide substitutions (dN) to synonymous substitutions (dS). ω is a widely used metric to assess the type and strength of selection acting on protein-coding genes. $\omega < 1$ indicates purifying selection, while $\omega = 1$ indicates purely neutral evolution. $\omega > 1$ would indicate strong positive selection. Even when a gene product is under strong positive selection, this selection typically affects only a small subset of amino acids, such as those near an enzyme's active site, while most others remain subject to purifying selection. For this reason, the average ω value for a gene rarely reaches or exceeds 1.

To compare ω values between branches with and without ecosystem gains, we used the RELAX algorithm from the HyPhy software package (Wertheim et al., 2015). RELAX is designed to assess whether the strength of natural selection on a gene family has been relaxed or intensified along a specified set of "test" branches (here defined as EG, ecosystem gain, branches) compared to "reference" branches (here, NEG branches). For each gene family, RELAX fits a branch-specific codon model involving three categories of omega values for each branch set. It also fits a global selection intensity parameter, K, which modifies the selection pressure on the test branches relative to the reference branches as an exponent of ω. $K > 1$ indicates intensified selection on the test branches, while $K < 1$ indicates relaxed selection. A likelihood ratio test is used to assess if a model with $K \neq 1$ explains the data better than a model assuming equal selection pressures on test and reference branches.

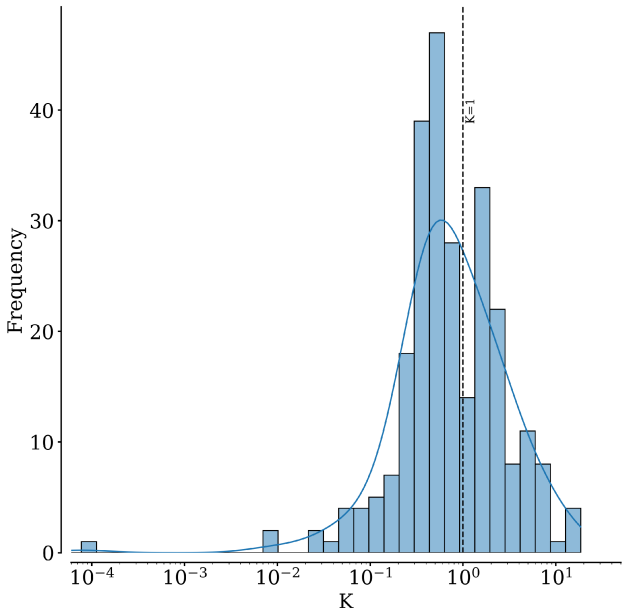


Fig. 1. Distribution of K, the exponent of $\omega = dN/dS$ on EG branches, for gene families with significant changes in selection intensity. $K < 1$ indicates a relaxation of natural selection on branches with ecosystem gains, while $K > 1$ indicates an intensification of selection. The skew toward values < 1 indicates that the majority of gene families show a decrease in the strength of selection during ecosystem gains. The median K value is 0.61.
ALT TEXT: Histogram of K values for gene families with significant changes in selection intensity at ecosystem gains.

For this analysis, we obtained nucleotide sequences from NCBI; for 3228 gene families, we could reliably map EggNOG taxon IDs to corresponding genomes across all relevant taxa using the NCBI Datasets API (Sayers et al., 2022) and successfully ran RELAX. As expected from a dominance of purifying selection, only a small percentage of amino acid sites show signs of positive selection ($\omega > 1$). Consistent with positive selection on a subset of environment-related genes, EG branches have a significantly higher percentage of such sites (median 2.40%) compared to NEG branches (median 1.63%; Mann-Whitney $U = 5709953.5$, two-sided $p = 9.71 \times 10^{-12}, N = 3228$).

RELAX detected a statistically significant change in the intensity of selection at ecosystem gains for 8.15% of gene families (263 out of 3228; adjusted p < 0.05 after correcting for multiple testing (Benjamini and Hochberg, 1995)). Of these, 62.36% (164 out of 263) showed a decrease in the strength of selection ($K < 1$), while the rest indicated intensified selection (Figure 1).

Thus, while the majority of gene families show no significant shift in selection intensity, those that do predominantly show signs of relaxed selection.

Bacteria transitioning to new ecosystems have smaller genomes and lower HGT rates. We next used the complete set of 8197 gene families to test the second prediction: that the genomes of strains that colonize new ecosystems tend to be large. Contrary to this expectation, branches with ecosystem gains tend to have smaller genomes at the ancestral node, with a median of 3,800 genes compared to 5,395 in branches without such gains (top marginal plot in Fig. 2; Mann-Whitney $U = 5884$, two-sided $p = 3.9 \times 10^{-12}, N_{EG} = 208, N_{NEG} = 108$).

Our third expectation was an increase in HGT rates at ecosystem gains. To robustly infer HGT events, we again used the complete gene set of 8197 gene families, complemented by the corresponding gene trees. A previous benchmark on empirical data demonstrated that Count's maximum parsimony approach outperforms alternative HGT inference methods, including explicit, maximum-likelihood-based phylogenetic methods (Mishra and Lercher, 2024). We estimated the *HGT rate* for each branch of the genome phylogeny by dividing the number of inferred HGT events by the branch length. Accordingly, HGT rates are expressed as the number of HGT events per point substitution per amino acid site.

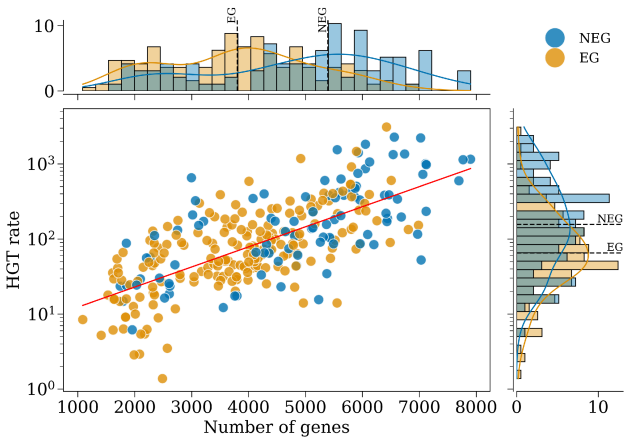


Fig. 2. Bacteria transitioning to new ecosystems tend to have small genomes and low HGT rates. As seen in the top marginal histogram, branches with ecosystem gains (EG, orange) tend to have a smaller genome size than branches with no ecosystem gains (NEG, blue). Branches with ecosystem gains also tend to have lower HGT rates, as shown in the right marginal histogram. Histograms show percentages; dotted lines show medians. Each point in the scatter plot represents a branch in the genome tree on which at least one gene was gained through HGT. The regression line is $log(y) = 6.65 + 6.16 \times 10^{-4} \cdot x$, with a Pearson correlation coefficient $r = 0.87$.
ALT TEXT: HGT rate is a function of genome size, as shown in the scatter plot. Marginal plots show the distribution of genome sizes (top) and HGT rates (right) for branches with and without ecosystem gains.

Contrary to the expectation that HGT rates would increase during transitions to new ecosystems, branches with ecosystem gains exhibited lower HGT rates than branches without (right marginal plot in Fig. 2; Mann-Whitney $U = 8306.5$, two-sided $p = 1.5 \times 10^{-4}, N_{EG} = 208, N_{NEG} = 108$). This trend persists when the analysis is restricted to single-copy gene families (Supplementary Figure S1).

Transitions to new ecosystems do not affect HGT rates when controlling for genome size. Thus, contrary to our expectations, we found that genomes undergoing transitions to new ecosystems tend to be small and experience low rates of HGTs. Given that smaller genomes generally exhibit lower rates of horizontal gene transfer (scatter plot in Fig. 2; Spearman $\rho = 0.530, p = 2.8 \times 10^{-24}$; see also Cordero and Hogeweg (2009)), it is conceivable that the lower HGT rates at ecosystem gains are simply a consequence of the smaller genome sizes. To test this conjecture, we fitted two linear models: a full model with HGT rate as a function of genome size and ecosystem gain (as a binary trait), and a reduced model that considered genome size only. A likelihood ratio test comparing these nested models showed no significant improvement when using the full model ($LRT = 1.62, p = 0.203, df = 1$), indicating that when accounting for genome size, transitions to new ecosystems do not significantly affect HGT rate. Note that this analysis considered all gene families together. It is conceivable that ecosystem changes influence the HGT rate for specific functional categories; this possibility is explored below.

HGT rates do not change at transitions to pathogenic lifestyles. To test for the robustness of the finding that transitions to new ecosystems are not associated with changes in HGT rates, we used a similar pipeline for the examination of an independent dataset related to a different type of ecological change: the transition of *Escherichia coli* and *Shigella* strains from a commensal to a pathogenic lifestyle. Bacterial pathogenicity has been associated with changes in environmental conditions, such as a transition of a non-host associated organism to a host environment, but also transitions between different ecosystems within the host (Nuss et al., 2016, Chin et al., 2018, Brown et al., 2006).

To analyze transitions to pathogenicity, we examined 151 *E. coli* and *Shigella* genomes retrieved from NCBI, comprising 98 commensal and 53 pathogenic strains. For these genomes, we identified 5,503 orthologous gene families with OrthoFinder (Emms and Kelly, 2019) and estimated HGT rates using Count. Based on binary labels indicating whether a given genome is pathogenic, we also used Count to infer the gain and loss of pathogenicity. Pathogenicity gains were detected in just 14 of 300 evolutionary branches, a notably low frequency compared to ecosystem gains. As seen in the top marginal histogram in Fig. 3, branches with pathogenicity gains did not have significantly smaller genomes (median 4,497 genes) than those without pathogenicity gains (median 4,721 genes; Mann-Whitney $U = 1489.5$, two-sided $p = 0.122, N_{PG} = 14, N_{NPG} = 287$).

Unlike in the ecosystem analyses, HGT rates did not increase significantly with genome size (Spearman $\rho = 0.075, p = 0.198$). Thus, whatever difference exists in HGT rates between branches with and without pathogenicity gains, they are unlikely to be caused by genome size effects.

As seen in the right marginal histogram in Fig. 3, median HGT rates are slightly lower in branches with pathogenicity gains (5.03×10^{-4}) than in those without (1.19×10^{-5}); however, this difference is not statistically significant (Mann-Whitney $U = 2137.0$, two-sided $p = 0.526, N_{PG} = 14, N_{NPG} = 287$). In sum, ecological shifts – whether through adaptations to pathogenic lifestyles or the colonization of new ecosystems – are not associated with elevated HGT rates.

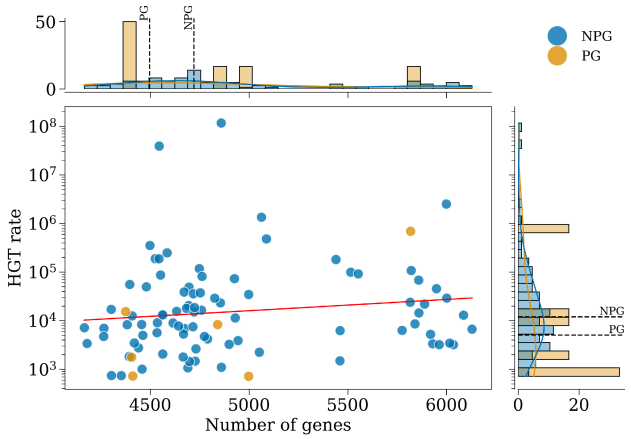


Fig. 3. Transitions to pathogenic lifestyles are not associated with increased HGT rates or smaller genomes. As seen in the top marginal histogram, branches with pathogenicity gains (PG, orange) do not have significantly smaller genomes than branches without pathogenicity gains (NPG, blue). Branches with pathogenicity gains also do not have significantly higher HGT rates, as shown in the right marginal histogram. Histograms show percentages; dotted lines show medians. Each point in the scatter plot represents a branch in the genome tree on which at least one gene was gained through HGT. The regression line is $log(y) = 1102.20 + 5.35 \times 10^{-4} \cdot x$, with a Pearson correlation coefficient $r = 0.03$.
ALT TEXT: HGT rate is a function of genome size, as shown here in the scatter plot. Marginal plots show the distribution of genome sizes (top) and HGT rates (right) for branches with and without pathogenicity gains.

Ecosystem transitions affect HGT rates in only two functional gene categories. We saw above that when all gene families are analyzed collectively, ecosystem gains have no significant effect on HGT rates after controlling for genome size. However, ecosystem gains may still influence HGT rates in specific functional categories. To investigate this possibility, we examined each functional category defined in the Clusters of Orthologous Groups (COG) database (Tatusov et al., 2000, Galperin et al., 2021) that included at least 20 gene families in our dataset. We performed likelihood ratio tests (LRTs) as described earlier, comparing a model that includes both genome size and ecosystem gains as linear predictors of HGT rates with a nested model that considers only genome size.

As shown in Fig. 4 (see also Table S1), the only HGT rates significantly affected by ecosystem gains are those for the COG categories 'Amino acid transport and metabolism' (adjusted $p = 3.1 \times 10^{-5}$ after correction for multiple testing (Benjamini and Hochberg, 1995)) and 'Transcription' (adjusted $p = 0.046$). In both cases, HGT rates are lower on branches with ecosystem gains.

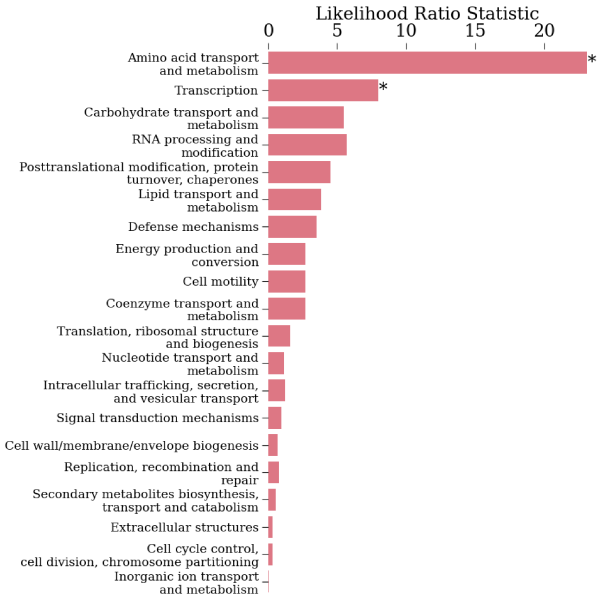


Fig. 4. Functional categories with significant ecosystem gain effects on HGT rates. This barplot displays Log Likelihood Ratio statistics from likelihood ratio tests (LRTs) comparing two models of HGT rates: one as a function of genome size and ecosystem gain, and another with genome size alone. Tests were performed separately for Clusters of Orthologous Groups (COG) functional categories. Asterisks (*) denote categories where ecosystem gain significantly influenced HGT rates ($p < 0.05$ after FDR correction via the Benjamini-Hochberg method). Notably, transitions to new ecosystems were associated with lower median HGT rates in all of the categories shown here.

Discussion

Our analyses reveal a slight reduction in the strength of purifying selection during transitions to new ecosystems. A global relaxation of purifying selection would be expected under the assumption that colonizing bacteria typically undergo population bottlenecks. Such bottlenecks may allow new mutations to escape elimination quickly, contributing to higher proportions of non-synonymous substitutions compared to synonymous ones. However, that only a small minority of genes appears to experience changes in selection pressures suggests that population bottlenecks – which would affect selection globally – may be less important at bacterial colonization events than widely assumed.

Contrary to expectations, we found that genomes transitioning to new ecosystems tend to be small. The observed trend might reflect a trade-off between adaptability, favored by large genomes, and efficiency, favored by small genomes. Although larger genomes could ease transitions by providing broader molecular toolkits or homologous gene landing pads for HGT (Maslov et al., 2009, Taylor et al., 2024), many genes may be irrelevant to the new environment, imposing maintenance costs for redundant genetic material (Stepkowski and Legocki, 2001). Accordingly, bacteria with smaller genomes may be able to replicate faster and use resources more efficiently, traits that might be favored during colonization of new environments.

Again contradicting our expectations, we found that neither

ecosystem gains nor gains of pathogenicity are associated with increased HGT rates. Instead, bacteria colonizing new ecosystems exhibit significantly reduced HGT rates – a phenomenon that is entirely attributable to their smaller genome sizes (Fig. 2, scatter plot) due to the established relationship between genome size and HGT (Cordero and Hogeweg, 2009). After controlling for genome size, only two functional categories of genes show an effect of ecosystem gains on HGT rates: 'Amino acid transport and metabolism' and 'Transcription' (which includes transcription factors). For these, HGT rates are actually lower at transitions to new ecosystem.

We would like to note two limitations of our study. First, the scopes of 'Ecosystem Type' labels in the GOLD database can vary substantially; e.g., the scope of 'Fermented vegetables' may not be comparable to that of 'Freshwater' aquatic environments. A broader ecosystem label may result in fewer inferred transitions, missing out on the narrower niches that may be a part of the broader ecosystem. Second, while sampling issues are not a significant concern in the gene families we analyzed, given the scale and dense taxon sampling of the dataset, the same cannot be said for the ecosystem labels. The number of genomes in the dataset for each ecosystem label is not uniform, and most genomes in GOLD are sampled in only one or a few ecosystems. On one hand, missing ecosystem labels may result in certain ecosystem gains not being inferred; on the other hand, missing labels for a subset of a given clade may also lead to the inference of a false positive ecosystem gain when in reality the clade's ancestor was already adapted to that ecosystem. While both limitations likely add noise to our dataset, its power to show biologically relevant signals is confirmed by our observations of (i) smaller genome sizes at ecosystem gains and (ii) a strong association between genome size and HGT rates. Moreover, both limitations are unlikely to bias our results in any systematic manner.

HGT is often considered the dominant mechanism of bacteria to adapt to new environments (Pang and Lercher, 2019, Pál et al., 2005, Arnold et al., 2021). Moreover, both a weakening of purifying selection and an increase of positive selection – as observed in the nucleotide substitution patterns – would be expected to be associated with higher HGT rates. How can these expectations be reconciled with our finding that HGT rates are not affected by transitions to new ecosystems (after controlling for genome size)? It is possible that a colonizing bacterium may occupy a different microniche than resident species due to different metabolic preferences, spatial segregation, or competitive exclusion, thereby reducing the physical interactions needed for HGT (Dmitrijeva et al., 2024, Polz et al., 2013, Foster and Bell, 2012). In addition, the colonizing bacterium's recombination machinery may not be compatible with resident mobile genetic elements (Johnson and Nolan, 2009). While these conjectures might potentially explain – at least in part – the lack of evidence for increased HGT rates at transitions to new ecosystems, more research will be needed to fully understand this surprising observation. Overall, our findings suggest that bacterial colonization events may be less driven by adaptive benefits of HGT than previously thought.

Methods

Data retrieval and processing for the analysis of ecosystem gains. We used extant bacterial genomic data obtained from the EggNOG database v6 (Hernández-Plaza et al., 2023), which clusters genes into non-supervised orthologous groups (NOGs), referred to as gene families in the main text. For the inference of transitions to new environments, we used the GOLD database, which contains metagenomic data from a variety of ecosystems. GOLD follows a hierarchical labeling system, and the 'Ecosystem Type' level provides the most fine-grained and diversified set of labels such that most genomes have a label at this level, unlike, for example, the 'Ecosystem Subtype' level. Our analysis was restricted to bacterial genomes in the GOLD database (v9) annotated with at least three Ecosystem Type labels (Mukherjee et al., 2023), leading to a dataset of 159 genomes across 80 diverse Ecosystem types.

We used the NCBI Datasets API to retrieve nucleotide sequences for each genome (O'Leary et al., 2024). Nucleotide sequences were used to prepare codon alignments for each NOG, translated from the amino acid multiple sequence alignments provided by EggNOG using PAL2NAL (Suyama et al., 2006).

COG functional categories were retrieved from the EggNOG database itself, which provides a mapping of NOGs to COG categories (Galperin et al., 2021).

We reconstructed the genome tree topology using ASTRAL-Pro 2 (Zhang and Mirarab, 2022) based on 233 single-copy gene families present in at least 95% of the genomes. The EggNOG database provides multiple sequence alignments for each NOG, which we used to estimate the branch lengths of the genome tree. The branch lengths were estimated with IQ-TREE 2 (Minh et al., 2020) using the `Q.pfam+I+R8` model, based on a concatenation of the multiple sequence alignments. The genome tree was rooted using the Minimum Ancestor Deviation method (Tria et al., 2017), which has been shown to be more accurate than other rooting methods (Wade et al., 2020).

Data retrieval and processing for the analysis of pathogenicity gains. Pathogenic and commensal *E.coli* and *Shigella* genomes were retrieved from NCBI RefSeq using the NCBI Datasets API. Strain genomes with at least 95% CheckM completeness and without suppressed assembly status were retained. CheckM estimates genome completeness and contamination based on the presence of lineage-specific single-copy marker genes (Parks et al., 2015). The pathogenicity of each genome was inferred from the 'Pathogenicity' label in the assembly attributes. OrthoFinder (Emms and Kelly, 2019) was used to infer orthologous groups of genes, and corresponding gene trees were estimated using IQ-TREE 2 with the 'JTT' substitution model. The genome tree topology and branch lengths were estimated using ASTRAL-Pro 2 and IQ-TREE 2, respectively, as described above for the ecosystem analysis.

Inference of HGT and environmental transitions. HGT was inferred using Count (Csűös, 2010) with asymmetric Wagner parsimony. For inferences of ecosystem gains and pathogenicity gains, the gain penalty ratio for Count was set to 1.0, which means that gains and losses are penalized equally. Based on a benchmark on empirical data, a gain penalty ratio of 7.0 was found to be optimal for HGT inference using Count, and we used this value for HGT inferences (Mishra and Lercher, 2024).

HGT rates were estimated as the number of HGT events per unit branch length, where the branch length is in units of average amino acid substitutions per site. This normalization allows for a fair comparison of HGT rates across branches of different lengths.

Inference of changes in selection intensities. We estimated the ratio of non-synonymous to synonymous substitutions ($\omega = dN/dS$) and the selection intensification/relaxation parameter (K) for 3,228 non-supervised orthologous groups (NOGs) using the HyPhy RELAX method (Wertheim et al., 2015). Our analysis was restricted to NOGs for which we could reliably retrieve nucleotide sequences for all associated taxa, with at least two branches in both the ecosystem gain (EG) and no ecosystem gain (NEG) branch sets, and for which the RELAX analysis completed successfully.

The RELAX algorithm fits a codon model to the nucleotide sequences of each NOG to estimate an ω value for every branch in the genome tree. The method then compares the distribution of ω values on EG branches with those on NEG branches, estimating the parameter K, which models the change in selection pressure. A K value greater than 1 indicates an intensification of selection pressure on EG branches, while a value less than 1 indicates a relaxation. Finally, a likelihood ratio test is used to assess whether a model allowing for different selection pressures ($K \neq 1$) provides a significantly better fit to the data than a null model assuming equal pressure on both branch sets ($K = 1$).

Likelihood ratio tests. Likelihood ratio tests (LRTs) were used to test for the effect of ecosystem gains, pathogenicity gains, as well as the effect of ecosystem gains on specific functional categories. These tests compared generalised linear models (GLMs) that used as predictors either only genome size (a continuous variable) or both genome size and the binary 'environmental gain' variable. The GLMs were fitted using the `GLM` function in Python Statsmodels (Seabold and Perktold, 2010) using a log-link function. The p-values from LRTs were adjusted for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). For the likelihood ratio test (LRT) of HGT rates in functional categories, we used the COG functional categories as defined in the COG database (Tatusov et al., 2000, Galperin et al., 2021). We performed LRTs for each COG category with at least 20 NOGs in our dataset.

Code and data availability

The code used in the analysis of ecosystem transitions is available in the following Gitlab repository: https://gitlab.cs.uni-duesseldorf.de/general/ccb/hgt_rates_and_ecosystem_transfers.
For the pathogenicity analysis, it is at: https://gitlab.cs.uni-duesseldorf.de/general/ccb/hgt_rates_and_pathogenicity_gains.
The data used or generated in this study can be retrieved from Zenodo: https://doi.org/10.5281/zenodo.14555204.

Competing interests

The authors have declared no competing interest.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1310 and by the Volkswagenstiftung through the "Life?" initiative.

Bibliography

Danping Zheng, Timur Liwinski, and Eran Elinav. Interaction between microbiota and immunity in health and disease. *Cell Research*, 30(6):492–506, June 2020. ISSN 1748-7838. doi: 10.1038/s41422-020-0332-7.

Tami D. Lieberman. Detecting bacterial adaptation within individual microbiomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1861):20210243, August 2022. doi: 10.1098/rstb.2021.0243.

Anne Richter, Felix Blei, Guohai Hu, Jan W. Schwitalla, Carlos N. Lozano-Andrade, Jiyu Xie, Scott A. Jarmusch, Mario Wibowo, Bodil Kjeldgaard, Surabhi Surabhi, Xinming Xu, Theresa Jautzus, Christopher B. W. Phippen, Olaf Tyc, Mark Arentshorst, Yue Wang, Paolina Garbeva, Thomas Ostenfeld Larsen, Arthur F. J. Ram, Cees A. M. van den Hondel, Gergely Maróti, and Ákos T. Kovács. Enhanced surface colonisation and competition during bacterial adaptation to a fungus. *Nature Communications*, 15(1):4486, May 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-48812-1.

Gabriella Caruso. Microbial Colonization in Marine Environments: Overview of Current Knowledge and Emerging Research Topics. *Journal of Marine Science and Engineering*, 8(2):78, February 2020. ISSN 2077-1312. doi: 10.3390/jmse8020078.

Sergei Maslov, Sandeep Krishna, Tin Yau Pang, and Kim Sneppen. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proceedings of the National Academy of Sciences*, 106(24):9743–9748, June 2009. doi: 10.1073/pnas.0903206106.

Balázs Szappanos, Jonathan Fritzemeier, Bálint Csörgő, Viktória Lázár, Xiaowen Lu, Gergely Fekete, Balázs Bálint, Róbert Herczeg, István Nagy, Richard A. Notebaart, Martin J. Lercher, Csaba Pál, and Balázs Papp. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nature Communications*, 7(1):11607, May 2016. ISSN 2041-1723. doi: 10.1038/ncomms11607.

Lucas Serra Moncadas, Cyrill Hofer, Paul-Adrian Bulzu, Jakob Pernthaler, and Adrian-Stefan Andrei. Freshwater genome-reduced bacteria exhibit pervasive episodes of adaptive stasis. *Nature Communications*, 15(1):3421, April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-47767-7.

Aidan J. Taylor, Koji Yahara, Ben Pascoe, Seungwon Ko, Leonardos Mageiros, Evangelos Mourkas, Jessica K. Calland, Santeri Puranen, Matthew D. Hitchings, Keith A. Jolley, Carolin M. Kobras, Sion Bayliss, Nicola J. Williams, Arnoud H. M. van Vliet, Julian Parkhill, Martin C. J. Maiden, Jukka Corander, Laurence D. Hurst, Daniel Falush, Paul Keim, Xavier Didelot, David J. Kelly, and Samuel K. Sheppard. Epistasis, core-genome disharmony, and adaptation in recombining bacteria. *mBio*, 15(6):e00581–24, April 2024. doi: 10.1128/mbio.00581-24.

Tin Yau Pang and Martin J. Lercher. Each of 3,323 metabolic innovations in the evolution of E. coli arose through the horizontal transfer of a single DNA segment. *Proceedings of the National Academy of Sciences*, 116(1):187–192, January 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1718997115.

Csaba Pál, Balázs Papp, and Martin J. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12):1372–1375, December 2005. ISSN 1546-1718. doi: 10.1038/ng1686.

Brian J. Arnold, I.-Ting Huang, and William P. Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, pages 1–13, November 2021. ISSN 1740-1534. doi: 10.1038/s41579-021-00650-4.

Jan Engelstädter and Danesh Moradigaravand. Adaptation through genetic time travel? Fluctuating selection can drive the evolution of bacterial transformation. *Proceedings of the Royal*

Society B: Biological Sciences, 281(1775):20132609, January 2014. doi: 10.1098/rspb.2013.2609.

Marija Dmitrijeva, Janko Tackmann, João Frederico Matias Rodrigues, Jaime Huerta-Cepas, Luis Pedro Coelho, and Christian von Mering. A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. *Nature Ecology & Evolution*, pages 1–13, March 2024. ISSN 2397-334X. doi: 10.1038/s41559-024-02357-0.

Laura C. Woods, Rebecca J. Gorrell, Frank Taylor, Tim Connallon, Terry Kwok, and Michael J. McDonald. Horizontal gene transfer potentiates adaptation by reducing selective constraints on the spread of genetic variation. *Proceedings of the National Academy of Sciences*, 117 (43):26868–26875, October 2020. doi: 10.1073/pnas.2005331117.

Ying-Xian Goh, Sai Manohar Balu Anupoju, Anthony Nguyen, Hailong Zhang, Monica Ponder, Leigh-Anne Krometis, Amy Pruden, and Jingqiu Liao. Evidence of horizontal gene transfer and environmental selection impacting antibiotic resistance evolution in soil-dwelling Listeria. *Nature Communications*, 15(1):10034, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54459-9.

Katayoun Dadeh Amirfard, Momoko Moriyama, Satoru Suzuki, and Daisuke Sano. Effect of environmental factors on conjugative transfer of antibiotic resistance genes in aquatic settings. *Journal of Applied Microbiology*, 135(6):lxae129, June 2024. ISSN 1364-5072. doi: 10.1093/jambio/lxae129.

Hande Acar Kirit, Mato Lagator, and Jonathan P. Bollback. Experimental determination of evolutionary barriers to horizontal gene transfer. *BMC microbiology*, 20(1):326, October 2020. ISSN 1471-2180. doi: 10.1186/s12866-020-01983-5.

Bram van Dijk, Paulien Hogeweg, Hilje M Doekes, and Nobuto Takeuchi. Slightly beneficial genes are retained by bacteria evolving DNA uptake despite selfish elements. *eLife*, 9:e56801, May 2020. ISSN 2050-084X. doi: 10.7554/eLife.56801.

Charles Coluzzi, Martin Guillemet, Fanny Mazzamurro, Marie Touchon, Maxime Godfroid, Guillaume Achaz, Philippe Glaser, and Eduardo P C Rocha. Chance Favors the Prepared Genomes: Horizontal Transfer Shapes the Emergence of Antibiotic Resistance Mutations in Core Genes. *Molecular Biology and Evolution*, 40(10):msad217, October 2023. ISSN 1537-1719. doi: 10.1093/molbev/msad217.

Supratim Mukherjee, Dimitri Stamatis, Cindy Tianqing Li, Galina Ovchinnikova, Jon Bertsch, Jagadish Chandrabose Sundaramurthi, Mahathi Kandimalla, Paul A Nicolopoulos, Alessandro Favognano, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy. Twenty-five years of Genomes OnLine Database (GOLD): Data updates and new features in v.9. *Nucleic Acids Research*, 51(D1):D957–D963, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac974.

Ana Hernández-Plaza, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, Thomas Rattei, Ivica Letunic, Lars J Jensen, Peer Bork, Christian von Mering, and Jaime Huerta-Cepas. eggNOG 6.0: Enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, 51(D1):D389–D394, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1022.

Chao Zhang and Siavash Mirarab. ASTRAL-Pro 2: Ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics*, 38(21):4949–4950, November 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac620.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa015.

Miklós Csűös. Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, August 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq315.

Joel O. Wertheim, Ben Murrell, Martin D. Smith, Sergei L. Kosakovsky Pond, and Konrad Scheffler. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Molecular Biology and Evolution*, 32(3):820–832, March 2015. ISSN 0737-4038. doi: 10.1093/molbev/msu400.

Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher Lanczycki, Stacy Lathrop, Zhiyong Lu, Francoise Thibaud-Nissen, Terence Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D20–D26, January 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1112.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Swastik Mishra and Martin J. Lercher. Horizontal Gene Transfer Inference: Gene presence-absence outperforms gene trees, December 2024.

Otto X. Cordero and Paulien Hogeweg. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of the National Academy of Sciences*, 106(51):21748–21753, December 2009. doi: 10.1073/pnas.0907584106.

Aaron Mischa Nuss, Franziska Schuster, Louisa Roselius, Johannes Klein, René Bücker, Katharina Herbst, Ann Kathrin Heroven, Fabio Pisano, Christoph Wittmann, Richard Münch, Johannes Müller, Dieter Jahn, and Petra Dersch. A Precise Temperature-Responsive Bistable Switch Controlling Yersinia Virulence. *PLOS Pathogens*, 12(12):e1006091, December 2016. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006091.

Chui Yoke Chin, Kyle A. Tipton, Marjan Farokhyfar, Eileen M. Burd, David S. Weiss, and Philip N. Rather. A high-frequency phenotypic switch links bacterial virulence and environmental survival in Acinetobacter baumannii. *Nature Microbiology*, 3(5):563–569, May 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0151-5.

Nat F. Brown, Mark E. Wickham, Brian K. Coombes, and B. Brett Finlay. Crossing the Line: Selection and Evolution of Virulence Traits. *PLOS Pathogens*, 2(5):e42, May 2006. ISSN 1553-7374. doi: 10.1371/journal.ppat.0020042.

David M. Emms and Steven Kelly. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1):1–14, December 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1832-y.

Roman L. Tatusov, Michael Y. Galperin, Darren A. Natale, and Eugene V. Koonin. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):33–36, January 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.33.

Michael Y. Galperin, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, January 2021. ISSN 1362-4962. doi: 10.1093/nar/gkaa1018.

T. Stepkowski and A. B. Legocki. Reduction of bacterial genome size and expansion resulting from obligate intracellular lifestyle and adaptation to soil habitat. *Acta Biochimica Polonica*, 48(2):367–381, 2001. ISSN 0001-527X.

Martin F. Polz, Eric J. Alm, and William P. Hanage. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in genetics: TIG*, 29(3):170–175, March 2013. ISSN 0168-9525. doi: 10.1016/j.tig.2012.12.006.

Kevin R. Foster and Thomas Bell. Competition, not cooperation, dominates interactions among culturable microbial species. *Current biology: CB*, 22(19):1845–1850, October 2012. ISSN 1879-0445. doi: 10.1016/j.cub.2012.08.005.

Timothy J. Johnson and Lisa K. Nolan. Pathogenomics of the Virulence Plasmids of Escherichia coli. *Microbiology and Molecular Biology Reviews : MMBR*, 73(4):750–774, December 2009. ISSN 1092-2172. doi: 10.1128/MMBR.00015-09.

Nuala A. O'Leary, Eric Cox, J. Bradley Holmes, W. Ray Anderson, Robert Falk, Vichet Hem, Mirian T. N. Tsuchiya, Gregory D. Schuler, Xuan Zhang, John Torcivia, Anne Ketter, Laurie Breen, Jonathan Cothran, Hena Bajwa, Jovany Tinne, Peter A. Meric, Wratko Hlavina, and Valerie A. Schneider. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Scientific Data*, 11(1):732, July 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03571-y.

Mikita Suyama, David Torrents, and Peer Bork. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl_2):W609–W612, July 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl315.

Fernando Domingues Kümmel Tria, Giddy Landan, and Tal Dagan. Phylogenetic rooting using minimal ancestor deviation. *Nature Ecology & Evolution*, 1(1):1–7, June 2017. ISSN 2397-334X. doi: 10.1038/s41559-017-0193.

Taylor Wade, L. Thiberio Rangel, Soumya Kundu, Gregory P. Fournier, and Mukul S. Bansal. Assessing the accuracy of phylogenetic rooting methods on prokaryotic gene families. *PLOS ONE*, 15(5):e0232950, May 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0232950.

Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, July 2015. ISSN 1549-5469. doi: 10.1101/gr.186072.114.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference 2010 (SciPy 2010), Austin, Texas, June 28 - July 3, 2010*, page 92. scipy.org, 2010. doi: 10.25080/MAJORA-92BF1922-011.

Supplementary Information

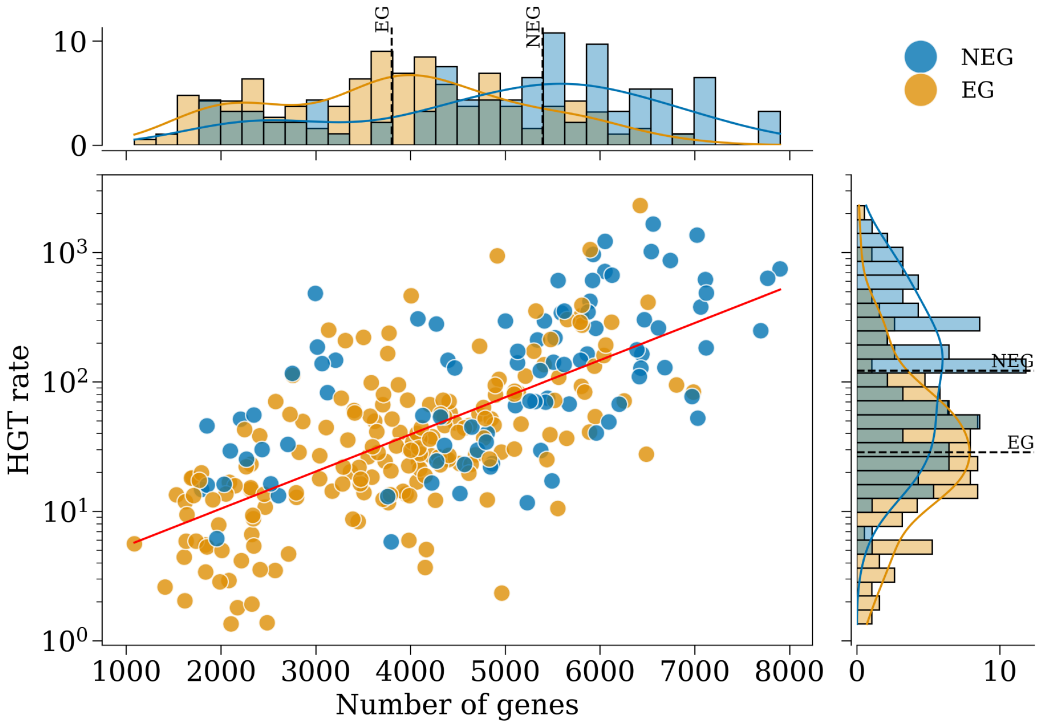


Fig. S1. Bacteria transitioning to new ecosystems tend to have small genomes and low HGT rates (single-copy NOGs). As seen in the top marginal histogram, branches with ecosystem gains (EG, orange) tend to have a smaller genome size than branches with no ecosystem gains (NEG, blue; Mann-Whitney $U = 5884.0$, two-sided $p = 3.9 \times 10^{-12}$, $N_{EG} = 208$, $N_{NEG} = 108$, median genome size on EG branches $= 3,800$ and on NEG branches $= 5,395$). Branches with ecosystem gains also tend to have lower HGT rates, as shown in the right marginal histogram (Mann-Whitney $U = 7622.5$, two-sided $p = 3 \times 10^{-6}$, $N_{EG} = 208$, $N_{NEG} = 108$, median HGT rate on EG branches $= 28.53$ and on NEG branches $= 122.14$). Histograms show percentages; dotted lines show medians. Each point in the scatter plot represents a branch in the genome tree on which at least one gene was gained through HGT. The regression line is $log(y) = 2.79 + 6.61 \times 10^{-4} \cdot x$, with a Pearson correlation coefficient $r = 0.47$.

ALT TEXT: HGT rate is a function of genome size, as shown in the scatter plot. Marginal plots show the distribution of genome sizes (top) and HGT rates (right) for branches with and without ecosystem gains.

COG category	Median HGT rate NEG	EG	Log Likelihood Ratio	p-value
Amino acid transport and metabolism	16.01	5.16	23.09	0.00
Transcription	26.94	6.08	8.00	0.05
Carbohydrate transport and metabolism	15.75	8.68	5.46	0.10
RNA processing and modification	11.42	3.68	5.71	0.10
Posttranslational modification, protein turnover, chaperones	16.47	5.47	4.53	0.13
Lipid transport and metabolism	9.84	6.02	3.83	0.17
Defense mechanisms	17.87	6.02	3.52	0.17
Energy production and conversion	13.75	7.94	2.69	0.20
Cell motility	14.45	5.62	2.71	0.20
Coenzyme transport and metabolism	13.02	7.51	2.72	0.20
Translation, ribosomal structure and biogenesis	12.83	5.81	1.59	0.38
Nucleotide transport and metabolism	6.36	5.74	1.18	0.43
Intracellular trafficking, secretion, and vesicular transport	7.61	5.13	1.24	0.43
Signal transduction mechanisms	17.26	10.12	0.98	0.46
Cell wall/membrane/envelope biogenesis	16.27	8.61	0.72	0.50
Replication, recombination and repair	12.71	7.00	0.79	0.50
Secondary metabolites biosynthesis, transport and catabolism	15.28	5.73	0.59	0.52
Extracellular structures	12.44	6.75	0.32	0.60
Cell cycle control, cell division, chromosome partitioning	16.84	8.05	0.32	0.60
Inorganic ion transport and metabolism	16.47	7.16	0.06	0.80

Table S1. COG categories with effect of ecosystem gain on HGT rates. All categories shown here have a lower median HGT rate at ecosystem gains (EG) than without ecosystem gains (NEG). The LRT p-values are corrected for multiple testing using the Benjamini-Hochberg method. They suggest that only the categories 'Amino acid transport and metabolism' and 'Transcription' are significantly affected by ecosystem gains.

Chapter 4

The two phases of losing horizontally acquired genes: rapid initial turnover followed by long-term persistence

At the time of submission of the thesis, this manuscript is under consideration at Molecular Biology & Evolution.
I designed the study, performed the analyses, and drafted the manuscript.

The fate of horizontally acquired genes: rapid initial turnover followed by long-term persistence

Swastik Mishra[1], **Käthe Weit**[1], **and Martin J. Lercher**[1] ✉

[1]Institute for Computer Science and Department of Biology, Heinrich Heine University, Düsseldorf, Germany

A major driver of bacterial evolution is horizontal gene transfer (HGT), the acquisition of genes from other strains or species. Transfers between closely related taxa are more likely to succeed, while the pervasive deletion bias of bacterial genomes drives frequent turnover of horizontally acquired genes. However, whether the rate of gene loss after acquisition is constant across lineages or time remains unclear. Here, we analyze a comprehensive dataset of bacterial genomes to infer the frequency, distribution, and retention of inter-phylum HGT events. The retention of inter-phylum gene transfers is highly skewed, with only a small subset of bacterial genomes accounting for the majority of such events. Most transferred genes are lost rapidly. Those genes that survive the initial purging are retained over long periods, are biased toward functions such as transport and metabolism, and have larger numbers of protein-protein interactions.

Introduction

Horizontal gene transfer (HGT) is a key driver of prokaryotic evolution (1, 2), even if its role in environmental adaptations may have been overestimated (3). Unlike vertical inheritance, HGT facilitates the exchange of genetic elements across phylogenetic boundaries, profoundly shaping bacterial evolution and ecological diversification.

Evidence indicates that most horizontally acquired genes originate from phylogenetically related donors rather than from distantly related taxa. This pattern likely reflects both selective pressures and mechanistic barriers to gene flow between divergent lineages, such as constraints imposed by genome architecture, functional compatibility, and integration with host regulatory networks (4–7). Although horizontal transfer of genes between bacteria from different phyla (inter-phylum HGT) is less frequent than within-phylum transfer, it has nonetheless been documented in notable cases, for instance, in the transfer of genes from archaea to bacteria at the origin of certain archaeal clades (8, 9). While such events are considered rare, their impact on recipient genomes can be substantial. Although it is difficult to estimate the frequency of inter-phylum HGT, Sheinman et al. (7) reported that approximately 8% of bacterial genomes harbor DNA segments identical to those in other phyla, indicating that inter-phylum gene transfer is not uncommon. Nevertheless, because these DNA segments may not always function as expressed genes, the functional relevance of such horizontally transferred sequences – and whether all bacterial clades are equally susceptible to inter-phylum HGT – remains unclear.

Furthermore, studies on sets of closely related species indicate that the retention of newly acquired genes is often transient; empirical studies indicate that the majority of horizontally transferred genes are rapidly lost following their initial acquisition (10–12). However, it is unclear if this pattern of rapid loss holds generally across bacteria.

Here, we employ a systematic, sequence-identity-based method for HGT inference to analyze patterns of gene loss across a comprehensive dataset of $> 33{,}000$ bacterial genomes. Our approach differentiates between very recent and slightly older inter-phylum HGT events and examines their persistence, functional categorization, and involvement in protein–protein interaction networks. The horizontal acquisition of potentially functional genes from other phyla is very rare; however, a quarter of the genomes that contain any inter-phylum gene acquisitions harbor genes from multiple gene families. We observe that horizontally acquired genes are initially lost at a high rate; those retained appear to be very stable afterwards and display distinct functional characteristics compared to genes eliminated early. Collectively, these findings support a two-phase model of post-transfer gene loss and broaden our understanding of the evolutionary forces shaping the retention of horizontally acquired genes.

Results and Discussion

Inference of recent inter-phylum gene transfers. We analyzed 35,439 gene families from 33,918 extant bacterial genomes, obtained from the EggNOG database v6 (13), which clusters genes into non-supervised orthologous groups (NOGs; referred to as gene families below). Horizontal gene transfer (HGT) was inferred by identifying, within a gene family, pairs of genes with high sequence identity that are found in two different phyla as defined by NCBI Taxonomy (14). In this framework, a horizontally acquired gene is represented as a pair of homologous genes from different taxa. Sequence identity — the percentage of identical amino acids – was calculated from the multiple sequence alignments (MSA) of EggNOG gene families. The distribution of such gene pairs across percent sequence identities is shown in Fig. 1. The direction of transfer (i.e., which gene in the pair is from the donor taxon and which from the recipient) was inferred by examining the outgroup of the most

recent common ancestor in the gene tree; if one phylum is represented in the outgroup and the other is not, directionality can be assigned, i.e., we can identify which genome is the recipient and which is the donor (see Methods). The vast majority of gene families contain no inter-phylum gene pairs with sequence identities above 80%. We find candidate inter-phylum HGT gene pairs in 796 gene families and 4,445 genomes for downstream analyses.

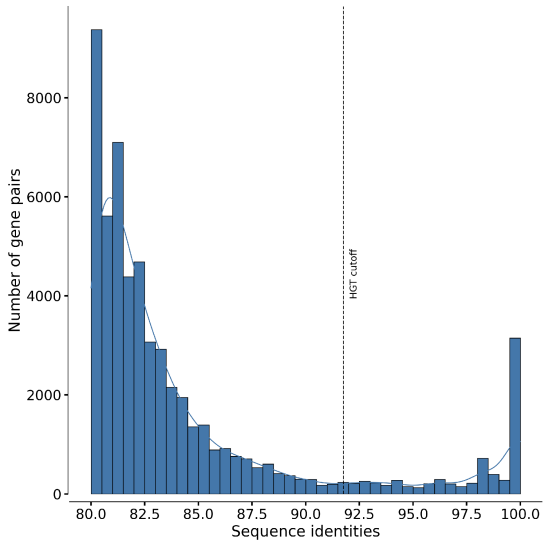


Fig. 1. Recent inter-phylum HGTs can be distinguished from vertically inherited genes from sequence identities. Distribution of sequence identities of gene pairs in a gene family such that the pair of genes are from two different phyla. The dotted line indicates the estimated cutoff for HGT inference, above which we consider the gene pairs to be horizontally transferred.

The overall distribution of inter-phylum gene pairs exhibits three major trends: a prominent peak at 100% sequence identity, an approximately uniform distribution at intermediate sequence identities, and a gradual increase as sequence identity decreases further. For comparison, we performed the same analysis to identify candidate inter-class and inter-order HGT pairs. For these, the increase in frequency begins at much higher sequence identities (Suppl. Fig. S1). This pattern suggests that gene pairs with lower sequence identity primarily result from vertical inheritance, as vertically inherited genes within the same family will be more similar among more closely related taxa. We conclude that pairs of genes from different taxonomic groupings with high sequence identities, including those at 100% sequence identity, are predominantly the results of HGT.

Accordingly, we applied a sequence identity cutoff of 91.75% to identify gene pairs from inter-phylum HGT. This threshold corresponds to the point at which the smoothed frequency distribution of sequence identities flattens when moving from lower to higher sequence identities (see Methods and Suppl. Fig. S2).

For two highly similar homologous genes found in two genomes from different phyla, sequence identity is an approximate measure of the number of substitutions since their last common ancestor, and can thus serve as a proxy for the time that has passed since the HGT event. Even without the assumption of a molecular clock, one can distinguish approximately between recent and older transfers by considering the inter-phylum gene pairs that are identical versus the ones that are not. Accordingly, we classified pairs with 100% sequence identity as *recent* transfers, whereas those with sequence identities below 100% but above the cutoff are regarded as *older* (albeit still relatively recent) transfers. The majority of inter-phylum gene pairs identified as HGTs (84.51% or 4,245 of 5,023) are older transfers.

Inferences of HGT from high sequence similarity between phyla are highly precise. Neighboring genes are frequently co-acquired (15–17), and inference methods that identify a higher fraction of such co-acquisitions are generally considered to infer fewer false positives. Accordingly, HGT inference methods can be benchmarked based on how often co-acquisitions were inferred for neighboring genes in the target genome (18). We used this strategy to check the accuracy of our sequence identity-based HGT inference method, which does not use position information of the genes for its inferences. It appears to perform on par with the best-performing methods in Mishra and Lercher (18); note, however, that the performances are not directly comparable, as our dataset differs from the one used in the previous benchmark.

Across all inter-phylum transfers, 3.83% of gene pairs inferred to be co-acquired within the same genome are neighbors in the target genome. Among co-transferred gene pairs (i.e., pairs of HGTs where both genes have the same donor and recipient genome, the majority (81.09%) are neighbors in the target genome.

We estimated the probability of finding a single identical gene pair in our dataset that results from vertical inheritance from the common ancestor of two phyla – the family-wise error rate (FWER) – to be approximately 2.81×10^{-10} (see Methods). This low rate of false positives, combined with the high percentage of neighboring co-acquisitions, suggests that the inferred inter-phylum HGT events are reliable, although the actual number of inter-phylum transfers may be higher than what we can detect in this dataset.

Inter-phylum gene acquisitions are concentrated in a few genomes. The distribution of the number of inter-phylum HGTs per genome is highly skewed, with most genomes (98.16%; 33294 of 33918) not hosting any inter-phylum HGTs that are recent enough to pass our filters. Only 0.57% (195 of 33918) of the genomes in our dataset (195 out of 33,918) are involved in any relatively recent inter-phylum HGT event for which we could infer the direction of transfer. Of these, only about a quarter (53 out of 195) show evidence of recent transfers, while most (180 out of 195) only harbour older transfers. Only about a quarter of the genomes involved in inter-phylum HGTs (26.7%, or 52 out of 195) received genes from more than one gene family, and only 3.1% (6 out of 195) received genes from more than 5 gene families (Fig. 2a) These genomes belong to diverse phyla, with no single phylum dominating the distribution of inter-phylum

HGTs (see Suppl. Table S1 for lineage information). Fig. 2b shows that recent inter-phylum HGTs are also concentrated in a handful of bacterial genomes.

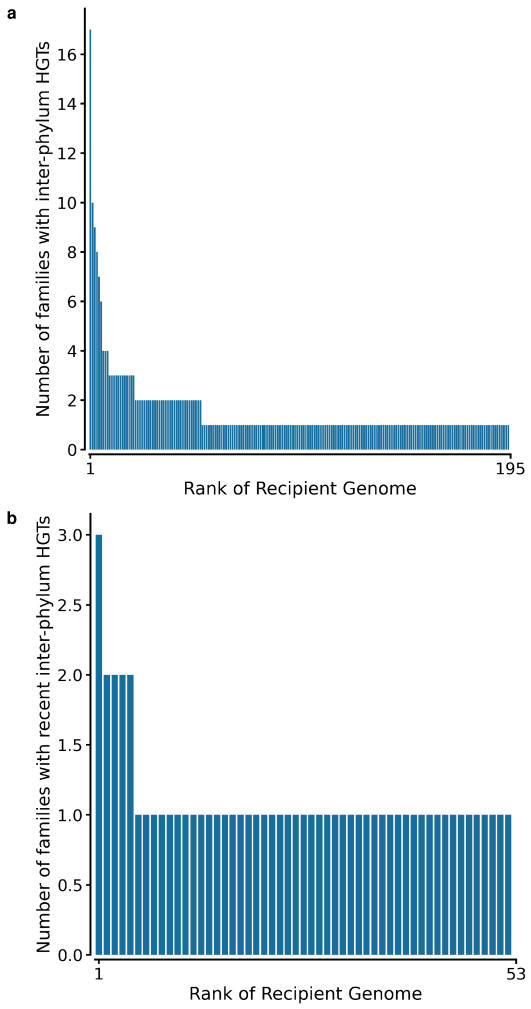


Fig. 2. The distribution of inter-phylum HGTs is strongly skewed. Rank plot of the number of inter-phylum gene acquisitions per genome. The x-axis shows the genomes ranked by the number of inter-phylum HGTs, while the y-axis shows the number of gene families with inter-phylum HGTs where the direction of transfer is known. **(a)** All inter-phylum HGTs. **(b)** Recent inter-phylum HGTs.

The loss rate of horizontally acquired genes shows two phases. The pronounced peak of inter-phylum gene pairs at 100% sequence identity in Fig. 1 indicates that a large proportion of horizontally acquired genes are lost shortly after their acquisition: only few of these genes are still present after enough time has passed for the gene pair to have accumulated even a small number of substitutions. The near-uniform distribution across lower identity levels implies that very few genes are lost at later stages. The pronounced peak at 100% sequence identity, immediately followed by a near-uniform distribution, is also seen for inter-order and inter-class HGT (Suppl. Fig. S1). Thus, gene loss following HGT appears to generally follow a two-phase dynamic: an early phase of rapid loss, followed by long-term retention of the remaining genes.

Recent and older transfers differ systematically in functional categories and connectivity. How do horizontally acquired genes that persist in their host genome differ from those eliminated in the early, high-turnover phase? To answer this question, we used Clusters of Orthologous Groups (COG) functional category annotations in EggNOG. We performed a Fisher's exact test to check if the proportion of HGT pairs in a given COG category differs between recent and older HGT pairs. After correcting for multiple hypothesis testing (19), we observed that compared to recent inter-phylum acquisitions, older transfers are significantly depleted in the COG categories of (J) translation, ribosomal structure and biogenesis, (K) transcription, (L) replication, recombination and repair, (U) intracellular trafficking, secretion, and vesicular transport, and (V) defense mechanisms. Conversely, we observed a significant enrichment in older gene acquisitions in a number of transport and metabolism categories: (E) amino acid transport and metabolism, (H) coenzyme transport and metabolism, and (P) inorganic ion transport and metabolism. This result is consistent with the complexity hypothesis (20), which posits that genes that are successfully transferred belong more frequently to operational functional categories, such as metabolism and transport, than to informational ones, such as transcription and translation.

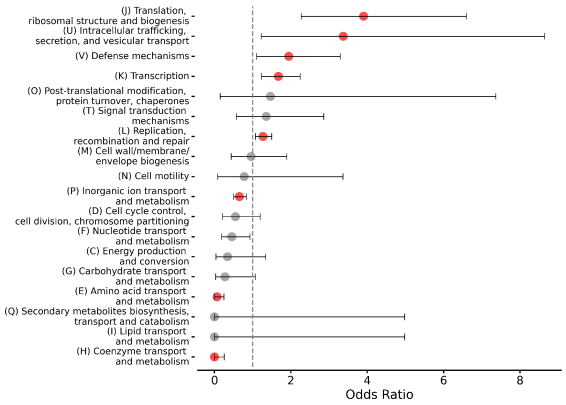


Fig. 3. Recent and older inter-phylum transfers differ systematically in their functions. Odds ratios (OR) of COG categories depleted (OR>1) and enriched (OR<1) in older transfers. Error bars show 95% confidence intervals. Red markers indicate COG categories that are significantly enriched (adjusted $p \leq 0.05$ after correcting for multiple testing). The dashed line at OR=1 indicates no difference between recent and older transfers.

The rationale behind the complexity hypothesis is that genes in the information-processing functional categories are more complex in the sense of having more interaction partners, and are thus less likely to integrate successfully into the interaction network of a new host (20, 21). One therefore expects that genes transferred more frequently have a lower connectivity in terms of protein-protein interactions (PPI), a prediction that was previously confirmed in a much more limited dataset of older transfers (21).

We used the STRING database v11 (22) to obtain the number of PPI for every gene in our HGT dataset based on STRING's own gene family clustering and a minimum STRING interaction score of 900 (max: 1000). Although we observe a

marginal negative correlation between the number of HGTs inferred for a gene family and the mean PPI of the gene family (Spearman's ρ = -0.14 and -0.11 for recent and older transfers, respectively), neither of these correlations is statistically significant ($p = 0.08$ and $p = 0.18$ respectively; see Suppl. Fig. S3).

However, we find that the mean PPI across all gene pairs involved in older transfers is significantly higher than that across all gene pairs in recent transfers (6.95 vs. 5.99, $p = 0.033$ from a permutation test). Thus, genes involved in older transfers have on average one more protein-protein interaction partner than those involved in recent transfers.

Discussion

Our analysis shows that inter-phylum transfers of potentially functional genes are rare: less than 2% of genomes have been involved – either as donor or recipient – in any inter-phylum HGTs that our methodology can infer. However, when focusing on those events for which we can infer the direction of transfer, a quarter of those genomes that contain any inter-phylum gene acquisition harbor acquisitions from multiple gene families. We observed that the probability of a randomly selected genome harboring an inter-phylum gene acquisition is 0.57%. Consequently, if all inter-phylum gene acquisitions occurred independently, the expected probability for a genome to acquire a second acquisition after the first would again be 0.57%, two orders of magnitude lower than the observed 26.7%. This observation suggests that once a genome has acquired a gene from another phylum, it becomes considerably more likely to acquire additional inter-phylum genes.

When inter-phylum gene acquisitions occur, the acquired genes follow a biphasic trajectory. A brief, intense purge removes most of the newcomers, while the remaining genes appear to be surprisingly stable and are retained over extended time periods. The fast, early attrition matches qualitatively the findings of Puigbò et al. (10), who reported a loss rate roughly three times higher than the rate of acquisition. Lineage-specific bursts of inter-phylum gene acquisitions have been described for individual species and genera (23, 24), lending further support to our findings.

The two phases of losing genes acquired through inter-phylum HGT likely reflect the same evolutionary processes that are responsible for the pronounced differences between intra-population polymorphisms, which are characterized by negative selection and drift, and inter-population substitutions, characterized by positive selection (25, 26). Thus, it appears likely that the most recent inter-phylum HGTs detected in our study are not fixed in the respective populations, but are polymorphism specific to the genomes that were deposited in the EggNOG database.

The complexity hypothesis (20) predicts that operational genes are more frequently transferred than informational ones. Our results align with this prediction: compared to very recent inter-phylum gene acquistions, older (yet still relatively recent) transfers are depleted in categories such as transcription, translation, replication/recombination/repair, intracellular trafficking, and defense – several of which represent informational functions. In contrast, genes that survived the early loss phase predominantly participate in operational roles, such as transport and metabolism.

The original formulation of the complexity hypothesis emphasized – without empirical support – the greater connectivity and functional entanglement of informational genes, a view that was supported by a more limited, earlier analysis of older transfers across different phylogenetic distances (21). However, our data reveal a different trend. Genes that undergo initial purging of horizontally acquired genes have, on average, more interaction partners than the most recent transfers. Thus, the observed functional biases likely reflect functional adaptations more than the sheer complexity initially proposed.

Our sequence similarity-based HGT inference approach offers precision that – based on neighboring co-acquisitions – is comparable to the best state-of-the-art methods at highest stringency, but focussed on relatively recent HGT events. However, it is inherently limited to detecting HGT between gene pairs covered by the dataset examined. Increasing both genome and gene family diversity would enable more accurate detection of gene similarities across clades, yielding a more complete picture of inter-phylum HGT. Sheinman et al. (7) observed that 8% of bacterial genomes share identical DNA segments with other phyla. That we observed a lower number of less than 2% likely reflects our focus on genes rather than DNA segments, as well as a more limited dataset. While the results of the two studies are not directly comparable, this discrepancy suggests that most inter-phylum DNA segment exchanges may not yield functional genes.

Together, our findings clarify the rarity and rapid attrition of inter-phylum horizontal gene transfers in bacteria, while highlighting the selective processes shaping their long-term retention. Future studies that combine broader genome sampling will be essential to map the ecological and molecular contexts in which such transfers persist.

Materials and Methods

Data and implicit phylogenetic method for HGT inference. Sequence identity, the number of substitutions, mismatches, and other relevant metrics were calculated for every gene pair within each NOG using multiple sequence alignments (MSA) from the EggNOG database. EggNOG also provides COG functional annotations for each gene family. Taxonomic information was retrieved from the NCBI Taxonomy via the Entrez API implemented through the Biopython library (14, 27). To infer the direction of HGT, gene trees available in EggNOG were used to identify appropriate outgroups. PPI data were obtained from STRING v11, considering only pairs of gene families with a minimum interaction score of 900 out of 1000 (22). Chromosomal positions of genes were also extracted from the STRING database.

To improve tractability and reduce noise, analyses were performed on a curated subset of the data. Specifically, only NOGs containing at least 10 taxa (to exclude very small groups) and a maximum of 2,000 genes were included. This subset comprised 33,918 genomes representing 35,439

NOGs across all bacterial phyla. For downstream analyses in this study, we further focused only on the 4,445 genomes containing 796 NOGs in which gene pairs exhibited sequence identity above 80% and originated from two different phyla. A subset of these gene pairs are inferred to be horizontally transferred based on our sequence identity cutoff.

Although recent transfers are defined as those with 100% sequence identity, to consider older transfers we identify a sequence identity cutoff based on the point where distribution of sequence identities flattens out. The distribution was smoothed using a rolling average with a window size of 5% sequence identity across bins of width 0.5%, excluding the bin for identical sequence matches. The cutoff was set at the point where the derivative of the smoothed distribution reached zero, indicating a plateau in the distribution.

Family-wise error rate for recent HGT. We assume that substitutions at each site follow a Poisson process. Then, the probability that a site undergoes no amino acid substitutions over a branch length d (the mean number of substitutions per site) is e^{-d}. For a gene of length L, the probability that all sites are unchanged is then

$$P_{\text{identical}} = e^{-dL}$$

Across M independent gene pairs, the probability that no pair is identical by chance is $(1 - P_{\text{identical}})^M$. Then, the probability that at least one gene pair is identical by chance – the family-wise error rate (FWER) – is one minus this probability, i.e.,

$$\text{FWER} = 1 - (1 - e^{-dL})^M \quad .$$

For our dataset, we estimate d as 4.94 substitutions per site, M as $\approx 10^9$ (number of gene pairs), and L as 267.41 (average gene length in number of amino acids). This leads to an estimate of FWER ≈ 0, i.e., a very low probability of false positives in our HGT inference.

Outgroup for inferring direction of transfer and co–transfers. For any potential HGT in a gene family, the pair of genes must have two different taxonomic groupings at the taxonomic level of interest. Let A and B be two taxonomic groupings represented in a gene family tree, with gene members A_1 and B_1. If the outgroup of A_1 and B_1 has genes from A and not from B, then the direction of transfer is from the taxon containing A_1 to the taxon containing B_1, and vice versa. If two pairs of genes are inferred to be transferred from the same donor taxon to the same recipient taxon, we define it as a co-transfer. If they have the same recipient taxon, but not necessarily the same donor taxon, we define it as a co-acquisition. We limit our analysis of co-transfers and co-acquisitions to the gene pairs where the direction of transfer could be inferred.

Functional categories for PPI analysis. We use the COG (Clusters of Orthologous Groups) functional categories as defined in NCBI COG (28), since EggNOG gene families already use these annotations for categorizing the NOGs. Fisher's exact test was performed with a contingency table comparing the number of recent vs older HGT gene pairs, in a COG vs those not in the COG. The reliance on gene pairs introduces a potential for overestimation: in cases where several genomes within the same family are highly similar and only one belongs to a different phylum, multiple gene pairs may be labeled as HGT even if they result from a single transfer event.

Permutation tests were performed with 10^4 iterations of randomly shuffling whether a gene pair is a recent or older transfer, and then calculating the mean PPI for the pair of genes. The p-value is calculated as the fraction of iterations where the mean PPI of the older transfers is greater than that of the recent transfers. These gene families across the two sets are not independent, since a gene family can have both recent and older transfers. However, the permutation test is still valid, since the null hypothesis is that the mean PPI of the two sets of gene pairs is equal.

Code and data availability

The code used for the analyses in this study is available on Gitlab (https://gitlab.cs.uni-duesseldorf.de/general/ccb/imli).

The corresponding data is available at Zenodo (https://doi.org/10.5281/zenodo.16745487).

Competing interests

The authors have declared no competing interest.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC 1310 and by the Volkswagenstiftung through the "Life?" initiative.

Bibliography

1. Brian J. Arnold, I.-Ting Huang, and William P. Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4):206–218, April 2022. ISSN 1740-1534. doi: 10.1038/s41579-021-00650-4.
2. Csaba Pál, Balázs Papp, and Martin J. Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12):1372–1375, December 2005. ISSN 1546-1718. doi: 10.1038/ng1686.
3. Swastik Mishra and Martin J Lercher. Streamlined genomes, not horizontal gene transfer, mark bacterial transitions to unfamiliar environments | bioRxiv. *bioRxiv*, 2024. doi: 10.1101/2024.12.27.630308.
4. Christina L Burch, Artur Romanchuk, Michael Kelly, Yingfang Wu, and Corbin D Jones. Empirical Evidence That Complexity Limits Horizontal Gene Transfer. *Genome Biology and Evolution*, 15(6):evad089, May 2023. ISSN 1759-6653. doi: 10.1093/gbe/evad089.
5. Robert G. Beiko, Timothy J. Harlow, and Mark A. Ragan. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14332–14337, October 2005. ISSN 0027-8424. doi: 10.1073/pnas.0504068102.
6. Heather L. Hendrickson, Dominique Barbeau, Robin Ceschin, and Jeffrey G. Lawrence. Chromosome architecture constrains horizontal gene transfer in bacteria. *PLoS Genetics*, 14(5):e1007421, May 2018. ISSN 1553-7390. doi: 10.1371/journal.pgen.1007421.
7. Michael Sheinman, Ksenia Arkhipova, Peter F Arndt, Bas E Dutilh, Rutger Hermsen, and Florian Massip. Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. *eLife*, 10:e62719, June 2021. ISSN 2050-084X. doi: 10.7554/eLife.62719.
8. William F. Martin and Filipa L. Sousa. Early Microbial Evolution: The Age of Anaerobes. *Cold Spring Harbor Perspectives in Biology*, 8(2):a018127, December 2015. ISSN 1943-0264. doi: 10.1101/cshperspect.a018127.
9. Shijulal Nelson-Sathi, Filipa L. Sousa, Mayo Roettger, Nabor Lozada-Chávez, Thorsten Thiergart, Arnold Janssen, David Bryant, Giddy Landan, Peter Schönheit, Bettina Siebers, James O. McInerney, and William F. Martin. Origins of major archaeal clades correspond to

gene acquisitions from bacteria. *Nature*, 517(7532):77–80, January 2015. ISSN 1476-4687. doi: 10.1038/nature13805.

10. Pere Puigbò, Alexander E. Lobkovsky, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Biology*, 12(1):1–19, December 2014. ISSN 1741-7007. doi: 10.1186/s12915-014-0066-4.
11. Weilong Hao and G. Brian Golding. The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Research*, 16(5):636–643, May 2006. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.4746406.
12. Emmanuelle Lerat, Vincent Daubin, Howard Ochman, and Nancy A. Moran. Evolutionary Origins of Genomic Repertoires in Bacteria. *PLOS Biology*, 3(5):e130, April 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030130.
13. Ana Hernández-Plaza, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, Thomas Rattei, Ivica Letunic, Lars J Jensen, Peer Bork, Christian von Mering, and Jaime Huerta-Cepas. eggNOG 6.0: Enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research*, 51(D1): D389–D394, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1022.
14. Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, January 2020. ISSN 1758-0463. doi: 10.1093/database/baaa062.
15. Tin Yau Pang and Martin J. Lercher. Supra-operonic clusters of functionally related genes (SOCs) are a source of horizontal gene co-transfers. *Scientific Reports*, 7(1):40294, January 2017. ISSN 2045-2322. doi: 10.1038/srep40294.
16. Alexander Dilthey and Martin J. Lercher. Horizontally transferred genes cluster spatially and metabolically. *Biology Direct*, 10(1):72, December 2015. ISSN 1745-6150. doi: 10.1186/s13062-015-0102-5.
17. Tin Yau Pang and Martin J. Lercher. Each of 3,323 metabolic innovations in the evolution of E. coli arose through the horizontal transfer of a single DNA segment. *Proceedings of the National Academy of Sciences*, 116(1):187–192, January 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1718997115.
18. Swastik Mishra and Martin J Lercher. Horizontal Gene Transfer Inference: Gene Presence–Absence Outperforms Gene Trees. *Molecular Biology and Evolution*, 42(7):msaf166, July 2025. ISSN 1537-1719. doi: 10.1093/molbev/msaf166.
19. Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x.
20. Ravi Jain, Maria C. Rivera, and James A. Lake. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences*, 96(7):3801–3806, March 1999. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.96.7.3801.
21. Ofir Cohen, Uri Gophna, and Tal Pupko. The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer. *Molecular Biology and Evolution*, 28(4):1481–1489, April 2011. ISSN 0737-4038. doi: 10.1093/molbev/msq333.
22. Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database Issue):D433–D437, January 2005. ISSN 0305-1048. doi: 10.1093/nar/gki005.
23. Hyaekang Kim, Woori Kwak, Sook Hee Yoon, Dae-Kyung Kang, and Heebal Kim. Horizontal gene transfer of Chlamydia: Novel insights from tree reconciliation. *PLOS ONE*, 13(4): e0195139, April 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0195139.
24. Alejandro Caro-Quintero and Konstantinos T Konstantinidis. Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria. *The ISME Journal*, 9(4):958–967, April 2015. ISSN 1751-7362. doi: 10.1038/ismej.2014.193.
25. Jun Gojobori, Hua Tang, Joshua M. Akey, and Chung-I Wu. Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. *Proceedings of the National Academy of Sciences*, 104(10):3907–3912, March 2007. doi: 10.1073/pnas.0605565104.
26. Austin L. Hughes, Robert Friedman, Pierre Rivailler, and Jeffrey O. French. Synonymous and Nonsynonymous Polymorphisms versus Divergences in Bacterial Genomes. *Molecular Biology and Evolution*, 25(10):2199–2209, October 2008. ISSN 0737-4038. doi: 10.1093/molbev/msn166.
27. Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, June 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp163.
28. Michael Y. Galperin, Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, 49(D1):D274–D281, January 2021. ISSN 1362-4962. doi: 10.1093/nar/gkaa1018.

Supplementary Information

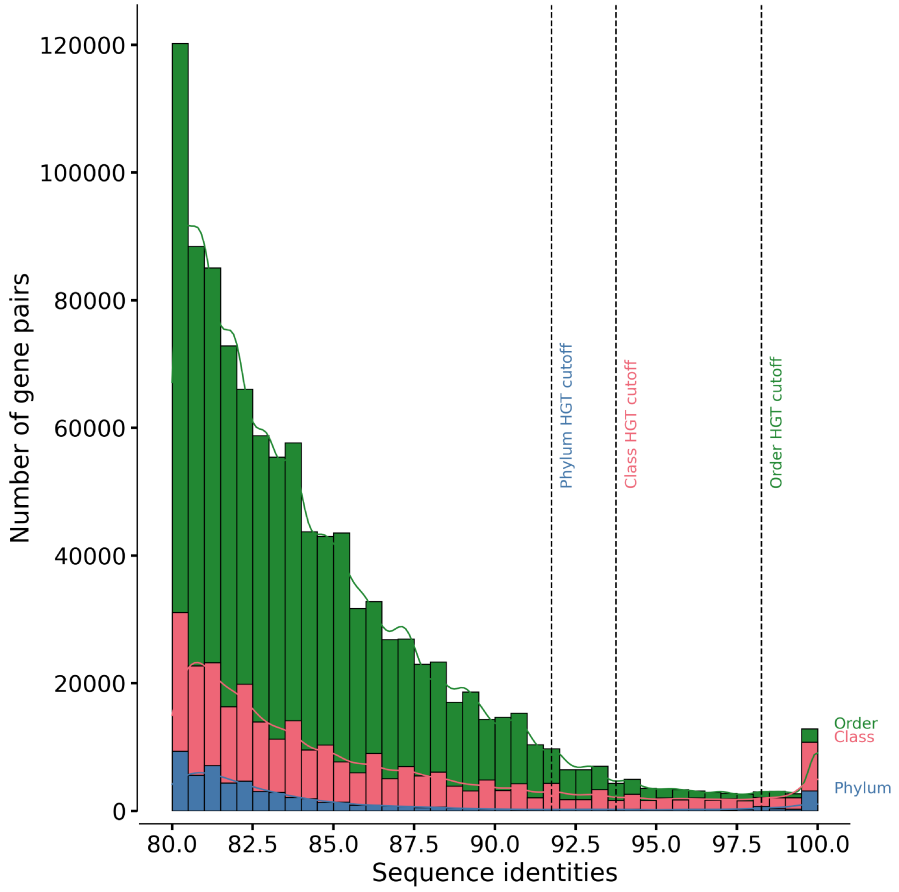


Fig. S1. Distributions of the number of highly similar inter-clade gene pairs at different sequence identities are qualitatively similar at inter-phylum, inter-class, and inter-order levels. The dashed lines indicate the estimated cutoff for HGT inference, above which we consider the gene pairs to be horizontally transferred.

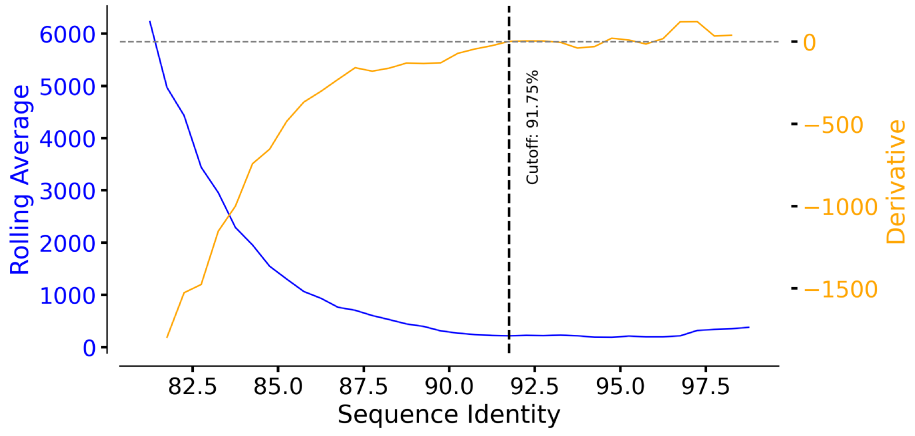


Fig. S2. The cutoff for purely horizontal gene acquisition can be determined from the derivative of the sequence identity histogram. The figure shows the rolling window average of the sequence identity histogram shown in Fig. 1 and its derivative. The dashed line indicates the cutoff for HGT inference, above which we consider the gene pairs to be horizontally transferred. The cutoff is chosen as the point where the rolling average flattens out, i.e., the derivative reaches zero for the first time when moving from lower to higher sequence identities.

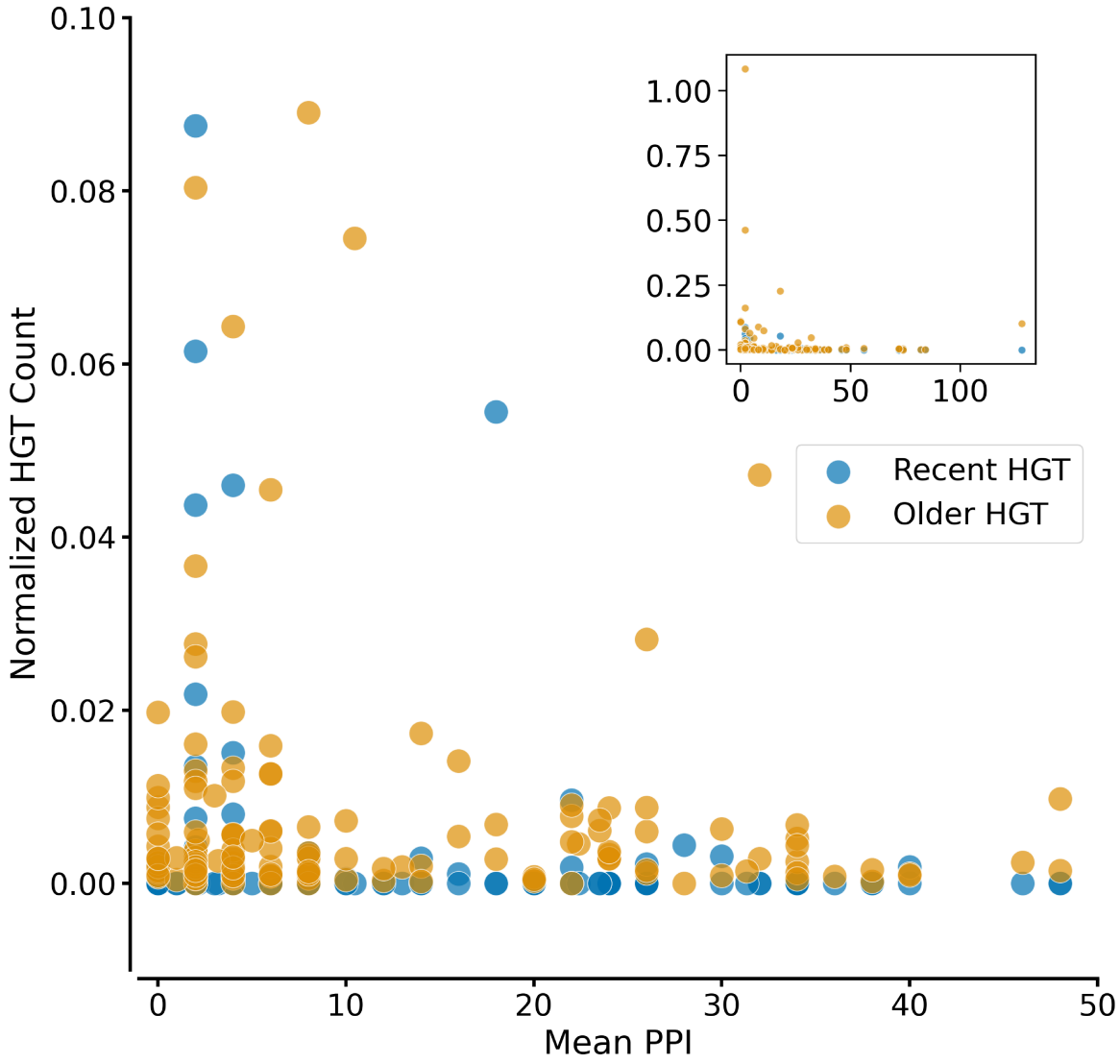


Fig. S3. HGT counts correlate at most weakly with the number of protein-protein interactions. Scatter plot of the normalized number of HGTs inferred for each gene family and its mean number of PPI (connectivity). The x-axis shows the mean PPI across all transferred genes in the gene family, and the y-axis shows the number of HGTs inferred for the gene family, divided by the total number of genes in the gene family for normalization. The inset shows the full figure, while the main figure shows a zoomed-in view of the data.

Recipient Taxon ID	Phylum	Class	Order	Family	Genus	Species/Strain
1120933	Actinobacteria	Actinomycetia	Actinomycetales	Actinomycetaceae	Actinotignum	Actinotignum urinale DSM 15805
330214	Nitrospirae	Nitrospira	Nitrospirales	Nitrospiraceae	Nitrospira	Nitrospira defluvii
469615	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	Fusobacterium gonidiaformans ATCC 25563
552811	Chloroflexi	Dehalococcoidia	NaN	NaN	Dehalogenimonas	Dehalogenimonas lykanthroporepellens BL-DC-9
653733	Chrysiogenetes	Chrysiogenetes	Chrysiogenales	Chrysiogenaceae	Desulfurispirillum	Desulfurispirillum indicum S5
938289	Firmicutes	Clostridia	Eubacteriales	NaN	Levyella	Levyella massiliensis

Table S1. Lineage information for the genomes receiving more than 5 gene families with inter-phylum HGTs. If strain information was missing, the species name is shown instead.

Chapter 5

Outlook

This chapter synthesizes the findings of the dissertation and discusses their implications for our understanding of prokaryotic genome evolution. The first part highlights areas of research required for methodological advances. The second part discusses the conceptual implications of the findings and raises key questions for future research.

The need for more accurate phylogenetic trees

Since direct observation of long-term evolutionary processes is impossible, as discussed in Chapter 2, benchmarking inference models on expected patterns in empirical data provides a practical way to assess their relative precision in inferring biologically meaningful events. The empirical benchmark presented in this thesis indicates that gene tree-based approaches generally underperform compared to presence-absence-based methods. In the presence of noise and inaccuracies in gene tree reconstructions (Than et al., 2008, 2007), even vast amounts of data may not lead to reliable inferences. The idea of simplicity-over-sophistication is also supported by the finding that the best-performing models are maximum parsimony methods, and not maximum-likelihood based reconciliation methods, even if maximum-likelihood methods are generally considered more accurate in other contexts such as tree estimation (Gadagkar and Kumar, 2005).

This does not mean that reconciliation-based methods are without merit; in particular, they may be essential when one needs to infer the donor of an HGT event, and the extra information of gene tree topology may be beneficial if accurate. Rather, these findings highlight the need for caution in interpreting the results of gene tree reconciliation. Improving the reliability of such methods will depend on advances in gene tree estimation, for example by using species tree-aware algorithms as described in Morel et al. (2020). However, this strategy introduces a conceptual challenge: species tree-aware approaches require an accurate species tree that is ideally independent of the gene trees themselves. Yet, state-of-the-art species tree estimations are typically constructed from the consensus of multiple gene trees or their subtrees (Davidson et al., 2015; Jiang et al., 2020; Rannala et al., 2020; Yang, 2014), making it difficult to avoid circularity.

Even in the absence of circularity, for example when using concatenated alignments for species tree estimation, there is no information in the species tree and the gene tree that can be used to distinguish between reconciliations based on inaccuracies in gene tree estimation and based on actual evolutionary events. This problem can potentially be solved by the use of additional information, such as ecological/environmental contexts or functional associations, helping to disentangle the processes that shape the gene trees and species trees.

The importance of traveling light

Given the higher phenotypic plasticity of bacteria with larger genomes and their larger "toolboxes" of molecular machines, it may seem intuitive that they can adapt more easily to new environments (Maslov et al., 2009). However, the results of Chapter 3 indicate that smaller genomes, which incur

lower genetic costs and generally show reduced HGT rates, are in fact more successful at adapting to novel habitats than larger genomes with a larger toolbox of functions. These streamlined genomes could be explained by long-term specialization in stable environments, but our findings also show that rapid loss of acquired genes plays a role, which is consistent with corresponding findings from experimental studies that focus on specific bacteria (Chaudhari et al., 2024; Lee and Marx, 2012).

While deletion bias is well documented in prokaryotic genomes (Mira et al., 2001) and genome streamlining has been described as the result of adaptation to stable, resource-limited niches, potentially decreasing adaptability (Giovannoni et al., 2014), our findings suggest a broader view. Streamlined genomes appear not only as evolutionary end products, but as features that may actively facilitate or enhance adaptation to new environments, presumably by reducing genetic "baggage". To what extent does such streamlining aid adaptation, and when does it instead restrict evolutionary flexibility? The answer may lie in the nature of genes acquired for specific functions in specific environmental contexts. If these genes are not contributing significantly to fitness in the new environment, they are likely to be lost quickly, and vice versa. The quality of the acquired genes in terms of their functional relevance (see Chapter 4) and the ecological context in which they are acquired or lost may therefore be much more important than the quantity of genes gained or lost.

The counterintuitive finding that smaller genomes with lower HGT rates tend to adapt more often to new environments questions the assumptions underlying our understanding of prokaryotic evolution. Even if the lower HGT rates are entirely accounted for by the smaller genome sizes, we found no evidence of any increase in HGT rates in environmental transitions. This means that at the very least, three key assumptions need to be revisited. First, HGT may not be the primary driver of adaptation; mutations or other mechanisms could play a more decisive role. Second, the magnitude of gene influx may not scale with adaptive challenges: rather than requiring a large repertoire of novel or optimized functions, adaptation – even to markedly different ecosystems – may instead be facilitated by genome streamlining reinforced by a few key gene acquisitions. A quality-over-quantity principle may thus apply at the genomic level. Third, transitioning to a new ecosystem may not inherently present a greater adaptive challenge than remaining in a seemingly stable one, either because such environments are less stable than presumed or because the ecological shift is not as drastic as traditionally assumed. In Chapter 3, we use ecosystem labels extracted from the GOLD database (Mukherjee et al., 2019) at the level of "Ecosystem Type"; perhaps a more fine-grained approach could reveal additional insights. Notably, our analysis does not incorporate possible influences of environmental fluctuations or other time-dependent ecological factors. Moreover, we find no strong evidence linking transitions to new ecosystems with population bottlenecks. Overall, the ecological context of gene gain and loss remains underexplored, emphasizing the need to expand our understanding of prokaryotic evolution by integrating eco-evolutionary frameworks. Such frameworks can, for instance, consider conserved genes across communities (Blanchet et al., 2023), or map functional traits such as metabolic strategies or stress tolerances onto phylogenies to directly link ecological functions with evolutionary history (Krause et al., 2014).

Taken together, several key questions regarding the biology of bacterial adaptation remain unresolved. To what extent is HGT central to prokaryotic adaptation? Are transitions to new ecosystems more challenging than remaining in stable ecosystems? How do the functional roles and ecological contexts of acquired genes shape their likelihood of retention or loss? The findings of this thesis underscore the necessity of systematically investigating the influence of ecological factors, and of rigorously incorporating ecological and functional information, to unravel the forces governing gene gain and loss and to clarify their significance for prokaryotic adaptation and long-term evolutionary patterns.

Bibliography

Acar Kirit, Hande, Mato Lagator, and Jonathan P. Bollback (Oct. 2020). "Experimental Determination of Evolutionary Barriers to Horizontal Gene Transfer". In: *BMC microbiology* 20.1, p. 326. ISSN: 1471-2180. DOI: 10.1186/s12866-020-01983-5 (cit. on p. 3).

Andreani, Nadia Andrea, Elze Hesse, and Michiel Vos (July 2017). "Prokaryote Genome Fluidity Is Dependent on Effective Population Size". In: *The ISME Journal* 11.7, pp. 1719–1721. ISSN: 1751-7370. DOI: 10.1038/ismej.2017.36. (Visited on 09/18/2022) (cit. on p. 1).

Arnold, Brian J., I.-Ting Huang, and William P. Hanage (Apr. 2022). "Horizontal Gene Transfer and Adaptive Evolution in Bacteria". In: *Nature Reviews Microbiology* 20.4, pp. 206–218. ISSN: 1740-1534. DOI: 10.1038/s41579-021-00650-4. (Visited on 08/07/2022) (cit. on pp. 1, 3).

Blanchet, Simon, Laura Fargeot, and Allan Raffard (2023). "Phylogenetically-Conserved Candidate Genes Unify Biodiversity–Ecosystem Function Relationships and Eco-Evolutionary Dynamics across Biological Scales". In: *Molecular Ecology* 32.16, pp. 4467–4481. ISSN: 1365-294X. DOI: 10.1111/mec.17043. (Visited on 08/11/2025) (cit. on p. 40).

Bradley, Patrick H., Stephen Nayfach, and Katherine S. Pollard (Aug. 2018). "Phylogeny-Corrected Identification of Microbial Gene Families Relevant to Human Gut Colonization". In: *PLOS Computational Biology* 14.8, e1006242. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006242. (Visited on 07/28/2025) (cit. on p. 1).

Caruso, Gabriella (Feb. 2020). "Microbial Colonization in Marine Environments: Overview of Current Knowledge and Emerging Research Topics". In: *Journal of Marine Science and Engineering* 8.2, p. 78. ISSN: 2077-1312. DOI: 10.3390/jmse8020078. (Visited on 04/10/2025) (cit. on p. 2).

Chaudhari, Narendrakumar M., Olga M. Pérez-Carrascal, Will A. Overholt, Kai U. Totsche, and Kirsten Küsel (Dec. 2024). "Genome Streamlining in Parcubacteria Transitioning from Soil to Groundwater". In: *Environmental Microbiome* 19.1. ISSN: 2524-6372. DOI: 10.1186/s40793-024-00581-6. (Visited on 08/07/2025) (cit. on p. 40).

Cordero, Otto X. and Paulien Hogeweg (Dec. 2009). "The Impact of Long-Distance Horizontal Gene Transfer on Prokaryotic Genome Size". In: *Proceedings of the National Academy of Sciences* 106.51, pp. 21748–21753. DOI: 10.1073/pnas.0907584106. (Visited on 05/16/2023) (cit. on p. 2).

Dadeh Amirfard, Katayoun, Momoko Moriyama, Satoru Suzuki, and Daisuke Sano (June 2024). "Effect of Environmental Factors on Conjugative Transfer of Antibiotic Resistance Genes in Aquatic Settings". In: *Journal of Applied Microbiology* 135.6, lxae129. ISSN: 1364-5072. DOI: 10.1093/jambio/lxae129. (Visited on 12/23/2024) (cit. on p. 3).

Davidson, Ruth, Pranjal Vachaspati, Siavash Mirarab, and Tandy Warnow (Oct. 2015). "Phylogenomic Species Tree Estimation in the Presence of Incomplete Lineage Sorting and Horizontal Gene Transfer". In: *BMC Genomics* 16.Suppl 10, S1. ISSN: 1471-2164. DOI: 10.1186/1471-2164-16-S10-S1. (Visited on 10/12/2020) (cit. on p. 39).

Dilthey, Alexander and Martin J. Lercher (Dec. 2015). "Horizontally Transferred Genes Cluster Spatially and Metabolically". In: *Biology Direct* 10.1, p. 72. ISSN: 1745-6150. DOI: 10.1186/s13062-015-0102-5. (Visited on 02/10/2021) (cit. on pp. 1, 2).

Dmitrijeva, Marija, Janko Tackmann, João Frederico Matias Rodrigues, Jaime Huerta-Cepas, Luis Pedro Coelho, and Christian von Mering (Mar. 2024). "A Global Survey of Prokaryotic Genomes Reveals the Eco-Evolutionary Pressures Driving Horizontal Gene Transfer". In: *Nature Ecology & Evolution* 8.5, pp. 986–998. ISSN: 2397-334X. DOI: 10.1038/s41559-024-02357-0. (Visited on 04/04/2024) (cit. on pp. 1, 3).

Doolittle, W. Ford (June 1999). "Phylogenetic Classification and the Universal Tree". In: *Science* 284.5423, pp. 2124–2128. DOI: 10.1126/science.284.5423.2124. (Visited on 07/28/2025) (cit. on p. 1).

Engelstädter, Jan and Danesh Moradigaravand (Jan. 2014). "Adaptation through Genetic Time Travel? Fluctuating Selection Can Drive the Evolution of Bacterial Transformation". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1775, p. 20132609. DOI: 10.1098/rspb.2013.2609. (Visited on 12/23/2024) (cit. on p. 3).

Felsenstein, Joseph (Oct. 2003). *Inferring Phylogenies*. Sinauer. ISBN: 978-0-87893-177-4 (cit. on p. 2).

Feyerabend, Paul (1993). *Against Method*. Verso. ISBN: 978-0-86091-646-8 (cit. on p. 2).

Frazão, Nelson, Ana Sousa, Michael Lässig, and Isabel Gordo (Sept. 2019). "Horizontal Gene Transfer Overrides Mutation in Escherichia Coli Colonizing the Mammalian Gut". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.36, pp. 17906–17915. ISSN: 1091-6490. DOI: 10.1073/pnas.1906958116 (cit. on p. 1).

Gadagkar, Sudhindra R. and Sudhir Kumar (Nov. 2005). "Maximum Likelihood Outperforms Maximum Parsimony Even When Evolutionary Rates Are Heterotachous". In: *Molecular Biology and Evolution* 22.11, pp. 2139–2141. ISSN: 0737-4038. DOI: 10.1093/molbev/msi212. (Visited on 01/29/2025) (cit. on p. 39).

Garud, Nandita R., Benjamin H. Good, Oskar Hallatschek, and Katherine S. Pollard (Jan. 2019). "Evolutionary Dynamics of Bacteria in the Gut Microbiome within and across Hosts". In: *PLOS Biology* 17.1, e3000102. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000102. (Visited on 07/28/2025) (cit. on p. 1).

Giovannoni, Stephen J., J. Cameron Thrash, and Ben Temperton (Aug. 2014). "Implications of Streamlining Theory for Microbial Ecology". In: *The ISME Journal* 8.8, pp. 1553–1565. ISSN: 1751-7370. DOI: 10.1038/ismej.2014.60. (Visited on 07/30/2025) (cit. on p. 40).

Goh, Ying-Xian, Sai Manohar Balu Anupoju, Anthony Nguyen, Hailong Zhang, Monica Ponder, Leigh-Anne Krometis, Amy Pruden, and Jingqiu Liao (Nov. 2024). "Evidence of Horizontal Gene Transfer and Environmental Selection Impacting Antibiotic Resistance Evolution in Soil-Dwelling Listeria". In: *Nature Communications* 15.1, p. 10034. ISSN: 2041-1723. DOI: 10.1038/s41467-024-54459-9. (Visited on 12/23/2024) (cit. on p. 3).

Hao, Weilong and G. Brian Golding (May 2006). "The Fate of Laterally Transferred Genes: Life in the Fast Lane to Adaptation or Death". In: *Genome Research* 16.5, pp. 636–643. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.4746406. (Visited on 08/09/2022) (cit. on p. 3).

Jiang, Xiaodong, Scott V. Edwards, and Liang Liu (July 2020). "The Multispecies Coalescent Model Outperforms Concatenation Across Diverse Phylogenomic Data Sets". In: *Systematic Biology* 69.4, pp. 795–812. ISSN: 1063-5157. DOI: 10.1093/sysbio/syaa008. (Visited on 08/09/2020) (cit. on p. 39).

Kapust, Nils, Shijulal Nelson-Sathi, Barbara Schönfeld, Einat Hazkani-Covo, David Bryant, Peter J. Lockhart, Mayo Röttger, Joana C. Xavier, and William F. Martin (May 2018). "Failure to Recover Major Events of Gene Flux in Real Biological Data Due to Method Misapplication". In: *Genome Biology and Evolution* 10.5, pp. 1198–1209. DOI: 10.1093/gbe/evy080. (Visited on 10/11/2019) (cit. on p. 2).

Kloesges, Thorsten, Ovidiu Popa, William Martin, and Tal Dagan (Feb. 2011). "Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths". In: *Molecular Biology and Evolution* 28.2, pp. 1057–1074. ISSN: 0737-4038. DOI: 10.1093/molbev/msq297. (Visited on 07/28/2025) (cit. on p. 1).

Krause, Sascha, Xavier Le Roux, Pascal A. Niklaus, Peter M. Van Bodegom, Jay T. Lennon, Stefan Bertilsson, Hans-Peter Grossart, Laurent Philippot, and Paul L. E. Bodelier (May 2014). "Trait-Based Approaches for Understanding Microbial Biodiversity and Ecosystem Functioning". In: *Frontiers in Microbiology* 5. ISSN: 1664-302X. DOI: 10.3389/fmicb.2014.00251. (Visited on 08/11/2025) (cit. on p. 40).

Lee, Ming-Chun and Christopher J. Marx (May 2012). "Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations". In: *PLoS Genetics* 8.5, e1002651. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1002651. (Visited on 08/07/2025) (cit. on p. 40).

Lerat, Emmanuelle, Vincent Daubin, Howard Ochman, and Nancy A. Moran (Apr. 2005). "Evolutionary Origins of Genomic Repertoires in Bacteria". In: *PLOS Biology* 3.5, e130. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030130. (Visited on 11/14/2022) (cit. on p. 3).

Lieberman, Tami D. (Aug. 2022). "Detecting Bacterial Adaptation within Individual Microbiomes". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1861, p. 20210243. DOI: 10.1098/rstb.2021.0243. (Visited on 04/10/2025) (cit. on p. 2).

Martin, William F. (2017). "Too Much Eukaryote LGT". In: *BioEssays* 39.12, p. 1700115. ISSN: 1521-1878. DOI: 10.1002/bies.201700115. (Visited on 10/11/2019) (cit. on p. 1).

Maslov, Sergei, Sandeep Krishna, Tin Yau Pang, and Kim Sneppen (June 2009). "Toolbox Model of Evolution of Prokaryotic Metabolic Networks and Their Regulation". In: *Proceedings of the National Academy of Sciences* 106.24, pp. 9743–9748. DOI: 10.1073/pnas.0903206106. (Visited on 04/10/2025) (cit. on pp. 3, 39).

Mira, Alex, Howard Ochman, and Nancy A. Moran (Oct. 2001). "Deletional Bias and the Evolution of Bacterial Genomes". In: *Trends in Genetics* 17.10, pp. 589–596. ISSN: 0168-9525. DOI: 10.1016/S0168-9525(01)02447-7. (Visited on 01/03/2023) (cit. on p. 40).

Morel, Benoit, Alexey M Kozlov, Alexandros Stamatakis, and Gergely J Szöllősi (Sept. 2020). "GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss". In: *Molecular Biology and Evolution* 37.9, pp. 2763–2774. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa141. (Visited on 07/30/2025) (cit. on p. 39).

Mukherjee, Supratim, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Hema Y. Katta, Alejandro Mojica, I.-Min A. Chen, Nikos C. Kyrpides, and T. B. K. Reddy (Jan. 2019). "Genomes OnLine Database (GOLD) v.7: Updates and New Features". In: *Nucleic Acids Research* 47.D1, pp. D649–D659. ISSN: 0305-1048. DOI: 10.1093/nar/gky977. (Visited on 10/11/2020) (cit. on p. 40).

Pál, Csaba, Balázs Papp, and Martin J. Lercher (Dec. 2005). "Adaptive Evolution of Bacterial Metabolic Networks by Horizontal Gene Transfer". In: *Nature Genetics* 37.12, pp. 1372–1375. ISSN: 1546-1718. DOI: 10.1038/ng1686. (Visited on 05/07/2020) (cit. on p. 3).

Pang, Tin Yau and Martin J. Lercher (Jan. 2017). "Supra-Operonic Clusters of Functionally Related Genes (SOCs) Are a Source of Horizontal Gene Co-Transfers". In: *Scientific Reports* 7.1, p. 40294. ISSN: 2045-2322. DOI: 10.1038/srep40294. (Visited on 07/21/2025) (cit. on pp. 1, 2).

– (Jan. 2019). "Each of 3,323 Metabolic Innovations in the Evolution of E. Coli Arose through the Horizontal Transfer of a Single DNA Segment". In: *Proceedings of the National Academy of Sciences* 116.1, pp. 187–192. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1718997115. (Visited on 10/05/2020) (cit. on pp. 1, 3).

Puigbò, Pere, Alexander E. Lobkovsky, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin (Dec. 2014). "Genomes in Turmoil: Quantification of Genome Dynamics in Prokaryote Supergenomes". In: *BMC Biology* 12.1. ISSN: 1741-7007. DOI: 10.1186/s12915-014-0066-4. (Visited on 07/22/2025) (cit. on p. 3).

Rannala, B., A. Leache, S. Edwards, and Z. Yang (2020). "The Multispecies Coalescent Model and Species Tree Inference". In: *In: Scornavacca, C and Delsuc, F and Galtier, N, (Eds.) Phylogenetics in the Genomic Era. (3.3:1-3.3:21). Self Published (2020)*. Ed. by C. Scornavacca, F. Delsuc, and N. Galtier. Self Published, 3.3:1–3.3:21. (Visited on 08/02/2023) (cit. on p. 39).

Ravenhall, Matt, Nives Škunca, Florent Lassalle, and Christophe Dessimoz (May 2015). "Inferring Horizontal Gene Transfer". In: *PLoS Computational Biology* 11.5. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.1004095. (Visited on 02/25/2020) (cit. on pp. 1, 2).

Richter, Anne, Felix Blei, Guohai Hu, Jan W. Schwitalla, Carlos N. Lozano-Andrade, Jiyu Xie, Scott A. Jarmusch, Mario Wibowo, Bodil Kjeldgaard, Surabhi Surabhi, Xinming Xu, Theresa Jautzus, Christopher B. W. Phippen, Olaf Tyc, Mark Arentshorst, Yue Wang, Paolina Garbeva, Thomas Ostenfeld Larsen, Arthur F. J. Ram, Cees A. M. van den Hondel, Gergely Maróti, and Ákos T. Kovács (May 2024). "Enhanced Surface Colonisation and Competition during Bacterial Adaptation to a Fungus". In: *Nature Communications* 15.1, p. 4486. ISSN: 2041-1723. DOI: 10.1038/s41467-024-48812-1. (Visited on 04/10/2025) (cit. on p. 2).

Serra Moncadas, Lucas, Cyrill Hofer, Paul-Adrian Bulzu, Jakob Pernthaler, and Adrian-Stefan Andrei (Apr. 2024). "Freshwater Genome-Reduced Bacteria Exhibit Pervasive Episodes of Adaptive Stasis". In: *Nature Communications* 15.1, p. 3421. ISSN: 2041-1723. DOI: 10.1038/s41467-024-47767-7. (Visited on 04/08/2025) (cit. on p. 3).

Sheinman, Michael, Ksenia Arkhipova, Peter F Arndt, Bas E Dutilh, Rutger Hermsen, and Florian Massip (June 2021). "Identical Sequences Found in Distant Genomes Reveal Frequent Horizontal Transfer across the Bacterial Domain". In: *eLife* 10. Ed. by Richard A Neher, Gisela Storz, and Richard A Neher, e62719. ISSN: 2050-084X. DOI: 10.7554/eLife.62719. (Visited on 07/21/2025) (cit. on p. 2).

Smillie, Chris S., Mark B. Smith, Jonathan Friedman, Otto X. Cordero, Lawrence A. David, and Eric J. Alm (Dec. 2011). "Ecology Drives a Global Network of Gene Exchange Connecting the Human Microbiome". In: *Nature* 480.7376, pp. 241–244. ISSN: 1476-4687. DOI: 10.1038/nature10571. (Visited on 12/23/2024) (cit. on p. 1).

Szappanos, Balázs, Jonathan Fritzemeier, Bálint Csörgő, Viktória Lázár, Xiaowen Lu, Gergely Fekete, Balázs Bálint, Róbert Herczeg, István Nagy, Richard A. Notebaart, Martin J. Lercher, Csaba Pál, and Balázs Papp (May 2016). "Adaptive Evolution of Complex Innovations through Stepwise Metabolic Niche Expansion". In: *Nature Communications* 7.1, p. 11607. ISSN: 2041-1723. DOI: 10.1038/ncomms11607. (Visited on 07/04/2025) (cit. on p. 3).

Taylor, Aidan J., Koji Yahara, Ben Pascoe, Seungwon Ko, Leonardos Mageiros, Evangelos Mourkas, Jessica K. Calland, Santeri Puranen, Matthew D. Hitchings, Keith A. Jolley, Carolin M. Kobras, Sion Bayliss, Nicola J. Williams, Arnoud H. M. van Vliet, Julian Parkhill, Martin C. J. Maiden, Jukka Corander, Laurence D. Hurst, Daniel Falush, Paul Keim, Xavier Didelot, David J. Kelly, and Samuel K. Sheppard (Apr. 2024). "Epistasis, Core-Genome Disharmony, and Adaptation in Recombining Bacteria". In: *mBio* 15.6, e00581–24. DOI: 10.1128/mbio.00581-24. (Visited on 04/08/2025) (cit. on p. 3).

Than, Cuong, Guohua Jin, and Luay Nakhleh (2008). "Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer". In: *Comparative Genomics*. Ed. by Craig E. Nelson and Stéphane Vialette. Berlin, Heidelberg: Springer, pp. 113–127. ISBN: 978-3-540-87989-3. DOI: 10.1007/978-3-540-87989-3_9 (cit. on pp. 2, 39).

Than, Cuong, Derek Ruths, Hideki Innan, and Luay Nakhleh (May 2007). "Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions". In: *Journal of Computational Biology* 14.4, pp. 517–535. ISSN: 1066-5277. DOI: 10.1089/cmb.2007.A010 (cit. on pp. 2, 39).

Treangen, Todd J. and Eduardo P. C. Rocha (Jan. 2011). "Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes". In: *PLoS Genetics* 7.1, e1001284. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1001284. (Visited on 07/28/2025) (cit. on p. 1).

Tria, Fernando D K and William F Martin (Oct. 2021). "Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes". In: *Genome Biology and Evolution* 13.10, evab224. ISSN: 1759-6653. DOI: 10.1093/gbe/evab224. (Visited on 07/28/2025) (cit. on p. 1).

Vos, Michiel and Adam Eyre-Walker (Dec. 2017). "Are Pangenomes Adaptive or Not?" In: *Nature Microbiology* 2.12, pp. 1576–1576. ISSN: 2058-5276. DOI: 10.1038/s41564-017-0067-5. (Visited on 07/28/2025) (cit. on p. 1).

Woods, Laura C., Rebecca J. Gorrell, Frank Taylor, Tim Connallon, Terry Kwok, and Michael J. McDonald (Oct. 2020). "Horizontal Gene Transfer Potentiates Adaptation by Reducing Selective Constraints on the Spread of Genetic Variation". In: *Proceedings of the National Academy of Sciences* 117.43, pp. 26868–26875. DOI: 10.1073/pnas.2005331117. (Visited on 12/23/2024) (cit. on p. 3).

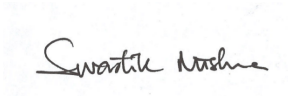
Yang, Ziheng (May 2014). "Coalescent Theory and Species Trees". In: *Molecular Evolution*. 1st ed. Oxford University PressOxford, pp. 308–360. ISBN: 978-0-19-960261-2 978-0-19-960260-5 978-0-19-178225-1. DOI: 10.1093/acprof:oso/9780199602605.003.0009. (Visited on 08/13/2024) (cit. on p. 39).

Zheng, Danping, Timur Liwinski, and Eran Elinav (June 2020). "Interaction between Microbiota and Immunity in Health and Disease". In: *Cell Research* 30.6, pp. 492–506. ISSN: 1748-7838. DOI: 10.1038/s41422-020-0332-7. (Visited on 04/10/2025) (cit. on p. 2).

Eidesstattliche Erklärung
laut §5 der Promotionsordnung vom 15.06.2018

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Düsseldorf, 20.08.2025
Ort, Datum



Swastik Mishra