Co-occurrence drives horizontal gene transfer among marine prokaryotes

Gavin M. Douglas[1,2,a,#], Nicolas Tromas[3,a,#], Marinna Gaudin[4,6], Patrick Lypaczewksi[5], Louis-Marie Bobay[1,2], B. Jesse Shapiro[5,b,#], Samuel Chaffron[4,6,b,#]

[1]Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA
[2]Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA
[3]Alpine Center for research on trophic networks and limnic ecosystems, INRAE - CARRTEL, France
[4]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
[5]Department of Microbiology & Immunology, McGill University, Montréal, Québec, Canada
[6]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France

[a]*Joint first author*
[b]*Joint senior author*
[#]*Correspondence: gavinmdouglas@gmail.com, nicolas.tromas@inrae.fr, jesse.shapiro@mcgill.ca, and samuel.chaffron@cnrs.fr*

Keywords: Horizontal gene transfer; lateral gene transfer; co-occurrence; phylogenetics; ocean ecology; prokaryotes; bacteria; metagenomics

Abstract

Understanding the drivers of horizontal gene transfer (HGT) is a key question in microbial evolution. While co-occurring taxa have long been appreciated to undergo HGT more often, this association is confounded with other factors, most notably their phylogenetic distance. To disentangle these factors, we analyzed 15,339 isolate and metagenome-assembled marine genomes. We identified HGT events across these genomes, and identified enrichments for functions previously shown to be prone to HGT. By mapping metagenomic reads from 1,862 ocean samples to these genomes, we also identified co-occurrence patterns and environmental associations. Although we observed an expected negative association between phylogenetic distance and HGT rates, we only detected the association between co-occurrence and phylogenetic distance when restricted to closely related taxa. This observation refines the previously reported trend to closely related taxa, rather than a consistent pattern across all taxonomic levels, at least within marine environments. In addition, we identified a significant association between co-occurrence and HGT, which remains even after controlling for phylogenetic distance and measured environmental variables. Overall, our findings demonstrate the significant influence of ecological associations in shaping marine bacterial evolution through HGT.

Introduction

Closely related prokaryotic taxa often display substantial differences in gene content, largely due to horizontal gene transfer (HGT; Innamorati et al., 2020). To better understand these dynamics, gene transfer networks have been investigated across diverse species (Beiko et al., 2005). A recurrent observation from this past work is that taxa from similar environments are associated with higher rates of HGT (Smillie et al., 2011; Fondi et al., 2016; Sheinman et al., 2021; Khedkar et al., 2022)

This observation is expected, for at least two reasons. First, all HGT mechanisms are dependent on physical proximity to some degree (Thomas & Nielsen, 2005). For instance, conjugation requires a physical interaction between cells for DNA transfer to occur. Second, taxa from the same environment can experience similar selective pressures and may potentially benefit from the same set of adaptive genes. This is clear within species, where strains experiencing similar pressures exhibit parallel adaptive point mutations (Turner et al., 2018). It is expected that similar dynamics occur between strains within the same environment based on horizontally-transferred genes (Niehus et al., 2015). Genes can also spread horizontally between more distantly related taxa within a community. This is most clearly seen for antibiotic and heavy metal resistance genes, which have the potential to confer adaptive benefits to many taxa in a contaminated environment (Pal et al., 2015).

Although these explanations for the association between HGT and co-occurrence are intuitive, there is also an important confounder to consider: phylogenetic distance. This point was recently raised in work disentangling the action of these factors in driving HGT inferred across reference genomes from isolates, and based on co-occurrence patterns of 16S rRNA sequences (Dmitrijeva et al., 2024). In this work, the authors found that the positive association between co-occurrence and horizontal gene transfer remained after controlling for phylogenetic distance, although it was decreased. This study is an important contribution, but was limited to 16S rRNA gene data used for identifying co-occurring genomes. As this is a coarse approach, it cannot reliably identify strains or species within a sample. In contrast, shotgun metagenomics provides improved resolution for identifying strains, given sufficient genome coverage (Sharpton, 2014). In addition, due to the scale of their work, the authors were not able to perform in-depth investigations of how particular environmental factors are associated with HGT and instead focused on general environment groupings.

Marine microbial communities are an appealing system for studying prokaryotic evolutionary dynamics. Marine prokaryotes are known to undergo widespread HGT, and to be stratified into communities based on depth and other environmental factors (Sunagawa et al., 2015; Martinez-Gutierrez & Aylward, 2022). In addition, understanding prokaryotic community dynamics and evolution is particularly important as marine microbes play key roles in global biogeochemical cycles (Azam & Malfatti, 2007), and are also experiencing heightened

environmental stress due to climate change (Heneghan et al., 2024). To help better understand these and other dynamics, numerous ocean samples have undergone metagenomics sequencing (Sunagawa et al., 2015; Biller et al., 2018).

Here, we used existing ocean metagenomes to investigate patterns of HGT and co-occurrence across the global ocean. Our key question focused on whether the relationship between HGT and co-occurrence in this dataset is robust to the inclusion of phylogenetic distance as a covariate. This question was also recently investigated by Dmitrijeva and colleagues (2024), but here we specifically focus on marine samples with shotgun metagenomics data. Another novel aspect of our work is the consideration of additional factors that could be driving the association between HGT and co-occurrence. In particular, by using in-depth environmental data associated with these metagenomics samples, we investigated to what degree shared environmental pressures can explain the observed association between HGT and co-occurrence. We also consider a wide range of methods for inferring HGT and co-occurrence to ensure our results are robust.

Methods

Acquiring genomes

We downloaded all 34,815 genomes from a prior metagenomic study of the global ocean microbiome (Paoli et al., 2022), which comprised whole genome sequencing from isolates, single-amplified genomes, and metagenome-assembled genomes (MAGs), along with pre-existing Prokka gene annotations (Seemann, 2014), taxonomic classification, and quality reports. These data correspond to version 1.0 of the Ocean Microbiomics Database, which we refer to as our focal dataset. The prior taxonomic classifications were obtained using the Genome Taxonomy Database Toolkit (Chaumeil et al., 2020). Using quality metrics provided with these data, we retained 15,339 genomes with mean completeness > 75% and mean contamination < 5%. This final set corresponded to 11,833 MAGs, 1,862 single-amplified genomes, and 1,644 genomes from whole genome sequencing of isolates.

Genome phylogenies

GToTree v1.7.00 (Lee & Ponty, 2019) was used on the 15,339 genomes using the Bacteria and Archaea HMM-set composed of 25 single copy genes to generate an alignment protein file with HMMER v3.3.2 (Eddy, 1998, 2011) and Muscle v5.1 (Edgar, 2022). Automated trimming was performed on the alignments using trimAl v1.4 (Capella-Gutiérrez et al., 2009), with the -automated1 and -phylip options. IQTREE2 v2.0.7 (Minh et al., 2020) was then used to construct a phylogenetic tree, using ModelFinder Plus and 1000 bootstraps. The best-fitting model, 'Q.yeast+I+R10', was automatically chosen based on a comparison of Bayesian Information Criterion values. Pairwise (patristic) distance between tips was computed using the castor R package v1.8.2 (Louca & Doebeli, 2018).

To compare our focal results (heavily based on MAGs) to those based on a set of higher-quality genomes, we also downloaded all representative aquatic genomes from the proGenomes database v3 (Fullam et al., 2023). Taxonomy for each genome was parsed based on TaxIDs using the taxize R package v0.9.102 (Chamberlain & Szöcs, 2013). Each genome in proGenomes is associated with a BioSample in the National Center for Biotechnology Information database, which we used to parse sample metadata using efetch v21.6 (https://www.ncbi.nlm.nih.gov/books/NBK179288/). All genomes linked to samples listed as metagenomics, or without any contigs greater than 100,000 bp were excluded, to help avoid mis-assembled genomes. This filtering resulted in a final set of 5,551 genomes for this database. The same workflow as above was used to produce a phylogenetic tree, based on the same best-fitting model, and pairwise distances.

Metagenomics samples preprocessing

To characterize the distribution of our filtered genomes across environments, we also prepared a metagenomics dataset representing diverse ocean samples that were then mapped to these genomes. To that end, we acquired all 1,865 ocean water metagenomics samples available in the OceanDNA resource (Nishimura & Yoshizawa, 2022). This raw metagenomics sequencing data was downloaded from the National Center for Biotechnology Information with Globus Connect (Foster, 2011). We filtered out low-quality reads with fastp v0.23.4 (Chen et al., 2018), with the options -l 70, --cut_mean_quality 20, --cut_front, --cut_tail, --cut_window_size=4, and --detect_adapter_for_pe (this last option only for paired-end FASTQ files). Files corresponding to technical replicates were concatenated. We then mapped all filtered metagenomics reads to the genomes with CoverM v0.6.1 (Aroney et al., 2024), using the options --exclude-supplementary, --min-covered-fraction 0, and --proper-pairs-only (for paired-end reads). We called genomes as present based on a >= 30% breadth of coverage cut-off (i.e., where at least 30% of the genomes was mapped by at least one read). For each sample, we set the reads per kilobase per million mapped reads values for genomes with breadth below the presence cut-off to 0. We excluded genomes not called as present in at least 10 samples, which resulted in 10,389 remaining genomes. We excluded three metagenomics samples where no genomes were called as present.

To distinguish metagenomics samples more likely to represent free-living microbes, compared to particle-attached microbes, we subset samples based on the size-filtration steps used during sample collection. A cut-off of 3 µm is commonly applied to distinguish free-living and particle-attached microbes by size (Crump et al., 1999; Urvoy et al., 2022). Accordingly, we categorized “free-living” samples as those size-selected at <= 3 µm, which corresponded to 906 samples. However, as this was not the initial focus of our analysis, we did not acquire metagenomics data equally split by size-selection criteria, and almost all samples had a lower size-selection filtration that would include some free-living bacteria. Thus, it is not possible to confidently categorize our analyzed metagenomics samples as “particle-attached”. Instead,

we opted to subset 846 “less-filtered” samples, which are those with an upper size-selection filter >= 20 µm. Our rationale for this category is that they are more likely to contain DNA from particle-attached prokaryotes compared to the “free-living” sample set. We then identified cases where genomes were predominately found in either putative “free-living” (small particle size) or “less-filtered” (large and small particle size) groups. We classified genomes as being associated with a given group if > 75% of the samples containing the genome were in that group. We also required that the median relative abundance of this genome, based on reads per kilobase per million mapped reads, was highest in this sample subgroup. Based on this criterion, 346 focal genomes could be classified as enriched in “less-filtered” and 2,470 as enriched in “free-living” metagenomics samples, respectively. We applied this same approach for the proGenomes database as well, where we could classify only 24 genomes as “less-filtered” and 137 as “free-living”.

The OceanDNA resource assembled several independent datasets. To ensure our results were not driven by technical differences between these datasets, we also subset metagenomics samples to be only those part of the *Tara* Oceans campaign (n=370) or the GEOTRACES campaign samples (n=610).

Clustering genes and identifying reciprocal best hits

Genes in the Ocean Microbiomics Database v1.0 that we analyzed were previously clustered based on sequence identity, with a representative gene indicated per clusters. However, these clusters included genes with high variation in length (e.g., genes ranging from 3 kb to 21 kb in the same cluster). Accordingly, we re-ran gene clustering to produce clusters of genes with high pairwise identity over their entire lengths.

To do so, we first parsed out all protein coding genes sequences from pre-existing GFF files per genome. We then clustered these genes with CD-HIT-EST v4.81 (Fu et al., 2012), with a sequence identity threshold (-c) of 0.95, a length difference cut-off (-s) of 0.95, alignment coverage settings (-aL and -aS) of 0.95, and in accurate mode (-g 1). Alignments were then built with MAFFT v7.525 (Katoh & Standley, 2013), with the options --retree 2 --maxiterate 2 to speed up processing, as these were generally trivial alignments. We then computed the pairwise hamming distance for all sequences per cluster in Python using the hammingdist package (v1.3.0). We identified gene pairs that were reciprocally the closest matches using this approach and between taxa that were at least in different genera (i.e., that were not in the same genus or species). We used a lower sequence identity cut-off of 95% as this allowed us to detect potential ancient HGT events, in addition to more recent events, which we assumed would be enriched among matches with sequence identity >= 99%. For all HGT-detection approaches, all genes (or hit regions, in the case of the BLASTn approach, described below) on contigs < 5000 bp were excluded after detecting HGT events.

203 ***Sensitive approach for identifying indirect gene flow with pairwise BLAST***
204 In addition to the clustering approach, we also performed pairwise identity comparisons
205 between all genomes using BLASTn, with BLAST+ v2.12.0 (Altschul et al., 1990). Only hits
206 >= 500 bp long and at least 95% identical were retained. To clarify, BLASTn matches were
207 generally based on short subsequences in contigs, rather than the entire contigs. Our analysis
208 for these data focused only on genomes within different genera or higher, to reduce the
209 number of false positives due to vertical inheritance rather than HGT. We ran bedtools v2.30.0
210 (Quinlan & Hall, 2010) to intersect gene annotations with BLASTn-identified matching regions.
211 For cases where multiple hits overlapped (for the same pairwise genome comparison) the
212 highest scoring hit was retained.
213
214 Importantly, BLASTn matches were called as putative HGT events regardless of whether they
215 were the reciprocal best hits in the dataset. Thus, this approach is not meant to identify direct
216 HGT events between genomes, and instead identifies regions between genomes that were
217 putatively horizontally acquired through a shared HGT network. We used this information for
218 complementary purposes, and expect these calls to be more sensitive, but also enriched for
219 more false positives, and less straightforward to interpret than the clustering method
220 described above.
221
222 ***Phylogenetic reconciliation within species***
223 We also identified putative gene transfer events within species with RANGER-DTL (Bansal et
224 al., 2018; Bansal, 2022). We included this approach as using sequence identity alone is
225 unreliable for detecting HGT between closely related genomes (due to high identity stemming
226 from a recent common ancestor). We also ran HoMeR v1.0 (Kloub et al., 2021) in conjunction
227 with RANGER-DTL to call multi-gene transfer events. Our custom code for running this
228 pipeline is located here: https://github.com/gavinmdouglas/HoMer_workflow.
229
230 We ran this approach over 190 species that had at least 10 genomes. We found that
231 topological differences between the core genome tree and accessory gene trees were
232 widespread: a mean of 91.6% (standard deviation: 0.06%) of tested genes were inferred to
233 have at least one transfer event across all species. We had little confidence in the accuracy of
234 individual events (see Discussion) and instead used these results only for supplementary
235 purposes. Our assumption here is that the relative ranks of inferred and real HGT event counts
236 between genome pairs are highly correlated, even if there is a baseline of noise.
237
238 ***Functional annotation and enrichment***
239 Clusters of Orthologous Genes (COG) annotations were taken from a previous run of eggNOG
240 (Cantalapiedra et al., 2021), which was distributed with the genomes. For each eggNOG
241 Orthologous Group, we parsed out all COG annotations in the definition and then re-mapped
242 them to higher-level COG categories. To determine the COG annotation associated with each

gene cluster, we considered the classifications of each gene in the cluster and used the majority rule to determine which COG (and COG categories) to assign. We also annotated all genes that matched to mobile genetic elements in the proMGE v1 database (Khedkar et al., 2022). We first translated all protein coding gene sequences to amino acids with the transeq EMBOSS program v6.6.0.0 (Rice et al., 2000), using the parameters -table 11, -trim, and -clean. We then ran HMMER v3.1b2 to identify matching genes.

We also ran genomAD v1.5.0 (Camargo et al., 2024) with database v1.1, and default parameters except --cleanup and --splits8 on the contigs of all studied genomes. This allowed us to identify contigs categorized as 'virus' or 'plasmid', as well as those that contain prophages.

The expected enrichments of COG categories in the HGT results (prior to cross-referencing with co-occurrence data) were taken from Dmitrijeva et al. (2024), based on prior studies (Nakamura et al., 2004; Cordero & Hogeweg, 2009; Cohen & Pupko, 2010; Cohen et al., 2011; Popa et al., 2017; Oliveira et al., 2017; Paquola et al., 2018; Song et al., 2019; Jeong et al., 2019; Sheinman et al., 2021). We defined categories expected to be depleted or enriched as those showing the consistent signal in all but one study (excluding studies with no result for a given category).

When testing for functional enrichments among clusters involved with HGT events vs. all other gene clusters, we only considered clusters assigned to a COG annotation. We also excluded rare clusters annotated as eukaryotic-specific COG categories (A, B, Y, and Z). We then ran Fisher's exact tests on the number of clusters in each COG category in at least one HGT relationship compared to all other clusters in that category (significance called at Benjamini–Hochberg-corrected (Benjamini & Hochberg, 1995) P-values < 0.05). We focused this analysis on clusters, rather than individual genes, so that clusters with many genes involved in HGT events would not be counted multiple times as independent observations.

Co-occurrence analyses

We used the backbone R package (Neal, 2022) to run hypergeometric distribution-based co-occurrence tests for all pairwise genomes. Note that the exact version 1.3.1 of this package needed to be used, as more recent versions no longer provide the method we required at scale. This allowed us to identify cases where genomes co-occurred more often than expected by chance (i.e., if they were randomly distributed across samples). We computed the effect size as the co-occurrence ratio, the ratio of observed vs. expected number of metagenomics samples containing both genomes. These values were used for comparison with phylogenetic distance, but we primarily focused on the binary classification of significantly co-occurring genome pairs. Significantly co-occurring genomes were identified based on Benjamini-Hochberg corrected P-values < 0.05 and co-occurrence ratios > 1.

To ensure our inferences were robust to this bioinformatic step, we also computed co-occurrence using two other approaches. First, we computed a simple overlap metric, as used previously (Dmitrijeva et al., 2024): $\frac{|samples_{Genome1} \cap samples_{Genome2}|}{min(|samples_{Genome1}|, |samples_{Genome2}|)}$. Note that this approach does not adequately correct for the total prevalence of both genomes being compared (unlike the hypergeometric approach above). For instance, if a genome is found across all samples, then it will trivially have a co-occurrence of 1.0 with all other genomes. The second additional approach we used was to compute co-abundance with the propr R package (Quinn et al., 2017) v5.1.4, which incorporates the relative abundance of genomes. This metric was computed using the NetCoMi R package (Peschel et al., 2021) v1.1.0, for convenience.

Statistical modelling

We ran logistic regression in R using the glm function. These models were based on pairwise comparisons of genomes, with predictors of co-occurrence, tip distance, and the difference in median environmental values of the samples each genome is found in. We also included categorical variables of which filter cut-off grouping the genome pair is within (see Results). For the focal model, the response was the binary value of whether each genome pair was inferred to have at least one HGT event. For models basing co-occurrence upon simple co-occurrence and propr-based co-abundance, these co-occurrence values were continuous rather than binary. All continuous variables were transformed using ordered quantile normalization (Peterson & Cavanaugh, 2020) using the bestNormalize R package v1.9.1. We computed the variance inflation factor for each fixed effect using the performance R package v0.12.4 (Lüdecke et al., 2021).

As the key data we analyzed was based on genome pairs, we also wanted to control for genome ID, to ensure that a small number of genomes (e.g., in numerous co-occurrence or HGT relationships) were not driving our results. To that end, we also built generalized linear mixed models with the R package glmmTMB v1.1.10 (Brooks et al., 2017), with genome ID as a random effect. For this supplementary analysis, we duplicated the input table for each model, so that each genome comparison was present twice, and assigned once to each genome being compared. Due to the table being duplicated, we do not have confidence in estimates of variability from these models, and used them simply to assess whether our overall interpretations shifted when controlling for genome ID.

Working environments

Custom code for processing HGT and co-occurrence result tables was written in Python (run with v.3.10). Visualizations and most statistical analyses were conducted in R v4.4.0 (R Core Team, 2024). Where possible, Linux commands (on Ubuntu 22.04.5 LTS) were parallelized

with GNU Parallel v20240322 (Tange, 2011) and tools were installed from Bioconda (Grüning et al., 2018).

Code and data availability

The accessions of all shotgun metagenomics data used in this study are present in the OceanDNA FigShare repository (Nishimura & Yoshizawa, 2022): https://doi.org/10.6084/m9.figshare.c.5564844.v1. All metagenome-assembled genomes and pre-existing annotations were taken from https://sunagawalab.ethz.ch/share/microbiomics/ocean/suppl_data/. Code and scripts are available on GitHub (https://github.com/gavinmdouglas/ocean_cooccur_hgt). Key data files are distributed through a Zenodo repository (https://doi.org/10.5281/zenodo.14867578).

Results

Identifying putative horizontal gene transfer events

To study prokaryotic genome diversity in the global ocean, we analyzed 15,339 genomes representing 3,270 species, which were previously constructed from marine metagenomics samples (Sunagawa et al., 2015; Nayfach et al., 2021; Paoli et al., 2022). These genomes are from diverse lineages (**Supplementary Figure 1**), primarily bacteria (90.9%). The most common phyla are Proteobacteria (47.9%) and Bacteroidota (12.0%) and the best represented archaeal phylum is Thermoplasmatota (5.92%).

There are several approaches for inferring HGT, which can produce different but complementary results (Douglas & Langille, 2019). Accordingly, we identified putative HGT events between these genomes using several methods to ensure our general results were robust (see Methods). The first two approaches were based on identifying regions with sequence similarity >= 95% between genomes from different genera (or that differ at higher taxonomic levels). In other words, we did not consider genome comparisons within the same genus or species. Our key assumption is that regions of high similarity are unlikely to occur (when of sufficient length) in the genomes of distantly related taxa due to these regions being present in a common ancestor (Sheinman et al., 2021). Instead, we take these observations to represent putative HGT events.

Based on this framework, our focal HGT-detection approach focused on clustering all genes across all genomes. We identified 9,046 putative HGT events based on this approach, 71.2% of which were >= 99% identical, and for which the majority were between genomes within the same family (**Figure 1a**). Although this is a relatively high number of putative HGT events, the number of events called between each pair of genomes was low (median of 1.0 and mean of 3.95 in genome pairs with at least one putative event).

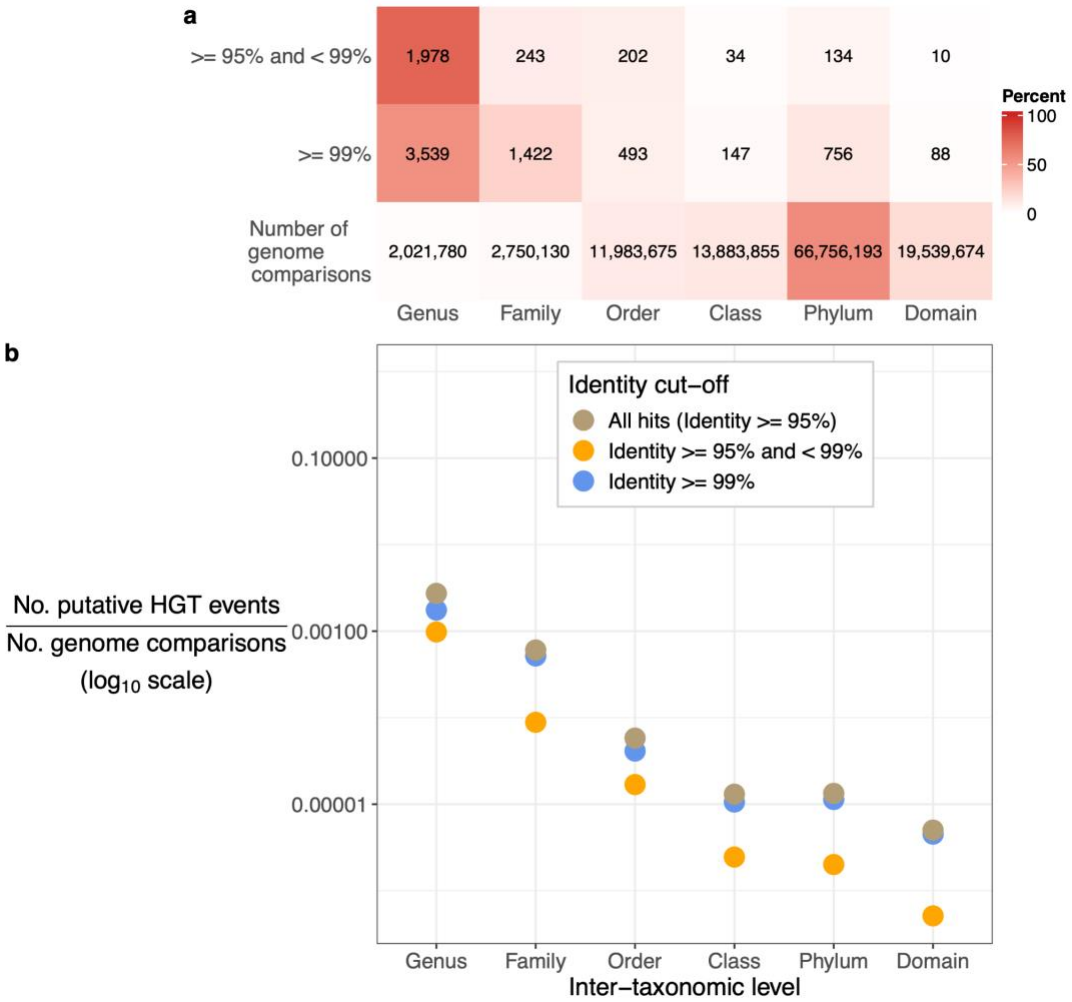


Figure 1: Summary of putative horizontal gene transfer events based on sequence identity cutoffs. (a) Counts of putative horizontal gene transfer (HGT) events based on the cluster-based approach, by the lowest taxonomic level at which the compared genomes differed. The total number of genome comparisons per level is also indicated. These cells are coloured by the percentage per column. (b) Proportion of putative horizontal gene transfer events per taxonomic level, after normalizing by the total number of genome comparisons per level.

We reasoned that the sequence identity of putative HGT hits, on average, would provide information on the age of the transfer event. One expectation of older HGT events is that mutations and recombination will result in a shorter region identifiable as a transfer event. We used a complementary approach based on BLASTn (see Methods) to evaluate how the lengths of matching regions vary by sequence identity. As expected, regions with high sequence identity (>= 99%) were longer and included more genes on average (mean: 6.6; standard deviation [SD]: 6.6) compared to those with lower sequence identity (>= 95% and < 99%; mean: 2.7; SD: 4.2).

379 We then investigated whether the expected negative relationship between evolutionary
380 divergence and HGT held for these results. There is a clear signal of decreasing HGT rates at
381 more distant taxonomic levels (**Figure 1b**), after normalizing for the number of pairwise
382 genomes compared at each level. This signal is similar for both the older (< 99%) and more
383 recent (>= 99%) putative HGT events.
384
385 ***Functional enrichments of horizontally transferred genes***
386 We next characterized whether the horizontally transferred genes we identified were
387 enriched in expected functional groups identified in previous studies (Dmitrijeva et al., 2024).
388 Many of the significantly enriched COG categories were in the expected directions (**Figure 2**).
389 For instance, mobile elements (X) and defense mechanisms (V) were strongly enriched among
390 HGT events. In addition, all significant categories that were expected to be depleted for HGT
391 were identified for putative transfers between identities of >= 95% and < 99%. However,
392 transfers at >= 99% identity showed a more mixed signal for categories expected to be
393 depleted for HGT. For instance, we observed a mixed signal of enrichment for lipid transport
394 and metabolism genes (I), and the unexpected enrichment of several other categories,
395 especially at higher taxonomic levels.
396
397 This observation highlights a unique aspect of our enrichment analyses, which is that we
398 conducted them for inferred HGT events at each taxonomic level separately (**Figure 2**). This
399 allowed us to identify cases where the same category is enriched in different directions at
400 different taxonomic levels, such as those involved in chromosome partitioning and cell
401 division (D) and amino acid transport and metabolism (E).
402
403 We also explored mobile genetic element-associated gene annotations and overall found
404 them enriched among horizontally transferred genes (Fisher's exact tests, $P < 0.001$).
405 Specifically, genes matching the proMGE database were 7.20-fold enriched, whereas genes
406 on contigs classified as a plasmid or virus were 8.69 and 3.32-fold enriched among HGT genes,
407 respectively. In contrast, genes on contigs classified as containing proviruses were 7.14-fold
408 more likely not to be found among the HGT set. With perhaps the exception of proviruses,
409 these functional enrichments agree with basic expectations of gene classes expected to be
410 enriched among HGT events in nature.
411

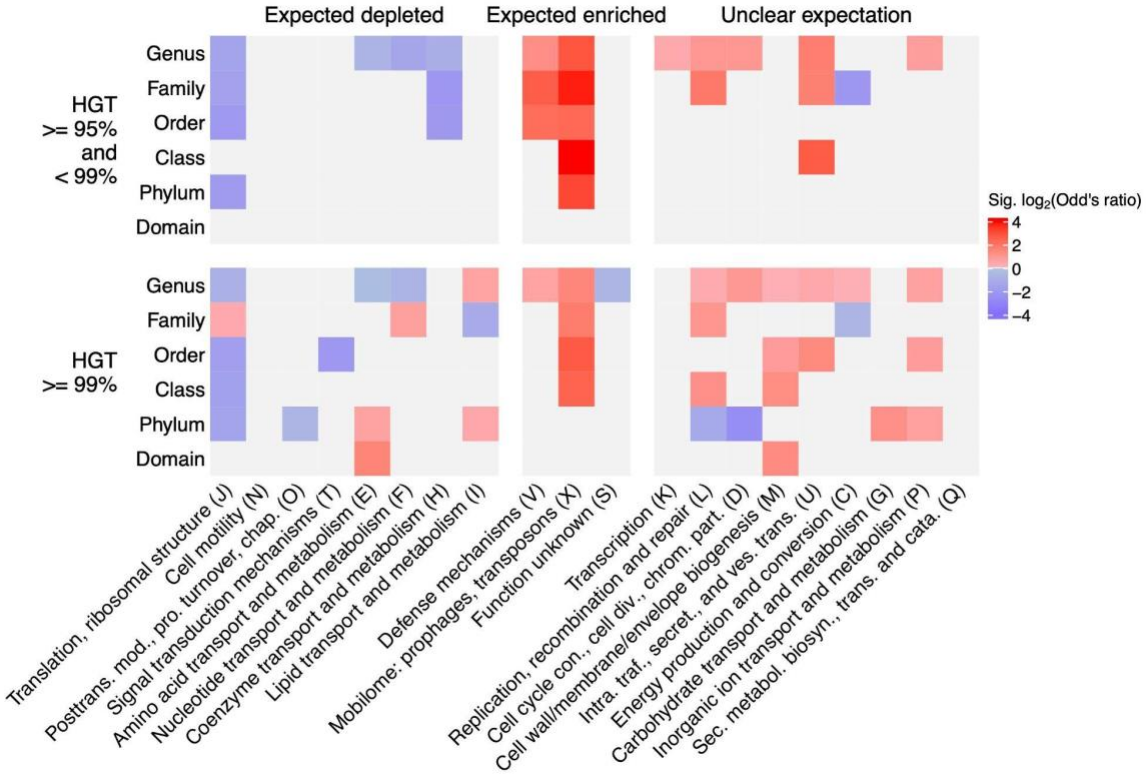


Figure 2: Cluster of Orthologous Genes (COG) categories enriched among horizontal gene transfer (HGT) hits based on sequence identity. HGT results used for this analysis were based on the cluster-based approach. Expected patterns of enrichment are indicated for COG categories at the top, based on (Dmitrijeva et al., 2024), and individual COG categories (in abbreviated form in some cases) are indicated at the bottom. The row groupings indicate the identity cut-off, in addition to the taxonomic level at which the genomes that encode the genes differ. Only significant enrichment Fisher's exact tests are displayed (Benjamini-Hochberg-corrected $P < 0.05$); light grey cells indicate non-significant tests.

436 ***Overall relationship between horizontal gene transfer and ecological variables***
437 We next compared how rates of HGT between genomes are associated with co-occurrence
438 and phylogenetic distance. To do so, we mapped metagenomics reads of previously published
439 ocean metagenomics samples against the 15,339 genomes to determine each genome's
440 relative abundance across samples. We identified 1,862 samples where at least one genome
441 was called as present. For these samples, a mean of 26,592,809 (SD: 36,857,948) reads were
442 mapped, which represented a mean of 23.9% (SD: 10.2%) of the reads per sample. A median
443 of 286 genomes were identified as present across these samples. We restricted our
444 subsequent analyses to 10,388 genomes that were called as present in at least 10
445 metagenomics samples, and then computed the pairwise co-occurrence of these genomes.
446
447 Importantly, this prevalence filtering step meant that 32.3% of genomes were excluded,
448 which resulted in the loss of 54.8% of genomes involved in HGT based on the clustering
449 approach (a drop from 2,295 to 1,037 unique genomes). This indicates that many HGT events
450 occurred in taxa for which we have insufficient data to study their co-occurrence relationships
451 within this dataset. Taxa in several poorly represented phyla tended to be especially less likely
452 to be found across at least 10 samples; >= 75% of genomes in SAR324, Gemmatimonadota,
453 Acidobacteriota, Dadabacteria, Latescibacterota, Poribacteria, and Chloroflexota_B were
454 filtered out at this step (**Supplementary Figure 1**).
455
456 As expected, we observed a clear association between taxa co-occurrence and HGT
457 (**Figure 3a**; odds ratio: 19.3; $P < 0.001$). In contrast, we did not observe the expected negative
458 association between phylogenetic distance and extent of co-occurrence based on the focal
459 genome pairs (**Figure 3b**), and instead there was a near-zero association (Spearman $\rho=0.056$;
460 $P < 0.001$; **Supplementary Figure 2a**). As a reminder, our analysis focused on genome pairs in
461 different genera or above, to more reliably infer HGT events, and these distributions (**Figure**
462 **3**) are based on this subset of genome pairs. However, when considering genome pairs within
463 the same genus or lower, which were excluded from our focal analysis, we do find the
464 expected negative association (Spearman $\rho=-0.325$; $P < 0.001$; **Supplementary Figure 2b**).
465 For the subset of genome pairs connected by HGT (**Figure 3b**, right panel) there is lower
466 phylogenetic distance compared to those with no evidence for HGT (mean difference: 0.900;
467 Wilcoxon test $P < 0.001$). In addition, genome pairs connected by HGT that do not co-occur
468 have especially low phylogenetic distance compared to those that do (mean
469 difference: 0.485; Wilcoxon test $P < 0.001$). This last observation could be due to collider bias
470 (i.e., a misleading signal driven by data subsetting; see Discussion). Overall, we find the
471 expected associations between HGT and both phylogenetic distance and co-occurrence, but
472 the negative association between co-occurrence and phylogenetic distance is only present for
473 closely related taxa.

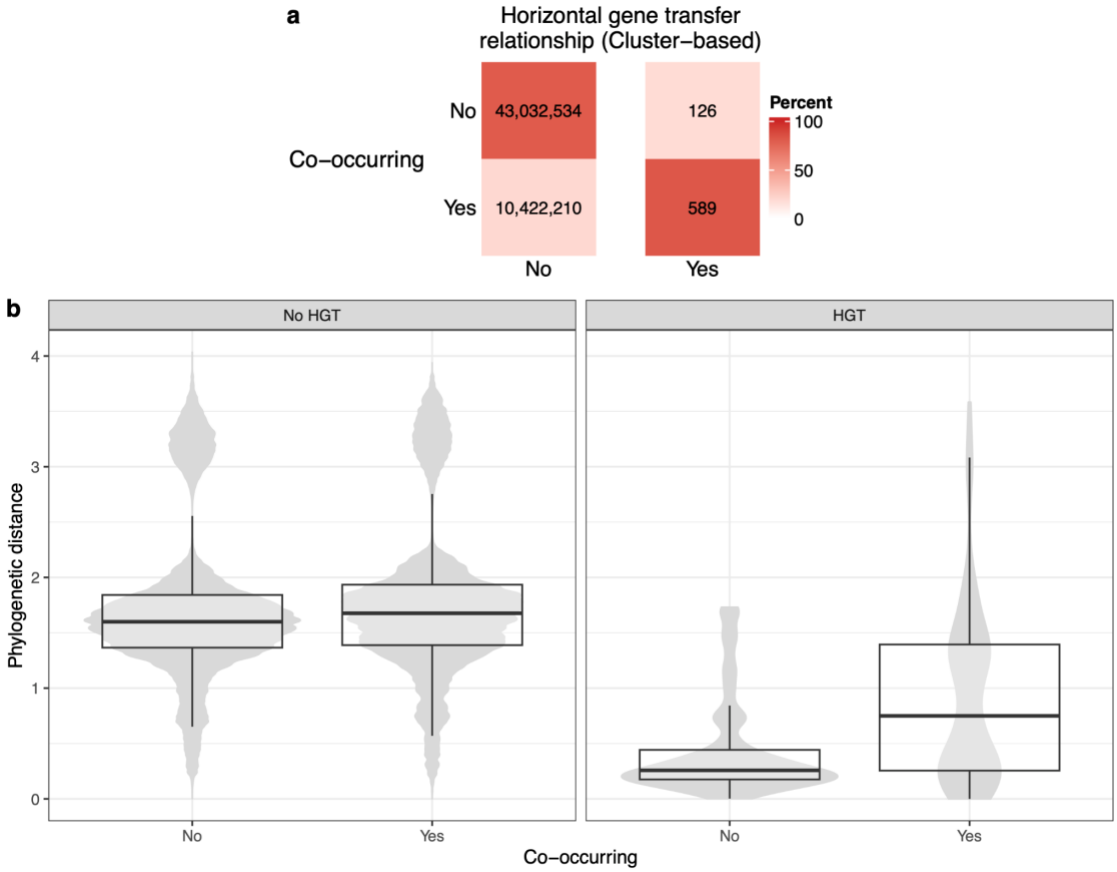


Figure 3: Breakdown of putative horizontally transferred genes identified, by co-occurrence and phylogenetic distance. (a) Tallies and percentages of pairwise genome comparisons, by whether they have at least one horizontal gene transfer (HGT) event inferred and whether they co-occur. (b) Boxplots over violin plots of all genomes tallied in panel a, displaying phylogenetic distance by HGT and co-occurrence. Note that the width of the violin plots does not reflect the total sample size for each category (see panel a for counts in each category).

Phylogenetic distance has previously been proposed as partially explaining the association between co-occurrence and HGT, but other factors may also constrain this link. An HGT event might be adaptive under specific environmental conditions, making it more likely to be retained and observed. Indeed, genome pairs connected by HGT tend to be sampled under more similar environmental conditions compared to genomes without evidence for HGT (**Figure 4**). This result highlights that including information on environmental factors is also needed, in addition to phylogenetic distance, to better interpret the relationship between co-occurrence and HGT.

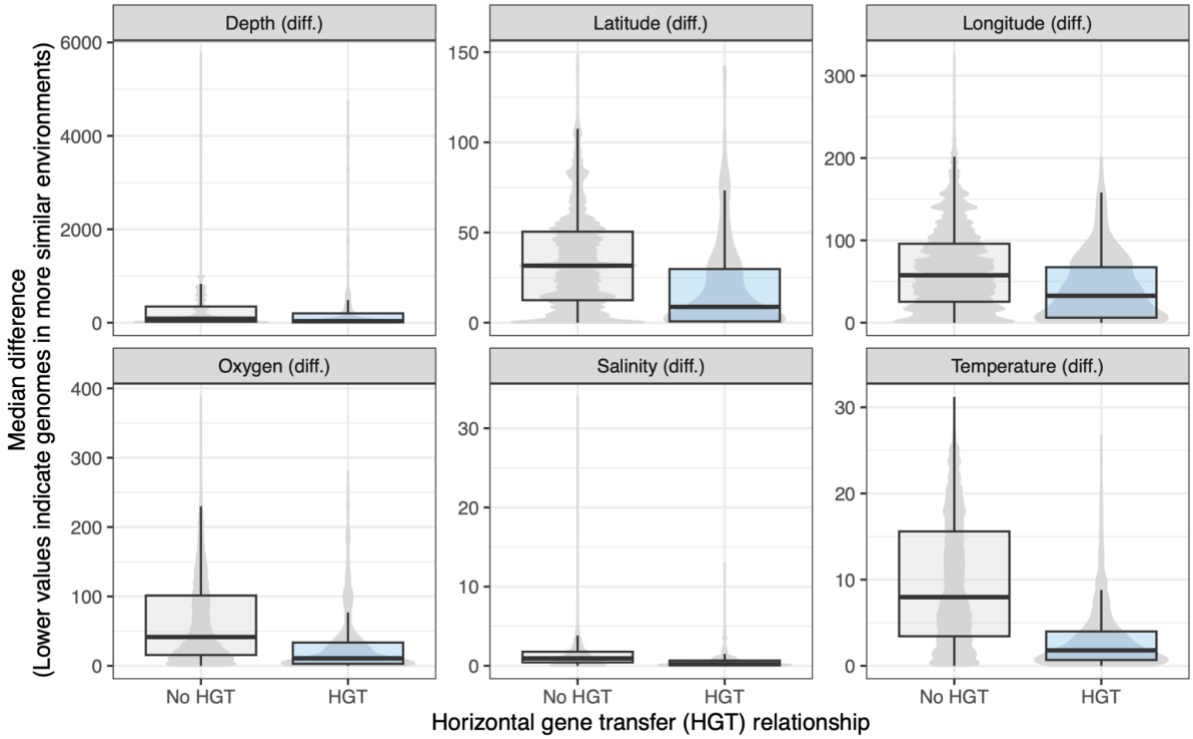


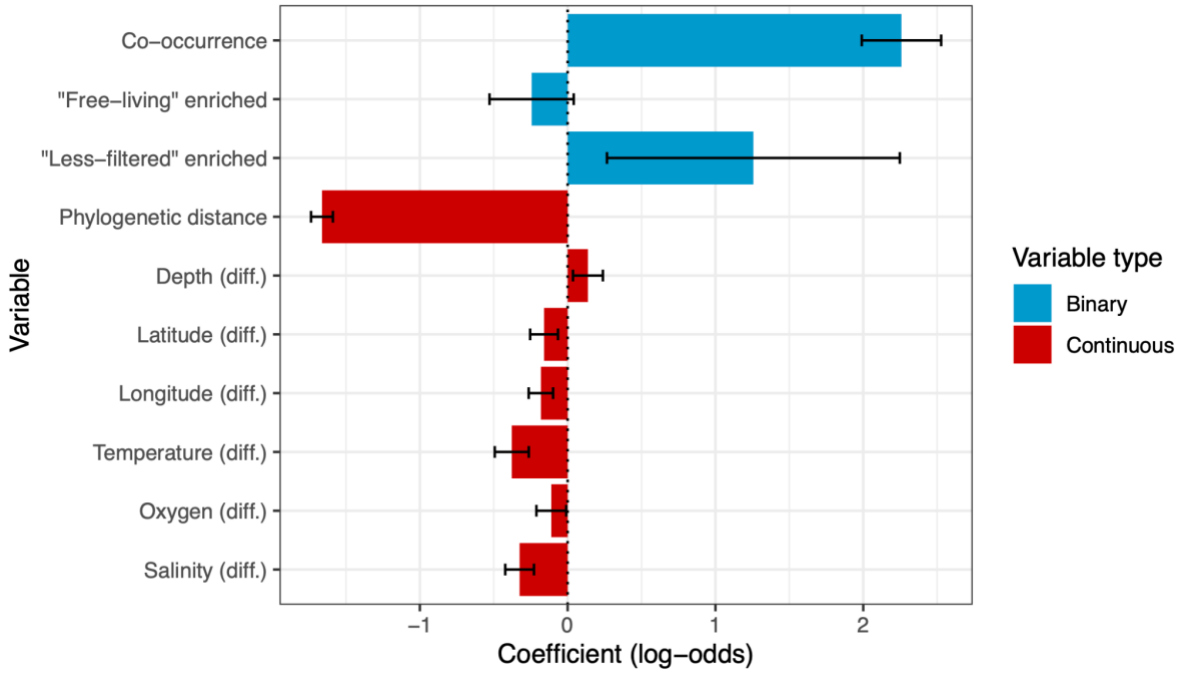
Figure 4: Distributions of the difference in median value of each environmental variable between genome pairs, split by genome pairs connected by HGT and not. The median value for each variable was computed for each genome based on all the metagenomics samples it was found within.

Testing for a relationship between horizontal gene transfer and co-occurrence, after correcting for phylogeny and environmental factors

We next investigated whether the relationship between HGT and co-occurrence remains, and to what degree, after controlling for phylogenetic distance and median environmental differences. To do so, we built a logistic regression model to predict the binary state of whether the genome pairs are connected by HGT or not. The predictor variables included whether the genome pairs are co-occurring or not, as well as the continuous phylogenetic distances and median environmental differences. Most of the measured environmental factors were abiotic (e.g., pH, temperature, depth), but we also included a biotic "lifestyle" factor as a proxy for genomes associated with free-living or potentially particle-associated niches, quantified by different filter sizes (see Methods). A smaller filter-cutoff enriches for free-living cells, whereas a larger filter can include particle-attached prokaryotes, which we refer to as "less-filtered" samples (although note these samples also include free-living microbes, and are not solely particle-attached). We hypothesized that genomes of particle-associated microbes are more likely to undergo HGT due to higher cell density and potentially higher encounter rates with other taxa.

514 Based on the model coefficients (**Figure 5**), the clearest predictor of HGT is co-occurrence
515 (2.26, $P < 0.001$). Since this is a binary predictor, this refers to the increased log-odds of being
516 connected by HGT when genomes co-occur, or 9.58-fold ($e^{2.26}$) increased odds. Phylogenetic
517 distance was also highly significant in the model (-1.66, $P < 0.001$). As this is a continuous
518 variable this means that for every one unit decrease in transformed phylogenetic distance
519 (which corresponds to one standard deviation due to the orderNorm transformation), there
520 is a 5.26-fold decrease in HGT. Although the mean differences in all environmental variables
521 were also significant, the absolute coefficients are much lower compared to phylogenetic
522 distance. Finally, pairs where both genomes are associated with "less-filtered" metagenomics
523 samples, which are more likely to include particle-attached prokaryotes, were positively
524 associated with HGT (1.26, $P=0.013$; although note the wide 95% confidence intervals).
525 Overall, this result is consistent with the relationship between co-occurrence and HGT
526 remaining robust after also including potential ecological drivers.
527
528 We fit this full logistic regression model, and other models based on subsets of predictors.
529 We found the full model fit best based on the Akaike Information Criterion (AIC; difference in
530 AIC of -3.46 compared to the next best model). In addition, although there is potential for
531 multicollinearity between these variables, the variance inflation factor was < 2 in all cases,
532 meaning that this is not a concern for this model.
533
534 Although the focal dataset we have highlighted throughout is based on robust methods for
535 determining co-occurrence and HGT, it is possible that our interpretations above could
536 change depending on the bioinformatic workflow and data subset. Accordingly, we re-ran this
537 logistic regression analysis based on two other methods for determining co-occurrence and
538 two other methods for determining HGT, as well as with genomes from the proGenomes
539 database (**Supplementary Figure 3**; see Methods). Although the differing methods have
540 similar motivations, they can result in highly different profiles. For instance, identifying HGT
541 with RANGER-DTL between strains of a species resulted in a huge number of putative events
542 identified (the positive intercept for these models indicating that more than half of genome
543 comparisons were called to have at least one HGT, or recombination, event). Nonetheless, in
544 almost all cases, the full models which included all tested variables produced the lowest AIC
545 values. Exceptions included the model based on cluster-based HGT and 'simple' co-
546 occurrence, as well as all models based on the proGenomes database, where the models
547 excluding the filter-cutoff groupings fit best. In the case of the proGenomes database models,
548 this is likely due to few genomes found to be associated with each grouping for this genome
549 database (see Methods).
550
551 Overall, this broader investigation resulted in the same major interpretations: phylogenetic
552 distance and co-occurrence are strongly associated with HGT as expected, even after
553 including additional factors. One exception to our general interpretation is a slight negative

554 association (-0.13, P < 0.001) between co-occurrence and HGT in the model based on BLASTn-
555 identified HGT events and propr-based co-abundance (**Supplementary Figure 3**; see
556 Discussion). Nonetheless, we also found overall concordance of our results based on an
557 independent genome database (proGenomes), from which we excluded metagenomics-
558 derived assemblies, and corresponded to higher quality genomes than our focal set (although
559 with fewer inferred HGT events).



560
561 **Figure 5**: Coefficients in logistic regression model based on all pairwise genome
562 comparisons. The response is whether those genomes have at least one called
563 horizontal gene transfer event (based on the cluster-based approach) between
564 them or not. Co-occurrence (based on the hypergeometric approach) is a binary
565 variable indicating whether those genomes significantly co-occur, and "free-
566 living" and "less-filtered" enriched indicate whether the genomes being
567 compared are both enriched in that sample-type. The other variables are ordered
568 quantile normalized continuous values. Variables with 'diff.' represent the median
569 inter-sample difference in each environmental variable for all samples encoding
570 the two genomes being compared. Coefficient values less than 0 indicate that
571 decreases in those variables are associated with increases in horizontal gene
572 transfer, while a positive value represents the opposite. The intercept for this
573 model was -14.35. Error bars represent 95% confidence intervals (1.96 x standard
574 error). Note that all variables are significant (P < 0.05) except for "free-living"
575 (enriched).
576
577 Finally, we also ran additional checks to ensure our interpretations were robust. We re-ran
578 several models on subsets of the overall OceanDNA metagenomics dataset corresponding to
579 the largest subsets: samples from *Tara* Oceans and GEOTRACES separately (**Supplementary**

580 **Figure 4**). Our goal here was to ensure that dataset substructure was not driving our results.
581 The model coefficients were qualitatively similar compared to the models on the entire
582 dataset (with one exception, but which included few HGT events and had high
583 multicollinearity). We also ran generalized linear mixed models to control for the genome IDs
584 in each comparison (i.e., to control for individual genomes which could be disproportionately
585 driving the observed signals). To do so, we duplicated the input table used for focal models,
586 to allow each genome ID to be associated with each comparison separately. Due to this
587 approach, the error estimates for these models are less reliable overall compared to the
588 original logistic regressions. Nonetheless, these models provide insight into how controlling
589 for genome ID can affect the models, and we found that our overall interpretations were
590 robust (**Supplementary Figure 4**). The exception was that the association trends for the filter-
591 cutoff groupings associated with genome pairs were less consistent (see Discussion). This is
592 not unexpected, as there were only four genome pairs associated with "less-filtered" samples
593 and called for HGT in the original analysis. These pairs consisted of only seven unique
594 genomes (**Supplementary Table 1**).
595
596 **Discussion**
597 We have shown that co-occurring environmental genomes in ocean metagenomics data are
598 more likely to show evidence for HGT, compared to genomes not co-occurring. In addition,
599 we have demonstrated that phylogenetic distance and measured environmental
600 characteristics are insufficient to fully explain this observation. Our work complements recent
601 analyses based on 16S rRNA gene sequencing profiles to assess co-occurring bacteria, and
602 reference genomes to assess putative HGT (Dmitrijeva et al., 2024). Based on metagenomics
603 data, we also found a clear negative association between phylogenetic distance and both co-
604 occurrence (at low phylogenetic distances) and HGT. In addition, in agreement with this
605 earlier study, we found that the association between co-occurrence and HGT holds after
606 controlling for phylogenetic distance. This result, and our finding that this association is also
607 robust to controlling for measured environmental factors, underlines the significant influence
608 of ecological preferences (biotic and abiotic) in shaping marine bacteria evolution.
609
610 Through our main workflow for identifying HGT events, we were able to stratify genes
611 putatively shared through HGT by taxonomic level and sequence identity (as a proxy for time
612 since transfer). The functional enrichment of these genes largely agreed with expectations
613 from the literature (**Figure 2**), which gave us confidence in our approach. However, we also
614 noted a few enrichment differences between taxonomic levels and sequence identity.
615 Particularly for enrichments based on genes at >= 99% identity, we found several functional
616 enrichments in opposite directions at different taxonomic levels. For instance, amino acid
617 transport and metabolism was depleted for transfer events at the genus level, but enriched
618 at the phylum and domain levels. Such signals, particularly for differences at higher taxonomic

levels for which there are fewer total HGT events, would be missed if functional enrichment were performed on all genes combined.

The association between HGT and both co-occurrence and phylogenetic distance is extremely clear in our data (**Figure 1** and **Figure 3**). However, the expected association between co-occurrence and phylogenetic distance was only present in genome pairs within the same genus or lower. This observation highlights that closely related taxa are especially driving this signal, rather than a consistent trend at all phylogenetic distances, at least in this marine dataset. In addition, we identified an intriguing trend where genome pairs that co-occur and show evidence for HGT were significantly more phylogenetically distant compared to those that do not co-occur yet undergo HGT (**Figure 3b**). This could indicate there are more complex associations between co-occurrence, HGT, and phylogenetic distance than previously appreciated. For instance, this observation could be driven by metabolic complementation across genomes, which is known to be associated with co-occurrence (Giordano et al., 2024), and could be enriched for distantly related genomes in our dataset. However, one technical explanation for this observation, which we think is more likely, is collider bias (Holmberg & Andersen, 2022). If both phylogenetic distance and co-occurrence independently affect the probability of genomes being connected by HGT, then when conditioning on one (e.g., co-occurrence) it is expected to see a different distribution of phylogenetic distances for those genome pairs. In other words, if genome pairs can be connected by HGT due to either being phylogenetically similar or due to co-occurring, then it is expected that genome pairs connected by HGT, but which are not co-occurring, will have lower phylogenetic distance on average.

There are at least two biological mechanisms that could underlie the general observation of an association between HGT and co-occurrence. First, taxa that are physically closer could simply have more opportunities for HGT, which must be true to some extent. For instance, taxa in completely isolated environments, with no gene flow or possibility of DNA transfer, cannot undergo HGT. In regions of the ocean stratified by geography, depth, and environmental factors, it is reasonable to expect that taxa would undergo HGT more frequently with other taxa sharing the same local environment.

This expectation was the motivation for classifying genomes by sampling filter cut-off, either in the range of free-living or potentially particle-attached taxa. Our expectation was that particle-attached taxa, being consistently physically proximal to other cells, would undergo higher rates of HGT. Our analysis focused on “less-filtered” metagenomics samples rather than strictly metagenomics samples in the range of particle-attached taxa (see Methods). We did find support for this expectation (**Figure 5**), but this was based on only seven unique genomes driving this signal. It is thus unsurprising that the signal was lost in models where genome ID was included as a random effect. In addition, the split of metagenomics samples

659 by what sampling filter cut-off was used is confounded by the underlying dataset: samples
660 within the free-living range are primarily *Tara* Oceans samples, while those which could
661 include particle-attached bacteria are primarily from the Geotraces project. Accordingly, it
662 could be driven by a few unrepresentative taxa and differences among datasets.
663
664 Another biological mechanism that could explain our overall finding of an association
665 between co-occurrence and HGT, is that adaptive genes spread through HGT could be
666 beneficial to taxa in similar environmental conditions. We investigated this hypothesis by
667 testing whether the difference in median environmental conditions of samples where each
668 compared taxon is present can account for the observed HGT levels. Importantly, where this
669 is the case, HGT would be more common among both taxa that co-occur and among those
670 that do not co-occur, but that nonetheless are found in samples where similar environmental
671 pressures are present. Although the differences in environmental factors were significant
672 predictors in our models (**Figure 5**), with higher differences usually associated with decreased
673 HGT, co-occurrence remained a strong predictor in our models. In addition, environmental
674 differences were much less predictive compared to phylogenetic distance (which is easiest to
675 compare to, as all continuous variables were normalized to have the same interpretation of
676 1 unit). This analysis is dependent upon the accuracy of the environmental data
677 measurements present in the OceanDNA resource. In addition, the environmental factors
678 considered are limited (e.g., salinity and temperature). More detailed environmental factors
679 might explain higher variation in HGT across genome pairs, and co-occurrence likely captures
680 many unmeasured environmental factors.
681
682 A novel aspect of our study is that we focused on metagenomic sequencing data, in a
683 complementary fashion to the approach of analyzing 16S ribosomal RNA gene sequencing
684 profiles and matching these to reference genomes (Dmitrijeva et al., 2024). However, most
685 genomes in our focal dataset were MAGs, which are known to have contamination and
686 misassembly issues (Shaiber & Eren, 2019; Majernik et al., 2024). In addition, depending on
687 species clonality within a given metagenomic sample, MAGs may represent a population
688 genome for a species, which may bias and preclude the detection of HGT across MAGs. To
689 address these problems, we conducted an additional analysis based on high quality genomes
690 from the proGenomes database, with MAGs excluded. Our overall interpretations were
691 similar (**Supplementary Figure 3**), suggesting that our inferences of HGT were not driven by
692 misassembled MAGs.
693
694 We also considered a variety of bioinformatic workflows for detecting HGT and determining
695 co-occurrence. One clear observation was that within-species inferences of HGT and
696 recombination with RANGER-DTL led to an extremely high number of putative hits. Since
697 phylogenetic reconciliation approaches are sensitive to errors in tree-building (Menet et al.,
698 2022), we cannot reject that this observation is simply driven by inaccuracies in the gene

trees. Indeed, we are skeptical that these hits are reliable inferences for exact HGT events, but nonetheless these results produced similar model outputs as our other approaches. In one model we found results that disagreed with the overall interpretation that co-occurrence and HGT are positively associated: a model based on BLASTn-inferred HGT events and propr co-abundance (**Supplementary Figure 3**). We focused on the cluster-based HGT events, as BLASTn-inferred events reflect a network of potential HGT partners. Specifically, all genomes that share genes above a certain identity cut-off were considered HGT connected based on the BLASTn results, even if a better match exists. In addition, unlike co-occurrence based on presence/absence of genomes, the propr results are based on genome relative abundances. There could be a bias where similar genes are difficult to independently assemble into separate MAGs when these genes are present in genomes at similar abundances in metagenomics data. In any case, all other models, including those based on more straightforward data, agreed with our general interpretation.

In conclusion, we have shown that the relationship between co-occurrence and HGT is robust to phylogeny and measured environmental factors. While we did observe increased HGT between genomes sampled under more similar abiotic conditions, and potentially among particle-associated genomes, these are relatively minor contributors to global patterns of HGT, which can largely be explained by co-occurrence. Future work disentangling the importance of physical proximity (at microscopic or larger scales) compared to more specific shared environmental pressures in driving HGT in the global ocean is needed.

Acknowledgements

We would like to thank all members of the Bobay, Chaffron, and Shapiro labs for feedback.

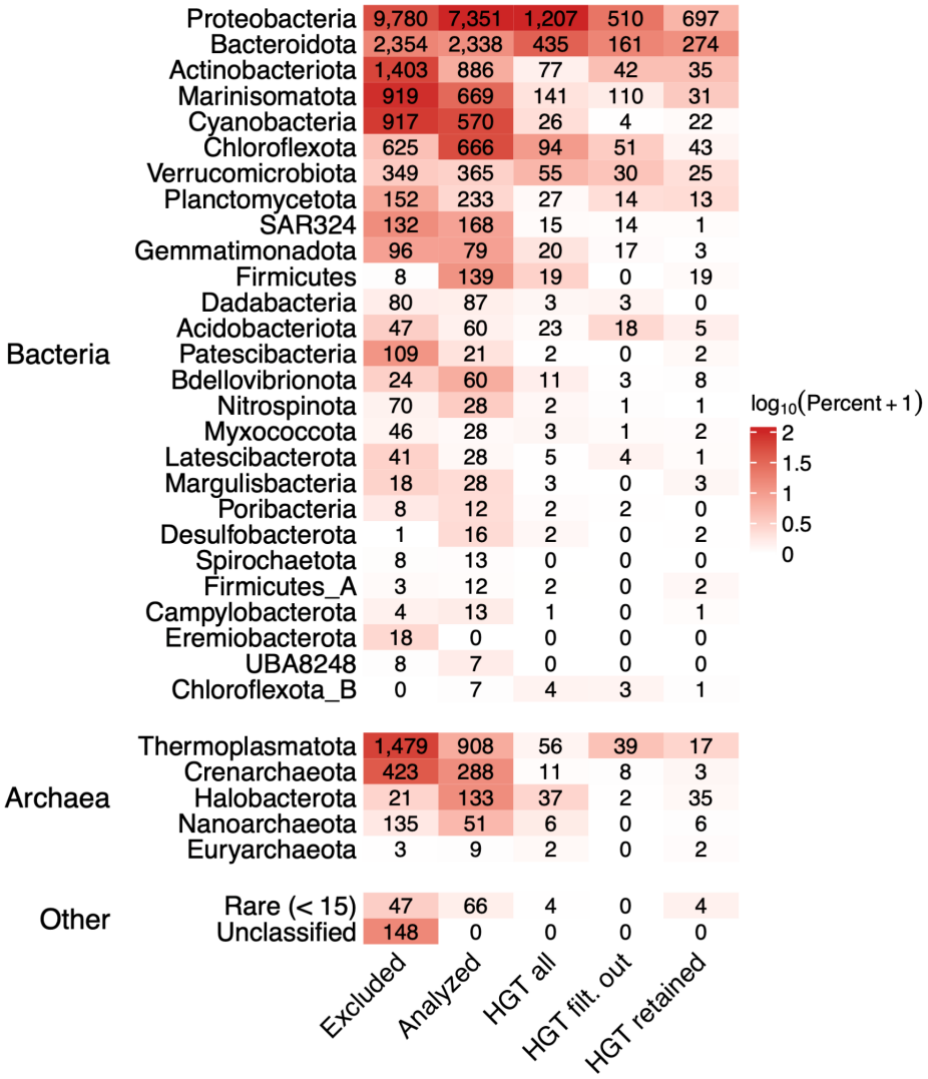
Funding

GMD is supported by a Banting Postdoctoral Fellowship from the Government of Canada. NT is supported as a Chaire de Professeur Junior from the French l'Institut national de recherche pour l'agriculture, l'alimentation et l'environnement. LMB is supported by the US National Institutes of Health NIGMS (R01GM132137). PL is supported by a Canadian Institutes of Health Research Postdoctoral Fellowship from the Government of Canada. This work was supported by a RFI ATLANSTIC2020 (ECOMOD grant to SC), the H2020 project AtlantECO (award number 862923), and a Natural Sciences and Engineering Research Council of Canada Discovery Grant to BJS. The computer infrastructure used for analyses was partially supported by the Canadian Foundation for Innovation's John R. Evans Leaders Fund and by the bioinformatics core facility of Nantes (BiRD - Biogenouest), Nantes Université, France.

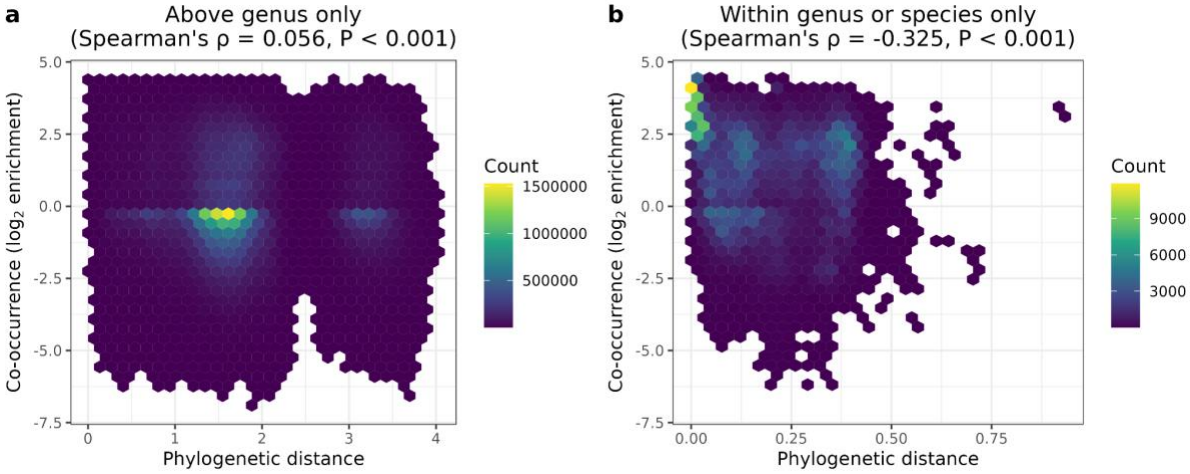
Conflict of interest disclosure

The authors state that they have no conflicts of interest with the content of this article.

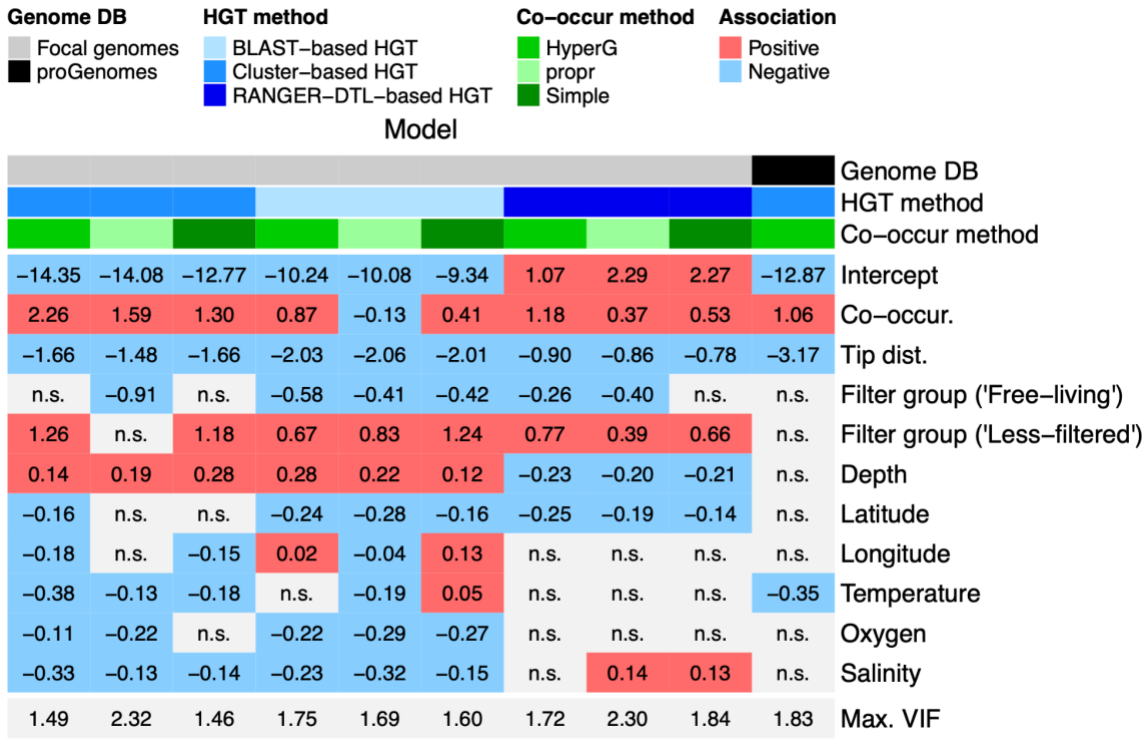
Supplementary Materials



Supplementary Figure 1: Counts of analyzed genomes by phylum, split into the bacterial and archaeal domains, and coloured by percentage. Only genomes from the focal genome database are shown. “Excluded”: genomes excluded due to being low quality. “Analyzed”: the set of all genomes analyzed in this study. “HGT all”: all genomes with at least one putative horizontal gene transfer (HGT) event identified based on the clustering approach (which are a subset of “Analyzed”). The “HGT filt. out” and “HGT retained” groupings represent genomes that were subsets of “HGT all”, and were either below or above the metagenomics sample prevalence cut-off, respectively. In other words, they represent genomes that were lost or retained after requiring genomes to be present in 10 samples based on mapping metagenomics reads to genomes with CoverM.

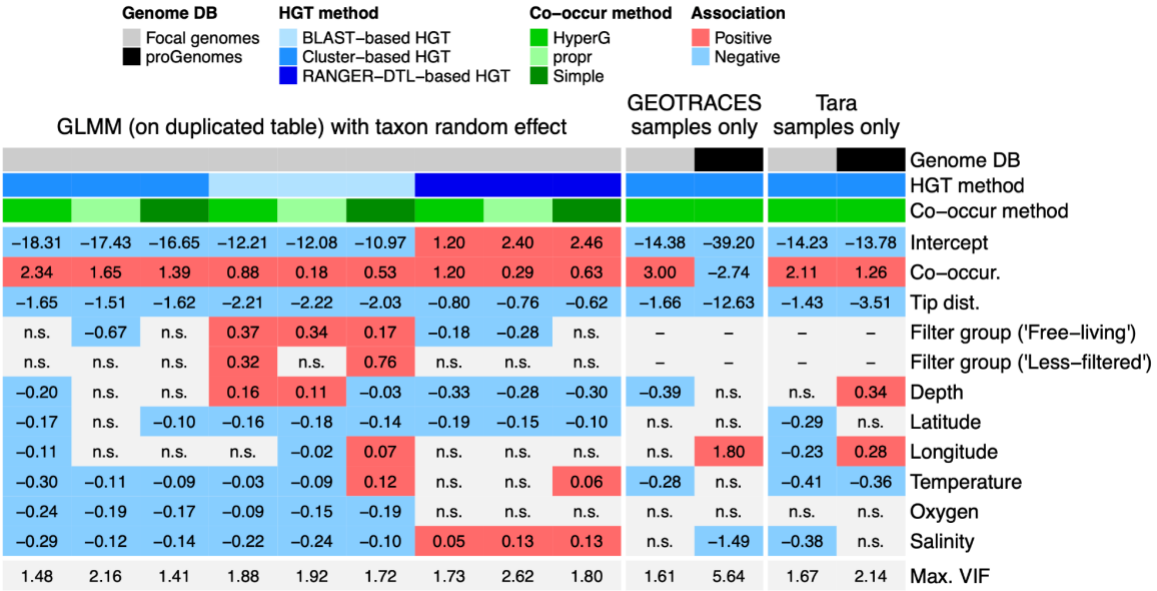


Supplementary Figure 2: Hex plots (coarse scatterplots with coloured hexes indicating counts of overlapping points) of the co-occurrence and phylogenetic distance for all pairwise genome comparisons in the focal dataset. (a) Restricted to genome comparisons between different genera and higher (which we focused on). (b) Restricted to genome comparisons between genomes in the same species or genus (which we excluded from all other analyses for our focal workflow). The plotted co-occurrence metric is: $\log_2((\text{Observed} + 1)/(\text{Expected} + 1))$, where "Observed" is the number of samples observed to share both genomes based on CoverM mapping, and "Expected" is the number expected by chance, based on the prevalence of each genome. Spearman correlation results are indicated above each panel. Note that for the Spearman correlation tests, simply the ratio of Observed/Expected was used as the co-occurrence metric.



−14.35	−14.08	−12.77	−10.24	−10.08	−9.34	1.07	2.29	2.27	−12.87	Intercept
2.26	1.59	1.30	0.87	−0.13	0.41	1.18	0.37	0.53	1.06	Co–occur.
−1.66	−1.48	−1.66	−2.03	−2.06	−2.01	−0.90	−0.86	−0.78	−3.17	Tip dist.
n.s.	−0.91	n.s.	−0.58	−0.41	−0.42	−0.26	−0.40	n.s.	n.s.	Filter group ('Free–living')
1.26	n.s.	1.18	0.67	0.83	1.24	0.77	0.39	0.66	n.s.	Filter group ('Less–filtered')
0.14	0.19	0.28	0.28	0.22	0.12	−0.23	−0.20	−0.21	n.s.	Depth
−0.16	n.s.	n.s.	−0.24	−0.28	−0.16	−0.25	−0.19	−0.14	n.s.	Latitude
−0.18	n.s.	−0.15	0.02	−0.04	0.13	n.s.	n.s.	n.s.	n.s.	Longitude
−0.38	−0.13	−0.18	n.s.	−0.19	0.05	n.s.	n.s.	n.s.	−0.35	Temperature
−0.11	−0.22	n.s.	−0.22	−0.29	−0.27	n.s.	n.s.	n.s.	n.s.	Oxygen
−0.33	−0.13	−0.14	−0.23	−0.32	−0.15	n.s.	0.14	0.13	n.s.	Salinity
1.49	2.32	1.46	1.75	1.69	1.60	1.72	2.30	1.84	1.83	Max. VIF

Supplementary Figure 3: Coefficients in logistic regression models based on all pairwise genome comparisons, as in Figure 5 but for a wide range of input data. The first column corresponds to the focal model, displayed in Figure 5. Each column is a different model, with input data described by the top three coloured strips: the genome database (DB), horizontal gene transfer (HGT)-detection method, and co-occurrence method. All coloured cells in the heatmap represent significant coefficients per model, and their direction. Non-significant coefficients (P >= 0.05) are indicated by "n.s.". The final row is the maximum variance inflation factor (Max. VIF) per model, indicating that the input variables are moderately, but not highly, correlated. Note that the variable types (binary vs. continuous) are the same as Figure 5, which affects the interpretation of these coefficients, except for co-occurrence in the simple and propr-based models, which was continuous. In addition, all environmental variables tested refer to the mean difference in environmental variables.



−18.31	−17.43	−16.65	−12.21	−12.08	−10.97	1.20	2.40	2.46	−14.38	−39.20	−14.23	−13.78	Intercept
2.34	1.65	1.39	0.88	0.18	0.53	1.20	0.29	0.63	3.00	−2.74	2.11	1.26	Co–occur.
−1.65	−1.51	−1.62	−2.21	−2.22	−2.03	−0.80	−0.76	−0.62	−1.66	−12.63	−1.43	−3.51	Tip dist.
n.s.	−0.67	n.s.	0.37	0.34	0.17	−0.18	−0.28	n.s.	–	–	–	–	Filter group ('Free–living')
n.s.	n.s.	n.s.	0.32	n.s.	0.76	n.s.	n.s.	n.s.	–	–	–	–	Filter group ('Less–filtered')
−0.20	n.s.	n.s.	0.16	0.11	−0.03	−0.33	−0.28	−0.30	−0.39	n.s.	n.s.	0.34	Depth
−0.17	n.s.	−0.10	−0.16	−0.18	−0.14	−0.19	−0.15	−0.10	n.s.	n.s.	−0.29	n.s.	Latitude
−0.11	n.s.	n.s.	n.s.	−0.02	0.07	n.s.	n.s.	n.s.	n.s.	1.80	−0.23	0.28	Longitude
−0.30	−0.11	−0.09	−0.03	−0.09	0.12	n.s.	n.s.	0.06	−0.28	n.s.	−0.41	−0.36	Temperature
−0.24	−0.19	−0.17	−0.09	−0.15	−0.19	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	Oxygen
−0.29	−0.12	−0.14	−0.22	−0.24	−0.10	0.05	0.13	0.13	n.s.	−1.49	−0.38	n.s.	Salinity
1.48	2.16	1.41	1.88	1.92	1.72	1.73	2.62	1.80	1.61	5.64	1.67	2.14	Max. VIF

Supplementary Figure 4: Similar to Supplementary Figure 3, but for additional model outputs. The first nine columns are results for generalized linear mixed models (GLMMs) with genome ID as the random effect. The last four columns are model coefficients for logistic regression models based on the subset of datasets with only GEOTRACES or *Tara* oceans samples. Non-significant coefficients (P >= 0.05) are indicated by "n.s.". The final row is the maximum variance inflation factor per model, which indicates that most models do not have concerning levels of multicollinearity, except for the proGenomes GEOTRACES-only model with a value > 5, indicating high multicollinearity. Note also the extremely low intercept for this model (-39.20), which reflects there being few HGT events identified for this subset.

826 **Supplementary Table 1: Taxonomy of seven metagenome-assembled genomes driving**
827 **signal of positive association between genomes that are 'less-filtered' and horizontal gene**
828 **transfer**

Study-assigned taxon ID	Taxonomy
Taxa_11628	d__Bacteria; p__Chloroflexota; c__Dehalococcoidia; o__UBA3495; f__UBA3495; g__UBA9611; s__; TARA_SAMEA2623756_METAG_DIFAIHGK
Taxa_3418	d__Bacteria; p__Actinobacteriota; c__Acidimicrobiia; o__Microtrichales; f__MedAcidi-G1; g__UBA9410; s__; MALA_SAMN05422166_METAG_OAKDMEED
Taxa_8602	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae; g__MED-G13; s__; TARA_SAMEA2621990_METAG_OMFNCAAD
Taxa_442	d__Bacteria; p__Chloroflexota; c__Dehalococcoidia; o__UBA3495; f__UBA3495; g__UBA3495; s__UBA3495 sp002716645; BGEO_SAMN07136539_METAG_DOCKDDBM
Taxa_86	d__Bacteria; p__Chloroflexota; c__Dehalococcoidia; o__UBA3495; f__UBA3495; g__UBA3495; s__UBA3495 sp002716645; BATS_SAMN07137087_METAG_CDOPEKMJ
Taxa_892	d__Bacteria; p__Gemmatimonadota; c__Gemmatimonadetes; o__SG8-23; f__UBA6960; g__UBA1138; s__UBA1138 sp003447875; BGEO_SAMN07136727_METAG_BJDLPBMF
Taxa_8654	d__Bacteria; p__Bacteroidota; c__Bacteroidia; o__Flavobacteriales; f__Flavobacteriaceae; g__MED-G14; s__; TARA_SAMEA2622074_METAG_COJFLNPO

829
830
831
832
833
834
835
836
837
838
839
840

841 **References**

842 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search
843 Tool. *Journal Molecular Biology*, **215**, 403–410. https://doi.org/10.1016/S0022-
844 2836(05)80360-2

845 Aroney STN, Newell RJP, Nissen J, Camargo AP, Tyson GW, Woodcroft BJ (2024) CoverM:
846 Read coverage calculator for metagenomics.
847 https://doi.org/10.5281/zenodo.10531254

848 Azam F, Malfatti F (2007) Microbial structuring of marine ecosystems. *Nature Reviews*
849 *Microbiology*, **5**, 782–791. https://doi.org/10.1038/nrmicro1747

850 Bansal MS (2022) Deciphering Microbial Gene Family Evolution Using Duplication-Transfer-
851 Loss Reconciliation and RANGER-DTL. In: *Environmental Microbial Evolution:*
852 *Methods and Protocols* (ed Luo H), pp. 233–252. Springer US, New York, NY.
853 https://doi.org/10.1007/978-1-0716-2691-7_11

854 Bansal MS, Kellis M, Kordi M, Kundu S (2018) RANGER-DTL 2.0: rigorous reconstruction of
855 gene-family evolution by duplication, transfer and loss. *Bioinformatics*, **34**, 3214–
856 3216. https://doi.org/10.1093/bioinformatics/bty314

857 Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *PNAS USA*,
858 **102**, 14332–14337. https://doi.org/10.1073/pnas.0504068102

859 Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and
860 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series*
861 *B (Methodological)*, **57**, 289–300.

862 Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, Hogle SL, Coe A, Bergauer
863 K, Bouman HA, Browning TJ, De Corte D, Hassler C, Hulston D, Jacquot JE, Maas EW,
864 Reinthaler T, Sintes E, Yokokawa T, Chisholm SW (2018) Marine microbial

865 metagenomes sampled across space and time. *Scientific Data*, **5**, 180176.
866 https://doi.org/10.1038/sdata.2018.176
867 Brooks M E, Kristensen K, Benthem K J ,van, Magnusson A, Berg C W, Nielsen A, Skaug H J,
868 Mächler M, Bolker B M (2017) glmmTMB Balances Speed and Flexibility Among
869 Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, **9**,
870 378–400. https://doi.org/10.32614/RJ-2017-066
871 Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC
872 (2024) Identification of mobile genetic elements with geNomad. *Nature*
873 *Biotechnology*, **42**, 1303–1312. https://doi.org/10.1038/s41587-023-01953-y
874 Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) eggNOG-
875 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction
876 at the Metagenomic Scale. *Molecular Biology and Evolution*, **38**, 5825–5829.
877 https://doi.org/10.1093/molbev/msab293
878 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated
879 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–
880 1973. https://doi.org/10.1093/bioinformatics/btp348
881 Chamberlain SA, Szöcs E (2013) taxize: taxonomic search and retrieval in R. *F1000 Research*.
882 https://doi.org/10.12688/f1000research.2-191.v2
883 Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH (2020) GTDB-Tk: a toolkit to classify
884 genomes with the Genome Taxonomy Database. *Bioinformatics*, **36**, 1925–1927.
885 https://doi.org/10.1093/bioinformatics/btz848
886 Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor.
887 *Bioinformatics*, **34**, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

888 Cohen O, Gophna U, Pupko T (2011) The Complexity Hypothesis Revisited: Connectivity
889 Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer. *Molecular*
890 *Biology and Evolution*, **28**, 1481–1489. https://doi.org/10.1093/molbev/msq333
891 Cohen O, Pupko T (2010) Inference and Characterization of Horizontally Transferred Gene
892 Families Using Stochastic Mapping. *Molecular Biology and Evolution*, **27**, 703–713.
893 https://doi.org/10.1093/molbev/msp240
894 Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on
895 prokaryotic genome size. *Proceedings of the National Academy of Sciences of the*
896 *United States of America*, **106**, 21748–21753.
897 https://doi.org/10.1073/pnas.0907584106
898 Crump BC, Armbrust EV, Baross JA (1999) Phylogenetic Analysis of Particle-Attached and
899 Free-Living Bacterial Communities in the Columbia River, Its Estuary, and the
900 Adjacent Coastal Ocean. *Applied and Environmental Microbiology*, **65**, 3192–3204.
901 https://doi.org/10.1128/AEM.65.7.3192-3204.1999
902 Dmitrijeva M, Tackmann J, Matias Rodrigues JF, Huerta-Cepas J, Coelho LP, von Mering C
903 (2024) A global survey of prokaryotic genomes reveals the eco-evolutionary
904 pressures driving horizontal gene transfer. *Nature Ecology & Evolution*, 1–13.
905 https://doi.org/10.1038/s41559-024-02357-0
906 Douglas GM, Langille MGI (2019) Current and promising approaches to identify horizontal
907 gene transfer events in metagenomes. *Genome Biology and Evolution*, **11**, 2750–
908 2766. https://doi.org/10.1093/gbe/evz184
909 Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
910 https://doi.org/10.1093/bioinformatics/14.9.755

911 Eddy SR (2011) Accelerated Profile HMM Searches. *PLOS Computational Biology*, **7**,
912 e1002195. https://doi.org/10.1371/journal.pcbi.1002195
913 Edgar RC (2022) Muscle5: High-accuracy alignment ensembles enable unbiased assessments
914 of sequence homology and phylogeny. *Nature Communications*, **13**, 6968.
915 https://doi.org/10.1038/s41467-022-34630-w
916 Fondi M, Karkman A, Tamminen MV, Bosi E, Virta M, Fani R, Alm E, McInerney JO (2016)
917 "Every Gene Is Everywhere but the Environment Selects": Global Geolocalization of
918 Gene Sharing in Environmental Samples through Network Analysis. *Genome Biology*
919 *and Evolution*, **8**, 1388–1400. https://doi.org/10.1093/gbe/evw077
920 Foster I (2011) Globus Online: Accelerating and Democratizing Science through Cloud-Based
921 Services. *IEEE Internet Computing*, **15**, 70–73. https://doi.org/10.1109/MIC.2011.64
922 Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: Accelerated for clustering the next-generation
923 sequencing data. *Bioinformatics*, **28**, 3150–3152.
924 https://doi.org/10.1093/bioinformatics/bts565
925 Fullam A, Letunic I, Schmidt TSB, Ducarmon QR, Karcher N, Khedkar S, Kuhn M, Larralde M,
926 Maistrenko OM, Malfertheiner L, Milanese A, Rodrigues JFM, Sanchis-López C,
927 Schudoma C, Szklarczyk D, Sunagawa S, Zeller G, Huerta-Cepas J, von Mering C, Bork
928 P, Mende DR (2023) proGenomes3: approaching one million accurately and
929 consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Research*, **51**,
930 D760–D766. https://doi.org/10.1093/nar/gkac1078
931 Giordano N, Gaudin M, Trottier C, Delage E, Nef C, Bowler C, Chaffron S (2024) Genome-
932 scale community modelling reveals conserved metabolic cross-feedings in epipelagic

933 bacterioplankton communities. *Nature Communications*, **15**, 2721.
934 https://doi.org/10.1038/s41467-024-46374-w
935 Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J
936 (2018) Bioconda: sustainable and comprehensive software distribution for the life
937 sciences. *Nature Methods*, **15**, 475–476. https://doi.org/10.1038/s41592-018-0046-7
938 Heneghan RF, Holloway-Brown J, Gasol JM, Herndl GJ, Morán XAG, Galbraith ED (2024) The
939 global distribution and climate resilience of marine heterotrophic prokaryotes.
940 *Nature Communications*, **15**, 6943. https://doi.org/10.1038/s41467-024-50635-z
941 Holmberg MJ, Andersen LW (2022) Collider Bias. *JAMA*, **327**, 1282–1283.
942 https://doi.org/10.1001/jama.2022.1820
943 Innamorati KA, Earl JP, Aggarwal SD, Ehrlich GD, Hiller NL (2020) The Bacterial Guide to
944 Designing a Diversified Gene Portfolio. In: *The Pangenome: Diversity, Dynamics and*
945 *Evolution of Genomes* (eds Tettelin H, Medini D), pp. 51–87. Springer, New York.
946 https://doi.org/10.1007/978-3-030-38281-0_3
947 Jeong H, Arif B, Caetano-Anollés G, Kim KM, Nasir A (2019) Horizontal gene transfer in
948 human-associated microorganisms inferred by phylogenetic reconstruction and
949 reconciliation. *Scientific Reports*, **9**, 5953. https://doi.org/10.1038/s41598-019-
950 42227-5
951 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
952 improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
953 https://doi.org/10.1093/molbev/mst010
954 Khedkar S, Smyshlyaev G, Letunic I, Maistrenko OM, Coelho LP, Orakov A, Forslund SK,
955 Hildebrand F, Luetge M, Schmidt TSB, Barabas O, Bork P (2022) Landscape of mobile

956 genetic elements and their antibiotic resistance cargo in prokaryotic genomes.
957 *Nucleic Acids Research*, **50**, 3155–3168. https://doi.org/10.1093/nar/gkac163
958 Kloub L, Gosselin S, Fullmer M, Graf J, Gogarten JP, Bansal MS (2021) Systematic Detection
959 of Large-Scale Multigene Horizontal Transfer in Prokaryotes. *Molecular Biology and*
960 *Evolution*, **38**, 2639–2659. https://doi.org/10.1093/molbev/msab043
961 Lee MD, Ponty Y (2019) GToTree: A user-friendly workflow for phylogenomics.
962 *Bioinformatics*, **35**, 4162–4164. https://doi.org/10.1093/bioinformatics/btz188
963 Louca S, Doebeli M (2018) Efficient comparative phylogenetics on large trees.
964 *Bioinformatics*, **34**, 1053–1055. https://doi.org/10.1093/bioinformatics/btx701
965 Lüdecke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D (2021) performance: An R
966 Package for Assessment, Comparison and Testing of Statistical Models. *Journal of*
967 *Open Source Software*, **6**, 3139. https://doi.org/10.21105/joss.03139
968 Majernik SN, Beaver L, Bradley PH (2024) Small amounts of misassembly can have
969 disproportionate effects on pangenome-based metagenomic analyses. *bioRxiv*,
970 2024.10.11.617902. https://doi.org/10.1101/2024.10.11.617902
971 Martinez-Gutierrez CA, Aylward FO (2022) Evolutionary Genomics of Marine Bacteria and
972 Archaea. In: *The Marine Microbiome* (eds Stal LJ, Cretoiu MS), pp. 327–354. Springer
973 International Publishing, Cham. https://doi.org/10.1007/978-3-030-90383-1_7
974 Menet H, Daubin V, Tannier E (2022) Phylogenetic reconciliation. *PLOS Computational*
975 *Biology*, **18**, e1010621. https://doi.org/10.1371/journal.pcbi.1010621
976 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear
977 R (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in

the Genomic Era. *Molecular Biology and Evolution*, **37**, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, **36**, 760–766. https://doi.org/10.1038/ng1381

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe-Fadrosh EA (2021) A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, **39**, 499–509. https://doi.org/10.1038/s41587-020-0718-6

Neal ZP (2022) backbone: An R package to extract network backbones. *PLOS ONE*, **17**, e0269137. https://doi.org/10.1371/journal.pone.0269137

Niehus R, Mitri S, Fletcher AG, Foster KR (2015) Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nature Communications*, **6**, 8924. https://doi.org/10.1038/ncomms9924

Nishimura Y, Yoshizawa S (2022) The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Scientific Data*, **9**, 305. https://doi.org/10.1038/s41597-022-01392-5

Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications*, **8**, 841. https://doi.org/10.1038/s41467-017-00808-w

1000 Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ (2015) Co-occurrence of resistance
1001 genes to antibiotics, biocides and metals reveals novel insights into their co-selection
1002 potential. *BMC Genomics*, **16**, 964. https://doi.org/10.1186/s12864-015-2153-5
1003 Paoli L, Ruscheweyh H-J, Forneris CC, Hubrich F, Kautsar S, Bhushan A, Lotti A, Clayssen Q,
1004 Salazar G, Milanese A, Carlström CI, Papadopoulou C, Gehrig D, Karasikov M,
1005 Mustafa H, Larralde M, Carroll LM, Sánchez P, Zayed AA, Cronin DR, Acinas SG, Bork
1006 P, Bowler C, Delmont TO, Gasol JM, Gossert AD, Kahles A, Sullivan MB, Wincker P,
1007 Zeller G, Robinson SL, Piel J, Sunagawa S (2022) Biosynthetic potential of the global
1008 ocean microbiome. *Nature*, **607**, 111–118. https://doi.org/10.1038/s41586-022-
1009 04862-3
1010 Paquola ACM, Asif H, Pereira CA de B, Feltes BC, Bonatto D, Lima WC, Menck CFM (2018)
1011 Horizontal Gene Transfer Building Prokaryote Genomes: Genes Related to Exchange
1012 Between Cell and Environment are Frequently Transferred. *Journal of Molecular*
1013 *Evolution*, **86**, 190–203. https://doi.org/10.1007/s00239-018-9836-x
1014 Peschel S, Müller CL, von Mutius E, Boulesteix A-L, Depner M (2021) NetCoMi: network
1015 construction and comparison for microbiome data in R. *Briefings in Bioinformatics*,
1016 **22**, bbaa290. https://doi.org/10.1093/bib/bbaa290
1017 Peterson RA, Cavanaugh JE (2020) Ordered quantile normalization: a semiparametric
1018 transformation built for the cross-validation era. *Journal of Applied Statistics*, **47**,
1019 2312–2327. https://doi.org/10.1080/02664763.2019.1630372
1020 Popa O, Landan G, Dagan T (2017) Phylogenomic networks reveal limited phylogenetic
1021 range of lateral gene transfer by transduction. *The ISME Journal*, **11**, 543–554.
1022 https://doi.org/10.1038/ismej.2016.116

1023 Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic
1024 features. *Bioinformatics*, **26**, 841–842.
1025 https://doi.org/10.1093/bioinformatics/btq033
1026 Quinn TP, Richardson MF, Lovell D, Crowley TM (2017) propr: An R-package for Identifying
1027 Proportionally Abundant Features Using Compositional Data Analysis. *Scientific*
1028 *Reports*, **7**, 16252. https://doi.org/10.1038/s41598-017-16520-0
1029 R Core Team (2024) R: A Language and Environment for Statistical Computing.
1030 Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open
1031 Software Suite. *Trends in Genetics*, **16**, 276–277. https://doi.org/10.1016/S0168-
1032 9525(00)02024-2
1033 Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–
1034 2069. https://doi.org/10.1093/bioinformatics/btu153
1035 Shaiber A, Eren AM (2019) Composite Metagenome-Assembled Genomes Reduce the
1036 Quality of Public Genome Repositories. *mBio*, **10**, e00725.
1037 https://doi.org/10.1128/mBio.00725-19
1038 Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. *Frontiers*
1039 *in plant science*, **5**, 209. https://doi.org/10.3389/fpls.2014.00209
1040 Sheinman M, Arkhipova K, Arndt PF, Dutilh BE, Hermsen R, Massip F (2021) Identical
1041 sequences found in distant genomes reveal frequent horizontal transfer across the
1042 bacterial domain (RA Neher, G Storz, RA Neher, Eds,). *eLife*, **10**, e62719.
1043 https://doi.org/10.7554/eLife.62719

1044 Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ (2011) Ecology drives a
1045 global network of gene exchange connecting the human microbiome. *Nature*, **480**,
1046 241–244. https://doi.org/10.1038/nature10571
1047 Song W, Wemheuer B, Zhang S, Steensen K, Thomas T (2019) MetaCHIP: Community-level
1048 horizontal gene transfer identification through the combination of best-match and
1049 phylogenetic approaches. *Microbiome*, **7**, 36. https://doi.org/10.1186/s40168-019-
1050 0649-y
1051 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G,
1052 Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S,
1053 Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G,
1054 Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M,
1055 Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky
1056 G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S,
1057 Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG,
1058 Bork P (2015) Structure and function of the global ocean microbiome. *Science*, **348**,
1059 1261359. https://doi.org/10.1126/science.1261359
1060 Tange O (2011) GNU Parallel: the command-line power tool. ;*login: The USENIX Magazine*,
1061 **36**, 42–47. https://doi.org/10.5281/zenodo.16303
1062 Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer
1063 between bacteria. *Nature Reviews Microbiology*, **3**, 711–721.
1064 https://doi.org/10.1038/nrmicro1234

1065 Turner CB, Marshall CW, Cooper VS (2018) Parallel genetic adaptation across environments
1066 differing in mode of growth or resource availability. *Evolution Letters*, **2**, 355–367.
1067 https://doi.org/10.1002/evl3.75
1068 Urvoy M, Gourmelon M, Serghine J, Rabiller E, L'Helguen S, Labry C (2022) Free-living and
1069 particle-attached bacterial community composition, assembly processes and
1070 determinants across spatiotemporal scales in a macrotidal temperate estuary.
1071 *Scientific Reports*, **12**, 13897. https://doi.org/10.1038/s41598-022-18274-w
1072