Co-occurrence is associated with horizontal gene transfer across marine bacteria independent of phylogeny

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

[1]Department of Biology, University of New Brunswick, Fredericton, New Brunswick, Canada
[2]Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA
[3]Department of Microbiology & Immunology, McGill University, Montréal, Québec, Canada
[4]Alpine Center for research on trophic networks and limnic ecosystems, INRAE - CARRTEL, France
[5]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France
[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: gavin.douglas@unb.ca, 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 and consequences of horizontal gene transfer (HGT) is a key goal of microbial evolution research. Although co-occurring taxa have long been appreciated to undergo HGT more often, this association is confounded with other factors, most notably their phylogenetic relatedness. To disentangle these factors, we analyzed 15,339 marine prokaryotic genomes (mainly bacteria) and their distribution in the global ocean. We identified HGT events across these genomes and 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 HGT rates and phylogenetic distance, we only detected an association between co-occurrence and phylogenetic distance for 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 here 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. In a subset of samples with extended environmental data, we identified higher HGT levels associated with particle-attached bacteria and associations of varying directions with specific environmental variables, such as chlorophyll *a* and photosynthetically available radiation. 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) [1]. To better understand these dynamics, gene transfer networks have been investigated across diverse species [2]. A recurrent observation from this work is that taxa from similar environments are associated with higher rates of HGT [3–6].

This observation is expected, for at least two reasons. First, all HGT mechanisms are dependent on physical proximity to some degree [7]. 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 [8]. It is expected that similar dynamics occur between strains within the same environment based on horizontally-transferred genes [9]. Genes can also spread horizontally between more distantly related taxa within a community. This is clearly seen for antibiotic and heavy metal resistance genes, which have the potential to confer adaptive benefits to many taxa in a contaminated environment [10].

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 gene sequences [11]. In this work, the authors found that the positive association between co-occurrence and horizontal gene transfer remained after controlling for phylogenetic distance, although the association was decreased. This study is an important contribution but was limited to 16S rRNA gene data used for identifying co-occurring taxa. 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 [12]. In addition, due to the broad and cross-environment focus of this study [11], the authors did not thoroughly investigate how particular environmental factors are associated with HGT.

Marine microbial communities are an appealing system for studying prokaryotic evolutionary dynamics. Marine prokaryotes undergo widespread HGT and are stratified into communities based on depth and other environmental factors [13–15]. In addition, understanding prokaryotic community dynamics and evolution is important as marine microbes play key roles in global biogeochemical cycles [16], and are also experiencing heightened environmental stress due to climate change [17]. To help better understand

these ecological and evolutionary dynamics, numerous ocean samples have undergone metagenomics sequencing [14, 18].

Here, we used publicly available ocean metagenomes to investigate patterns of HGT and co-occurrence across the global ocean. Our key question focused on whether the link between HGT and co-occurrence in this dataset is robust to the inclusion of phylogenetic distance as a covariate. While this question was also recently investigated across biomes [11], 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 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 [19]. This included whole genome sequencing from isolates, single-amplified genomes, and metagenome-assembled genomes (MAGs), along with pre-existing Prokka gene annotations [20], 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. We used the existing taxonomic classifications [19], which were previously obtained using the Genome Taxonomy Database Toolkit (GTDB-Tk) [21]. Using quality metrics provided with these data, we retained 15,339 genomes with mean completeness > 75% and mean contamination < 5%. The contamination cut-off was chosen based on the established cut-off used for defining high-quality draft genomes [22]. False positive HGT events would be driven primarily by contaminated, rather than incomplete, genomes, so we used a more relaxed completeness cut-off of > 75% (rather than the typical > 90% for high quality genomes [22]) to include more genomes in our analyses. 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 [23], with default parameters, was run on the 15,339 genomes using the Bacteria and Archaea Hidden Markov Model-set composed of 25 single copy genes to generate an alignment protein file with HMMER v3.3.2 [24, 25] and Muscle v5.1 [26]. Automated trimming was performed on the alignments using trimAl v1.4 [27], with the -automated1 and -phylip options. IQTREE2 v2.0.7 [28] was then used to construct a phylogenetic tree, using ModelFinder Plus and 1,000 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 [29].

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 [30]. Taxonomy for each genome was parsed based on TaxIDs using the taxize R package v0.9.102 [31]. 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 [32]. All genomes linked to metagenomic samples or without any contigs greater than 10,000 bp were excluded, to help avoid misassembled genomes. This filtering resulted in a final set of 5,551 genomes (out of 11,060) 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.

Metagenomic sample 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 listed in the OceanDNA resource [33]. This raw metagenomics sequencing data was downloaded from the National Center for Biotechnology Information with Globus Connect [34]. We filtered out low-quality reads with fastp v0.23.4 [35], 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 [36], 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 a genome 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 10,388 remaining genomes. We excluded three metagenomics samples where no genomes were called as present.

To distinguish metagenomics samples more likely to represent free-floating microbes, compared to particle-attached microbes, we partitioned samples based on the size-filtration steps used during sample collection, which are commonly referred to as size fractions. A cut-off of 3 μm is commonly applied to distinguish free-floating (small particles) and particle-attached (large particle-associated) microbes by size [37–39].

This approach is a proxy for habitat preferences, which could correspond to more planktonic (free-floating) or particle-associating lifestyles, but this filtering process is imperfect and should not be taken as absolute classification. We categorized small particle 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 did not exclude small particles. Thus, it is not possible to categorize our initially analyzed metagenomics samples as primarily particle-attached. Instead, we opted to categorize 846 samples as "less-filtered", 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 samples filtered to enrich for small particles. We then identified cases where genomes were predominately found in either small particle fraction 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 small particle 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 small particle-associated.

The OceanDNA resource lists several independent datasets. To ensure our results were not driven by technical differences between these datasets, we also subdivided metagenomics samples to be only those part of the *Tara* Oceans campaign (n=368) or the GEOTRACES campaign [40] samples (n=610).

After our initial analysis we conducted a follow-up with 160 additional ocean metagenomes from the *Tara Oceans* campaign that were processed to have size-selection filters ranging from 3 - 20 µm (referred to as large particle samples). These samples were originally meant to capture the protist-enriched fraction of ocean water [41, 42], but also include the size range for particle-attached bacteria. These metagenomic samples were pre-processed and mapped to the genome database with CoverM using the same steps as above, where only samples with at least 10 genomes called as present were retained. We then subdivided to 43 biological samples for which there was a match between the small (0.22 – 3 µm) and large (3 – 20 µm) particle-associated metagenomics sample. All small particle-associated metagenomics samples were already included in the subset of the OceanDNA dataset described above. We focused on this paired-sample approach (specifically for this analysis) to minimize the impact of environmental and technical confounding factors.

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 cluster. 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 [43], 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 with accurate mode (-g 1). Alignments were then built with MAFFT v7.525 [44], 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 potentially detect 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.

Sensitive approach for identifying indirect gene flow with pairwise BLAST

In addition to the clustering approach, we also performed pairwise identity comparisons between all genomes using BLASTn, with BLAST+ v2.12.0 [45]. Only hits >= 500 bp long and at least 95% identical were retained. To clarify, BLASTn matches were generally based on short subsequences in contigs, rather than the entire contigs. Our analysis for these data focused only on genomes within different genera or higher (i.e., above the genus-level as for the cluster-based HGT approach), to reduce the number of false positives due to vertical inheritance rather than HGT. We ran bedtools v2.30.0 [46] to intersect gene annotations with BLASTn-identified matching regions. For cases where multiple hits overlapped (for the same pairwise genome comparison) the highest scoring hit was retained.

BLASTn matches were called as putative HGT events regardless of whether they were the reciprocal best hits in the dataset. Thus, this approach is not meant to identify direct HGT events between genomes, and instead identifies regions between genomes that were putatively horizontally acquired through a shared HGT network. We used this

information for complementary purposes, and expected these calls to be more sensitive, but also enriched for more false positives, and less straightforward to interpret than the clustering method described above.

Phylogenetic reconciliation within species

We also identified putative gene transfer events within species with RANGER-DTL [47, 48]. We included this approach as using sequence identity alone is unreliable for detecting HGT between closely related genomes (due to high identity stemming from a recent common ancestor). We also ran HoMer v1.0 [49] in conjunction with RANGER-DTL to call multi-gene transfer events.

We ran this approach over 190 species that had at least 10 genomes. We found that topological differences between the core genome tree and accessory gene trees were widespread: a mean of 91.6% (standard deviation: 0.06%) of tested genes were inferred to have at least one transfer event across all species. We had little confidence in the accuracy of individual events (see Discussion) and instead used these results only for supplementary purposes. Our assumption here is that the relative ranks of inferred and real HGT event counts between genome pairs are highly correlated, even if there is a baseline of noise.

Functional annotation and enrichment

Clusters of Orthologous Genes (COG) annotations were taken from a previous run of eggNOG [50], which was distributed with the genomes. For each eggNOG Orthologous Group, we parsed out all COG annotations in the definition and then re-mapped 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 [6]. We first translated all protein coding gene sequences to amino acids with the transeq EMBOSS program v6.6.0.0 [51], using the parameters -table 11, -trim, and -clean. We then ran HMMER v3.1b2 to identify matching genes. These represent a range of transposase and recombinase protein-coding genes associated with diverse MGEs, such as insertion sequences and integrative conjugative elements.

We also ran geNomad v1.5.0 [52] with database v1.1, and default parameters except --cleanup and --splits8 on the contigs of all studied genomes. This analysis was run independently of the proMGE analysis, and returns MGE-associated contigs rather than specific genes. This allowed us to identify contigs categorized as ‘virus’ or ‘plasmid’, as well as those that contain prophages. We ran all enrichment tests for the MGE categories on genes rather than clusters due to this focus on contig-level annotations.

The expected enrichments of COG categories in the HGT results (prior to cross-referencing with co-occurrence data) were taken from a previous study [11], which itself based these expectations upon results across multiple studies [5, 53–61]. 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, relative to the background of all clusters in other COG categories (significance called at Benjamini–Hochberg-corrected [62] P-values < 0.05). We focused the COG category functional enrichment 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 [63] to run hypergeometric distribution-based co-occurrence tests for all pairwise genomes. 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 [11]: $\frac{|samples_{Genome1} \cap samples_{Genome2}|}{min(|samples_{Genome1}|, |samples_{Genome2}|)}$. 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 [64] v5.1.4, which incorporates the relative abundance of genomes. This metric was computed using the NetCoMi R package [65] 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 [66] 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 [67].

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 [68], 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.

Computing prevalence of genomes with horizontal gene transfer or related features

In addition to our logistic regression models, which focused on genome pairs, we calculated summary metrics that provide insight into the extent of HGT and other related characteristics per metagenomic sample. We did so by identifying all genomes called as present in a metagenome and counting how many had at least one instance of putative HGT called (based on the focal cluster-based approach). We similarly tallied genomes with at least one gene that matched the proMGE database, that were identified as plasmid or virus-related with geNomad, or for which geNomad identified at least one contig as containing a provirus. For each feature, this count was divided by the number of genomes in the sample, to return the proportion of genomes with at least one instance of the feature of interest.

In-depth environmental data analysis

To explore a wider range of environmental factors, beyond those present in the OceanDNA dataset, we acquired all environmental depth sensor data for *Tara* Oceans samples from the PANGAEA database [69]. For environmental variables with multiple summary statistics, we retained only the median value per sample and based on locally-calibrated estimates where available. We restricted to data intersecting with *Tara* Oceans samples we acquired (e.g., ignored metatranscriptomic data), and then retained only a single representative of highly redundant variables (defined as those with Spearman correlation coefficients >= 0.95). Based on this approach, we excluded: beta470 and bb470 (redundant with bbp470); absolute PAR (redundant with percent PAR); conductivity and potential temperature (redundant with temperature). We also excluded bac660 and bacp due to the difficulty of interpreting these optical properties, which are primarily used as intermediate measures to calculate other properties.

After this pre-processing, we focused on these variables:

- Latitude (degrees)
- Longitude (degrees)
- Depth (metres)
- Temperature (°C)
- Potential density (kg/m^3; Sigma-theta)
- Salinity
- Nitrate (NO_3^-; µmol/L)
- Oxygen (µmol/kg; Calibrated using World Ocean Atlas 2009 climatology)
- Chlorophyll *a* (mg/m^3; Nonphotochemical fluorescence quenching-corrected and calibrated using water samples). This is a measure of phytoplankton biomass, so it is a biotic variable although we group it with environmental variables herein.
- Photosynthetically available radiation (PAR [%]; Based on diffuse attenuation coefficient and average of eight-day satellite readings)
- fluorescent Coloured Dissolved Organic Matter (fCDOM; Parts per billion quinine sulfate equivalents)
- Backscattering coefficient of particles (per metre) at 470 nm (bbp470). This measure is related to particulate organic carbon levels [70–72]

We focused this analysis on 131 *Tara* Oceans samples that were all from the small particle-associated size fraction, with at least 10 genomes present based on CoverM mapping and with no missing environmental data. We then computed the Spearman correlation coefficients between all environmental variable pairs and computed pairwise distance as |1 – Spearman's ρ|. We performed hierarchical clustering with the average method and visualized the dendrogram of variable groupings and clustered

heatmap of standard scores across all samples with the ComplexHeatmap v2.20.0 R package [73].

We ran a random forest model [74] with 10,000 trees with the ranger R package v0.17.0 [75] to predict HGT prevalence in a metagenome based on these environmental variables. The number of variables to subsample at each split of each tree (the mtry parameter) was set to the default (3, given 12 predictors). Variable importance was computed based on the permutation method. To determine which variables were significantly predictive in the random forest, we re-ran random forests for 1,000 replicates with the dependent variable (HGT prevalence) shuffled. We used the inferred variable importance for each variable across these replicates to build null distributions of the variable importance under no signal. We calculated P-values based on the proportion of simulation replicates that resulted in variable importance equal to or higher than the observed importance for each variable (and called significance at Benjamini-Hochberg-corrected $P < 0.05$).

We also confirmed our overall variable importance findings were not affected by parameter settings. Specifically, we confirmed that changing the mtry parameter (ranging from 1-11) did not result in substantial changes in the order of variable importance. We also confirmed that including just one of a highly correlated pair of variables (e.g., depth and photosynthetically available radiation) did not result in the other variable's relative importance substantially changing (i.e., ensured variables were not masking each other). This analysis was conducted post-hoc to confirm model robustness, not for model selection.

We undertook several steps to make our random forest model more interpretable [76], using the iml R package [77] v0.11.4. This included Individual Conditional Expectation plots, which show how the prediction for each sample would change for different values of a predictor (an environmental variable in this case). We also computed pairwise interactions between significant variables based on the H-statistic [78]. This metric represents the proportion of the joint effect of both variables that is explained by their interaction.

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 [79]. Where possible, Linux commands (on Ubuntu 22.04.5 LTS) were parallelized with GNU Parallel v20240322 [80] and tools were installed from Bioconda [81]. Claude large language models were used to help debug code and to check for errors (in code and manuscript).

441
442 ***Code and data availability***
443 The accessions of most shotgun metagenomics data used in this study are present in
444 the OceanDNA FigShare repository [33]:
445 https://doi.org/10.6084/m9.figshare.c.5564844.v1. The exception is for the
446 protist/particle-attached fraction samples, which were taken from the BioProjects
447 PRJEB4352 and PRJEB9691. All metagenome-assembled genomes and pre-existing
448 annotations were taken from
449 https://sunagawalab.ethz.ch/share/microbiomics/ocean/suppl_data/. All code and key
450 data files are distributed through a Zenodo repository
451 (https://doi.org/10.5281/zenodo.14867578), which includes the filtered sample
452 metadata and genome information.
453
454 **Results**
455 ***Identifying putative horizontal gene transfer events***
456 To study prokaryotic genome diversity in the global ocean, we analyzed 15,339
457 genomes representing 3,270 species, which were previously constructed from marine
458 metagenomics samples [14, 19, 82]. These genomes are from diverse lineages
459 (**Supplementary Figure 1**), primarily bacteria (90.9%). The most common phyla are
460 Proteobacteria (47.9%) and Bacteroidota (15.2%) and the best represented archaeal
461 phylum is Thermoplasmatota (5.9%).
462
463 There are several approaches for inferring HGT, which can produce different but
464 complementary results [83]. Accordingly, we identified putative HGT events between
465 these genomes using several methods to ensure the overall results remain unchanged
466 (see Methods). The first two approaches were based on identifying regions with
467 sequence similarity >= 95% between genomes from different genera (or that differ at
468 higher taxonomic levels). In other words, we did not consider genome comparisons
469 within the same genus or species. Our key assumption is that regions of high similarity
470 are unlikely to occur (when of sufficient length) in the genomes of distantly related taxa
471 due to these regions being present in a common ancestor [5]. Instead, we take these
472 observations to represent putative HGT events.
473
474 Based on this framework, our focal HGT-detection approach focused on clustering all
475 genes across all genomes (**Supplementary Figure 2**). We identified 9,046 putative HGT
476 events based on this approach, 71.2% of which were >= 99% identical, and for which
477 the majority were between genomes within the same family (**Figure 1a**). Although this is
478 a relatively high number of putative HGT events, the number of events called between
479 each pair of genomes was low (median of 1.0 and mean of 3.95 in genome pairs with at
480 least one putative event).

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).

We then investigated whether the expected negative relationship between evolutionary divergence and HGT held in our marine dataset. There is a clear signal of decreasing HGT rates at more distant taxonomic levels (**Figure 1b**), after normalizing for the number of pairwise genomes compared at each level. This signal is similar for both the older (< 99%) and more recent (>= 99%) putative HGT events.

Functional enrichments of horizontally transferred genes

We next asked whether the inferred horizontally transferred genes were enriched in expected functional groups identified in previous studies [11]. Most COG categories were significantly enriched in expected directions (**Figure 2**). For instance, mobile elements (X) and defense mechanisms (V) were strongly enriched among HGT events. In addition, all significant categories that were expected to be depleted for HGT were identified for putative transfers between identities of >= 95% and < 99%. However, transfers at >= 99% identity showed a more mixed signal for categories expected to be depleted for HGT. For instance, we observed a mixed signal of enrichment for lipid transport and metabolism genes (I), and the unexpected enrichment of several other categories, especially at higher taxonomic levels.

This observation highlights a unique aspect of our enrichment analyses, which is that we conducted them for inferred HGT events at each taxonomic level separately (**Figure 2**). This allowed us to identify cases where the same category is enriched in different directions at different taxonomic levels, such as those involved in chromosome partitioning and cell division (D) and amino acid transport and metabolism (E).

We also explored mobile genetic element-associated gene annotations and overall found them enriched among horizontally transferred genes (Fisher's exact tests, $P < 0.001$). Specifically, genes matching the proMGE database were 7.20-fold enriched, whereas genes on contigs classified as a plasmid or virus were 8.69 and 3.32-fold enriched among HGT genes, respectively. In contrast, genes on contigs classified as containing proviruses were 7.29-fold more likely not to be found among the HGT set.

521 With perhaps the exception of proviruses, these functional enrichments agree with
522 basic expectations of gene classes expected to be enriched among HGT events in
523 nature.
524
525 ***Overall relationship between horizontal gene transfer and ecological variables***
526 We next compared how rates of HGT between genomes are associated with co-
527 occurrence and phylogenetic distance (**Supplementary Figure 2**). To do so, we
528 mapped metagenomic reads of previously published ocean metagenomics samples
529 against the 15,339 genomes to determine each genome's relative abundance across
530 samples. We identified 1,862 samples where at least one genome was called as
531 present. For these samples, a mean of 26,592,809 (SD: 36,857,948) reads were
532 mapped, which represented a mean of 23.9% (SD: 10.2%) of reads per sample. A
533 median of 286 genomes were identified as present across these samples. We restricted
534 our subsequent analyses to 10,388 genomes that were called as present in at least 10
535 metagenomics samples, and then computed the pairwise co-occurrence of these
536 genomes. This prevalence filtering step meant that 32.3% of genomes were excluded,
537 which resulted in the loss of 54.8% of genomes involved in HGT based on the clustering
538 approach (a drop from 2,295 to 1,037 unique genomes). This indicates that many HGT
539 events occurred in taxa for which we have insufficient data to study their co-occurrence
540 relationships within this dataset (**Supplementary Figure 1**).
541
542 As expected, we observed a clear association between taxa co-occurrence and HGT
543 (**Figure 3a**; odds ratio: 19.3; $P < 0.001$). In contrast, we did not observe the expected
544 negative association between phylogenetic distance and extent of co-occurrence
545 based on the focal genome pairs (**Figure 3b**), and instead there was a near-zero
546 association (Spearman $\rho=0.056$; $P < 0.001$; **Supplementary Figure 3a**). As a reminder,
547 our analysis focused on genome pairs in different genera or above, to more reliably infer
548 HGT events, and these distributions (**Figure 3**) are based on this subset of genome
549 pairs. However, when considering genome pairs within the same genus or lower, which
550 were excluded from our focal analysis, we did find the expected negative association
551 (Spearman $\rho=-0.325$; $P < 0.001$; **Supplementary Figure 3b**). For the subset of genome
552 pairs connected by HGT (**Figure 3b**, right panel) there is lower phylogenetic distance
553 compared to those with no evidence for HGT (mean difference: 0.900;
554 Wilcoxon test $P < 0.001$). In addition, genome pairs connected by HGT that do not co-
555 occur have especially low phylogenetic distance compared to those that do (mean
556 difference: 0.485; Wilcoxon test $P < 0.001$). This last observation could be due to
557 collider bias (i.e., a misleading signal driven by data subdividing; see Discussion).
558 Overall, we find the expected associations between HGT and both phylogenetic
559 distance and co-occurrence, but the negative association between co-occurrence and
560 phylogenetic distance is only present for closely related taxa.

561
562 Phylogenetic distance has previously been proposed as partially explaining the
563 association between co-occurrence and HGT, but other factors may also constrain this
564 link. An HGT event might be adaptive under specific environmental conditions, making
565 it more likely to be retained and observed. Indeed, genome pairs connected by HGT
566 tend to be sampled under more similar environmental conditions compared to
567 genomes without evidence for HGT (**Figure 4**). This result highlights that including
568 information on environmental factors is also needed, in addition to phylogenetic
569 distance, to better interpret the relationship between co-occurrence and HGT.
570
571 ***Testing for a relationship between horizontal gene transfer and co-occurrence,***
572 ***after correcting for phylogeny and environmental factors***
573 We next investigated whether the relationship between HGT and co-occurrence
574 remains, and to what degree, after controlling for phylogenetic distance and median
575 environmental differences. To do so, we built a logistic regression model to predict the
576 binary state of whether the genome pairs are connected by HGT or not. The predictor
577 variables included whether the genome pairs are co-occurring or not, as well as the
578 continuous phylogenetic distances and median environmental differences. Most of the
579 measured environmental factors were abiotic (e.g., temperature, depth), but we also
580 included a biotic size fraction factor as a proxy for genomes associated with free-
581 floating or potentially particle-associated niches, quantified by different filter sizes (see
582 Methods). A lower filter-cutoff enriches for smaller free-floating cells, whereas a larger
583 filter can include particle-attached prokaryotes, which we refer to as "less-filtered"
584 samples (these samples also include free-floating microbes, and are not solely
585 particle-attached). We hypothesized that genomes of particle-associated microbes are
586 more likely to undergo HGT due to higher cell density and potentially higher encounter
587 rates with other taxa.
588
589 Based on the model coefficients (**Figure 5**), the clearest predictor of HGT is co-
590 occurrence (2.26, $P < 0.001$). As this is a binary predictor, this refers to the increased
591 log-odds of being connected by HGT when genomes co-occur, or 9.58-fold ($e^{2.26}$)
592 increased odds. Phylogenetic distance was also highly significant in the model (-1.66, P
593 < 0.001). As this is a continuous variable this means that for every one unit increase in
594 transformed phylogenetic distance (which corresponds to one standard deviation due
595 to the orderNorm transformation), there is a 5.26-fold decrease in HGT. Although the
596 median differences in all environmental variables were also significant, the absolute
597 coefficients are much lower compared to phylogenetic distance. Finally, pairs where
598 both genomes are associated with "less-filtered" metagenomics samples, which are
599 more likely to include particle-attached prokaryotes, were positively associated with
600 HGT (1.26, P=0.013; with wide 95% confidence intervals). Overall, this result is

consistent with the relationship between co-occurrence and HGT remaining robust after also including potential ecological drivers.

We fit this full logistic regression model, and other models based on subsets of predictors. We found the full model fit best based on the Akaike Information Criterion (AIC; difference in AIC of -3.46 compared to the next best model). In addition, although there is potential for multicollinearity between these variables, the variance inflation factor was < 2 in all cases, meaning that this is not a concern for this model.

The above analyses focused on results from our focal bioinformatics workflow, but we also ensured our overall results did not change if differing methods for determining co-occurrence or HGT were used, or if different datasets were analyzed. Accordingly, we re-ran this logistic regression analysis based on two other methods for determining co-occurrence and two other methods for determining HGT, as well as with genomes from the proGenomes database (**Supplementary Figure 4**; see Methods). Although the differing HGT-identification methods have similar motivations, they can result in highly different profiles. For instance, identifying HGT with RANGER-DTL between strains of a species resulted in a huge number of putative events identified (the positive intercept for these models indicating that more than half of genome comparisons were called to have at least one HGT, or recombination, event). Nonetheless, in almost all cases, the models which included all tested variables produced the best-fitting models. Exceptions included the model based on cluster-based HGT and 'simple' co-occurrence, as well as all models based on the proGenomes database, where the models excluding the particle size groupings fit best. In the case of the proGenomes database models, this is likely due to few genomes found to be associated with each grouping for this genome database (see Methods).

Overall, this broader investigation resulted in the same major interpretations: phylogenetic distance and co-occurrence are strongly associated with HGT, even after including additional factors. One exception to our general interpretation is a slight negative association (coefficient of -0.13) between co-occurrence and HGT in the model based on BLASTn-identified HGT events and propr-based co-abundance (**Supplementary Figure 4**; see Discussion). Nonetheless, we also found overall concordance of our results based on an independent genome database (proGenomes), from which we excluded metagenomics-derived assemblies, and corresponded to higher quality genomes than our focal set (although with fewer inferred HGT events).

Finally, we also ran additional checks to help ensure our interpretations are not driven by technical artifacts. We re-ran several models on the largest subsets of the overall OceanDNA metagenomics dataset: samples from *Tara* Oceans and GEOTRACES

separately (**Supplementary Figure 5**). Our goal here was to ensure that dataset substructure was not driving our results. The model coefficients were qualitatively similar compared to the models on the entire dataset (with one exception, but which included few HGT events and had high multicollinearity). We also ran generalized linear mixed models to control for the genome IDs in each comparison (i.e., to control for individual genomes which could be disproportionately driving the observed signals). To do so, we duplicated the input table used for focal models, to allow each genome ID to be associated with each comparison separately. Due to this approach, the error estimates for these models are less reliable compared to the original logistic regressions. Nonetheless, these models provide insight into how controlling for genome ID can affect the models, and we found that our overall interpretations remained consistent (**Supplementary Figure 5**). The exception was that the association trends for the filter-cutoff groupings associated with genome pairs were less consistent (see Discussion). This is not unexpected, as there were only four genome pairs associated with "less-filtered" samples and called for HGT in the original analysis. These pairs consisted of only seven unique genomes (**Supplementary Table 1**).

In-depth environmental associations with horizontal gene transfer

Our main analysis focused on which factors are associated with HGT occurrence between pairs of genomes. This was useful for investigating how co-occurrence and phylogenetic distance relate to HGT, but the relationship to environmental factors was less clear. We thus conducted follow-up analyses, focused on samples rather than genome pairs, to investigate the association with more specific environmental factors. We did so by computing a per-sample measure of the prevalence of HGT: the proportion of genomes in each sample with at least one putative HGT event identified.

To characterize the extent of association between broader oceanographic environmental factors and HGT prevalence per sample, we acquired detailed environmental data for a subset of our *Tara* Oceans samples (see Methods). We focused on 12 environmental factors across 131 metagenomic samples (all small particle size samples; see Methods). We first characterized the variation in these variables across the samples, and identified clear clustering of environmental variables, such as depth with photosynthetically available radiation (**Supplementary Figure 6**). We then built a random forest model to predict HGT prevalence based on these variables ($R^2 = 0.687$). Based on a permutation approach we assessed the significance of variable importance (**Figure 6a**), and identified chlorophyll *a*, photosynthetically available radiation (PAR), bbp470, nitrate, oxygen, fluorescent Coloured Dissolved Organic Matter (fCDOM), and depth all as significant predictors. This was expected as these variables show

clear associations with horizontal gene transfer prevalence individually (**Supplementary Figure 7**). In particular, there is a discrete clustering of samples by low and high PAR, which are associated with higher and lower HGT prevalence, respectively (**Figure 6b**).

For improved insight into each variable's effect in the model, we also investigated the Individual Conditional Expectation plots (**Supplementary Figure 8a**). These highlight that each variable has only a small impact on the predicted HGT prevalence (i.e., there is a small range in how much the predicted HGT prevalence varies when each environmental variable is systematically altered). In addition, there are clear signals of the predictive value for chlorophyll *a*, PAR, and bbp470 mainly being driven by the difference between values near 0 (with high predicted HGT prevalence) and larger values for these variables (with low predicted HGT prevalence). Overall, there is a consistent effect on all samples' predictions as the variables are altered, indicating that these variables primarily have independent impacts on predicted HGT prevalence. Nonetheless, there are interactions between variables that impact HGT prevalence (**Supplementary Figure 8b**), particularly among oxygen, fCDOM, and nitrate, for which pairwise interactions explained on average 10.6% of their joint effects.

Paired comparison of small and large particle enriched samples

Our models focusing on genome pairs found a weak signal that genomes in 'less-filtered' samples undergo more HGT. However, such samples also include free-floating bacteria and are confounded by dataset divisions (e.g., almost all 'less-filtered' samples included were from the GEOTRACES dataset). Accordingly, we conducted a similar follow-up on whether HGT prevalence, and related features, show clear differences across samples processed with differing size fractions. We did so by focusing on additional *Tara* Oceans metagenomics samples that matched a subset of those analyzed above, but which were restricted to a higher size fraction (with small particles excluded, unlike the earlier 'less-filtered' grouping; see Methods). Since these are the same collected samples, this allowed us to evaluate the effect of filter size fraction while implicitly controlling for environmental factors (**Figure 7**). We found a clear signal of higher HGT prevalence in higher size fraction samples (a 2.09-fold increase on average), with similar signals for genomes with hits to the proMGE database or to genes matched to plasmids or viruses, but no significant difference for genomes containing proviruses.

Discussion

720 We have shown that co-occurring environmental genomes in ocean metagenomics
721 data are more likely to show evidence for HGT. In agreement with an earlier study using
722 lower-resolution 16S rRNA gene amplicon sequencing data [11], we found that the
723 association between co-occurrence and HGT holds after controlling for phylogenetic
724 distance and several measured environmental variables. This result underlines the
725 strong association between co-occurrence and HGT.
726
727 Our work also agrees with a survey of antibiotic resistance genes in metagenomes
728 across a wide range of environments, which similarly identified decreased HGT with
729 phylogenetic and environmental distance [84]. In this recent study, the authors
730 identified not only phylogenetic distance as negatively associated with HGT, but
731 specifically genetic incompatibility, which was based on several factors, including the
732 difference in 5-mer content between genomes. Intriguingly, cases were identified where
733 genomes with low genetic incompatibility displayed high rates of HGT despite being
734 phylogenetically distant. Although we did not specifically investigate genetic
735 incompatibility here, this will be an important area for future investigation to determine
736 whether similar trends are found to explain general HGT patterns, and not when
737 restricted to antibiotic resistance genes.
738
739 Through our main workflow for identifying HGT events, we were able to stratify genes
740 putatively shared through HGT by taxonomic level and sequence identity (as a proxy for
741 time since transfer). The functional enrichment of these genes largely agreed with
742 expectations from the literature (**Figure 2**). However, several functional enrichments
743 were in the opposite direction from previously reported expectations. Particularly for
744 enrichments based on genes at >= 99% identity, we found several functional
745 enrichments in opposite directions at different taxonomic levels. For instance, amino
746 acid transport and metabolism was depleted for transfer events at the genus level, but
747 enriched at the phylum and domain levels. These taxonomic level-specific enrichments
748 would be missed if functional enrichment were performed on all genes combined, and
749 could reflect differing recurrent selective pressures for HGT at varying taxonomic levels.
750
751 The association between HGT and both co-occurrence and phylogenetic distance is
752 extremely clear in our data (**Figure 1** and **Figure 3**). However, the expected association
753 between co-occurrence and phylogenetic distance was only present in genome pairs
754 within the same genus or lower. This observation highlights that closely related taxa are
755 primarily driving this signal, rather than a consistent trend at all phylogenetic distances,
756 at least in this marine dataset. Additionally, we identified an intriguing trend where
757 genome pairs that co-occur and show evidence for HGT were significantly more
758 phylogenetically distant compared to those that do not co-occur yet undergo HGT
759 (**Figure 3b**). This could indicate 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 [85], and could be enriched for distantly related genomes in our dataset. However, a technical explanation for this observation, which we think is more likely, is collider bias [86]. When both phylogenetic distance and co-occurrence independently affect the probability of genomes being connected by HGT, conditioning on one factor (e.g., co-occurrence) creates an expected difference in the distribution of the other factor (e.g., phylogenetic distance) among genome pairs. In other words, if genome pairs can be connected by HGT either due to being phylogenetically similar or due to co-occurring, then we would expect that genome pairs connected by HGT that are not co-occurring would 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. Second, an alternative causal explanation is that HGT of habitat-specific genes allows species sharing those genes to thrive in that habitat, and thus to co-occur. In this alternative explanation, co-occurrence is a proxy for unmeasured environmental variables that define cryptic microbial habitats. Distinguishing between these explanations for the observed associations is an avenue for future research, potentially harnessing time-series sampling to infer causal relationships.

Although we cannot formally distinguish between these explanations, our results are consistent with both physical proximity driving HGT and, potentially, HGT contributing to shared niche preferences. We consistently found evidence for increased HGT among genomes in larger particle-associated filter fractions. This suggests that physical proximity on particles, in contrast to reduced encounter rates among planktonic bacteria enriched on the small filter fractions, could drive HGT. The larger filter fraction could also share common selective pressures, associated with a particle-attached lifestyle, which may also explain the association with HGT independently of physical proximity. Consistent with environmental drivers of HGT (or that HGT of niche-adaptive genes allows survival in similar environments), we found that similarity in a few measured environmental factors was associated with higher rates of HGT in logistic regression models. However, the effects of these environmental factors on HGT was relatively small compared to the larger effects of co-occurrence and phylogeny.

Based on a more in-depth analysis for a small subset of samples with additional environmental data we identified clear environmental factors associated with HGT. Disentangling the causal factors underlying these associations, such as for chlorophyll *a*, which is a marker of phytoplankton biomass, generally higher at depths with optimum nutrient and light conditions (e.g., deep chlorophyll maxima), is difficult. However, we can infer that low-light, nitrate-rich environments are enriched for HGT. This work agrees with the previously identified association between increasing depth (or at least variables associated with increased depth) and transposase genes [87, 88]. In both our analyses and a previous study [81], transposase genes were also associated with genomes sampled from the large particle enriched filter fraction. This highlights the challenges of disentangling the effects of physical proximity and environmental similarity on HGT rates. It may also explain why co-occurrence – which effectively captures both the effects of proximity and niche similarity in both measured and unmeasured niche dimensions – is so strongly associated with HGT.

A novel aspect of our study is that we focused on metagenomic sequencing data, in a complementary fashion to the approach of analyzing 16S ribosomal RNA gene sequencing profiles and matching these to reference genomes [11]. However, most genomes in our focal dataset were MAGs, which have contamination and misassembly issues [89, 90], including chimeric MAGs which might not be detected based on standard quality controls. In addition, depending on species clonality within a given metagenomic sample, MAGs may represent a population genome for a species, which may bias and preclude the detection of HGT across MAGs. To address these problems, we conducted an additional analysis based on high quality genomes from the proGenomes database, with MAGs excluded. Our overall interpretations were similar (**Supplementary Figures 4 and 5**), suggesting that our inferences of HGT were not driven by misassembled MAGs.

We also considered a variety of bioinformatic workflows for detecting HGT and determining co-occurrence. One clear observation was that within-species inferences of HGT and recombination with RANGER-DTL led to an extremely high number of putative hits. Because phylogenetic reconciliation approaches are sensitive to errors in tree-building [91], we cannot reject that this observation is simply driven by inaccuracies in the gene tree topologies. 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 4**). 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 broad environmental factors. However, we also observed increased HGT between genomes sampled under more similar environmental conditions, and among particle-associated genomes. 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.

875 **Figure legends**

876 **Figure 1**: Summary of putative horizontal gene transfer events based on sequence
877 identity cutoffs. (a) Counts of putative horizontal gene transfer (HGT) events based
878 on the cluster-based approach, by the lowest taxonomic level at which the
879 compared genomes differed. The total number of genome comparisons per level
880 is also indicated. These cells are coloured by the percentage per row. (b)
881 Proportion of putative horizontal gene transfer events per taxonomic level, after
882 normalizing by the total number of genome comparisons per level.
883
884 **Figure 2**: Cluster of Orthologous Genes (COG) categories enriched among
885 horizontal gene transfer (HGT) hits based on sequence identity. HGT results used
886 for this analysis were based on the cluster-based approach. Expected patterns of
887 enrichment are indicated for COG categories at the top, and individual COG
888 categories (in abbreviated form in some cases) are indicated at the bottom. The
889 row groupings indicate the identity cut-off, in addition to the taxonomic level at
890 which the genomes that encode the genes differ. Results are only shown for
891 significant Fisher's exact tests (Benjamini-Hochberg-corrected $P < 0.05$); light grey
892 cells indicate non-significant tests.
893
894 **Figure 3**: Breakdown of putative horizontally transferred genes identified, by co-
895 occurrence and phylogenetic distance. (a) Tallies and percentages of pairwise
896 genome comparisons, by whether they have at least one horizontal gene transfer
897 (HGT) event inferred and whether they co-occur. (b) Boxplots over violin plots of
898 all genomes tallied in panel a, displaying phylogenetic distance by HGT and co-
899 occurrence. The width of the violin plots does not reflect the total sample size for
900 each category (see panel a for counts in each category).
901
902 **Figure 4**: Distributions of the difference in median value of each environmental
903 variable between genome pairs, split by genome pairs connected by HGT and not.
904 The median value for each variable was computed for each genome based on all
905 the metagenomics samples it was found within. The violin plots for the "no HGT"
906 relationship show more detailed variation likely because there are many more
907 genome pairs in this category.
908
909 **Figure 5**: Coefficients in logistic regression model based on all pairwise genome
910 comparisons. The response is whether those genomes have at least one called
911 horizontal gene transfer event (based on the cluster-based approach) between
912 them or not. Co-occurrence (based on the hypergeometric approach) is a binary
913 variable indicating whether those genomes significantly co-occur, and small
914 particle enriched and "less-filtered" enriched indicate whether the genomes being

compared are both enriched in samples filtered to enrich for small particles (small), or those with little or no size filters ("less-filtered"). The other variables are ordered quantile normalized continuous values. Variables with 'diff.' represent the median inter-sample difference in each environmental variable for all samples encoding the two genomes being compared. Coefficient values less than 0 indicate that decreases in those variables are associated with increases in horizontal gene transfer, whereas a positive value represents the opposite. The intercept for this model was -14.35. Error bars represent 95% confidence intervals (1.96 x standard error). All variables were significant ($P < 0.05$) except for "small particle enriched".

Figure 6: Key associations between environmental variables and horizontal gene transfer prevalence. (a) Variable importance of tested environmental variables for predicting horizontal gene transfer prevalence. Variable colour indicates the direction of the association (based on Spearman correlation, for context), while grey indicates variables that did not have significant variable importance. (b) Boxplots of horizontal gene transfer prevalence for samples split into different groupings of PAR (lower and above 20%, which is the approximate centre of a large range with no samples).

Figure 7: Breakdown of horizontal gene transfer (HGT) and related feature prevalence across 43 *Tara* Oceans samples. The same biological samples are matched based on whether they were enriched for small or large particles based on the fraction filter (i.e., to enrich for free-floating or particle-attached bacteria). The y-axis for all panels is the proportion of genomes in a sample with at least one instance of that category (e.g., the proportion of genomes in a sample with at least one gene identified to be either acquired or transferred through HGT). Asterisks indicate significant comparisons (adjusted $P < 0.05$) based on paired Wilcoxon tests.

References

1. Innamorati KA et al. The Bacterial Guide to Designing a Diversified Gene Portfolio. In: Tettelin H, Medini D (eds), The Pangenome: Diversity, Dynamics and Evolution of Genomes. New York: Springer, 2020, 51–87.
2. Beiko RG, Harlow TJ, Ragan MA. Highways of gene sharing in prokaryotes. *PNAS USA* 2005;**102**:14332–14337. https://doi.org/10.1073/pnas.0504068102

951 3. Smillie CS et al. Ecology drives a global network of gene exchange connecting the
952 human microbiome. *Nature* 2011;**480**:241–244.
953 https://doi.org/10.1038/nature10571
954 4. Fondi M et al. "Every Gene Is Everywhere but the Environment Selects": Global
955 Geolocalization of Gene Sharing in Environmental Samples through Network
956 Analysis. *Genome Biology and Evolution* 2016;**8**:1388–1400.
957 https://doi.org/10.1093/gbe/evw077
958 5. Sheinman M et al. Identical sequences found in distant genomes reveal frequent
959 horizontal transfer across the bacterial domain. *eLife* 2021;**10**:e62719.
960 https://doi.org/10.7554/eLife.62719
961 6. Khedkar S et al. Landscape of mobile genetic elements and their antibiotic
962 resistance cargo in prokaryotic genomes. *Nucleic Acids Research* 2022;**50**:3155–
963 3168. https://doi.org/10.1093/nar/gkac163
964 7. Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer
965 between bacteria. *Nature Reviews Microbiology* 2005;**3**:711–721.
966 https://doi.org/10.1038/nrmicro1234
967 8. Turner CB, Marshall CW, Cooper VS. Parallel genetic adaptation across
968 environments differing in mode of growth or resource availability. *Evolution Letters*
969 2018;**2**:355–367. https://doi.org/10.1002/evl3.75
970 9. Niehus R et al. Migration and horizontal gene transfer divide microbial genomes
971 into multiple niches. *Nature Communications* 2015;**6**:8924.
972 https://doi.org/10.1038/ncomms9924

973 10. Pal C et al. Co-occurrence of resistance genes to antibiotics, biocides and metals
974 reveals novel insights into their co-selection potential. *BMC Genomics*
975 2015;**16**:964. https://doi.org/10.1186/s12864-015-2153-5
976 11. Dmitrijeva M et al. A global survey of prokaryotic genomes reveals the eco-
977 evolutionary pressures driving horizontal gene transfer. *Nat Ecol Evol* 2024;1–13.
978 https://doi.org/10.1038/s41559-024-02357-0
979 12. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data.
980 *Frontiers in plant science* 2014;**5**:209. https://doi.org/10.3389/fpls.2014.00209
981 13. Stepanauskas R et al. Net rate of lateral gene transfer in marine prokaryoplankton.
982 *ISME J* 2025;**19**:wraf159. https://doi.org/10.1093/ismejo/wraf159
983 14. Sunagawa S et al. Structure and function of the global ocean microbiome. *Science*
984 2015;**348**:1261359. https://doi.org/10.1126/science.1261359
985 15. Martinez-Gutierrez CA, Aylward FO. Evolutionary Genomics of Marine Bacteria and
986 Archaea. In: Stal LJ, Cretoiu MS (eds), The Marine Microbiome. Cham: Springer
987 International Publishing, 2022, 327–354.
988 16. Azam F, Malfatti F. Microbial structuring of marine ecosystems. *Nat Rev Microbiol*
989 2007;**5**:782–791. https://doi.org/10.1038/nrmicro1747
990 17. Heneghan RF et al. The global distribution and climate resilience of marine
991 heterotrophic prokaryotes. *Nat Commun* 2024;**15**:6943.
992 https://doi.org/10.1038/s41467-024-50635-z
993 18. Biller SJ et al. Marine microbial metagenomes sampled across space and time. *Sci*
994 *Data* 2018;**5**:180176. https://doi.org/10.1038/sdata.2018.176

995 19. Paoli L et al. Biosynthetic potential of the global ocean microbiome. *Nature*
996 2022;**607**:111–118. https://doi.org/10.1038/s41586-022-04862-3
997 20. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*
998 2014;**30**:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
999 21. Chaumeil P-A et al. GTDB-Tk: a toolkit to classify genomes with the Genome
1000 Taxonomy Database. *Bioinformatics* 2020;**36**:1925–1927.
1001 https://doi.org/10.1093/bioinformatics/btz848
1002 22. Bowers RM et al. Minimum information about a single amplified genome (MISAG)
1003 and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature*
1004 *Biotechnology* 2017;**35**:725–731. https://doi.org/10.1038/nbt.3893
1005 23. Lee MD, Ponty Y. GToTree: A user-friendly workflow for phylogenomics.
1006 *Bioinformatics* 2019;**35**:4162–4164. https://doi.org/10.1093/bioinformatics/btz188
1007 24. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–763.
1008 https://doi.org/10.1093/bioinformatics/14.9.755
1009 25. Eddy SR. Accelerated Profile HMM Searches. *PLOS Computational Biology*
1010 2011;**7**:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
1011 26. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased
1012 assessments of sequence homology and phylogeny. *Nat Commun* 2022;**13**:6968.
1013 https://doi.org/10.1038/s41467-022-34630-w
1014 27. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated
1015 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
1016 2009;**25**:1972–1973. https://doi.org/10.1093/bioinformatics/btp348

1017 28. Minh BQ et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
1018 Inference in the Genomic Era. *Molecular Biology and Evolution* 2020;**37**:1530–
1019 1534. https://doi.org/10.1093/molbev/msaa015
1020 29. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees.
1021 *Bioinformatics* 2018;**34**:1053–1055. https://doi.org/10.1093/bioinformatics/btx701
1022 30. Fullam A et al. proGenomes3: approaching one million accurately and consistently
1023 annotated high-quality prokaryotic genomes. *Nucleic Acids Research*
1024 2023;**51**:D760–D766. https://doi.org/10.1093/nar/gkac1078
1025 31. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. *F1000*
1026 *Research* 2013. https://doi.org/10.12688/f1000research.2-191.v2
1027 32. Kans J. Entrez Direct: E-utilities on the Unix Command Line. Entrez Programming
1028 Utilities Help [Internet]. Bethesda, Maryland: National Center for Biotechnology
1029 Information (US), 2025.
1030 33. Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000
1031 prokaryotic genomes originated from various marine environments. *Sci Data*
1032 2022;**9**:305. https://doi.org/10.1038/s41597-022-01392-5
1033 34. Foster I. Globus Online: Accelerating and Democratizing Science through Cloud-
1034 Based Services. *IEEE Internet Computing* 2011;**15**:70–73.
1035 https://doi.org/10.1109/MIC.2011.64
1036 35. Chen S et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*
1037 2018;**34**:i884–i890. https://doi.org/10.1093/bioinformatics/bty560
1038 36. Aroney STN et al. CoverM: Read coverage calculator for metagenomics. 2024.
1039 Zenodo, 2024.

1040 37. Crump BC, Armbrust EV, Baross JA. Phylogenetic Analysis of Particle-Attached and
1041 Free-Living Bacterial Communities in the Columbia River, Its Estuary, and the
1042 Adjacent Coastal Ocean. *Applied and Environmental Microbiology* 1999;**65**:3192–
1043 3204. https://doi.org/10.1128/AEM.65.7.3192-3204.1999
1044 38. Shapiro BJ et al. Population genomics of early events in the ecological
1045 differentiation of bacteria. *Science* 2012;**335**:48–51.
1046 https://doi.org/10.1126/science.1218198
1047 39. Urvoy M et al. Free-living and particle-attached bacterial community composition,
1048 assembly processes and determinants across spatiotemporal scales in a
1049 macrotidal temperate estuary. *Sci Rep* 2022;**12**:13897.
1050 https://doi.org/10.1038/s41598-022-18274-w
1051 40. Biller SJ et al. Marine microbial metagenomes sampled across space and time. *Sci*
1052 *Data* 2018;**5**:180176. https://doi.org/10.1038/sdata.2018.176
1053 41. Carradec Q et al. A global ocean atlas of eukaryotic genes. *Nat Commun*
1054 2018;**9**:373. https://doi.org/10.1038/s41467-017-02342-1
1055 42. Guérin N et al. Genomic adaptation of the picoeukaryote Pelagomonas calceolata
1056 to iron-poor oceans revealed by a chromosome-scale genome sequence.
1057 *Commun Biol* 2022;**5**:983. https://doi.org/10.1038/s42003-022-03939-z
1058 43. Fu L et al. CD-HIT: Accelerated for clustering the next-generation sequencing data.
1059 *Bioinformatics* 2012;**28**:3150–3152. https://doi.org/10.1093/bioinformatics/bts565
1060 44. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
1061 improvements in performance and usability. *Mol Biol Evol* 2013;**30**:772–780.
1062 https://doi.org/10.1093/molbev/mst010

1063 45. Altschul SF et al. Basic Local Alignment Search Tool. *Journal Molecular Biology*
1064 1990;**215**:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
1065 46. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic
1066 features. *Bioinformatics* 2010;**26**:841–842.
1067 https://doi.org/10.1093/bioinformatics/btq033
1068 47. Bansal MS et al. RANGER-DTL 2.0: rigorous reconstruction of gene-family
1069 evolution by duplication, transfer and loss. *Bioinformatics* 2018;**34**:3214–3216.
1070 https://doi.org/10.1093/bioinformatics/bty314
1071 48. Bansal MS. Deciphering Microbial Gene Family Evolution Using Duplication-
1072 Transfer-Loss Reconciliation and RANGER-DTL. In: Luo H (ed.), Environmental
1073 Microbial Evolution: Methods and Protocols. New York, NY: Springer US, 2022,
1074 233–252.
1075 49. Kloub L et al. Systematic Detection of Large-Scale Multigene Horizontal Transfer in
1076 Prokaryotes. *Molecular Biology and Evolution* 2021;**38**:2639–2659.
1077 https://doi.org/10.1093/molbev/msab043
1078 50. Cantalapiedra CP et al. eggNOG-mapper v2: Functional Annotation, Orthology
1079 Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology*
1080 *and Evolution* 2021;**38**:5825–5829. https://doi.org/10.1093/molbev/msab293
1081 51. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open
1082 Software Suite. *Trends in Genetics* 2000;**16**:276–277.
1083 https://doi.org/10.1016/S0168-9525(00)02024-2
1084 52. Camargo AP et al. Identification of mobile genetic elements with geNomad. *Nat*
1085 *Biotechnol* 2024;**42**:1303–1312. https://doi.org/10.1038/s41587-023-01953-y

1086 53. Nakamura Y et al. Biased biological functions of horizontally transferred genes in
1087 prokaryotic genomes. *Nat Genet* 2004;**36**:760–766.
1088 https://doi.org/10.1038/ng1381
1089 54. Cordero OX, Hogeweg P. The impact of long-distance horizontal gene transfer on
1090 prokaryotic genome size. *Proceedings of the National Academy of Sciences of the*
1091 *United States of America* 2009;**106**:21748–21753.
1092 https://doi.org/10.1073/pnas.0907584106
1093 55. Cohen O, Pupko T. Inference and Characterization of Horizontally Transferred
1094 Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution*
1095 2010;**27**:703–713. https://doi.org/10.1093/molbev/msp240
1096 56. Cohen O, Gophna U, Pupko T. The Complexity Hypothesis Revisited: Connectivity
1097 Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer. *Molecular*
1098 *Biology and Evolution* 2011;**28**:1481–1489.
1099 https://doi.org/10.1093/molbev/msq333
1100 57. Popa O, Landan G, Dagan T. Phylogenomic networks reveal limited phylogenetic
1101 range of lateral gene transfer by transduction. *ISME J* 2017;**11**:543–554.
1102 https://doi.org/10.1038/ismej.2016.116
1103 58. Oliveira PH et al. The chromosomal organization of horizontal gene transfer in
1104 bacteria. *Nat Commun* 2017;**8**:841. https://doi.org/10.1038/s41467-017-00808-w
1105 59. Paquola ACM et al. Horizontal Gene Transfer Building Prokaryote Genomes: Genes
1106 Related to Exchange Between Cell and Environment are Frequently Transferred. *J*
1107 *Mol Evol* 2018;**86**:190–203. https://doi.org/10.1007/s00239-018-9836-x

1108 60. Song W et al. MetaCHIP: Community-level horizontal gene transfer identification
1109 through the combination of best-match and phylogenetic approaches.
1110 *Microbiome* 2019;**7**:36. https://doi.org/10.1186/s40168-019-0649-y
1111 61. Jeong H et al. Horizontal gene transfer in human-associated microorganisms
1112 inferred by phylogenetic reconstruction and reconciliation. *Sci Rep* 2019;**9**:5953.
1113 https://doi.org/10.1038/s41598-019-42227-5
1114 62. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and
1115 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*
1116 *Series B (Methodological)* 1995;**57**:289–300.
1117 63. Neal ZP. backbone: An R package to extract network backbones. *PLOS ONE*
1118 2022;**17**:e0269137. https://doi.org/10.1371/journal.pone.0269137
1119 64. Quinn TP et al. propr: An R-package for Identifying Proportionally Abundant
1120 Features Using Compositional Data Analysis. *Sci Rep* 2017;**7**:16252.
1121 https://doi.org/10.1038/s41598-017-16520-0
1122 65. Peschel S et al. NetCoMi: network construction and comparison for microbiome
1123 data in R. *Briefings in Bioinformatics* 2021;**22**:bbaa290.
1124 https://doi.org/10.1093/bib/bbaa290
1125 66. Peterson RA, Cavanaugh JE. Ordered quantile normalization: a semiparametric
1126 transformation built for the cross-validation era. *J Appl Stat* 2020;**47**:2312–2327.
1127 https://doi.org/10.1080/02664763.2019.1630372
1128 67. Lüdecke D et al. performance: An R Package for Assessment, Comparison and
1129 Testing of Statistical Models. *Journal of Open Source Software* 2021;**6**:3139.
1130 https://doi.org/10.21105/joss.03139

68. Brooks M E et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 2017;**9**:378–400. https://doi.org/10.32614/RJ-2017-066

69. Consortium CTO, Expedition PTO. Registry of all samples from the Tara Oceans Expedition (2009-2013). 2017. PANGAEA, 2017.

70. Cetinić I et al. Particulate organic carbon and inherent optical properties during 2008 North Atlantic Bloom Experiment. *Journal of Geophysical Research: Oceans* 2012;**117**. https://doi.org/10.1029/2011JC007771

71. Stramski D et al. The role of seawater constituents in light backscattering in the ocean. *Progress in Oceanography* 2004;**61**:27–56. https://doi.org/10.1016/j.pocean.2004.07.001

72. Martinez-Vicente V et al. Optical backscattering is correlated with phytoplankton carbon across the Atlantic Ocean. *Geophysical Research Letters* 2013;**40**:1154–1158. https://doi.org/10.1002/grl.50252

73. Gu Z. Complex heatmap visualization. *iMeta* 2022;**1**:e43. https://doi.org/10.1002/imt2.43

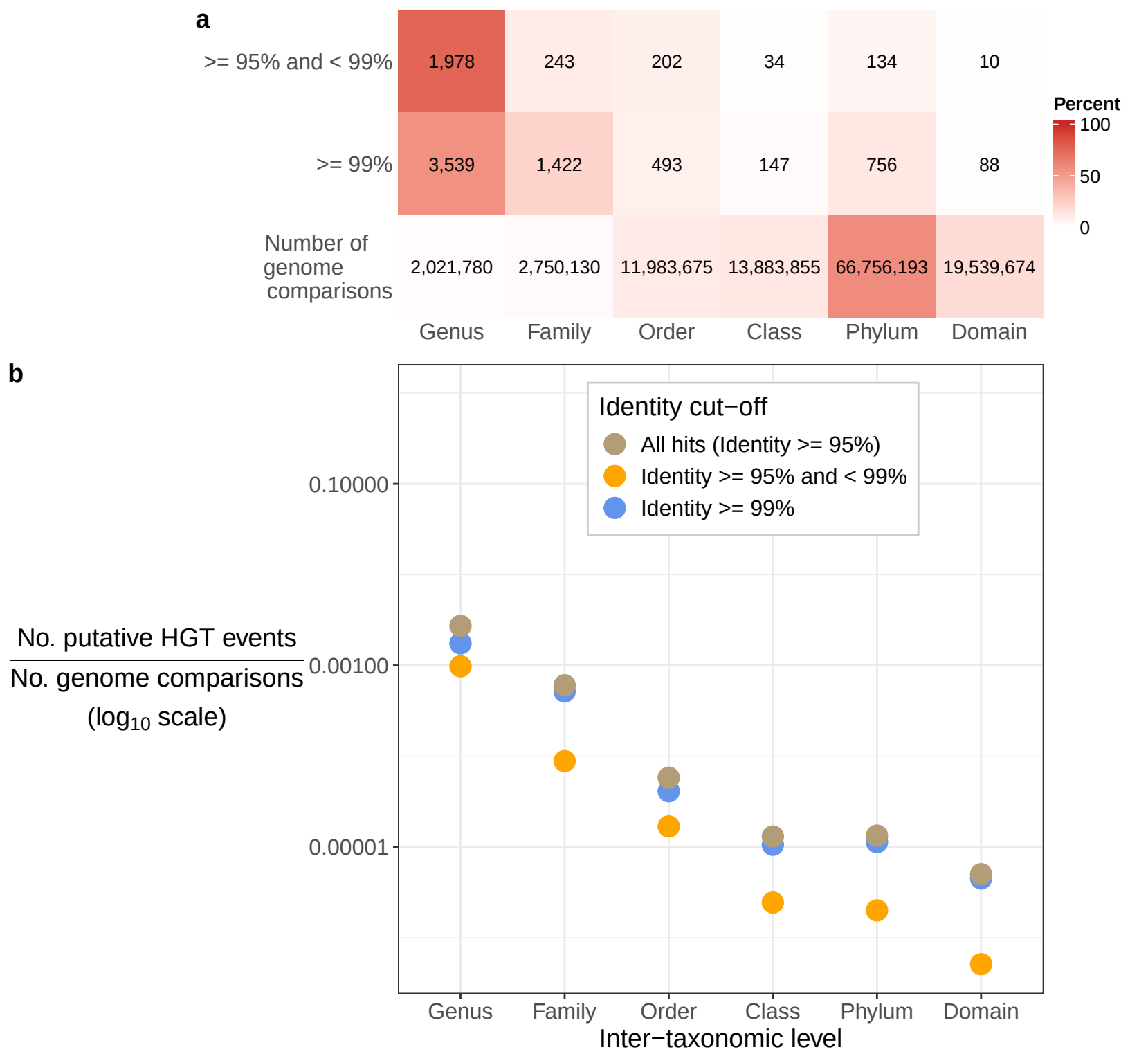
74. Breiman L. Random Forests. *Machine Learning* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324

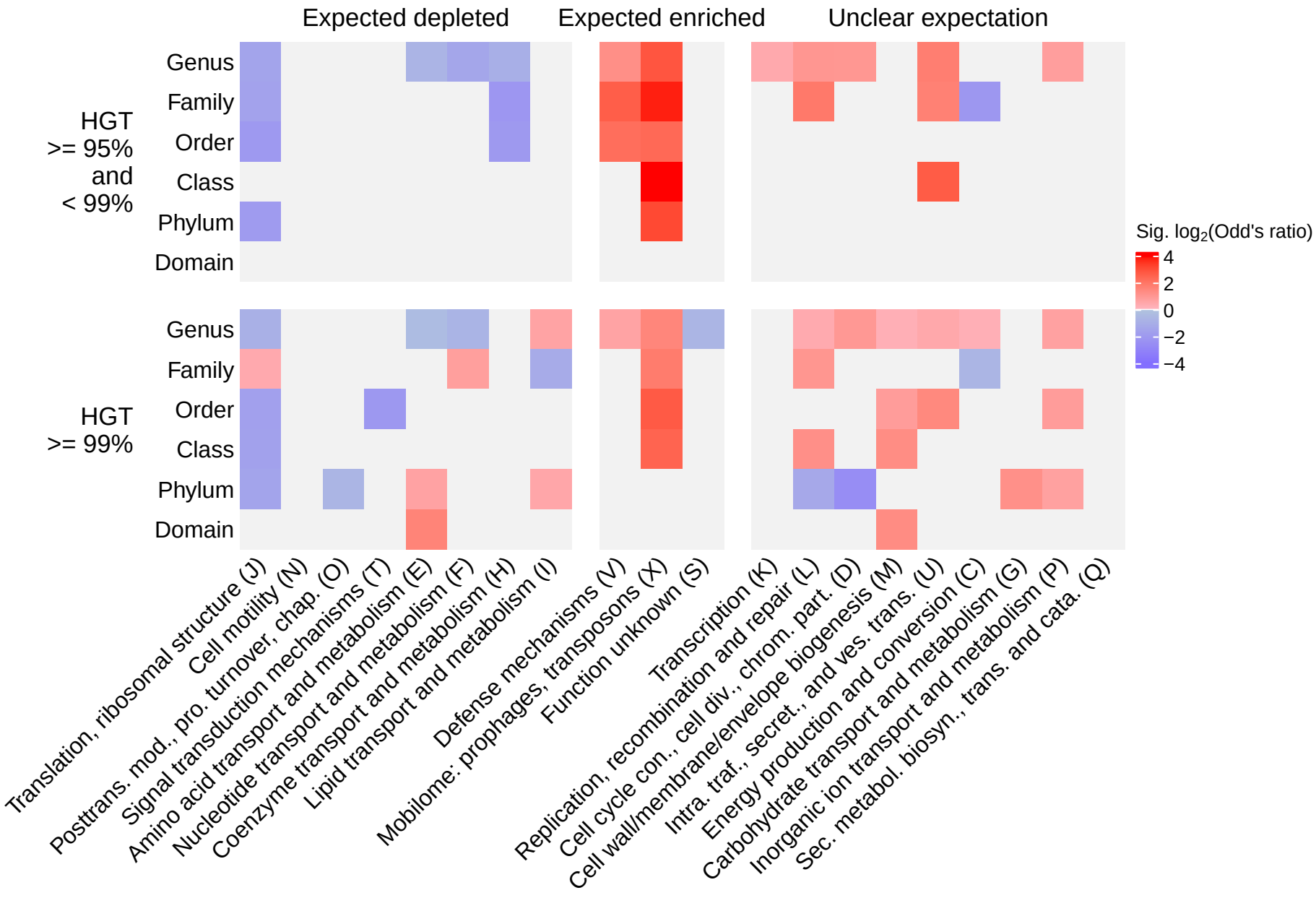
75. Wright MN, Ziegler A. ranger: A fast implementation of Random Forests for high dimensional data in C++ and R. *Journal of Statistical Software* 2017;**77**:1–17. https://doi.org/10.18637/jss.v077.i01

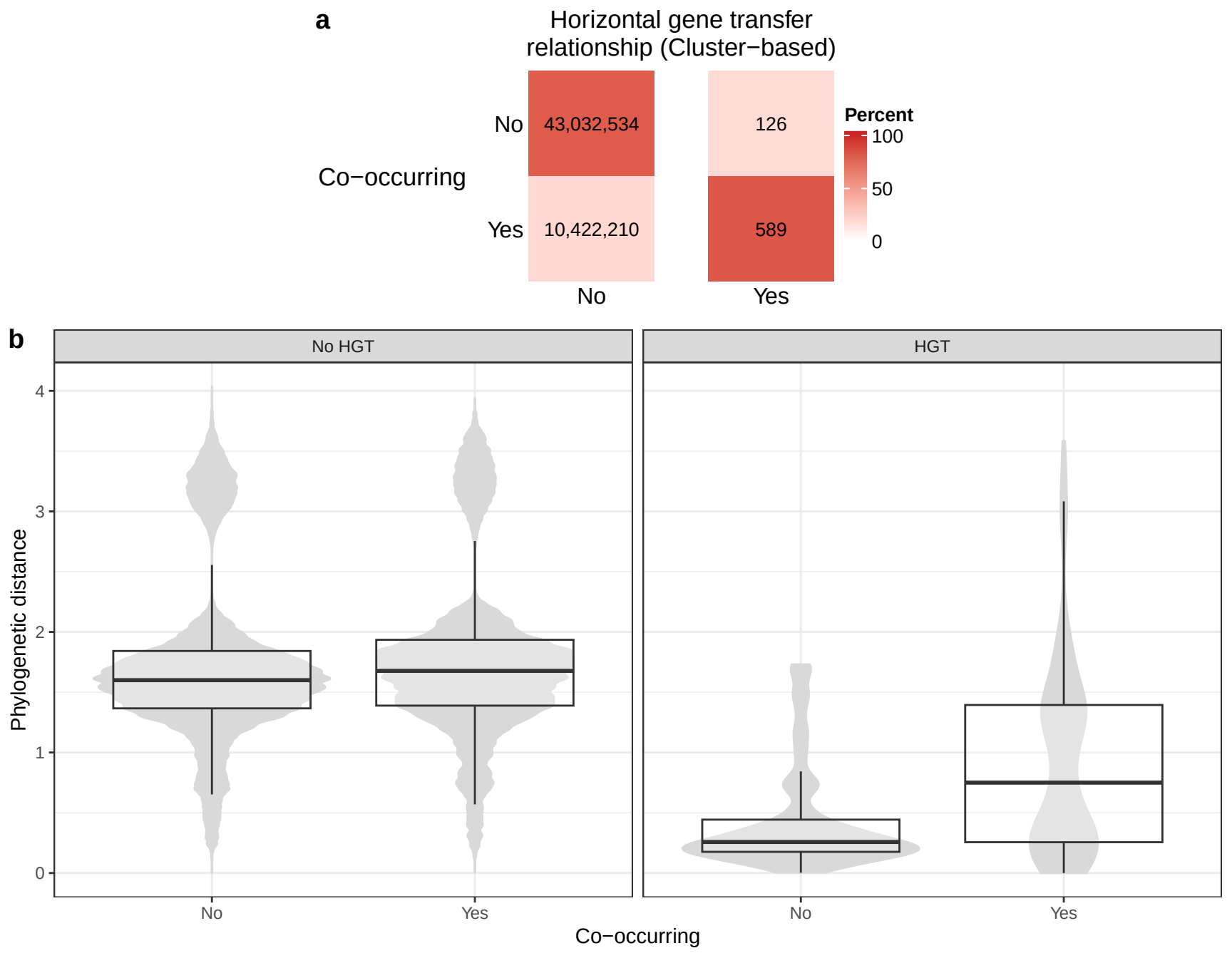
76. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 3rd edn. 2025.

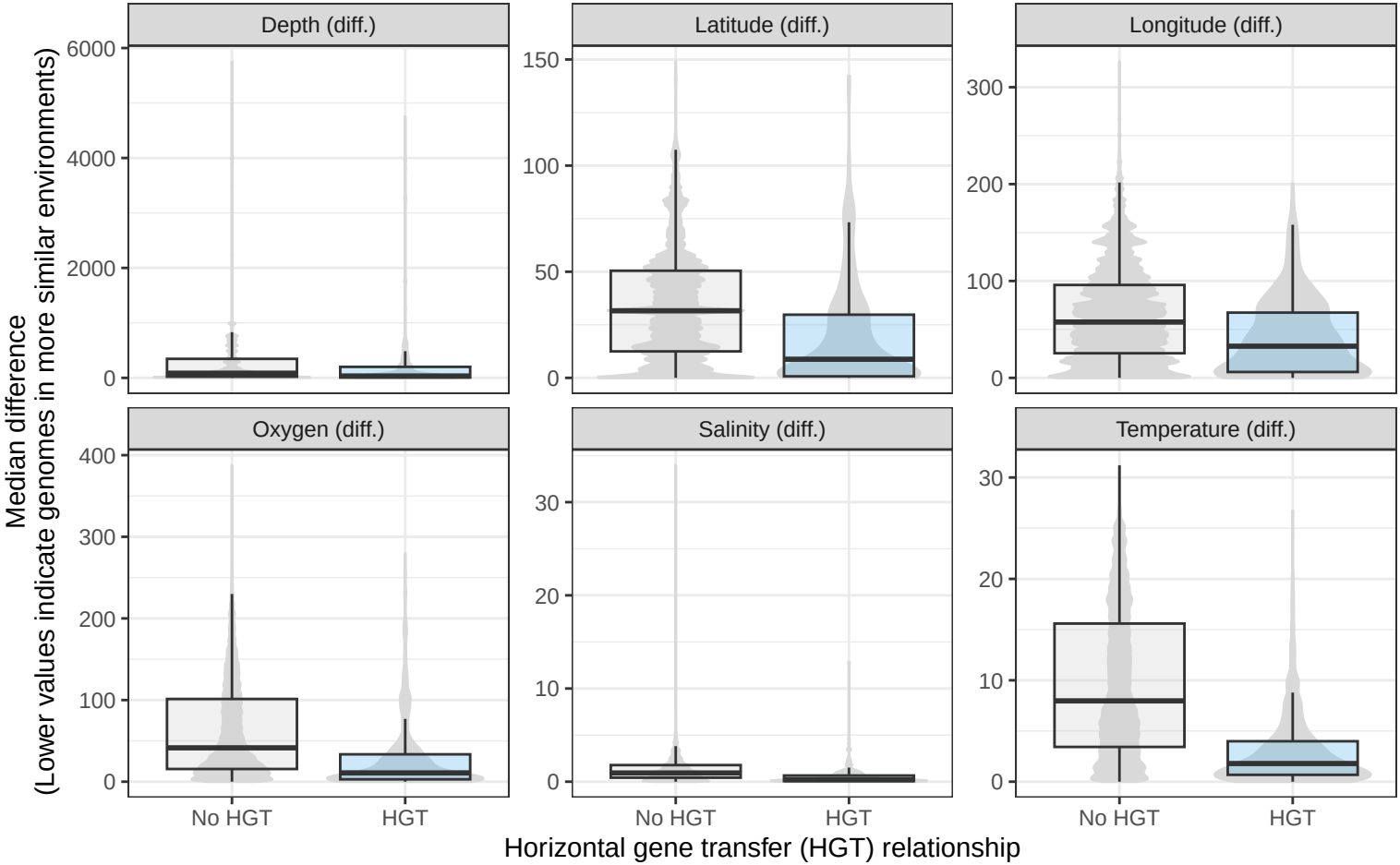
1154 77. Molnar C, Casalicchio G, Bischl B. iml: An R package for Interpretable Machine
1155 Learning. *Journal of Open Source Software* 2018;**3**:786.
1156 https://doi.org/10.21105/joss.00786
1157 78. Friedman JH, Popescu BE. Predictive Learning via Rule Ensembles. *The Annals of*
1158 *Applied Statistics* 2008;**2**:916–954. https://doi.org/10.1214/07-AOAS148
1159 79. R Core Team. R: A Language and Environment for Statistical Computing. 2024.
1160 Vienna, Austria: R Foundation for Statistical Computing, 2024.
1161 80. Tange O. GNU Parallel: the command-line power tool. ;*login: The USENIX Magazine*
1162 2011;**36**:42–47. https://doi.org/10.5281/zenodo.16303
1163 81. Grüning B et al. Bioconda: sustainable and comprehensive software distribution
1164 for the life sciences. *Nat Methods* 2018;**15**:475–476.
1165 https://doi.org/10.1038/s41592-018-0046-7
1166 82. Nayfach S et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol*
1167 2021;**39**:499–509. https://doi.org/10.1038/s41587-020-0718-6
1168 83. Douglas GM, Langille MGI. Current and promising approaches to identify
1169 horizontal gene transfer events in metagenomes. *Genome Biology and Evolution*
1170 2019;**11**:2750–2766. https://doi.org/10.1093/gbe/evz184
1171 84. Lund D et al. Genetic compatibility and ecological connectivity drive the
1172 dissemination of antibiotic resistance genes. *Nat Commun* 2025;**16**:2595.
1173 https://doi.org/10.1038/s41467-025-57825-3
1174 85. Giordano N et al. Genome-scale community modelling reveals conserved
1175 metabolic cross-feedings in epipelagic bacterioplankton communities. *Nat*
1176 *Commun* 2024;**15**:2721. https://doi.org/10.1038/s41467-024-46374-w

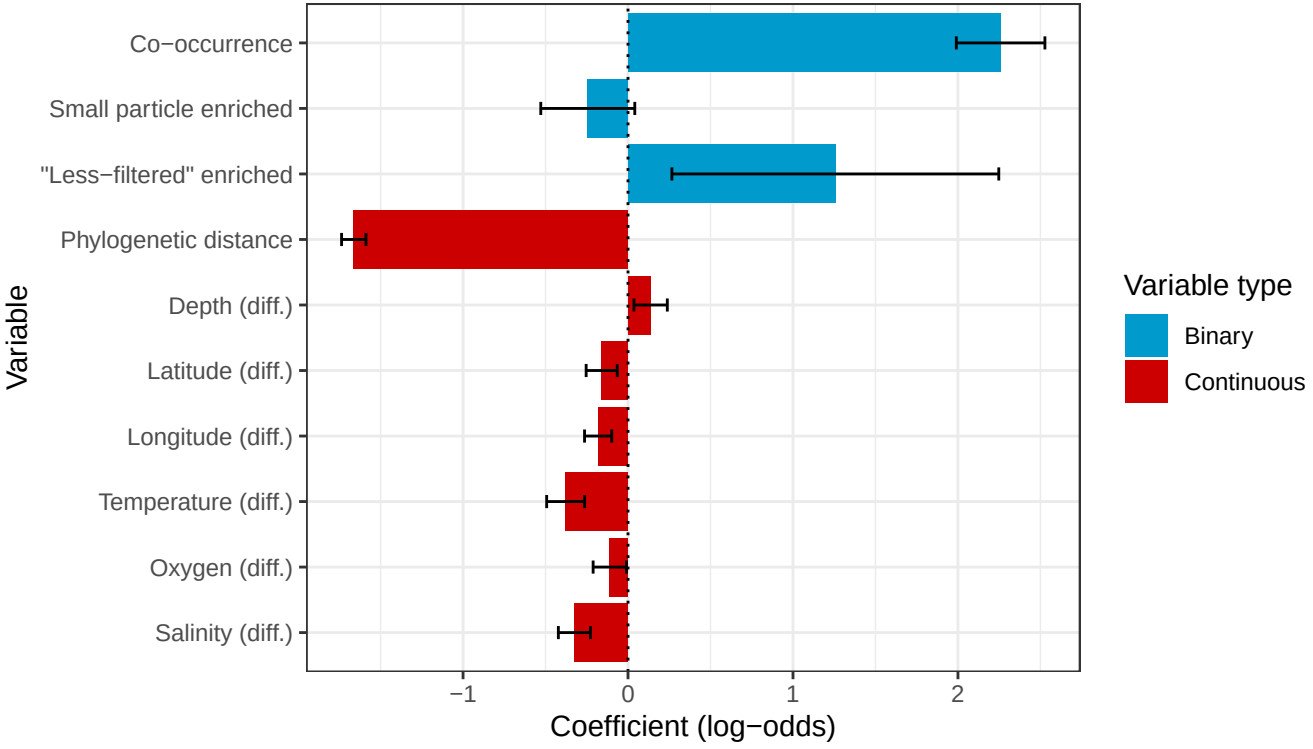
1177 86. Holmberg MJ, Andersen LW. Collider Bias. *JAMA* 2022;**327**:1282–1283.
1178 https://doi.org/10.1001/jama.2022.1820
1179 87. DeLong EF et al. Community Genomics Among Stratified Microbial Assemblages in
1180 the Ocean's Interior. *Science* 2006;**311**:496–503.
1181 https://doi.org/10.1126/science.1120250
1182 88. Zhong J et al. Increasing transposase abundance with ocean depth correlates with
1183 a particle-associated lifestyle. *mSystems* 2024;**9**:e00067-24.
1184 https://doi.org/10.1128/msystems.00067-24
1185 89. Shaiber A, Eren AM. Composite Metagenome-Assembled Genomes Reduce the
1186 Quality of Public Genome Repositories. *mBio* 2019;**10**:e00725.
1187 https://doi.org/10.1128/mBio.00725-19
1188 90. Majernik SN, Beaver L, Bradley PH. Small amounts of misassembly can have
1189 disproportionate effects on pangenome-based metagenomic analyses. *bioRxiv*
1190 2024;2024.10.11.617902. https://doi.org/10.1101/2024.10.11.617902
1191 91. Menet H, Daubin V, Tannier E. Phylogenetic reconciliation. *PLOS Computational*
1192 *Biology* 2022;**18**:e1010621. https://doi.org/10.1371/journal.pcbi.1010621
1193

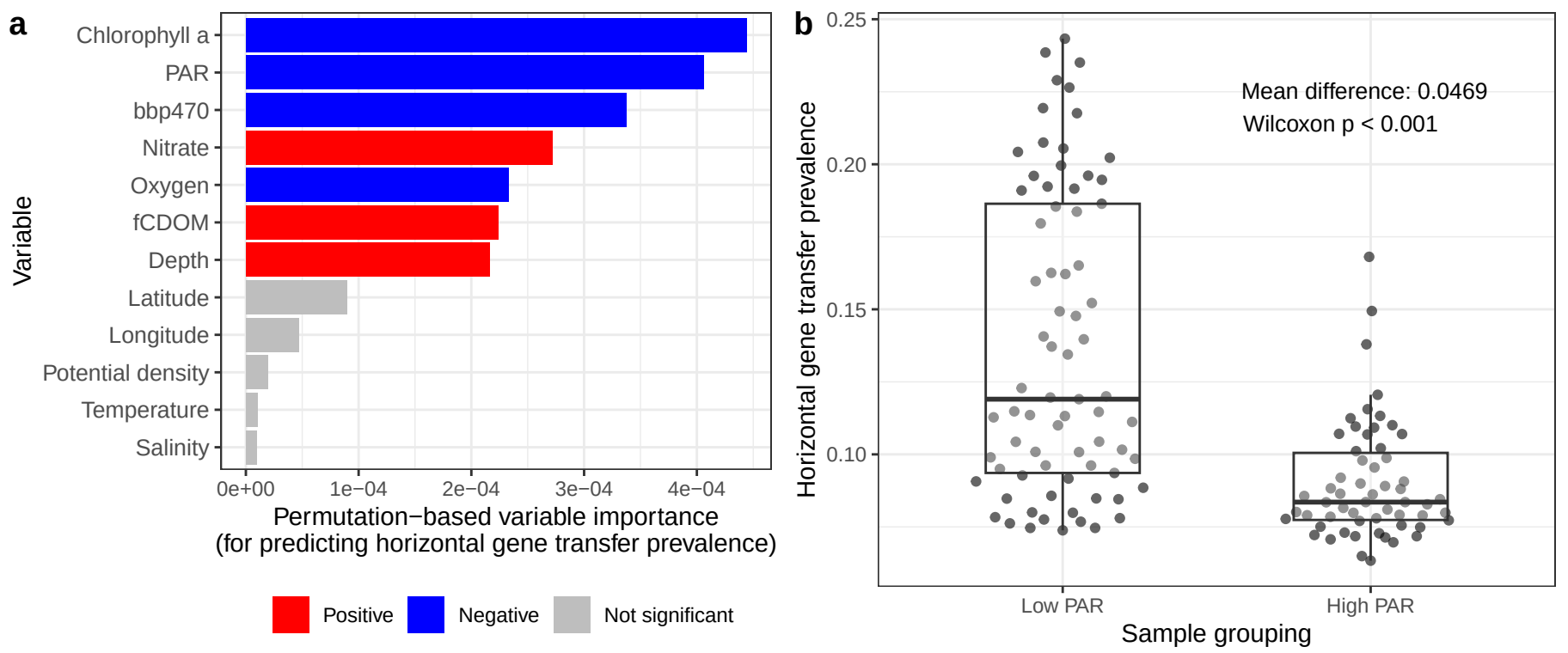

a
>= 95% and < 99%
>= 99%
Number of genome comparisons
1,978
243
202
34
134
10
3,539
1,422
493
147
756
88
2,021,780
2,750,130
11,983,675
13,883,855
66,756,193
19,539,674
Genus
Family
Order
Class
Phylum
Domain
Percent
100
50
0
b
Identity cut−off
All hits (Identity >= 95%)
Identity >= 95% and < 99%
Identity >= 99%
0.10000
0.00100
0.00001
No. putative HGT events
No. genome comparisons
(log10 scale)
Genus
Family
Order
Class
Phylum
Domain
Inter−taxonomic level

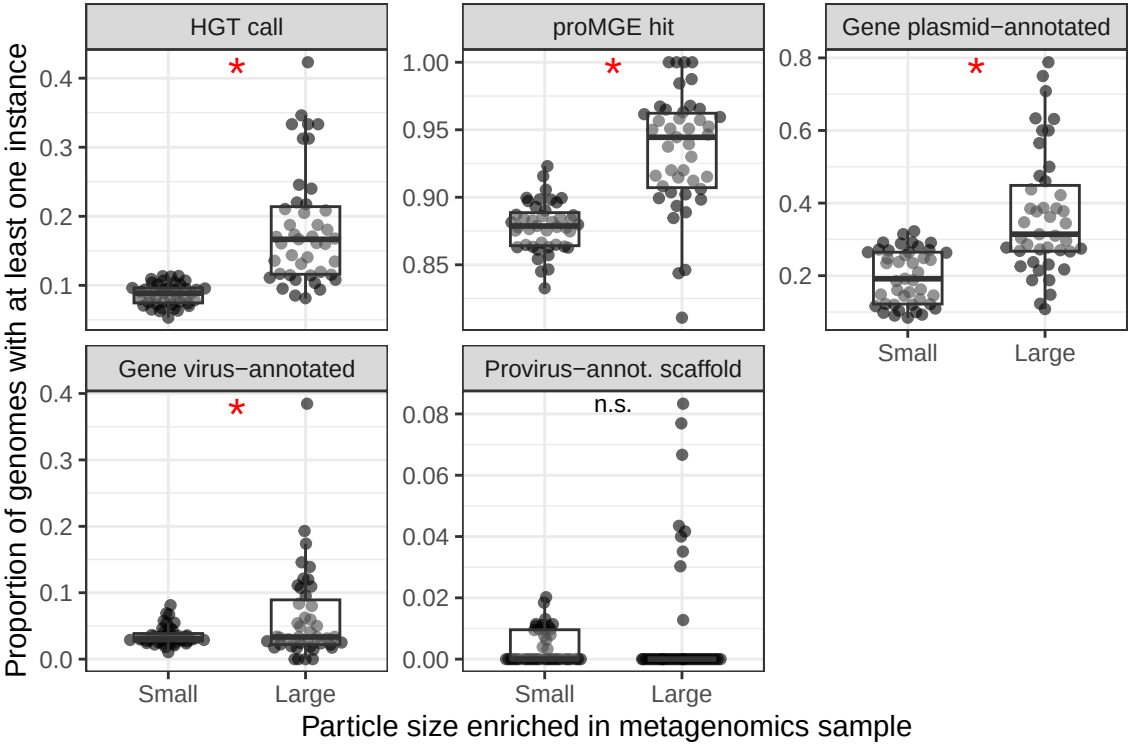

Expected depleted
Expected enriched
Unclear expectation
HGT >= 95% and < 99%
HGT >= 99%
Genus
Family
Order
Class
Phylum
Domain
Sig. log2(Odd's ratio)
4
2
0
−2
−4
Translation, ribosomal structure (J)
Cell motility (N)
Posttrans. mod., pro. turnover, chap. (O)
Signal transduction mechanisms (T)
Amino acid transport and metabolism (E)
Nucleotide transport and metabolism (F)
Coenzyme transport and metabolism (H)
Lipid transport and metabolism (I)
Defense mechanisms (V)
Mobilome: prophages, transposons (X)
Function unknown (S)
Transcription (K)
Replication, recombination and repair (L)
Cell cycle con., cell div., chrom. part. (D)
Cell wall/membrane/envelope biogenesis (M)
Intra. traf., secret., and ves. trans. (U)
Energy production and conversion (C)
Carbohydrate transport and metabolism (G)
Inorganic ion transport and metabolism (P)
Sec. metabol. biosyn., trans. and cata. (Q)


a
Horizontal gene transfer relationship (Cluster−based)
Co−occurring
No
Yes
43,032,534
126
10,422,210
589
No
Yes
Percent
100
50
0
b
No HGT
HGT
Phylogenetic distance
0
1
2
3
4
No
Yes
No
Yes
Co−occurring


Depth (diff.)
Latitude (diff.)
Longitude (diff.)
Oxygen (diff.)
Salinity (diff.)
Temperature (diff.)
Median difference
(Lower values indicate genomes in more similar environments)
No HGT
HGT
Horizontal gene transfer (HGT) relationship


Co−occurrence
Small particle enriched
"Less−filtered" enriched
Phylogenetic distance
Depth (diff.)
Latitude (diff.)
Longitude (diff.)
Temperature (diff.)
Oxygen (diff.)
Salinity (diff.)
Variable
−1
0
1
2
Coefficient (log−odds)
Variable type
Binary
Continuous


a
Chlorophyll a
PAR
bbp470
Nitrate
Oxygen
fCDOM
Depth
Latitude
Longitude
Potential density
Temperature
Salinity
Variable
0e+00
1e−04
2e−04
3e−04
4e−04
Permutation−based variable importance
(for predicting horizontal gene transfer prevalence)
Positive
Negative
Not significant
b
0.25
0.20
0.15
0.10
Horizontal gene transfer prevalence
Mean difference: 0.0469
Wilcoxon p < 0.001
Low PAR
High PAR
Sample grouping


HGT call
proMGE hit
Gene plasmid−annotated
Gene virus−annotated
Provirus−annot. scaffold
*
*
*
*
n.s.
Proportion of genomes with at least one instance
Small
Large
Particle size enriched in metagenomics sample