Limited codiversification of the gut microbiota with humans

Benjamin H. Good[1,2,†]

[1] *Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[2] *Chan Zuckerberg Biohub, San Francisco, CA 94158, USA*
[†] *bhgood@stanford.edu*

A recent study by Suzuki & Fitzstevens *et al* (1) argued that dozens of species of gut bacteria have codiversified with modern human populations. Further analysis of this data reveals that the correlations between human and microbial phylogenies are substantially weaker than unlinked human chromosomes, and that similar correlations can arise through geographic structure alone. This suggests that traditional codiversification has been a limited phenomenon within humans.

Main Text

In a recent study, Suzuki & Fitzstevens *et al* (1) assembled a large collection of paired gut metagenomes and human genomes from diverse human populations. They used these data to argue that dozens of species of gut bacteria have evolved in parallel (or codiversified) with specific human populations. This intimate relationship would be remarkable if true, since it would create extensive opportunities for coevolution between bacterial and human genomes.

While codiversification has previously been observed between host species (2, 3), it is more challenging to detect on the shorter evolutionary timescales separating modern human populations (4). In many species of gut bacteria, the average genetic divergence between strains from different continents is comparable to the diversity among strains from the same geographic location (5). Horizontal transmission is also widespread: previous work has shown that unrelated hosts from different continents can sometimes share nearly identical bacterial strains (6–8), while even the most closely related hosts – identical twins – are often colonized by highly diverged strains over their lifetimes (6).

To identify signatures of codiversification in these settings, Suzuki & Fitzstevens *et al* (1) compared genome-wide phylogenies constructed from human and microbial genomes. Care must be taken when interpreting these trees on shorter within-species timescales: recombination within both humans (9) and gut bacteria (6, 8) produces mosaic genomes that are rarely captured by a single phylogenetic tree. However, recent work has shown that the genome-wide phylogenies estimated from these species can still encode information about the historical patterns of gene flow among the ancestors of a sample (10). In these settings, similarities between the human and bacterial trees will reflect the migration patterns of larger ancestral populations (11), rather than the genealogical fidelity of individual microbial lineages.

Suzuki & Fitzstevens *et al* (1) used a phylogenetic congruence test (PACo) to quantify the similarities between human and bacterial phylogenies. By permuting host labels, they found that 36 of 59 bacterial species were more similar to the human phylogeny than expected by chance ($q<0.05$). While these significant q-values show that the human and bacterial phylogenies are not completely independent, they do not directly demonstrate the evolutionary parallelism implied by traditional models of codiversification (12). For example, theoretical arguments suggest that small amounts of geographic structure can also produce significant PACo scores even in the absence of

codiversification (Fig. 1). Additional information is therefore needed to interpret the PACo scores observed by Suzuki & Fitzstevens *et al* (1).

In principle, independent regions of the human genome could provide a positive control for genomes that have truly codiversified with each other. While mitochondria are strictly maternally inherited, different halves of the nuclear genome typically trace their ancestry back to hundreds of distinct genetic ancestors within a few hundred years, similar to a large random sample from the same local population (13). The residual correlations between these chromosomes can therefore provide a natural null model for codiversification that does not rely on strict maternal inheritance. However, the same PACo analysis performed on separate halves of the human genome yields an effect size of ES$\approx$0.7 – more than an order of magnitude greater than the largest effect sizes reported by Suzuki & Fitzstevens *et al* (1). This gap suggests that the parallelism observed between human and bacterial phylogenies is substantially lower than expected under traditional models of codiversification.

These results are not specific to the PACo test. Direct examination of the phylogenetic distance matrices reveals that two halves of the human genome are substantially more correlated with each other than with any gut bacterial genomes (Fig. 2). For species with the strongest reported codiversification scores in Ref. (1) (e.g. *Prevotella copri*, Fig. 2A), one can observe a small but significant decrease in bacterial relatedness with increasing host distance ($P < 10^{-3}$; Mantel test). However, these small average differences are still dwarfed by the variation within individual countries (Fig. 2A, right, orange) or among the most closely related hosts (grey). Similar behavior can be observed in other species with significant PACo scores (e.g. *Bacteroides vulgatus*, Fig. 2C), though the subtle shifts in mean relatedness can be more difficult to discern by eye ($P < 10^{-3}$, Kolmogorov-Smirnov test). In both cases, the data show that even the most closely related hosts contain bacteria that span nearly the full range of bacterial genetic relatedness. Conversely, hosts from different continents frequently share strains that are as similar as those from the most closely related hosts. These findings are consistent with previous observations in adult twins (6), suggesting that they do not depend on phylogenetic inference scheme employed here.

These results suggest that traditional codiversification is unlikely to have occurred among modern human populations. Gut bacteria may still have adapted to specific human populations, but their genome-wide diversity does not strongly mirror the evolutionary history of their host populations. This distinction has important evolutionary implications, and also potential practical consequences. Widespread codiversification might suggest that probiotic therapies should be sourced based on the geographic origin or genetic ancestry of the host (1). In contrast, the large range of phylogenetic distances in Fig. 2 indicates that such therapies might better informed by directly genotyping the host's microbiota (14), or by identifying specific functional pathways with strong geographic differentiation (15). In light of these results, it interesting to ask why the bacterial phenotypes that were most associated with codiversification in Ref. (1) (e.g. oxygen sensitivity or poor growth at 27°C, Fig. 2D) have not lead to stronger genetic differences in species like *P. copri*, which would seem to have a limited ability to survive outside their hosts. Understanding the mechanisms that allow these bacteria to spread between people – and the signatures this leaves in their genomes – will be critical for furthering our understanding of how humans and their microbiota have coevolved.

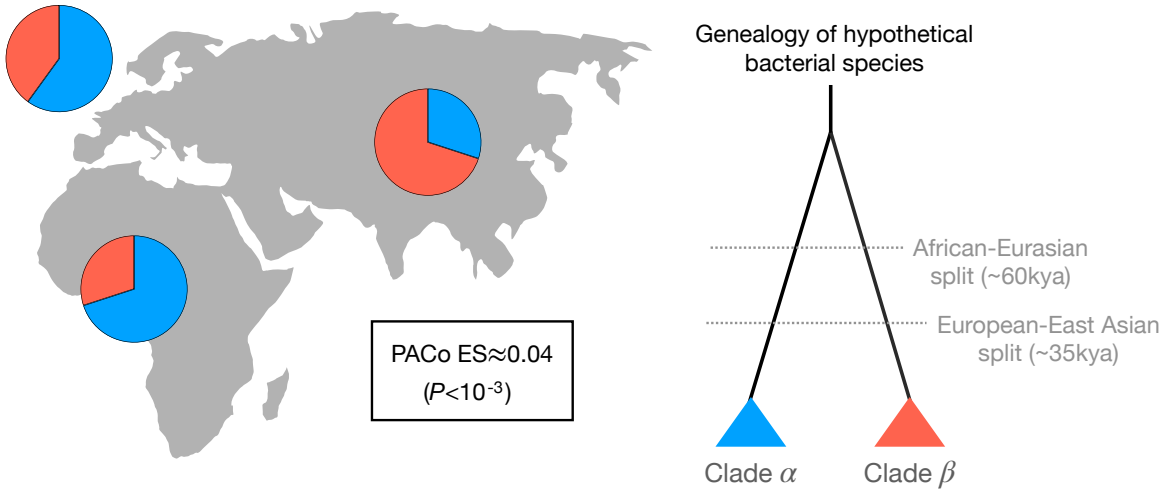

Figure 1: Geographic structure can produce statistically significant PACo scores even in the absence of codiversification. An example of a hypothetical bacterial species that did not evolve in parallel with modern human populations. Two major clades originated before the out-of-Africa migration (right), and are currently shared among all modern human populations. Small shifts in the frequencies of the clades across continents (unrelated to human genetics) can lead to spurious correlations with the observed human phylogeny (left) that are comparable to the largest effect sizes observed in Ref. (1); PACo scores were calculated using the observed human phylogeny for the subset of hosts with *Prevotella copri* strains.

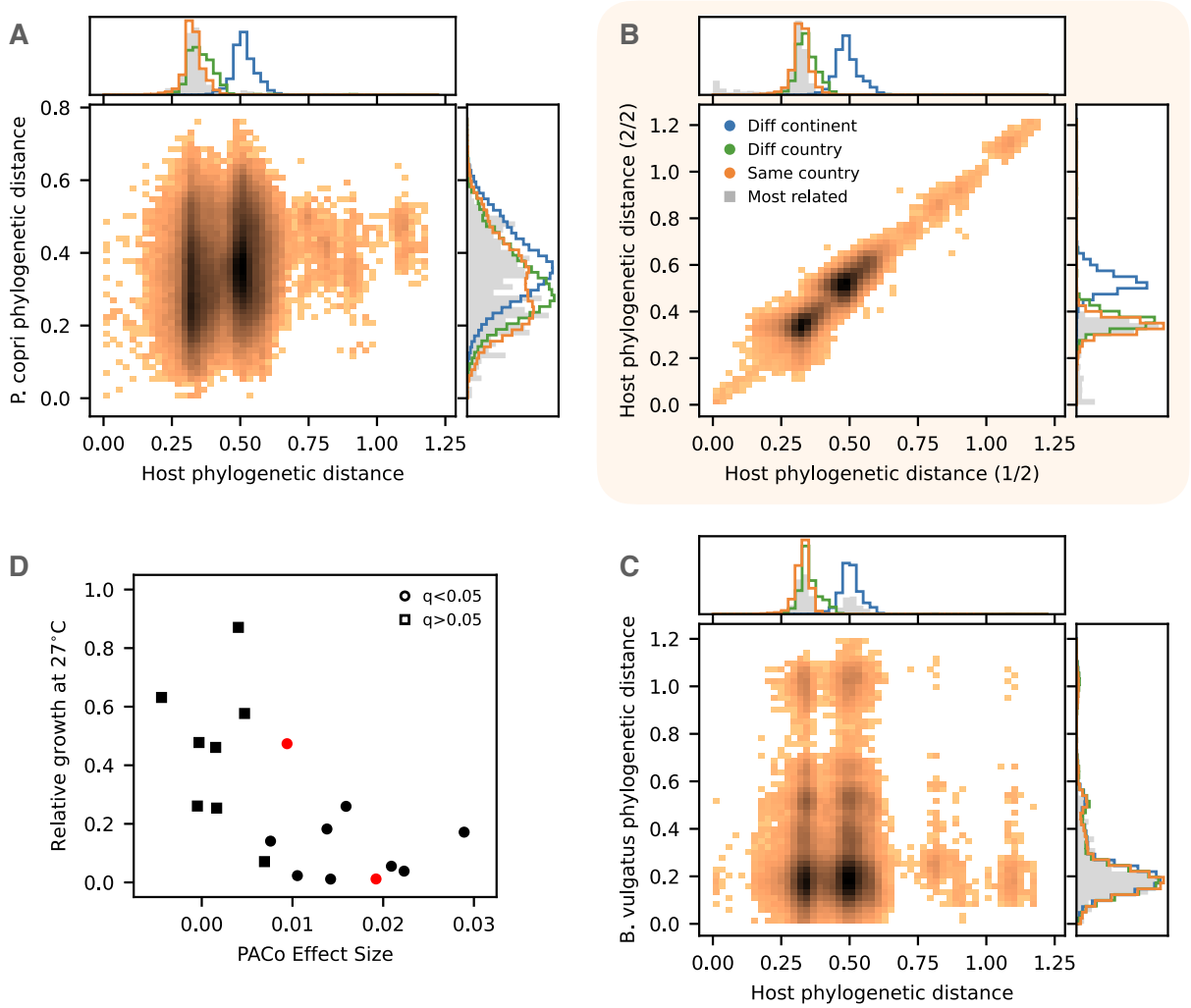

Figure 2: Host genetic similarity is not strongly predictive of bacterial genetic similarity. (a) Correlation between human and bacterial phylogenetic distance for *Prevotella copri*. Heat map shows all pairwise comparisons between adult individuals sequenced by Suzuki & Fitzstevens *et al* (1). Right histogram panel shows the distribution of bacterial distances for hosts in different geographic categories (blue, orange, green), or in the next most closely related hosts (grey); top histogram panel shows analogous distributions for host phylogenetic distance, using the next most closely related *P. copri* strain to construct the grey distribution. (b,c) Analogous figures constructed for separate halves of the human genome (b) and for *Bacteroides vulgatus* (c). (d) Correlation between PACo effect size and relative growth at 27°C as measured by Suzuki & Fitzstevens *et al* (1). *P. copri* and *B. vulgatus* are indicated in red.

References

[1] T. A. Suzuki, *et al.*, *Science* **377**, 1328 (2022).

[2] A. H. Moeller, *et al.*, *Science* **353**, 380 (2016).

[3] N. D. Youngblut, *et al.*, *Nature Communications* **10**, 1 (2019).

[4] A. Bergström, C. Stringer, M. Hajdinjak, E. M. Scerri, P. Skoglund, *Nature* **590**, 229 (2021).

[5] S. Schloissnig, *et al.*, *Nature* **493**, 45 (2013).

[6] N. R. Garud, B. H. Good, O. Hallatschek, K. S. Pollard, *PLoS Biology* **17**, e3000102 (2019).

[7] D. T. Truong, A. Tett, E. Pasolli, C. Huttenhower, N. Segata, *Genome Research* **27**, 626 (2017).

[8] Z. Liu, B. H. Good, *bioRxiv* p. 2022.08.24.505183 (2022).

[9] J. D. Wall, J. K. Pritchard, *Nature Reviews Genetics* **4**, 587 (2003).

[10] T. Sakoparnig, C. Field, E. van Nimwegen, *Elife* **10**, e65366 (2021).

[11] D. Falush, *et al.*, *Science* **299**, 1582 (2003).

[12] M. Groussin, F. Mazel, E. J. Alm, *Cell Host & Microbe* **28**, 12 (2020).

[13] J. Kelleher, A. Etheridge, A. Véber, N. H. Barton, *Theoretical population biology* **108**, 1 (2016).

[14] Z. J. Shi, B. Dimitrov, C. Zhao, S. Nayfach, K. S. Pollard, *Nature Biotechnology* **40**, 507 (2022).

[15] N. Karcher, *et al.*, *Genome biology* **21**, 1 (2020).

Acknowledgments
I thank T. Suzuki, N. Garud, J. Lopez, and other members of the Good lab for useful discussions. This work was supported in part by the Alfred P. Sloan Foundation grant FG-2021-15708 and NIH NIGMS Grant No. R35GM146949. B.H.G. is a Chan Zuckerberg Biohub Investigator.

Competing interests: None declared.

Data and materials availability: Phylogenies and alignments from Ref. (1) were downloaded from the Dryad repository provided in that work; host and bacterial metadata were obtained from the Supplementary Materials. All analysis code used in this work is available on Github (`https://github.com/bgoodlab/codiversification_microbiome`).

Supplemental Figures

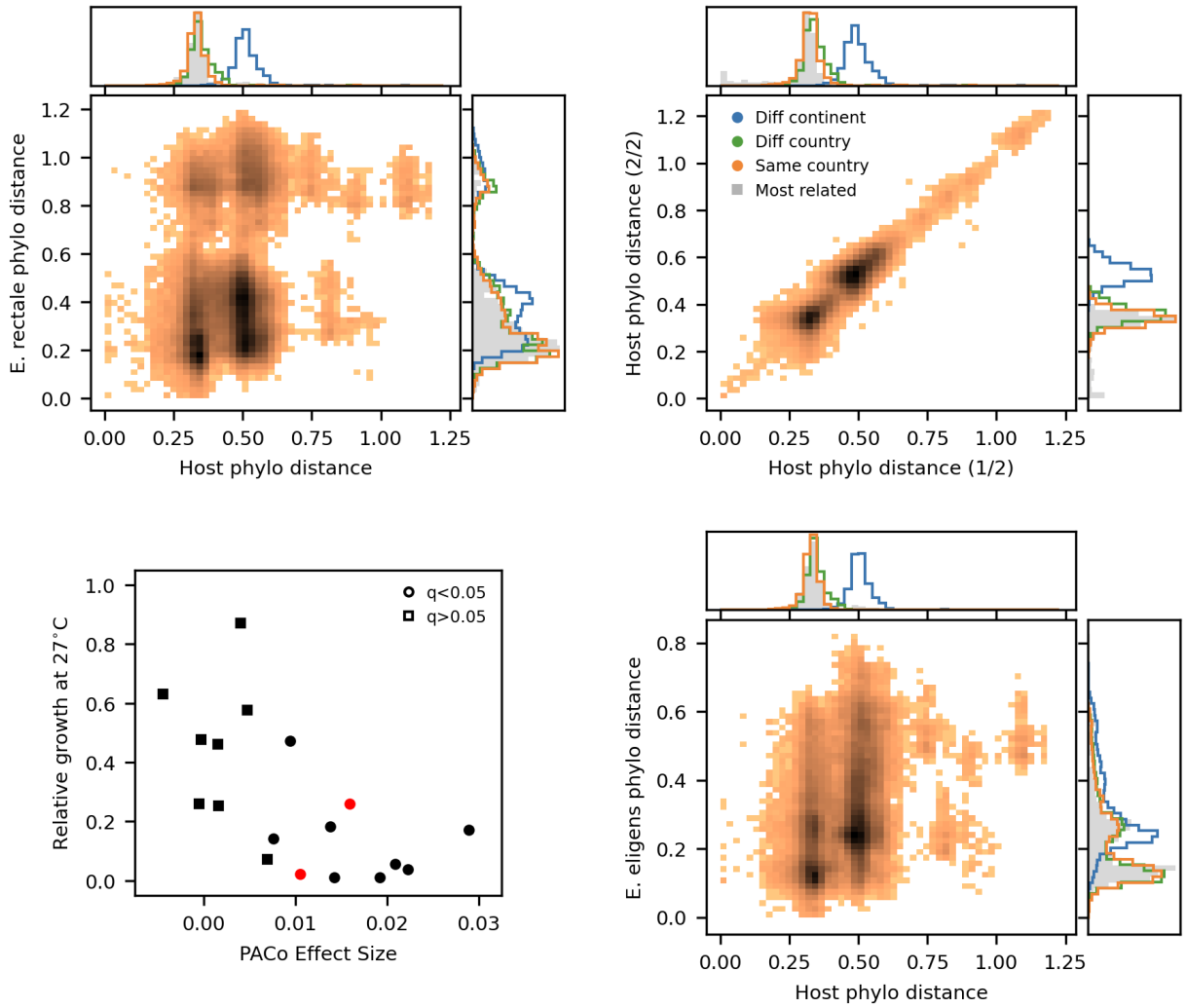

Figure S1: Analogous version of Fig. 2 for two *Eubacterium* species, *Eubacterium rectale* and *Eubacterium eligens*, both of which had highly significant PACo scores. The behavior in both species is largely similar to *P. copri* in Fig. 2A.

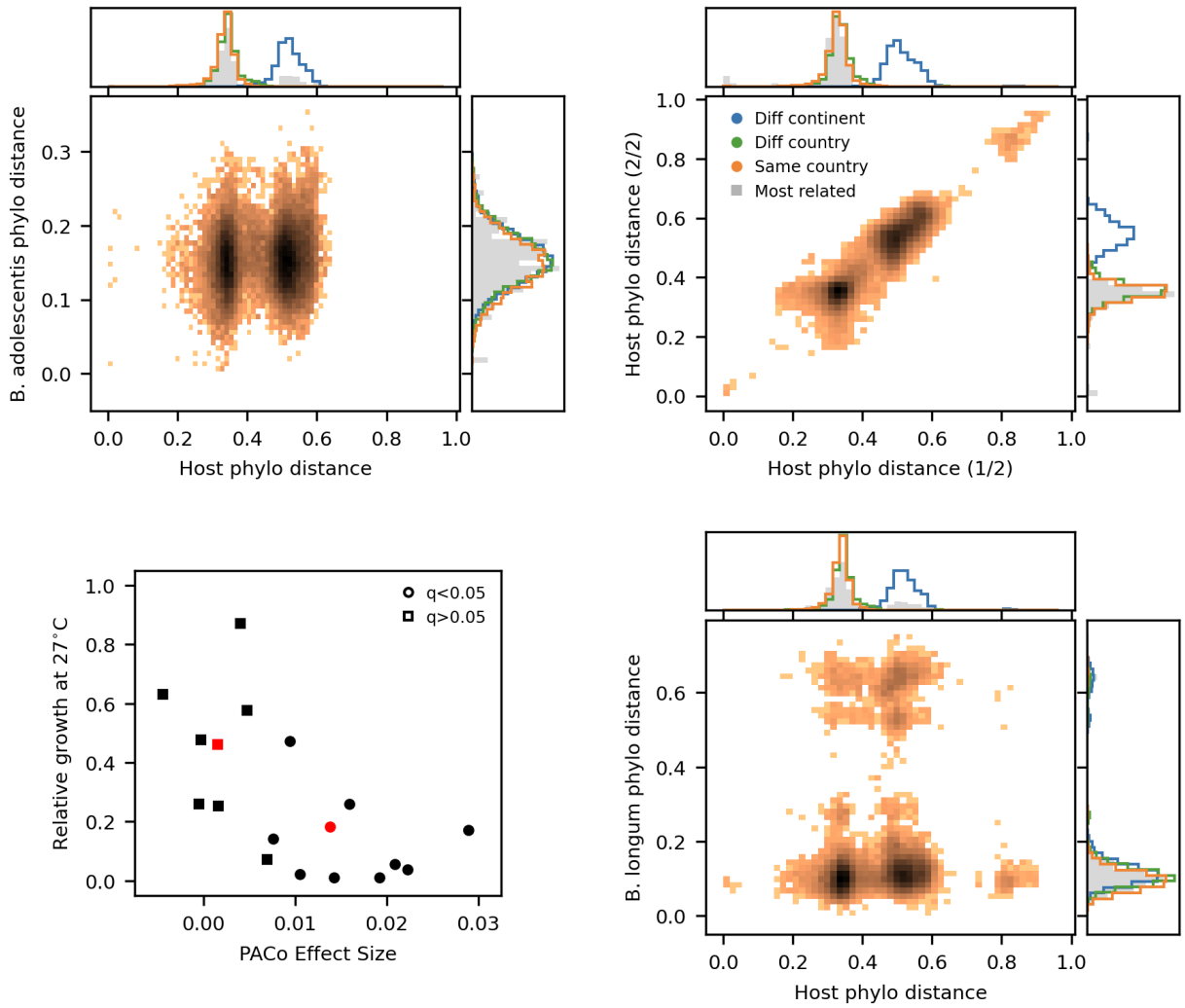

Figure S2: Analogous version of Fig. 2 for two *Bifidobacterium* species. *B. adolescentis* has a significant PACo score that is comparable to *E. rectale*; *B. longum* has a non-significant PACo score that is close to zero. Despite these differences – as well as differences in their relative growth at 27°C – their phylogenetic correlations are similar to each other, and resemble the *B. vulgatus* strains in Fig. 2C.

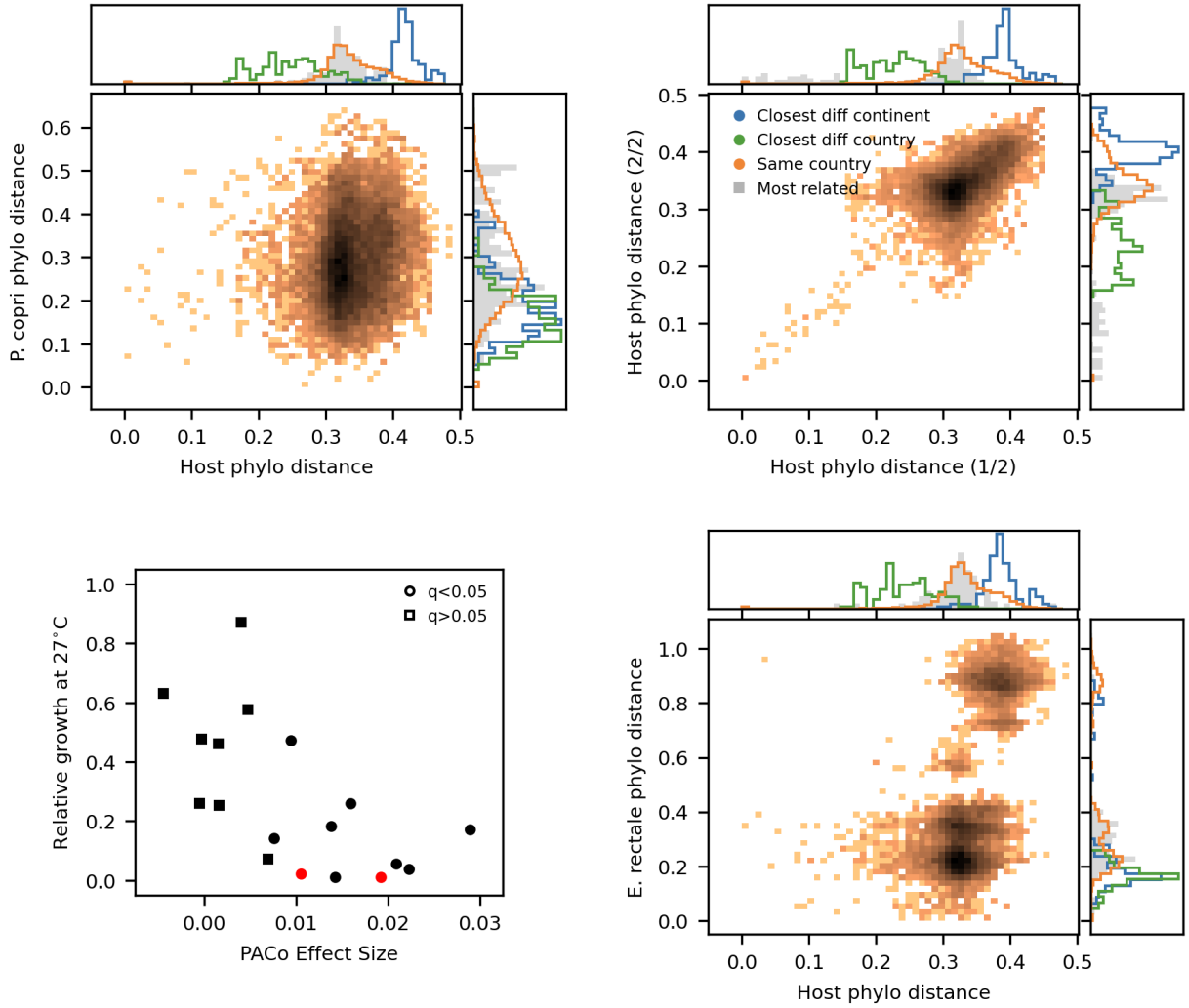

Figure S3: Analogous version of Fig. 2 for *P. copri* and *E. rectale* strains from Gabon. In contrast to Fig. 2, the blue and green distributions in the right panels now refer to the closest bacterium from different continents or different countries, while the top panels show the closest host from different continents and countries. *P. copri* was detected to have a significant within-country codiversification signal by Suzuki & Fitzstevens *et al* (1), while *E. rectale* had an uncorrected P-value <0.05. The present figures show that while the correlations within Gabon are highly significant in both species ($P < 10^{-3}$, Mantel test), they are inconsistent with codiversification, because the diverged bacterial clades within Gabon both have close analogues in other countries and continents.

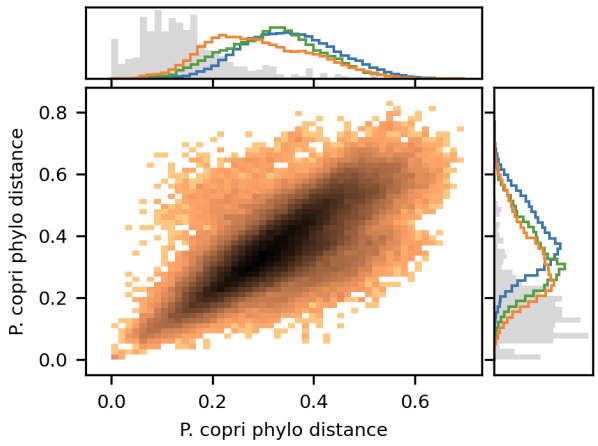

Figure S4: Analogous version of Fig. 2 comparing phylogenetic trees constructed from two halves of the *P. copri* genome. Similar to human genes, the correlation within *P. copri* is much higher than between *P. copri* and humans.

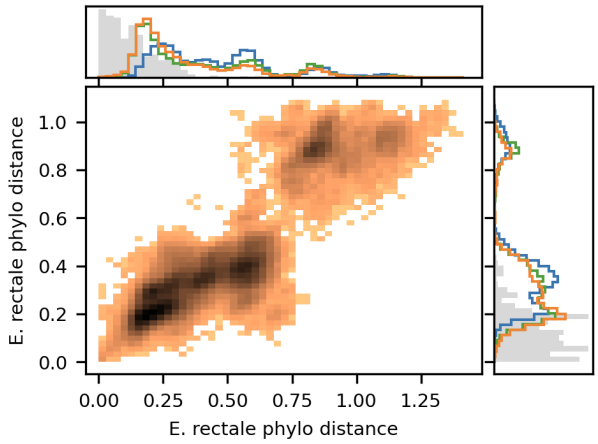

Figure S5: Analogous version of Fig. S4 for *E. rectale.*