Limited codiversification of the gut microbiota with humans

Benjamin H. Good[1,2,3†]

[1] *Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Biology, Stanford University, Stanford, CA 94305, USA*
[3] *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. Reanalysis of their 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 codiversification has been limited within humans.

Main Text

In a recent study, Suzuki & Fitzstevens *et al* (1) assembled a large collection of gut metagenomes and paired 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 (1). An accompanying perspective (2) took this argument one step further, suggesting that gut bacteria have kept fidelity to human lineages for thousands of human generations. Such intimate relationships would be remarkable if true, since they would create extensive opportunities for coevolution between human and bacterial genomes.

While striking examples of codiversification have been observed between host species (3, 4), they are more challenging to detect on the shorter evolutionary timescales separating modern human populations. The underlying genetic distance scales are roughly compatible. Estimates of the mutation accumulation rate (5, 6) suggest that isolated bacterial lineages should have diverged by $\sim$1-10% in the $\sim$60,000 years since the out-of-Africa expansion (7). These differences are comparable to the observed genetic diversity within many species of human gut bacteria (8). However, much like our own genomes, the diversity of the gut microbiota is highly intermingled across space and time. Previous work has shown that the divergence between strains from different continents is often comparable to the diversity among strains from the same geographic location (8, 9). Horizontal transmission is also widespread. Unrelated hosts from different continents can sometimes share nearly identical bacterial strains (10–12), while even the most closely related hosts – identical twins – are often colonized by highly diverged strains over their lifetimes (11).

To identify signatures of codiversification against this backdrop, Suzuki & Fitzstevens *et al* (1) compared genome-wide phylogenies constructed from human and microbial genomes. Care must be taken when interpreting such trees on shorter within-species timescales: recombination within both humans and gut bacteria (11, 12) 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 some information about the historical patterns of gene flow within a sample (13). In these settings, similarities between the human and bacterial trees will reflect the migration patterns of larger ancestral populations (9), 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 the human and bacterial phylogenies. By permuting the links between hosts and bacteria, 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 (14). For example, theoretical arguments show that small amounts of geographic structure can also produce significant PACo scores even in the absence of codiversification (Fig. 1), since human genetic diversity is already correlated with geography (15). The presence of these confounding factors suggests that additional information is 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 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 (16). 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 $ES\approx0.7$ – more than an order of magnitude greater than the largest effect sizes reported by Suzuki & Fitzstevens *et al* (1) (Fig. 2D). Similar values are obtained when the host labels are randomized within countries ($ES\approx0.6$) or continents ($ES\approx0.54$), confirming that they are driven by the largest geographic scales. This large gap in effect sizes 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 distances 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 (e.g. *Prevotella copri*, Fig. 2B), 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 dwarfed by the strain variation within individual countries (Fig. 2B, 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 each of these cases, even the most closely related hosts contain bacteria that span nearly the full range of bacterial relatedness, while hosts from different continents frequently share strains that are as similar as those from the most closely related hosts (Fig. 2E, Figs. S1 & S2). These findings are consistent with previous observations in adult twins (11), suggesting that they do not sensitively depend on the phylogenetic inference scheme employed here (Figs. S3 & S4).

These results suggest that codiversification of the gut microbiota has been limited among modern human populations. Gut bacteria may still have adapted to specific populations or geographic locales, 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. Strong forms of codiversification would imply that it might be beneficial to source probiotic therapies or other microbiome treatments based on the geographic origin or genetic ancestry of the host. In contrast, the large range of phylogenetic distances in Fig. 2 suggests that such therapies might be better informed by directly genotyping the host's microbiota, or by identifying specific microbial genes with strong geographic differentiation (17).

In light of these results, it is interesting to ask why the bacterial phenotypes that were most associated with codiversification in Ref. (1) (e.g. genome size or poor growth at 27°C; Fig. 2D) have not led to stronger genetic differences in species like *P. copri*, given their dramatically reduced ability to survive outside their hosts. One can also find species with smaller genomes (e.g. *Bifidobacterium longum*; Fig. 2D, top) or similar temperature sensitivities (e.g. *Parabacteroides merdae*; Fig. 2D, bottom) that exhibit some of the lowest correlations with their host genomes. These countervailing examples suggest that subtle differences in the rates of gene flow could be a major driver of these patterns, even if they are not directly related to the historical process of codiversification. Understanding the mechanisms that allow these bacteria to spread between hosts – and the signatures this leaves in their genomes – will be critical for furthering our understanding of how humans and their microbiota have coevolved.

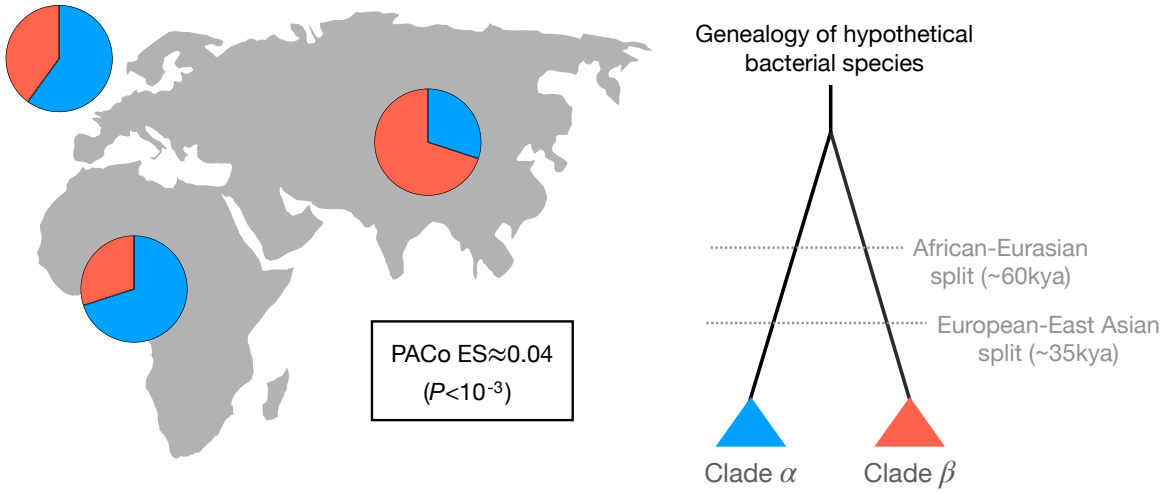

Figure 1: Geographic structure can produce statistically significant PACo scores even in the absence of codiversification. An example of a hypothetical bacterial species that did not evolve in parallel with modern human populations. Two major clades originated before the out-of-Africa migration (right), 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) (Fig. 2D); PACo scores were calculated using the observed human phylogeny for the subset of hosts with *Prevotella copri* strains.

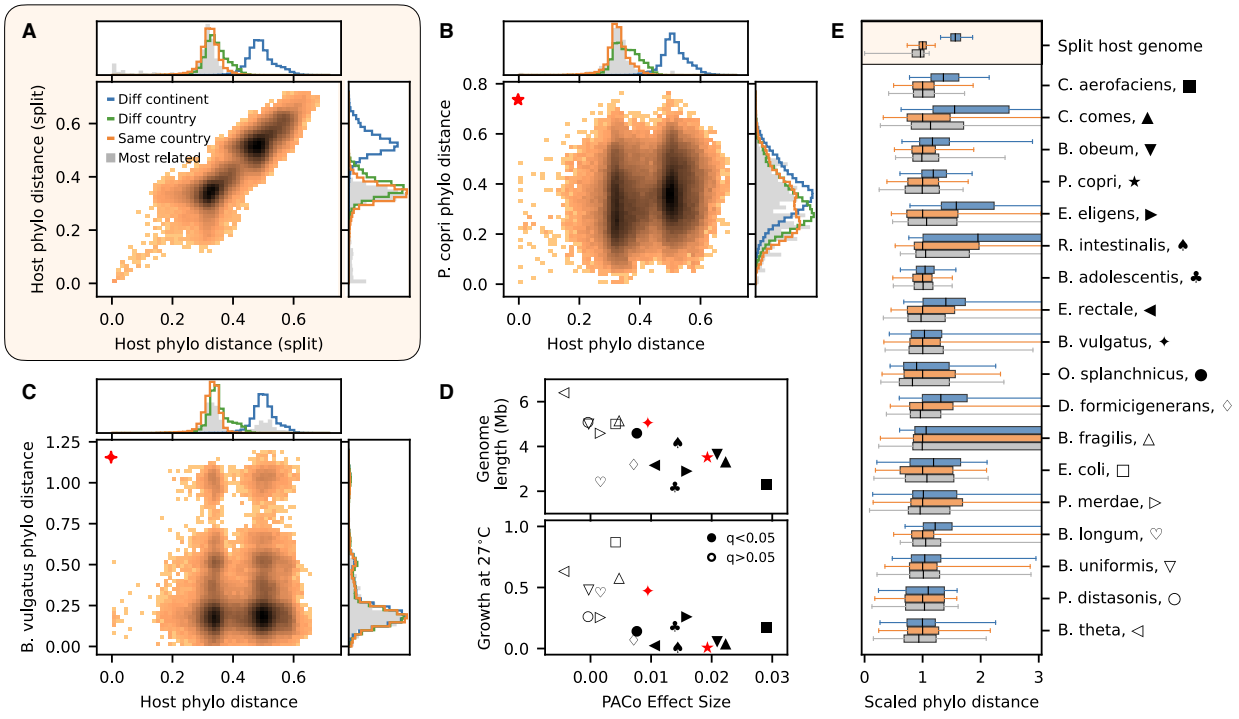

Figure 2: Host genetic similarity is not strongly predictive of bacterial genetic similarity. (A-C) Phylogenetic correlations for separate halves of the human genome (A), *Prevotella copri* (B), and *Bacteroides vulgatus* (C). Heat map shows estimated phylogenetic distances between pairs of adult individuals sequenced by Suzuki & Fitzstevens *et al* (1); three samples with anomalously large host distances (Fig. S2) were removed so that they would not bias the correlation in panel A. 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 bacterial strain to construct the grey distribution. (D) Correlation between PACo effect size and genome length (top) or relative growth at 27°C (bottom) as measured by Suzuki & Fitzstevens *et al* (1). *P. copri* and *B. vulgatus* are indicated in red. (E) Summarized versions of the right histograms in panels A-C for all of the species in panel D. Box plots show medians and interquartile ranges (boxes) and 95% confidence intervals (whiskers); phylogenetic distances are rescaled by the median within-country distance for each species.

References

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

[2] A. H. Moeller, *Science* **377**, 1263 (2022).

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

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

[5] M. Ghalayini, *et al.*, *Appl. Environ. Microbiol.* **84**, e02377 (2018).

[6] S. Zhao, *et al.*, *Cell Host & Microbe* **25**, 656 (2019).

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

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

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

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

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

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

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

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

[15] B. M. Peter, D. Petkova, J. Novembre, *Molecular Biology and Evolution* **37**, 943 (2020).

[16] J. Kelleher, A. Etheridge, A. Véber, N. H. Barton, *Theoretical Population Biology* **108**, 1 (2016).

[17] 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

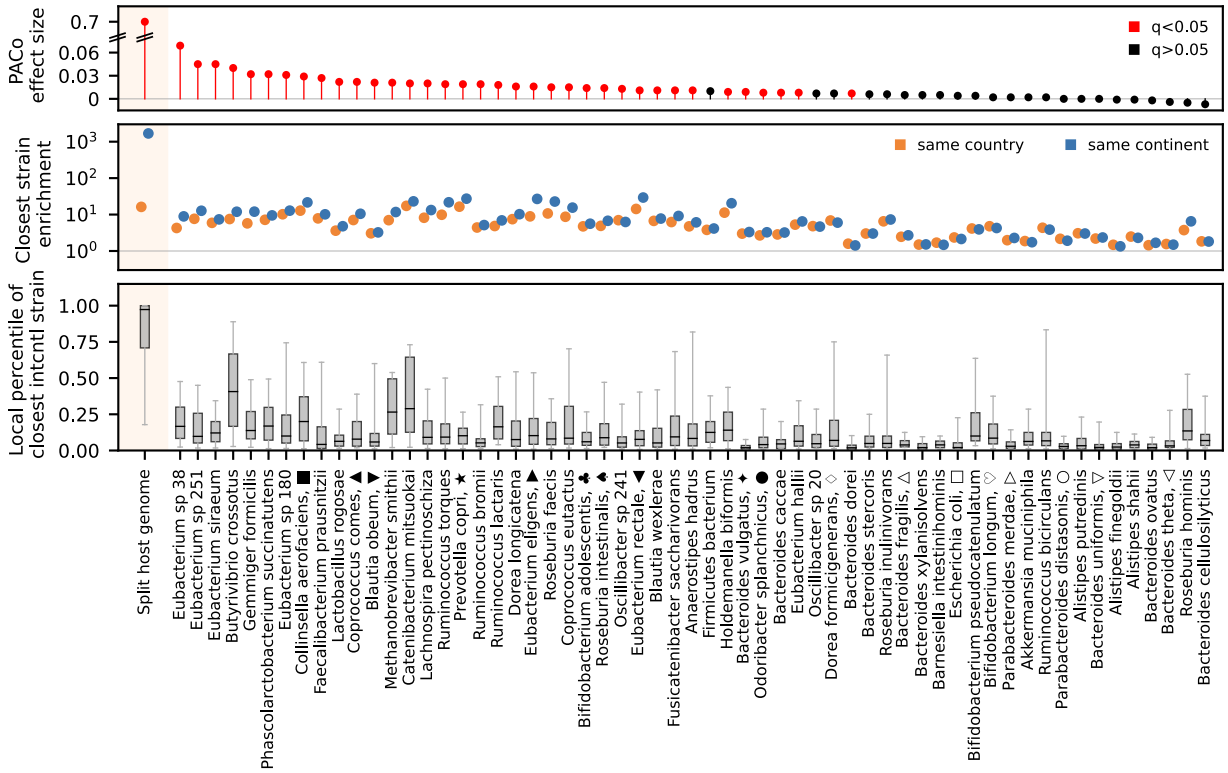

Figure S1: Closely related bacterial strains within and across continents. Top: PACo effect sizes for all of the species analyzed by Suzuki & Fitzstevens *et al* (1); species with significant q-values are highlighted in red. Note the large discontinuity between the human and bacterial effect sizes (hatches). Middle: Enrichment of the probability that the closest relative of a strain derives from the same country (orange) or continent (blue). The enrichment factor is calculated as $e = f(1-f_0)/(1-f)f_0$, where f is observed fraction of strains whose closest relative is located in the same country (or continent) and f_0 is the expected fraction if the strains were randomly permuted across countries. These data show that closest bacterial strains are more likely to come from the same country or continent, but less frequently than their human hosts. Bottom: the phylogenetic distance percentile of the closest inter-continental strain, relative to the set of strains from the same country. These data show that most strains have close relatives on other continents (e.g. in the top quartile of the within-country distribution).

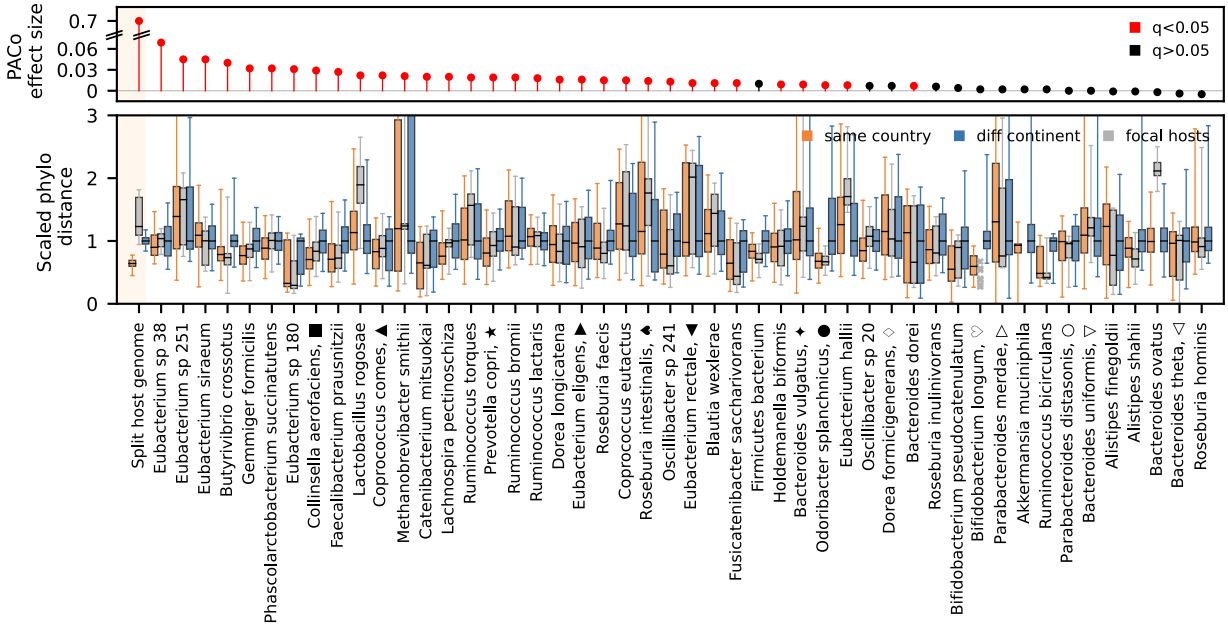

Figure S2: Limited signatures of codiversification in the most highly diverged hosts. Three hosts from Cameroon (BEB30, BEN18, BEN43) were highly diverged from other Cameroonians and from the rest of the human population (bottom, left). The bottom panel shows that for most species, the divergence of their gut bacteria was comparable to the local strain diversity within Cameroon, regardless of the species' PACo effect size (top). To aid visualization, the phylogenetic distances in the bottom panel are scaled by the median inter-continental distance within each species.

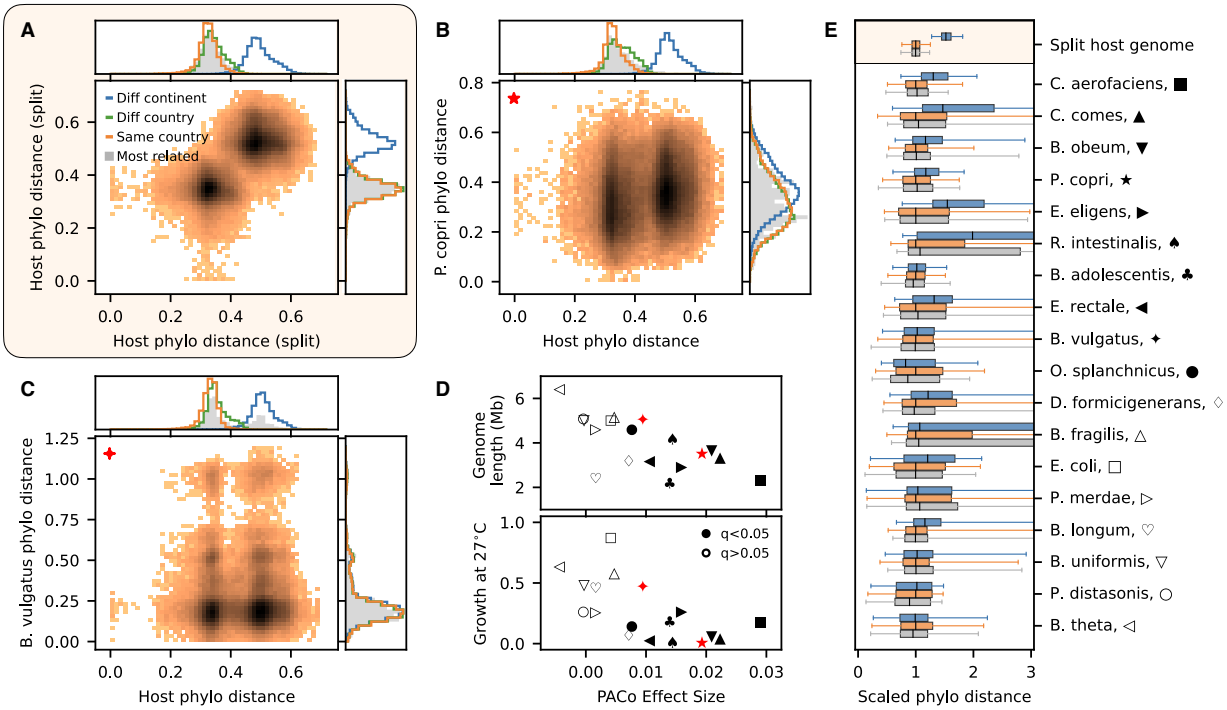

Figure S3: Analogous version of Fig. 2 computed after randomizing host labels within continents.

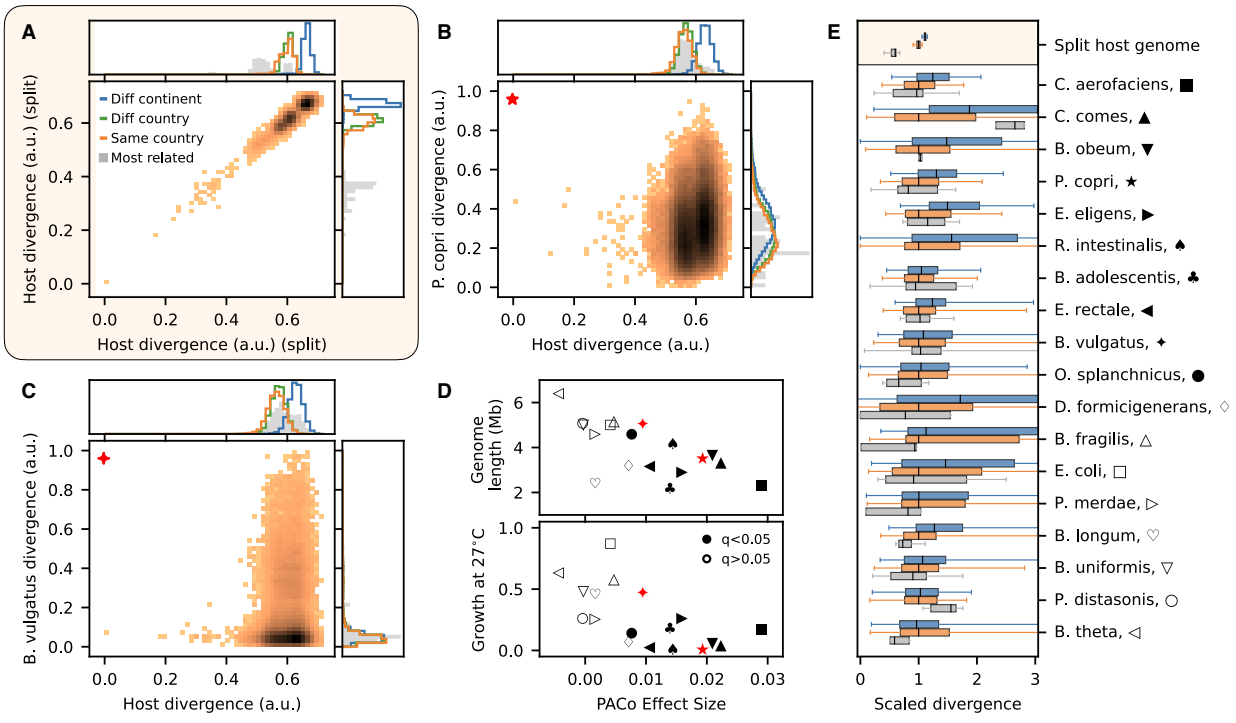

Figure S4: Analogous version of Fig. 2 computed using raw genetic divergences rather than inferred phylogenetic distances.