nature microbiology

Article

https://doi.org/10.1038/s41564-023-01388-w

Widespread extinctions of co-diversified primate gut bacterial symbionts from humans

In the format provided by the authors and unedited

Table of Contents

Supplementary Discussion

Effects of sampling effort on co-diversification statistics

We conducted post-hoc analyses to assess the relationship between sampling effort (*i.e.*, number of hosts or symbiont MAGs sampled) and the observed strength and significance of co-diversification. Across all nodes tested, *p*-values displayed a negative association with the number of host species with MAGs represented present in the clade ($R^2 = 0.038$), but this relationship was weaker than that between Mantel's r and the number of host species with MAGs represented present in the clade ($R^2 = 0.0031$). Conversely, the negative relationship between *p*-values and the number of MAGs present in a clade was weaker ($R^2 = 0.011$) than that between Mantel's r and the number of MAGs present in the clade ($R^2 = 0.0219$). These post-hoc analyses support the combined use of Mantel's r and *p*-values to identify co-diversifying symbiont clades, as the combined use of these statistics appeared to be less biased by sampling effort than the use of either Mantel's r or *p*-values alone.

In addition, we conducted down-sampling analyses to test the sensitivity of our results to the inclusion of each host species. In these analyses, individual host species and their MAGs were removed from the dataset and scans for co-diversification were performed. Thus, these analyses allowed us to test how removal of individual host species affected the degree of co-diversification detected. In every case, the removal of a host species and their MAGs decreased the number of co-diversifying clades detected (2–71 fewer clades detected), as expected. The strength of this effect was proportional to the number of MAGs available for that host species. For instance, removing humans (*Homo sapiens*) or chimpanzees (*Pan troglodytes)*—the two most well-represented host species in the dataset—had the largest negative effect on the number of co-diversifying clades detected (68 or 71 fewer clades detected, respectively). In contrast, removal of lemur species, which were poorly sampled relative to other NHPs and humans, had

the smallest effects on the number of co-diversifying clades detected. These results support that reducing the number of host species sampled did not increase the spurious detection of co-diversifying gut bacterial clades and highlight the positive effect of sampling additional host species on ability of these analyses to detect co-diversification events.

Significant evidence for widespread co-diversification after accounting for tree structure and pseudoreplication

The repeated sampling of host species introduced pseudoreplication into tests for co-diversification between bacterial and host species/subspecies lineages[70]. To assess whether significantly more strongly co-diversifying clades (p-value < 0.05; r > 0.75) were detected than expected by chance while controlling for tree structure and pseudoreplication, we conducted an additional analysis in which the species labels on the host phylogeny were permuted and the scan for co-diversifying clades was repeated. This analysis did not alter the bacterial genome-to-host mapping assignments, such that any pseudoreplication at the level of host-species/subspecies/individual was retained. Results showed that the real data contained more examples of significantly co-diversifying clades than expected based on the permutation tests. For example, results from the scan that permuted the host tips revealed on average 73.92 (standard deviation = 27.9) significant clades, whereas the scan based on the real data revealed 219 (z-score = 4.73). None of the 100 permutations yielded a larger number of significant clades than the real data (219) (non-parametric p-value < 0.05). These results indicate that the primate gut microbiota phylogeny displayed more instances of significant evidence for co-diversification than expected by chance after controlling for tree structure and pseudoreplication. In addition,

these results suggest that >64% of the nodes identified as displaying significant evidence of co-diversification were true positives.

PACo and Hommola tests for co-cladogenesis based on de-replicated symbiont clades

We also assessed the significance of co-diversification based on a subsampling approach in which MAGs from each co-diversifying clade were downsampled to one MAG per host species. For each of the co-diversifying clades (Mantel's r > 0.75, non-parametric *p*-value < 0.01) identified by the scan of the symbiont phylogeny (*i.e.*, results shown in Fig. 1B), we tested for co-cladogenesis between symbionts and host-species (or subspecies) lineages by subsampling each clade to a single MAG per host species (or subspecies) and conducting two additional tests. Specifically, we performed on the subsampled clades both Mantel and Procrustes (PACo) tests[23,71] for association with the host phylogeny. Overall, we observed an elevated frequency of *p*-values <0.05 from these tests relative to that expected under the null hypothesis of no co-cladogenesis despite host-species specificity (Supplementary Table 3), including 11 clades with highly significance *p*-values < 0.001. The significance of co-diversification increased with the number of host species represented by MAGs in symbiont clades. Of the 147 clades containing MAGs from >4 host species, 108 displayed significant evidence of co-diversification in the tests based on the dereplicated data (*p*-value < 0.05). This proportion increased when considering clades containing MAGs from >5 host species (58/68) or >6 host species (22/27). Results of these analyses are presented in Supplementary Table 3. Given that the tests on subsampled clades ignored information about host-species specificity of clades (i.e., subsampling clades eliminated opportunities to falsify co-diversification by observing non-monophyly of MAGs recovered from host species). Thus, these results further indicate that multiple lineages of gut bacteria have diversified in concert with primate host-species.

Mapping analyses corroborate gut bacterial extinctions from humans

The observation that co-diversifying clades of bacterial MAGs ancestral to the Hominini (i.e., present in humans and/or *Pan* and at least one outgroup to the Hominini) are missing representatives from humans is consistent with extinction of these lineages from humans. However, an alternative explanation that cannot be ruled out by analyses of MAGs alone is that these lineages are present in human metagenomes but failed to assemble into MAGs. We reasoned that if lineages identified as absent from humans were in fact present in human metagenomic data, then the mappability of human metagenomic reads should be improved by mapping reads to MAGs from co-diversifying clades ancestral to the Hominini but lacking human-derived representatives compared to mapping reads to human MAGs alone. Adding bacterial genomes to the mapping reference database is expected to improve mappability of metagenomic reads even if the genomes are in fact not present in the metagenomes. However, substantially increased read mappability following the addition of MAGs from co-diversifying clades putatively missing from humans would provide evidence against the extinction of these clades form humans, particularly if MAGs from these clades were present at 1x or more coverage in metagenomic reads.

For these analyses we mapped reads from gut metagenomes from the Human Microbiome Project Healthy Human Subjects (HHS) cohort and Hadza hunter gatherers to the combined set of human-derived MAGs included in Fig. 1B. These analyses showed that >95% of reads were mappable to these genomes, consistent with previous results reported by Pasolli et al. (2019)[19]. Next, we mapped the unmapped reads from the previous analysis to the set of 935 MAGs from co-diversifying clades ancestral to the Hominini but missing from humans. This analysis showed that, in both the HSS and Hadza datasets, >99% of the reads that failed to map to human MAGs also failed to map to MAGs from co-diversifying clades ancestral to the Hominini but missing from humans. For example, 5.169Gb of Hadza reads failed to map to

human-derived MAGs, and 5.093Gb of these reads also failed to map to the MAGs from co-diversifying clades ancestral to the Hominini but missing from humans. Assuming an average bacterial genome size of ~3Mb for the 935 MAGs from co-diversifying clades ancestral to the Hominini but missing from humans, this result indicates a mean per-human host coverage of these MAGs in the Hadza metagenome dataset of ~0.0010. The mean per-host coverage of these MAGs in the HHS dataset was lower at ~6.3e-4. In addition, no individual MAG from co-diversifying clades ancestral to the Hominini but missing from humans was detected in these reads at >1x coverage. Considered together, these results corroborate the conclusion that the co-diversifying clades ancestral to the Hominini identified as absent from humans are either indeed not present in these human populations or present at abundances below the detection threshold afforded by available metagenomic data.

Proportion of ancestral clades absent from humans did not differ among bacterial phyla

Fisher's exact tests were employed to test whether the proportion of clades present in *Pan* and another NHP (i.e., ancestral to the Hominini) that were missing from humans varied significantly among bacterial phyla. These analyses revealed that no two pairs of phyla displayed significant differences in the proportion of co-diversifying clades extinct from humans. For example, Bacteroidota contained 29 ancestral clades not missing from humans and 21 missing from humans, whereas Firmicutes contained 54 not missing and 23 missing from humans (Fisher's exact test *p*-value = 0.1843). For some pairs of comparisons, the small number of clades within phyla yielded low power to detect significant differences of rates of extinction from humans between phyla. For example, Spirochaetota contained the highest proportion of ancestral clades retained in humans (7/8), and Actinobacteriota the second highest (9/12), but these proportions did not differ significantly from those of any other phyla (Fisher's exact test *p*-value > 0.05 in each comparison).

Extinction events from humans were supported by analyses based on the UHGG catalog

To further test whether the co-diversifying clades identified as extinct were absent from available human data, we reperformed phylogenetic analyses using the species-level representative genomes from the UHGG catalog[35]. These analyses recapitulated the monophyly of all co-diversifying clades identified through analyses of MAGs from Pasolli *et al.* (2019)[19], and, moreover, confirmed the absence of human-derived lineages from these clades. Each of the co-diversifying clades inferred to be ancestral to the African apes and missing representatives from humans based on analyses of data from Pasolli *et al.* (2019)[19] was also missing representatives from humans based on analyses of the UHGG catalog (Supplementary Table 3). These results further support the extinction of multiple ancestral clades of co-diversifying gut bacteria from human populations.

Supplementary References

70. Mazel, F., Davis, K. M., Loudon, A., Kwong, W. K., Groussin, M. & Parfrey, L. W. Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? *mSystems* **3** (2018).

71. Balbuena, J.A., Míguez-Lozano, R. & Blasco-Costa, I. PACo: a novel procrustes application to cophylogenetic analysis. *PloS ONE **8***, e61048 (2013).

Supplementary Table Captions

Supplementary Table 1. Metadata for chimpanzee and bonobo samples.

Supplementary Table 2. Assembly statistics, taxonomy, and other metadata for human and NHP MAGs.

Supplementary Table 3. Trees and statistics for co-diversifying and non-co-diversifying clades; statistics for genes enriched in co-diversifying clades. The uncorrected p-values shown for the co-diversification scan results were derived from Mantel tests. The uncorrected p-values shown for functions in enriched in co-diversifying MAGs were derived from phylogenetic ANOVA.

Supplementary Table 4. Molecular clock analyses of co-diversifying bacterial genomes.

Supplementary Table 5. Functional enrichment of COG functions, categories, and pathways between Pan-derived MAGs from co-diversifying clades missing from humans versus present in humans. Enrichment scores were calculated as the Rao test statistic for equality of proportions as implemented in Anvi'o anvi-compute-functional-enrichment.

Supplementary Table 6. Statistics for non-codiversifying clades (r < 0) detected in Pan and another NHP. The uncorrected p-values shown for the co-diversification scan results were derived from Mantel tests.

Supplementary Table 7. Annotations, taxonomy, and dN/dS values for CORFs shared by representatives of co-diversifying clades from humans and chimpanzees.