Vast differences in strain-level diversity in the gut microbiota of two closely related honey bee species

Kirsten M Ellegaard\textsuperscript{a}, Shota Suenami\textsuperscript{b}, Ryo Miyazaki\textsuperscript{bct}, Philipp Engel\textsuperscript{a}#

\textsuperscript{a}Department of Fundamental Microbiology, University of Lausanne, 1015 Lausanne, Switzerland
\textsuperscript{b}Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 305-8566 Tsukuba, Japan
\textsuperscript{c}Computational Bio Big Data Open Innovation Laboratory (CBBD-OIL), AIST, 169-8555 Tokyo, Japan

#equally contributed senior authors.

Correspondence:
Prof. Philipp Engel or Kirsten M. Ellegaard
Department of Fundamental Microbiology
University of Lausanne, CH-1015 Lausanne, Switzerland
Tel.: +41 (0)21 692 56 12
e-mail: philipp.engel@unil.ch; kirsten.ellegaard@unil.ch
**Abstract**

Most bacterial species encompass strains with vastly different gene content. Strain-level diversity in host-associated microbial communities is therefore considered to be of functional importance. Yet, little is known about strain-level diversity differences in phylotypes shared among related host species, or the underlying mechanisms that may constrain and maintain diversity. Here, we used shotgun metagenomics to characterize and compare the gut microbiota of two closely related honey bee species, *Apis mellifera* and *Apis cerana*. While both host species are colonized by largely overlapping bacterial 16S rRNA phylotypes, our metagenomic analysis revealed that their communities are host-specific at the species- and strain-level with few between-host transfers being detected. Strikingly, *A. mellifera* displayed a much higher extent of strain-level diversity and functional gene content in the microbiota than *A. cerana*, per colony and per individual bee. In particular, the gene repertoire for polysaccharide degradation was massively expanded in the microbiota of *A. mellifera* relative to *A. cerana*. Our results illustrate that the gut microbiota of two closely related animal hosts, with similar ecology and overlapping geographic distribution, can undergo dramatic changes in composition and diversity within a timespan of only 6 million years of evolution, with likely consequences for gut microbiota functioning and host-symbiont interactions. Human bee management, divergent ecological adaptation, or habitat size may have contributed to the observed differences in microbiota composition of these key pollinator species. Further studies are needed to evaluate both the underlying causes and functional consequences for honey bee health and bee management.
Significance

Little is known about how gut bacterial communities evolve and adapt across closely related host species, because it requires genome-wide approaches that go beyond classical marker gene analysis and provide insights at the resolution of individual strains and genomic content. We show that the gut microbiota of two closely related honey bee species has undergone major changes in strain-level diversity and functional gene content, despite the relatively recent common ancestry. Western honey bees harbor much more diversity per bee and per colony than Eastern honey bees, with a larger and more diverse repertoire of functions for sugar breakdown. This is likely to have consequences for the functioning of the gut microbiota, raising important questions about the impact of the gut microbiota on pollinator health.
Introduction

Most bacteria live in complex communities, composed of hundreds to thousands of species, which in turn encompass strains with highly variable gene content (1, 2). Also in host-associated bacterial communities, strain-level diversity can be substantial, despite the general assumption that genetic diversity destabilizes mutualistic interactions (3). For example, multiple strains of a sulfur-oxidizing endosymbiont were found to co-colonize individual hosts of deep-sea mussels, presumably because they encode complementary functions (4). In contrast, strains of the human gut microbiota have been shown to segregate among individuals, resulting in host-specific genetic profiles (5-7). However, despite the increased awareness of the existence and functional importance of strain-level diversity in host-associated bacterial communities, little is known about differences in strain-level diversity across host species, or the underlying mechanisms which constrain and maintain diversity within and among hosts. This is largely due to the technical challenges associated with the deep-sequencing of bacterial genomes from complex natural communities, and the generation of comparable datasets across host organisms.

Unlike most animals, eusocial corbiculate bees (honey bees, stingless bees, and bumble bees) have been shown to harbor a specialized gut microbiota with a simple and highly consistent taxonomic composition, consisting of up to 10 phylotypes, as based on 16S rRNA gene analyses (8). Considering the common occurrence of these phylotypes across bees, it is plausible that they were acquired around the time when eusociality evolved in the bees (8). Interestingly, previous studies on bacterial isolates have demonstrated an impressive amount of genomic diversity within
phylotypes, where strains isolated from different hosts often represent divergent phylogenetic sub-lineages (9-12). Moreover, a recent metagenomic analysis of the gut microbiota of the Western honey bee, *Apis mellifera*, found that even within the same host species divergent sub-lineages can be present (5). We refer to these sub-lineages as sequence-discrete populations (SDPs), since metagenomic data demonstrated that they were discrete from each other, in addition to being sufficiently divergent to be considered as different bacterial species (13). Given these results, it is possible that bee species with similar phylotype-level composition in their gut microbiota, including very closely related species of honey bees, harbor very different communities. However, comparative community-wide analyses, based on data with strain- and genome-level resolution, are still lacking.

In the current study, we perform a comparative metagenomic analysis of the gut microbiota of two closely related species of honey bees, *Apis mellifera* and *Apis cerana*. Based on molecular data, their last common ancestor has been dated to approximately 6 million years ago (14-17), and previous 16S rRNA-based studies have shown that they harbor largely overlapping gut microbiota phylotypes (>97% 16S rRNA identity) (8). With the exception of *A. mellifera*, all extant species of honey bees (genus *Apis*) are confined to Asia, pointing towards an Asian origin of the *Apis* genus (15). Based on molecular analysis, *A. mellifera* expanded into its native range (Africa, Europe, and Western Asia) approximately 300,000 years ago (14). However, *A. mellifera* was recently re-introduced to Asia by humans (18, 19), thereby bringing not only the bees, but also their associated bacterial communities into close proximity, potentially resulting in a homogenization of their gut microbiota.
In order to compare the composition, diversity, and evolution of the gut microbiota of *A. mellifera* and *A. cerana* at the strain-level, we analyzed shotgun metagenomes of individual bees, using a common DNA extraction protocol and comparable sequencing depth. We find that each host species harbors a highly distinct bacterial community, largely composed of different SDPs, with occasional transfers among sympatric bees. Quantitative analysis revealed that the gut microbiota diversity of *A. mellifera* is much higher than for *A. cerana*, resulting in a larger metabolic flexibility at both the individual and colony level. These results represent the first comparative genome-wide analysis of strain-level diversity between related host-associated microbial communities, raising new questions regarding underlying mechanisms and functional consequences.
Results

Metagenomic data reveals that the gut microbiota of *A. mellifera* and *A. cerana* are distinct.

A total of 40 shotgun metagenome samples were collected from individual bees, with 20 bees per host. Two colonies were sampled from each host, all from different apiaries, no more than 100km apart, close to Tsukuba (Japan). In order to perform a metagenomic characterization of the community composition across samples, we first established a genomic database of isolated strains (Dataset S1) representative of both hosts. In total, 10 new genomes isolated from *A. cerana* were sequenced and added to a previously established, non-redundant honey bee gut microbiota database (5), together with previously published genomes isolated from more distantly related bee species. Approximately 90% of the host-filtered reads mapped to the new database, regardless of the host affiliation of the samples, indicating that the database is highly and equally representative of the gut microbiota of both hosts (Fig. S1).

To make an initial broad comparison of the gut microbiota composition, the relative abundance of all phylotypes in the database was quantified, based on mapped read coverage to single-copy core genes. Consistent with previous studies employing amplicon sequencing of the 16S rRNA gene (8), the five phylotypes constituting the core microbiota of corbiculate bees were found to colonize both hosts, although the relative abundance profiles were distinct between the hosts (Fig. S2). Other phylotypes associated with *A. mellifera* (*Bartonella apis*, *Frischella perrara*, *Commensalibacter* sp.) were not detected in any of the *A. cerana* samples.
Conversely, while *Apibacter* was not detected in any of the *A. mellifera* samples, it appeared to be a prevalent member for *A. cerana* (Fig. S2, yellow color).

To determine whether the five core phylotypes colonizing both hosts are distinct at the SDP level, candidate SDPs were inferred from isolate genomes in the database, based on core genome phylogenies (Fig. 1A-C,G,H) and pairwise average nucleotide identities (ANI) (Dataset S2). Subsequent metagenomic validation (Fig. S3-S4) confirmed three new SDPs within the *Gilliamella* phylotype (Fig. 1A), and one new SDP within the *Lactobacillus* Firm5 phylotype (Fig. 1B), all of which were represented exclusively by *A. cerana*-derived isolates. A new SDP was also confirmed for *Snodgrassella* (Fig. 1C), based on isolates from *A. cerana*, *A. andreiformis* and *A. florea*, suggesting that this SDP may be shared among other species of honey bees than *A. mellifera*. In contrast, for the two remaining core phylotypes *Lactobacillus* Firm4 and *Bifidobacterium*, no new candidate SDPs were inferred (Fig. 1G-H), since the *A. cerana*-derived genome isolates had ANI values of up to 95% and 90% to the *A. mellifera* isolates, thus falling within the range of ANI values observed among the *A. mellifera* isolates (Dataset S2).

To further validate the host specificity of the novel SDPs and detect eventual cross-host transfers, we quantified the relative abundance of each SDP across the metagenomic samples, including samples previously collected in Switzerland. All SDPs found to be host-specific in the genomic database displayed a clear host preference across the metagenomic samples, but a small number of transfers were nevertheless detected among the Japanese samples within the *Lactobacillus* Firm5 and *Gilliamella* phylotypes (Fig. 1D-F). These results therefore suggest that honey
bees do get exposed to non-native SDPs, at least in Japan, occasionally resulting in colonization. For the SDPs previously described for A. mellifera (5), the abundance patterns were very similar between Swiss and Japanese samples, with co-occurrence of SDPs within individuals being the norm for all core phylotypes except Gilliamella (Fig. 1D-F, Fig. S5). Thus, these distributions likely represent conserved patterns of co-existence.

For the two core phylotypes harboring SDPs with genome isolates from both hosts (Fig. 1G-H), rooting of the core genome phylogenies with related SDPs showed that the A. cerana-derived isolates diverged prior to the A. mellifera-derived isolates, as would be expected if a more recent host-specialization had occurred. Therefore, to determine whether the communities differ at the strain-level, the fraction of shared single-nucleotide variants was calculated for all sample pairs, and visualized using principal coordinate analysis (Fig. 1I and J). Remarkably, the samples were found to cluster strongly by host, indicating that each host is colonized by a distinct population of strains. In contrast, there was no clustering by country or colony affiliation (Fig. 1I-J). For other community members, clustering by country or colony was observed for only a subset of SDPs (Dataset S3), indicating that strains are not necessarily geographically specialized.

Finally, to obtain a database-independent estimate of the divergence between the gut microbiota of A. mellifera and A. cerana, we de novo assembled the metagenomes and compared the gene contents by clustering all predicted ORFs (open reading-frames) by sequence identity. Metagenomic assemblies and ORFs were inferred individually for each sample, using a subset of 20 million paired-end host-filtered
reads per sample, and clustering was done using a range of thresholds (80-95% nucleotide identity). Only a very minor fraction of the clusters contained sequences from both hosts (Fig. 1K) further corroborating the small overlap of the gut microbiota between the two honey bee species.

In conclusion, although the gut microbiota of A. mellifera and A. cerana appear similar when the community is characterized at the phylotype level, metagenomic analysis clearly demonstrates that the communities are highly distinct, being composed of divergent SDPs and strains, with a low frequency of cross-host transfers.

The diversity of the gut microbiota is higher in A. mellifera compared to A. cerana.

As shown in Fig. 1K, the clustering analysis of the metagenomic ORFs resulted in a much larger number of clusters for A. mellifera, compared to A. cerana. This difference may in part be explained by A. mellifera housing a bacterial community composed of more SDPs. While 90% of the host-filtered reads map to the database for both hosts (Fig. S1), this fraction represents 12 SDPs for the five core phylotypes in A. mellifera (plus up to 4 non-core members), but only 7 in A. cerana (plus Apibacter sp., and occasionally Lactobacillus kunkeei) (Fig. 1 and Fig. S1). Indeed, the number of gene clusters and the total genome assembly size per bee were both approximately twice as large for A. mellifera when using comparable subsets of host-filtered reads (Fig. 2A and Fig. S6). However, the sharp drop in the number of clusters from 95% to 90% sequence identity observed only for A. mellifera (Fig. 1K) indicates that strain-level diversity is also a contributing factor.
Previous analysis of strain-level diversity in *A. mellifera* showed that strains tend to segregate among individuals within colonies (5), in contrast to the SDPs, which mostly co-exist ([Fig.1D-F, Fig. S5](#)). Therefore, sequence clusters occurring only in a subset of bees are likely to represent strain-level diversity. To determine how the sequence clusters distribute across bees, we plotted the number of clusters relative to the number of samples, using the 95% nucleotide sequence identity threshold. From the cumulative curves ([Fig. 2B](#)), it is evident that the gene content harbored within individual bees represent a minor fraction of the total gene content present across hosts. Moreover, the number of sequence clusters increased more rapidly with sample size for *A. mellifera* as compared to *A. cerana*, and did not appear to have reached saturation with the current sampling size. This difference was not related to diversity between colonies or countries for *A. mellifera* ([Fig. S7](#)). Rather, the gene content of the gut microbiota is highly variable among bees, also within colonies, and much more so for *A. mellifera* compared to *A. cerana*. Taken together with the consistent taxonomic profile among bees ([Fig. 1D-F and Fig.S5](#)), these results indicate that a major fraction of the variation in gene content is related to strain-level diversity.

To quantify the extent of strain-level diversity, the total fraction of single nucleotide variants (SNVs) within core genes was determined for all SDPs. At the level of individual bees, only two of the SDPs colonizing *A. cerana* were found to harbor more than 2% SNVs in any of the samples (“Firm5-7”, “Bifido-1”) ([Fig. 2C](#)). In contrast, nearly all the SDPs colonizing *A. mellifera* had more than 2% SNVs per bee in a major fraction of the samples ([Fig. 2C](#)). For example, “Bifido-1”, an SDP shared
between both bee species, had on average 9.6% SNVs per individual bee in *A. mellifera*, while in *A. cerana* the average percentage SNVs was as low as 0.8%.

To quantify diversity at the colony level, we again generated cumulative curves of SNVs as a function of the number of analyzed bees (**Fig. 2D-E, Fig. S8**). Interestingly, the two SDPs occurring in both hosts, “Bifido-1” and “Firm4-1” (**Fig. 1G** and **H**), displayed a clear difference in strain-level diversity between the host species (**Fig. 2D-E**). This difference was not explained by the choice of reference genome, since the pattern persisted after swapping the reference genome with an isolate from the alternate host (**Fig. S8**).

Based on these results, we conclude that strain-level diversity in *A. mellifera* is substantially higher than in *A. cerana*, both in individual bees and within colonies.

**The gut microbiota of *A. mellifera* encodes more diverse enzymes for polysaccharide degradation.**

The observed differences in SDP composition and strain-level diversity raise the question whether the gut microbiota is functionally distinct between the two hosts. To address this question, we annotated metagenomic ORFs using the COG and CAzyme databases. Furthermore, the analysis was restricted to the Japanese metagenomes (20 samples per host), using metagenome assemblies with 20 million host-filtered paired-end reads per sample, in order to facilitate quantitative comparisons.
Although the number of ORFs per sample was approximately twice as high for *A. mellifera* compared to *A. cerana*, the relative COG profiles were indistinguishable among hosts (Fig. 3A). Consistent with previous studies (20), carbohydrate metabolism and transport” (COG category “G”) was abundant across the metagenomic samples of both host species. In order to identify possible differences within this important functional category, metagenomic ORFs encoding glycoside hydrolases (GHs) and polysaccharide lyases (PLs) were annotated with dbCAN2 (21), and the number of annotations per family were counted for each sample. Notably, when calculating the mean number of annotations per family for each host species, the relative abundance of each GH/PL family was highly similar between the hosts, as evidenced by the linear correlation between the mean counts per family (Fig. 3B). However, samples from *A. mellifera* harbored approximately twice as many genes per family, as evidenced from the slope of the correlation (Fig. 3B).

Although the annotation of GH/PL families is based on sequence homology, a wide range of specificities have been reported within families (22). Therefore, to estimate the diversity among the genes within GH/PL families, they were clustered separately for each family, using a highly conservative threshold (50% amino acid sequence identity). Even at this threshold, the clustering resulted in an average of 10.4 clusters per family, indicating that a major fraction of the genes annotated to the same family is likely to have distinct functions or substrate specificities. A total of 52 out of 62 GH/PL families contained host-specific clusters (Fig. 3C). However, only 17 families contained clusters specific to *A. cerana*, whereas all 52 contained clusters specific to *A. mellifera* (Fig. 3C). Moreover, the mean number of host-specific clusters per family was 6.0 for *A. mellifera*, but only 3.1 for *A. cerana*. Taken together, these results
therefore indicate that A. mellifera has a much larger host-specific repertoire of GH/PL families as compared to A. cerana.

In order to gain further insights into the origin of the GH/PL families, all sequences were blasted against the honey bee gut microbiota database. Overall, 97% of the sequences had significant hits to the database (e-value < 10e-05, > 80% query coverage), with 79% having a close hit (>95% amino acid identity). Among the sequences with close hits, the vast majority of hits were to genomes of the “Bifido-1” SDP, followed by other SDPs of the Lactobacillus Firm4 and Firm5 phylotypes (Fig. 3D, upper panel). For the host-specific GH/PL clusters, only 52% of the sequences falling within A. mellifera-specific clusters had a close hit to the database, whereas 67% of the A. cerana-specific clusters had close hits. Thus, although the current database contains fewer genome isolates from A. cerana compared to A. mellifera, it is more representative of A. cerana in terms of GH/PL families, likely as a consequence of GH/PL genes being at least partly associated with strain-level diversity. Among the sequences corresponding to host-specific GH/PL clusters, the majority of blast hits were once again to “Bifido-1” (Fig. 3D, lower panel). Strikingly, they all originated from A. mellifera metagenomes, suggesting that GH/PL families encoded by “Bifido-1” in A. cerana represent a subset of those present in A. mellifera. A similar pattern was observed for "Firm4-1", the other SDP shared between the two host species. While GH/PL family genes matching "Firm-4" sequences were found in the gut microbiota of both honey bees, almost all host-specific sequences came from A. mellifera. Instead, for the clusters specific to A. cerana, most of the hits were to Apibacter sp. and Lactobacillus “Firm5-7”, suggesting that diversity occurring at the
In conclusion, despite the similarity in the general functional profiles, *A. mellifera* harbors a much more diverse repertoire of GH/PL families, with many more host-specific GH/PL clusters compared to *A. cerana*. Moreover, the majority of GH/PL sequences were associated with the “Bifido-1” SDP, which is much more diverse in *A. mellifera* compared to *A. cerana*, suggesting that strain-level diversity in *A. mellifera* is a major contributor to functional differences between the two host species.

* *A. mellifera* and *A. cerana* differ in bacterial community size within individual bees.*

According to neutral theory, diversity is expected to correlate with habitat size and population size. Therefore, we sought to determine whether *A. mellifera* and *A. cerana* differ in terms of the spatial niche they provide for their gut bacterial communities. Based on wet-weight, the hindgut, where most of the bacteria reside, was not significantly different between the hosts (Fig. 4A). However, the bacterial community size, as estimated from quantitative real-time PCR with universal 16S rRNA primers (normalized to the copy number of the host gene actin) was still found to be significantly larger for *A. mellifera* compared to *A. cerana* (p < 0.001, Mann-Whitney U test) (Fig. 4B).
Discussion

In the current study, we carried out a community-wide metagenomic characterization of the gut microbiota of two closely related honey bee species, A. mellifera and A. cerana. From this analysis, three key results emerged. Firstly, we found that the gut bacterial communities of the two host species were highly divergent, consisting of different SDPs and strains, despite having a very similar phylotype-level composition. Secondly, the two host species displayed major differences in the magnitude of strain-level diversity within their bacterial communities. And thirdly, the gut microbiota of A. mellifera harbored a much larger repertoire of enzymes related to polysaccharide breakdown. Thus, in the time since their last common ancestor, approximately 6 million years ago (14, 17), the gut bacterial communities of A. mellifera and A. cerana have undergone substantial changes in composition, genomic diversity, and functionality, with likely consequences for the interaction with their hosts.

Based on amplicon sequencing of the 16S rRNA gene, multiple studies have shown that gut microbiota composition is influenced by host phylogeny (23, 24) with the overall observation that closely related host species harbor more similar gut bacterial communities at the phylotype level than more distant ones (25). However, the slow evolutionary rate of the 16S rRNA gene does not permit evolutionary analysis of phylotypes that are shared across related hosts. Therefore, there is currently little data providing insights into the evolution of the gut microbiota for closely related animal hosts. Targeting the fast-evolving gyrA gene for three bacterial families colonizing hominids, Moeller et al. found evidence of co-diversification in the Bacteroidaceae and Bifidobacteriaceae, but not the Lachnospiraceae (26). Similarly,
for honey bees and bumble bees, amplicon-sequencing of the minD gene uncovered both host-specific and more generalist clades within the core phylotype Snodgrassella (27). Consistently, comparative genome analyses of bacterial isolates have also uncovered several examples of host-specific lineages (9-12). However, the current study is the first to report community-wide patterns of host-specialization using metagenomic data.

Interestingly, while we found evidence of host-specialization for all of the five core phylotypes colonizing both hosts, the extent of divergence differed widely among them. For three of the phylotypes (Lactobacillus Firm5, Gilliamella, and Snodgrassella), each host was found to be colonized by different SDPs, i.e. bacterial lineages which are sufficiently divergent to be classified as different species (13). In contrast, the host-specialization of the Bifidobacterium and Lactobacillus Firm4 phylotypes only became evident when analyzed with strain-level resolution, consistent with the comparatively short branch lengths observed for the core genome phylogenies. Contemplating a 16S rRNA gene divergence rate of about 1% per 50 million years (28, 29), it seems unlikely that any of the SDPs could have emerged within the 6 million years separating A. mellifera and A. cerana. More likely, the current SDP composition represents a selection of pre-existing SDPs, with secondarily evolved traits resulting in the currently observed host preference. For example, we found that all SDPs within the Lactobacillus Firm5 phylotype contributed to highly divergent host-specific glycoside hydrolases, which could potentially allow for dietary specialization and host adaptation. In contrast, the comparatively little sequence divergence between the host-specialized strains of Lactobacillus Firm4 and the Bifidobacteria could match a time span of 6 million years, and thus be a
product of co-diversification. However, a larger genomic dataset from multiple host species will be needed to properly test this hypothesis.

Remarkably, we also found that the two host species differed widely in terms of the extent of strain-level diversity across the gut microbiota. Whether such differences are commonplace among related animals is unknown, since comparable data is not yet available. Despite a rapid increase in metagenomic studies, quantification of strain-level diversity is still technically challenging and quantitative comparisons across hosts are therefore rare. Differences in sampling, DNA extraction, and sequencing depth can have a strong impact on both composition and diversity, making cross-comparisons between samples and studies particularly difficult (30). For example, for the human gut microbiota, it was estimated that a cumulative genome coverage of at least 1000x would be required to obtain a representative sampling of the strain-level diversity (31), a number which is still well beyond the current norm. Even in simple communities, like the bacterial endosymbionts colonizing deep-sea mussels, quantification of strains can be sensitive to sequencing depth (4). In the current study, technical biases were limited by using a common sampling and DNA extraction protocol, and a comparable sequencing depth. Given the simple taxonomic composition of the honey bee gut microbiota, we were able to obtain deep sequencing of the individual community members, with comparable mapping and de novo metagenome assembly efficiencies in both hosts. Furthermore, normalization by sampling and sequencing depth was applied in all analyses. Taken together, we are therefore confident that the observed community-wide differences in strain-level diversity are not due to technical biases.
Several factors could potentially have resulted in the observed difference in strain-level diversity, which are not mutually exclusive (Fig. 4C). Firstly, although *A. cerana* has long been used for honey production in Asia, *A. mellifera* differs from all other extant honey bee species by having been extensively transported around the globe by humans, for hundreds of years (19). Thus, it is possible that humans have contributed to a mixing of locally adapted strains, thereby increasing diversity within colonies (Fig. 4C, “Human influence”). As of yet, large-scale studies on the distribution of strains are still lacking for honey bees, and previous studies have reported mixed results (8, 27). Likewise, in the current study, ordination plots based on shared SNVs clustered by country for only a subset of SDPs. Future studies on wild honey bee populations should provide further insights into this question.

Interestingly, if the high strain-level diversity in *A. mellifera* is caused by human interference, it raises the possibility that the gut microbiota of *A. mellifera* is sub-optimal for local conditions, potentially resulting in reduced colony fitness.

However, based on molecular data, *A. mellifera* has had a large and varied geographic range long before human interference (14, 15). Moreover, despite the occurrence of colony collapse disorder in managed colonies around the world (32-34), *A. mellifera* has been found to successfully establish feral colonies even outside its native range (i.e. the New World), indicative of a remarkable ability to survive under highly varied conditions (35). Thus, it is possible that *A. mellifera* is more of a generalist than *A. cerana*, for example by having a wider foraging range, and therefore is able to maintain a more varied bacterial community and larger repertoire of sugar breakdown functions (Fig. 4C, “Dietary specialization”). Indeed, diet has been shown to have an impact on diversity in the gut microbiota in multiple studies,
with for example a reduction of diversity reported for the human gut microbiota as a consequence of westernized diet (36, 37). As of yet, large-scale systematic studies on the foraging preferences of A. cerana and A. mellifera have not been conducted, but they appear to have largely overlapping foraging ranges in Asia (38, 39). Further studies are therefore needed to determine whether differences in strain-level diversity are related to dietary differences.

Finally, it is also possible that the difference in strain-level diversity between A. mellifera and A. cerana is driven by neutral processes (Fig. 4C, “Species-area relationship”). Specifically, the “species-area relationship”, posit a positive correlation between habitat size and diversity, and is widely held to constitute one of the few laws in ecology (40-42). However, it is still debated whether this relationship also applies to bacteria (43-45). The gut microbiota represents an attractive model system for testing the hypothesis, since bacterial populations in this case can be easily delineated, due to the host-association. Interestingly, A. mellifera and A. cerana exhibit spatial differences at multiple levels (Fig. 4C, “Species-area relationship”). Firstly, we found that the size of the bacterial communities within individuals was significantly larger for A. mellifera compared to A. cerana, consistently with the findings of a previous study (8). Secondly, A. mellifera is known to form larger colonies than A. cerana (16). And thirdly, the native range of A. mellifera is also larger compared to A. cerana (19, 46, 47). Our analysis clearly shows that the diversity within individual bees is much lower than the diversity found within their colonies. Thus, it seems plausible that competition among strains is alleviated when strains colonize different hosts. If so, it also follows that larger colonies should be able to support more strain-level diversity, by providing more colonization
opportunities for strains that would otherwise compete. However, given that the
diversity was also consistently higher for individual bees of *A. mellifera* as compared
to *A. cerana*, the species-area relationship may be equally applicable at this level. In
contrast, the inclusion of multiple colonies had very little impact on diversity,
suggesting that host geographic range is of less importance for maintenance of
strain-level diversity in honey bees.

In conclusion, the results of the current study brought several fundamental questions
regarding the evolution and maintenance of diversity in host-associated bacterial
communities to the foreground. While the term “diversity” has an inherently positive
connotation, it is not obvious whether diversity in host-associated bacterial
communities should be beneficial, and if so, in what sense (48). For example, high
strain-level diversity in the gut microbiota of *A. mellifera* may provide more metabolic
flexibility, facilitating foraging on more diverse pollen sources, and thereby faster
adaptation to changing environmental conditions. On the other hand, high strain-level
diversity could also lead to increased competition within the gut microbiota, with
resources being diverted towards inter-bacterial warfare, rather than host-symbiont
mutualistic interactions. It is also possible that a less diverse gut microbiota
consisting of strains adapted to local conditions would be more beneficial than a
more diverse one. These possibilities can be experimentally tested in honey bees,
thereby providing novel insights into the functional relevance of strain-level diversity
in host-associated bacterial communities and for honey bee health.
Methods

Detailed protocols are available in the Supplementary Information (Supplementary Materials and Methods). All bioinformatic tasks were accomplished with custom scripts (perl, R, Bash) unless otherwise indicated. Metagenomic samples were collected from individual honey bees, with 20 samples for each host species (A. mellifera, A. cerana), each represented by two colonies. All colonies were located in different apiaries, no more than 100km apart, close to Tsukuba, Japan. DNA extraction was done as described previously (5), using a custom protocol for enrichment of bacterial cells. Sequencing was done with the Nextera XT library kit, on an Illumina Hiseq 2500 instrument. A honey bee gut microbiota database was generated for the current study, including previously published genomes and novel genomes isolated from A. cerana, in order to be representative of both targeted communities. Candidate sequence-discrete populations (SDPs) were inferred from isolate genomes in the database (based on core genome phylogenies and average nucleotide identities), and subsequently validated by recruitment of metagenomic ORFs to core gene alignments. Quantification of SDPs within and across samples was done as described previously (5), based on mapped read coverage to single-copy core gene families. Profiling of single-nucleotide variants (SNVs) was done with freebayes (49), using a reduced version of the database, with a single representative genome per SDP, and a strict filtering pipeline. Metagenome assemblies were generated from host-filtered reads, based complete datasets and subsets of 20 million paired-end reads per sample, using SPAdes (50). Annotation of metagenomic ORFs was done with the eggnog (51) and dbcan2 databases (21). To estimate bacterial loads and wet-weights of honey bee hindguts, new samples were collected, from which total DNA was extracted with a CTAB-based protocol. Bacterial load
quantification was done with qPCR, targeting the V3-V4 region of the 16S rRNA gene, and normalized based on the actin gene of the host.

Data Availability

The metagenomic data has been deposited under Bioproject PRJNA59809.
Acknowledgments

We would like to thank Yoshiko Sakamoto, Shohei Kanari, Nao Tsuyuki, Yoshihiro Saito, Hiroki Kohno, and Takeo Kubo for their help during the sampling of honey bees. We also thank Kohei Fukuda for the technical assistance of molecular work.

P.E. is supported by the European Research Council ERC-StG ‘MicroBeeOme’ (714804) and the Swiss National Science Foundation SNFS project grant 31003A_179487. R.M. and P.E. are supported by the Human Frontier Science Program HFSP Young Investigator grant RGY0077/2016.
**Figures**

**Figure 1.** The gut microbiota of *A. mellifera* and *A. cerana* are composed of divergent SDPs and strains. (A-C, G-H) Core genome phylogenies of the five shared core phylotypes colonizing *A. mellifera* and *A. cerana*. Confirmed SDPs are indicated by the labels of the clades. Genomes are highlighted with blue shades for isolates from *A. mellifera*, and red shades for isolates from *A. cerana*. Grey shades indicate isolates from other honey bee species. Bars correspond to 0.1 substitutions per site (D-F) Barplots displaying relative abundance of the confirmed SDPs shown in panels A-C, across metagenomes from Japan (see Fig. S5 for the same plots including metagenome samples from Switzerland). (I-J) PCoA ordination plots based
on the pairwise fractions of shared SNVs (jaccard distance) (see Dataset S3 for plots of other SDPs). Dots represent individual samples, color-coded by host and colony origin as indicated by the legend. (K) Number of host-specific and mixed sequence clusters generated from metagenomic ORFs at different clustering thresholds for the complete dataset (Swiss and Japanese samples).
Figure 2. Major differences in strain-level diversity in the gut microbiota of *A. mellifera* and *A. cerana*. (A) Number of sequence clusters per sample for each host, based on metagenomic ORFs from assemblies generated with 20 million paired-end host-filtered reads. (B) Cumulative number of sequence clusters for each host, relative to the number of bee samples. (C) Fraction of single nucleotide variants (SNVs) within core genes in each sample, for SDPs corresponding to core phylotypes, plus *Apibacter* sp.. SDP labels correspond to Fig. 1. (D-E) Cumulative fractions of SNVs within core genes relative to the number of samples, for the “Bifido-1” and “Firm4-1” SDPs. Blue-green shades represent different *A. mellifera* colonies, whereas the *A. cerana* samples were pooled due to the smaller number of samples. See Fig. S8 for curves corresponding to other SDPs and mappings to alternative reference genomes.
Figure 3. Functional comparison of the gut microbiota of *A. mellifera* and *A. cerana*. (A) Relative abundance of COG annotations according to general functional COG categories. (B) Mean number of ORFs assigned to each CAzyme family, calculated separately for each host. As shown by the blue regression line, there was a linear correlation between the counts, indicating that the CAzyme families display similar relative abundance patterns in both hosts. But, the total number of ORFs assigned to each CAzyme family is larger in *A. mellifera* samples as compared to *A. cerana* samples as indicated by the deviation from the black line. (C) Number of sequence clusters within each CAzyme family (clustered at 50% amino acid sequence identity), with colors indicating the subsets of clusters specific to each host, and clusters containing ORFs derived from both hosts. (D) SDP affiliation of blast hits for all ORF sequencess annotated as GH/PL CAzyme families, in the honey bee gut microbiota database. Results are only shown for close hits (>95% amino acid sequence identity). The upper panel includes all ORFs passing this threshold, the lower panel includes the subset of these occurring within host-specific clusters. All
analyses were carried out on only the Japanese metagenome samples, to facilitate quantitative comparisons (20 samples per host, with assemblies based on 20 million paired-end host-filtered reads per sample). SDP labels correspond to Fig. 1. Species abbreviations: Ap - *Apibacter* sp., Fp - *Frischella perrara*, Co: *Commensalibacter* sp., Lk - *Lactobacillus kunkeei*.
Figure 4. Possible factors explaining the difference in gut microbiota diversity between *A. mellifera* and *A. cerana*. (A and B) Wet-weight of the hindgut (where most of the bacteria reside) and qPCR results for estimation of bacterial community size (targeting the 16S rRNA gene, and normalized by copy-number of the host gene actin). Statistical significance was calculated using a Mann-Whitney *U* test (ns - not significant, ***p<0.001). (C) Schematic illustration of three possible factors explaining differences in diversity. “Human influence”: transportation and mixing of *A. mellifera* colonies and genotypes around the world by beekeepers, resulting in mixing of strains from different geographic origins, and thereby increasing strain-level diversity in *A. mellifera*. “Dietary Specialization”: *A. mellifera* may have a more generalist diet.
(here illustrated by pollen grain diversity) as compared to *A. cerana*, and thereby be able to sustain a more diverse community. “Species-Area relationship”: Although previously applied to species-level diversity in animals, this concept may also apply to strain-level diversity in bacteria. For the honey bee gut microbiota, spatial differences are applicable at three levels: The size of the bacterial community within individual bees, the size of honey bee colonies, and the size of the geographic range.
References

1. Kashtan N, et al. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* 344(6182):416-420.

2. Cordero OX & Polz MF (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* 12(4):263-273.

3. Mitri S & Foster KR (2013) The genotypic view of social interactions in microbial communities. *Annu Rev Genet* 47:247-273.

4. Ansorge R, et al. (2019) Functional diversity enables multiple symbiont strains to coexist in deep-sea mussels. *Nat Microbiol* 4(12):2487-2497.

5. Ellegaard KM & Engel P (2019) Genomic diversity landscape of the honey bee gut microbiota. *Nat Commun* 10(1):446.

6. Zhu A, Sunagawa S, Mende DR, & Bork P (2015) Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol* 16:82.

7. Greenblum S, Carr R, & Borenstein E (2015) Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160(4):583-594.

8. Kwong WK, et al. (2017) Dynamic microbiome evolution in social bees. *Sci Adv* 3(3):e1600513.

9. Steele MI, Kwong WK, Whiteley M, & Moran NA (2017) Diversification of Type VI Secretion System Toxins Reveals Ancient Antagonism among Bee Gut Microbes. *MBio* 8(6).

10. Kwong WK, Engel P, Koch H, & Moran NA (2014) Genomics and host specialization of honey bee and bumble bee gut symbionts. *Proc Natl Acad Sci U S A* 111(31):11509-11514.

11. Zheng H, et al. (2016) Metabolism of Toxic Sugars by Strains of the Bee Gut Symbiont Gilliamella apicola. *MBio* 7(6).
12. Ellegaard KM, et al. (2015) Extensive intra-phylotype diversity in lactobacilli and bifidobacteria from the honeybee gut. *BMC Genomics* 16:284.

13. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, & Aluru S (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9(1):5114.

14. Wallberg A, et al. (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet* 46(10):1081-1088.

15. Han F, Wallberg A, & Webster MT (2012) From where did the Western honeybee (*Apis mellifera*) originate? *Ecol Evol* 2(8):1949-1957.

16. Park D, et al. (2015) Uncovering the novel characteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC Genomics* 16:1.

17. Chen C, et al. (2018) Population genomics provide insights into the evolution and adaptation of the eastern honey bee (*Apis cerana*). *Mol Biol Evol*.

18. Nanzan Takamura YT, Shigeru Nakano, Mitsuyuki Horikiri (1966) Yoho no Seisei. *Chikusan Hattatsu Shi*, ed Livestock Industry Bureau MoAaF (Chuo Kouron Jigyo, Tokyo), pp 1313-1334.

19. Requier F, et al. (2019) The Conservation of Native Honey Bees Is Crucial. *Trends Ecol Evol* 34(9):789-798.

20. Engel P, Martinson VG, & Moran NA (2012) Functional diversity within the simple gut microbiota of the honey bee. *Proc Natl Acad Sci U S A* 109(27):11002-11007.

21. Zhang H, et al. (2018) dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 46(W1):W95-W101.
22. Helbert W, et al. (2019) Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space. Proc Natl Acad Sci U S A 116(13):6063-6068.

23. Nishida AH & Ochman H (2018) Rates of gut microbiome divergence in mammals. Mol Ecol 27(8):1884-1897.

24. Ley RE, et al. (2008) Evolution of mammals and their gut microbes. Science 320(5883):1647-1651.

25. Groussin M, et al. (2017) Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat Commun 8:14319.

26. Moeller AH, et al. (2016) Cospeciation of gut microbiota with hominids. Science 353(6297):380-382.

27. Powell E, Ratnayeke N, & Moran NA (2016) Strain diversity and host specificity in a specialized gut symbiont of honeybees and bumblebees. Mol Ecol 25(18):4461-4471.

28. Ochman H, Elwyn S, & Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96(22):12638-12643.

29. Moran NA, Munson MA, Baumann P, & Ishikawa H (1993) A Molecular Clock in Endosymbiotic Bacteria Is Calibrated Using the Insect Hosts. P Roy Soc B-Biol Sci 253(1337):167-171.

30. Reese AT & Dunn RR (2018) Drivers of Microbiome Biodiversity: A Review of General Rules, Feces, and Ignorance. MBio 9(4).

31. Schloissnig S, et al. (2013) Genomic variation landscape of the human gut microbiome. Nature 493(7430):45-50.

32. Cox-Foster DL, et al. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318(5848):283-287.
33. Stokstad E (2007) Entomology. The case of the empty hives. *Science* 316(5827):970-972.

34. Oldroyd BP (2007) What's killing American honey bees? *PLoS Biol* 5(6):e168.

35. Seeley T (2019) *The Lives of Bees: The Untold Story of the Honey Bee in the Wild* (Princeton University Press, New Jersey, USA).

36. Sonnenburg ED, *et al.* (2016) Diet-induced extinctions in the gut microbiota compound over generations. *Nature* 529(7585):212-215.

37. Fragiadakis GK, *et al.* (2019) Links between environment, diet, and the hunter-gatherer microbiome. *Gut Microbes* 10(2):216-227.

38. Tatsuno M & Osawa N (2016) Flower visitation patterns of the coexisting honey bees *Apis cerana japonica* and *Apis mellifera* (Hymenoptera: Apidae). *Entomological Science* 19(3):255-267.

39. Suryanarayana M, Mohana Rao G, & Singh T (1992) Studies on pollen sources for *Apis cerana* Fabr and *Apis mellifera* L bees at Muzaffarpur, Bihar, India. *Apidologie* 23.

40. Rosenzweig M (1995) *Species diversity in space and time* (Cambridge University Press, Cambridge, UK).

41. Arrhenius O (1921) Species and Area. *Journal of Ecology* 9:95-99.

42. Gleason H (1922) On the relation between species and area. *Ecology* 3(2):158-162.

43. Logue JB, *et al.* (2012) Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *ISME J* 6(6):1127-1136.

44. Bell T, *et al.* (2005) Larger islands house more bacterial taxa. *Science* 308(5730):1884.
45. Koskella B, Hall LJ, & Metcalf CJE (2017) The microbiome beyond the horizon of ecological and evolutionary theory. *Nat Ecol Evol* 1(11):1606-1615.

46. Koetz AH (2013) Ecology, Behaviour and Control of Apis cerana with a Focus on Relevance to the Australian Incursion. *Insects* 4(4):558-592.

47. Ruttner F (1988) *Biogeography and Taxonomy of Honeybees* (Springer).

48. Shade A (2017) Diversity is the question, not the answer. *ISME J* 11(1):1-6.

49. Garrison E & Marth G (2012) Haplotype-based variant detection from short-read sequencing. *bioRxiv*.

50. Bankevich A, *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19(5):455-477.

51. Huerta-Cepas J, *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47(D1):D309-D314.