Genomic changes underlying host specialization in the bee gut symbiont *Lactobacillus* Firm5

Ellegaard KM[1], Brochet S[1], Bonilla-Rosso G[1], Emery O[1], Glover N[2], Hadadi N[1], Jaron KS[2,4], van der Meer JR[1], Robinson-Rechavi M[2,4], Sentchilo V[1], Tagini F[3], SAGE class 2016-17, Engel P[1]*

[1]Department of Fundamental Microbiology, University of Lausanne, 1015 Lausanne, Switzerland

[2]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

[3]Institute of Microbiology, Department of Laboratory Medicine, University of Lausanne & Lausanne University Hospital, Lausanne, Switzerland

[4]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

***Correspondence:**

Prof. Philipp Engel

Department of Fundamental Microbiology

University of Lausanne, CH-1015 Lausanne, Switzerland

Tel.: +41 (0)21 692 56 12

e-mail: philipp.engel@unil.ch

Abstract

Bacteria that engage in longstanding associations with particular hosts are expected to evolve host-specific adaptations that limit their capacity to thrive in other environments. Consistent with this, many gut symbionts seem to have a limited host range, based on community profiling and phylogenomics. However, few studies have experimentally investigated host specialization of gut symbionts and underlying mechanisms have largely remained elusive. Here, we studied host specialization of a dominant gut symbiont of social bees, *Lactobacillus* Firm5. We show that Firm5 strains isolated from honey bees and bumble bees separate into deep-branching phylogenetic lineages. Despite their divergent evolution, colonization experiments show that bumble bee strains are capable of colonizing the honey bee gut. However, they were less successful than honey bee strains, and competition with honey bee strains completely abolished their colonization, whereas honey bee strains were able to coexist. This suggests that both host selection and interbacterial competition play important roles for host specialization. Using comparative genomics of 27 Firm5 isolates, we identified candidate genomic changes underlying host specialization. We found that honey bee strains harbored a larger and more diverse gene pool of carbohydrate-related functions than bumble bee strains. As dietary-derived carbohydrates are the main energy source for strains of the Firm5 phylotype, the metabolic flexibility of honey bee strains may give these bacteria a competitive advantage over bumble bee strains in colonizing the gut niche and hence contribute to host specialization.

Introduction

Symbiotic relationships between bacteria and eukaryotes are pervasive and range from loose associations to obligate interdependencies (McFall-Ngai *et al.* 2013; Kostic *et al.* 2013). The evolution of a host-associated lifestyle is typically accompanied by the loss of generalist characteristics, limiting a symbionts' capacity to compete and survive in other environments. This in turn results in host specialization (Bobay & Ochman 2017; Sriswasdi *et al.* 2017). In particular, bacteria with longstanding associations are often host-specific and undergo marked genomic changes (Toft & Andersson 2010). Among the most extreme examples are primary endosymbionts of plant-sap feeding insects. These obligate mutualists reside within host cells, are vertically inherited through the germ-line, and have experienced extreme genome reduction due to population bottlenecks and genetic drift (McCutcheon & Moran 2012).

Based on phylogenetic analyses, host specialization has also been inferred for many gut symbionts, as bacterial lineages are frequently found to be exclusively associated with particular hosts (Ley *et al.* 2008; Oh *et al.* 2010; Ochman *et al.* 2010; Eren *et al.* 2015; Moeller *et al.* 2016; Kwong *et al.* 2017). This is remarkable considering that gut symbionts are horizontally transmitted, and are exposed at least at some point to the environment outside the host, which in principle provides opportunities for host switching. This leads to the question whether the realized niche of gut symbionts (the conditions where the bacteria actually live) is different from their fundamental niche (the conditions where they can live). If there is a difference, the next question is which factors limit niche realization, such as

70 competition with other bacteria, dispersion, or host selection (Macarthur & Levins
71 2015; Hutchinson 1957).
72
73 In the case of the gut symbiont *Lactobacillus reuteri*, strains isolated from mice are
74 capable of colonizing the mouse gut, whereas those from humans, pigs, or chickens
75 are not, suggesting that host association in this case has resulted in the restriction of
76 the fundamental niche (Oh *et al.* 2010; Frese *et al.* 2011). In contrast, in another
77 study it was shown that bacterial communities from diverse habitats can colonize
78 and persist in the mouse gut (despite the fact that these species naturally do not
79 occur in the mouse gut), suggesting that a species' realized niche is frequently more
80 restricted than its fundamental niche (Seedorf *et al.* 2014). A notable difference
81 between the two studies is that host specificity of *L. reuteri* was tested in mice that
82 were free of *Lactobacilli*, but otherwise harbored a conventionalized microbiota,
83 whereas in the second study, most experiments were carried out in microbiota-free
84 mice. Hence, interbacterial competition may be an important factor defining the
85 realized niche (Seedorf *et al.* 2014).
86
87 Given that both experimental and phylogenetic evidence is required to determine
88 host specialization, our understanding of host specialization is limited for most gut
89 symbionts. Moreover, little is known about the underlying mechanisms and the
90 genomic changes accompanying host-specific evolution of gut symbionts. Selective
91 forces acting on gut symbionts may differ between hosts due to varying degrees of
92 population bottlenecks during transmission, or due to differences in dietary

93 preferences, gut structure or host physiology, resulting in distinct evolutionary
94 patterns.
95
96 A good model to study host specialization of bacterial inhabitants is the gut
97 microbiota of corbiculate bees (Kwong & Moran 2015). Most species of honey bees,
98 bumble bees, and stingless bees share a specialized core gut microbiota that is
99 composed of five phylotypes (strains sharing ≥97% 16S rRNA sequence identity as
100 estimated from amplicon sequencing studies): the gammaproteobacterium
101 *Gilliamella*, the betaproteobacterium *Snodgrassella alvi*, two Lactobacilliales (Firm5
102 and Firm4), and a Bifidobacterium (Cox-Foster *et al.* 2007; Moran *et al.* 2012; Corby-
103 Harris *et al.* 2014). These phylotypes are likely to have been acquired in a last
104 common ancestor of the corbiculate bees, as they are widely distributed among
105 contemporary species of honey bees, bumble bees, and stingless bees (Kwong *et al.*
106 2017). Moreover, there is evidence for host specialization and coevolution, because
107 strains isolated from the three groups of corbiculate bees separate into divergent
108 sublineages for most phylotypes (Koch *et al.* 2013; Kwong *et al.* 2014; Ellegaard *et*
109 *al.* 2015; Zheng *et al.* 2016; Kwong *et al.* 2017; Steele *et al.* 2017).
110
111 The best-studied member of the bee gut microbiota with respect to host
112 specialization is *S. alvi* (Kwong & Moran 2015). Reciprocal mono-colonization
113 experiments of microbiota-depleted bees showed that *S. alvi* isolates from honey
114 bees (*Apis mellifera*) colonize poorly the gut of bumble bees (*Bombus impatiens*),
115 and vice versa, suggesting that the host-specific evolution of these isolates has led to
116 specialization (Kwong *et al.* 2014). Based on the comparison of three *S. alvi*

117 genomes, it was suggested that bumble bee isolates tend to have smaller genomes
118 and contain larger amounts of mobile elements than honey bee isolates. Genomic
119 differences were also identified among isolates from different host groups (honey
120 bees and bumble bees) for the phylotype *Gilliamella*: honey bee isolates encoded
121 more carbohydrate-related functions than bumble bee isolates (Kwong *et al.* 2014).
122 However, recent genome sequencing of a larger number of *Gilliamella* strains
123 revealed that some isolates from honey bees have genomes as small as those from
124 bumble bees, suggesting that a large metabolic repertoire is not necessarily needed
125 for colonization of the honey bee gut (Zheng *et al.* 2016; Ludvigsen *et al.* 2017;
126 Steele *et al.* 2017).
127
128 For the other phylotypes of the bee gut microbiota, little is known about the link
129 between phylogeny, host range, and genome features (Ellegaard *et al.* 2015; Kwong
130 *et al.* 2017). One of the most widely distributed and abundant phylotypes of the bee
131 gut microbiota is *Lactobacillus* Firm5. In the gut of honey bees (*Apis mellifera*), four
132 deep-branching monophyletic sublineages have been identified for this phylotype,
133 and different species names have been proposed (Olofsson *et al.* 2014). Strains
134 belonging to these four sublineages typically co-occur in individual bees *(Ellegaard*
135 *and Engel, in revision)*, and can vary by up to 40% in gene content (Ellegaard *et al.*
136 2015), suggesting that these sublineages may have adapted to distinct metabolic or
137 spatial niches within the honey bee gut. Interestingly, Firm5 strains isolated from
138 other corbiculate bees seem to belong to different sublineages than the honey bee
139 isolates, as indicated by single gene phylogenies (Kwong *et al.* 2017). Moreover, a
140 divergent Firm5 strains from bumble bees has been isolated and described as a new

141 species, *Lactobacillus bombicola* (Praet *et al.* 2015). Given that Firm5 strains can be
142 cultured, and that the honey bee is amenable to experimental colonization, this
143 phylotype represents an excellent opportunity to study evolutionary trajectories of
144 host adaptation and the consequences for the fundamental and realized niche of this
145 gut symbiont.
146
147 Here, we used gnotobiotic bee experiments, genome sequencing, and comparative
148 genomics to address host specialization in the Firm5 phylotype. First, we show that
149 isolates from honey bees and bumble bees belong to distinct sublineages suggesting
150 longstanding host-specific associations. Second, we provide experimental evidence
151 that both host selection and interbacterial competition contribute to host
152 specialization. Third, our comparative genome analysis reveals marked differences
153 in carbohydrate utilization capacities between honey bee and bumble bee isolates
154 suggesting that metabolic flexibility gives honey bee isolates a competitive
155 advantage over bumble bee isolates to establish in the honey bee gut.

Results

Bumble bee and honey bee isolates belong to separate sublineages of the Firm5 phylotype.

We sequenced the genomes of 15 new isolates of the Firm5 phylotype. Five isolates were obtained from the honey bee *Apis mellifera* and ten isolates from three different bumble bee species (five from *Bombus pascuorum*, four from *Bombus bohemicus*, and one from *Bombus terrestris*). All bees were collected in Western Switzerland (**Table S1**). We also included 12 previously sequenced isolates (one from a bumble bee, the others from honey bees) to be more comprehensive in our analyses. All 27 isolates shared >95% sequence identity across the full-length 16S rRNA gene (**Table S2**), suggesting that they all belong to the Firm5 phylotype. Isolates from conspecific hosts tended to have higher 16S rRNA sequence identities. The draft genomes of the 15 newly sequenced isolates consisted of 11-24 contigs with total lengths of 1.63-2.11 Mb, which is in the range of the previously sequenced Firm5 strains (**Table S1**). While the genomes of the bumble bee isolates tended to be smaller (1.63-1.70 Mb) than those of the honey bee isolates (1.68-2.15 Mb), genome synteny was largely conserved across the entire Firm5 phylotype (**Figure S1 and S2**).

To assess the evolutionary relationship between the 27 sequenced Firm5 strains, we inferred a genome-wide phylogeny (including 15 close and three more distant outgroup strains, see methods) (**Figure 1**), and calculated pairwise average nucleotide identities (ANI) (**Figure S3, Table S3**). This analysis showed that the

179 Firm5 strains fall into six monophyletic sublineages with >91% ANI for within-
180 lineage divergence and <86% ANI for between-lineage divergence. Four of these six
181 sublineages consisted of only honey bee isolates and corresponded to the previously
182 identified Firm5 sublineages (Ellegaard *et al.* 2015). The two other sublineages
183 consisted of only bumble bee isolates and formed a monophyletic clade within
184 Firm5 (**Figure 1**). One sublineage comprised isolates from three different bumble
185 bee species (*B. lapidarus, B. terrestris, B. pascuorum*) including the previous isolate
186 described as species *L. bombicola* (Praet *et al.* 2015). The other sublineage
187 comprised exclusively isolates from *B. bohemicus*. Based on its deep divergence
188 from the other sublineages (ANI <80%, **Figure S3**), this second sublineage of
189 bumble bee isolates is likely to represent a novel species, for which we propose the
190 *Candidatus* name '*Lactobacillus bohemicus*'.
191
192 Out of the 27 Firm5 isolates included in the current study, five isolates (ESL0262,
193 ESL0234, ESL0236, ESL0245, ESL0247) from three different sublineages were
194 identical or almost identical to other isolates (ANI >99.99%, **Table S3**). In all cases,
195 the nearly identical isolates were obtained from the same individual. Hence, they
196 were excluded from all subsequent analyses to avoid biases due to repeated
197 sampling of the same genotype.
198 In summary, our phylogenetic analysis of the Firm5 phylotype revealed a pattern
199 suggesting host specialization, because strains of each of the six deep-branching
200 sublineages were exclusively associated with either honey bees or bumble bees.
201

Bumble bee strains can colonize microbiota-depleted honey bees, but are outcompeted by honey bee strains.

To test if the differences in the realized niche of Firm5 isolates are due to host specialization, we experimentally tested the ability of bumble bee strains to colonize the honey bee gut. In the absence of competitors (i.e. when microbiota-depleted bees were mono-colonized), we found that all bumble bee strains were able to colonize the honey bee hindgut (**Figure 2A**). The number of recovered bacterial cells at day 5 post colonization (10^6 – 10^8 CFUs per gut) was substantially higher than in the inoculum (**Figure S5A**) indicating active growth of the bumble bee strains in the honey bee gut. Thus, we can conclude that the fundamental niche of the bumble bee strains includes the honey bee gut. However, the colonization levels were slightly lower compared to mono-colonizations with honey bee strains, which reached 10^8 – 10^9 CFUs per gut (**Figure 2A**). Moreover, the percentage of successfully colonized bees varied between strains and was overall also lower for bumble bee (10-80%) than for honey bee strains (80-100%).

To test if the colonization success depends on the number of cells in the inoculum, we colonized microbiota-depleted honey bees with different inocula of the bumble bee Firm5 strain ESL0228 (**Figure S5B**). With the lowest inoculum, no colonization was obtained at day 5 post colonization (n=10), while with the highest inoculum, all bees were colonized, yielding again between 10^6 – 10^8 CFUs per gut (**Figure 2B**). The relatively high number of bacteria that was needed to achieve a robust colonization suggests that stronger selection is at play on bumble bee than on honey bee Firm5 strains for gut colonization.

226

227 To test for the effect of competitive exclusion between strains, we co-colonized
228 microbiota-depleted bees with the bumble bee Firm5 strain ESL0228 and a mix of
229 four honey bee strains (ESL0183, ESL0184, ESL0185, and ESL0186), each from one
230 of the four divergent sublineages. We kept the number of bacteria in the inoculum
231 constant for the honey bee strains (1:1:1:1), but provided the bumble bee strain at
232 ratios of 1:1, 10:1, or 100:1 relative to the honey bee strains (**Figure S5B**). All bees
233 in the experiment (n=30, n=10 per treatment) were successfully colonized by the
234 Firm5 phylotype and the total numbers of CFUs per gut were in the same range as
235 for the mono-colonizations (10^8 – 10^9 CFUs). We used amplicon sequencing of a
236 short fragment of a conserved housekeeping gene that allowed us to determine the
237 relative abundance of the five Firm5 strains tested in the community (see Methods).
238 This analysis revealed that overall all four honey bee strains successfully colonized
239 and coexisted in the gut, except for strain ESL0184 which was absent from a few
240 samples (**Figure 2C**). In contrast, the bumble bee strain ESL0228 was detected in
241 only a few bees and at very low relative abundance (<0.1%), even when inoculated
242 with a ratio of 100:1.

243

244 Collectively, these experiments show that bumble bee strains of the Firm5
245 phylotype are capable of colonizing the honey bee gut, but consistent colonization
246 can only be achieved with a relatively high inoculum and when honey bee strains
247 are absent. Therefore, we conclude that both host selection and interbacterial
248 competition contribute to the restriction of the realized niche of bumble bee strains.

249

Firm5 strains harbor a large gene pool of phylotype-specific functions of which few are conserved.

In order to identify genomic characteristics that may contribute to host specialization among Firm5 strains, we carried out a detailed comparative genome analysis. We first determined the distribution of the entire pan genome across the analyzed Firm5 strains. We included 15 divergent outgroup strains in this analysis (i.e. strains not belonging to the Firm5 phylotype, see **Figure 1** and methods) to identify Firm5-specific gene families that could play a role in adaptation to the bee gut environment. In total, 8,248 pan genome gene families were identified across the 37 genomes (i.e. gene families present in at least one genome), of which 2,131 gene families were specific to the Firm5 strains. Of those, 571 and 1,222 gene families were only found in bumble bee and honey bee strains, respectively, and 338 gene families were shared across the two hosts (**Figure 3A**).

Despite this relatively large gene pool of Firm5-specific functions, few gene families were conserved (5.6% of the shared gene families and 1.4% and 1.6% of the host-specific gene families, **Figure 3A**). Among the gene families conserved across all Firm5 strains, we found an ABC transporter system for branched chain amino acids and two putative adhesin genes (DUF4097). Amino acid transporter genes were also present among the few conserved gene families specific to the honey bee strains, while those specific to the bumble bee strains were almost all annotated as either hypothetical proteins or transcriptional regulators, providing few clues about their functional roles for host adaptation (**Dataset S1**). Altogether, this analysis revealed

very few phylotype- or host-specific gene functions as potential candidates for general determinants of host adaptation across the analyzed Firm5 strains.

Firm5-specific gene content is restricted to sublineages

As strains from the same host can belong to divergent sublineages, it is possible that sublineage-specific gene functions are involved in host specialization, e.g. by adaptation to different metabolic niches within the gut. Such genes could also explain the ability of the four honey bee sublineages of Firm5 to coexist in bees *(Ellegaard and Engel, in review)*.

Indeed, we found that a relatively large number of the Firm5-specific pan gene content is confined to strains of the same sublineage (840 of 1,222 for honey bee strains, 532 of 571 for bumble bee strains). However, as for the previous analysis, relatively few of these sublineage-specific gene families were conserved, i.e. present in all strains of a given sublineage (**Figure 3B**). In the honey bee sublineage Firm5-2 (*L. helsingborgensis*) 53 gene families were conserved (i.e. 34% of the sublineage-specific gene content), including several sugar transporter genes and a genomic island for the breakdown of rhamnogalacturonan, a major polysaccharide of pectin (**Dataset S1**). This genomic island was also found in a recent metagenomic study to correlate in abundance with core genes of this sublineage *(Ellegard and Engel, in review)*, suggesting that rhamnogalacturonan utilization is a conserved function of strains belonging to Firm5-2 (*L. helsingborgensis*). For the other three honey bee sublineages, only 9-20 gene families (0.3%-1.1%) were conserved (**Figure 3B**). In sublineage Firm5-3 (*L. melliventris*), functions for rhamnose utilization were found

297 to be conserved, whereas the annotations of the gene families in the other two
298 sublineages provided little functional insights (**Dataset S1**). The same was the case
299 for the conserved gene families found in the two bumble bee sublineages (9 and 59
300 gene families, **Figure 3B**), of which most were annotated as hypothetical proteins
301 (**Dataset S1**). Overall, these results indicate a high degree of gene content variability
302 among strains from the same host and from the same sub-lineage.
303
304 **Honey bee strains harbor a larger diversity of carbohydrate-related functions**
305 **than bumble bee strains.**
306 The high degree of gene content plasticity within the Firm5 phylotype prompted us
307 to look at the functional composition of the entire Firm5-specific gene pool. This
308 analysis revealed marked differences between honey bee and bumble bee strains
309 with respect to carbohydrate-related functions. While 'Carbohydrate transport and
310 metabolism' was by far the most dominant COG category (COG 'G') among the gene
311 families specific to the honey bee strains (184 gene families, 50% of those with COG
312 annotation), this category was nearly absent among the gene families specific to the
313 bumble bee strains (4 gene families, 10% of those with COG annotation) (**Figure**
314 **3C**). In fact, most gene families specific to bumble bee strains had no COG annotation
315 at all. Analysis of the sublineage-specific gene content revealed a similar pattern. For
316 three of the four honey bee sublineages, 'Carbohydrate transport and metabolism'
317 was by far the most abundant COG category among the sublineage-specific gene
318 content, while for the two bumble bee sublineages this category was much less
319 prominent (**Figure 3D**).
320

This trend was confirmed by analyzing the overall abundance of carbohydrate-related functions per Firm5 strain. Both relative and total number of genes assigned to COG category 'G' was higher for most honey bee strains compared to bumble bee strains or outgroup strains (**Figure 4A, Figure S6)**. However, the Firm5-1 sublineage represented an exception to this pattern. All three strains of this sublineage encoded fewer COG category 'G' genes than other honey bee strains. A large proportion of the genes assigned to COG category 'G' encoded phosphotransferase systems (PTSs), i.e. transporters involved in sugar. Correspondingly, these gene families showed a similar distribution as the COG category 'G' genes across the Firm5 strains, with most honey bee strains harboring a much larger number of PTS genes than bumble bee strains (**Figure 4B)**.

PTS transport systems are often genetically linked to glycoside hydrolases (GHs), which mediate the cleavage of sugar residues from polysaccharides or other glycosylated compounds. To assess if bee gut bacteria harbor a specific arsenal of these sugar-cleaving enzymes, we identified all GH genes in the analyzed genomes. As for COG category 'G' and PTS transporters, we found a larger number of GH genes for honey bee strains compared to bumble bee strains (**Figure 4C, Dataset S3**). Some honey bee strains harbored twice as many GH genes than bumble bee strains. However, there was remarkable variation in the number of GH genes among the honey bee strains, both within and across sublineages. Specifically, all strains of sublineages Firm5-3 and Firm5-4 harbored a relatively high number of GH genes, while strains of sublineage Firm5-1 varied substantially in the number of GH genes, and those of sublineage Firm5-2 were consistently low.

345 The identified GH genes belonged to 79 different gene families (**Figure 4D, Dataset**
346 **S3**), of which 43 were specific to the Firm5 phylotype. Most of these (67%) were
347 only detected among honey bee strains, 19% were shared, and only 14% were
348 specific to bumble bee strains. Moreover, honey bee strains also shared more GH
349 gene families with the outgroup strains than bumble bee strains (18 vs 1 gene
350 families).
351 While the substrate specificity of GH gene families cannot be unambiguously
352 inferred from sequence data, many of the Firm5-specific GH gene families included
353 glucosidases, fucosidases, mannosidases, xylosidases, and arabinofuranosidases
354 (e.g. GH29, GH38, GH39, GH43, and GH51), as based on the CAZY (Carbohydrate-
355 Active enZYmes) database classification (**Figure 4E).** A similarity search against the
356 publicly available non-redundant database NCBI nr (NCBI Resource Coordinators
357 2018) revealed that many of the Firm5-specific GH families, especially those
358 exclusively present among the honey bee strains, have best hits to other taxonomic
359 groups than lactobacilliales (**Figure S7**). While these gene families may have been
360 acquired by horizontal gene transfer or secondarily lost in other lactobacillus, their
361 limited distribution among lactobacilliales suggests specific functions in the bee gut
362 environment.
363
364 Overall, the analysis of the carbohydrate-related gene content shows that Firm-5
365 strains from honey bees harbor a larger diversity of PTS transporters and glycoside
366 hydrolases than bumble bee strains. However, differences in the type and
367 abundance of these functions between strains and sublineages suggest that honey
368 bee strains have diversified in their ability to utilize different sugar resources.

369

370 **Firm5 strains from bumble bees, but not from honey bees harbor class II**
371 **bacteriocins**

372 Most gene families specific to the bumble bee strains were annotated as
373 hypothetical proteins (**Figure 3C and D**), providing no insights about the possible
374 genetic basis of adaptation to the bumble bee gut environment. However, we found
375 several short open-reading frames encoding putative class II bacteriocins.
376 Bacteriocins are small peptide toxins that act against closely related bacterial
377 strains (Cotter *et al.* 2003). In the case of class II bacteriocins, an ABC-like
378 transporter usually facilitates toxin secretion, and a dedicated immunity protein
379 provides self-protection. Except for strain ELS0228, all bumble bee strains harbored
380 at least one class II bacteriocin gene with homology to lactococcin 972, described to
381 inhibit septum formation (Martínez *et al.* 2000). Consistent with the genetic
382 organization of lactococcin loci in other species (Letzel *et al.* 2014), putative
383 immunity proteins and ABC transporter genes were encoded downstream of the
384 bacteriocin gene **(Figure 5)**. We identified four distinct genomic regions with this
385 genetic organization. All four regions exhibited a high degree of genomic plasticity,
386 with many non-conserved open reading frames close by (**Figure 5 and Figure S8**).
387 Each bacteriocin locus was specific to one of the two bumble bee sublineages and
388 only present in a subset of the analyzed strains. In sublineage Firm5-5, one region
389 encoded two adjacent bacteriocin loci, and in several instances, one of the immunity
390 protein or toxin genes was pseudogenized **(Figure 5)**. Strikingly, none of these
391 genomic regions were present in the honey bee strains analyzed, suggesting that
392 this genetic feature may be specific to bumble bee strains. However, we found

393 homologs of genes for helveticin-J in honey bee strains, another protein with known
394 bactericidal activity against related bacteria. This gene family was conserved in all
395 strains of Firm5 as well as in some of the outgroup strains **(Figure S9**).
396 In summary, while the two bumble bee sublineages of Firm5 harbored a large pool
397 of host-specific gene families, bacteriocins were the only conserved genes with
398 annotated functions, and thus the only identified candidates to play a role for niche
399 specialization in the present state of our knowledge.

Discussion

In this study, we combined honey bee colonization experiments with comparative genomics to investigate host specialization of Lactobacillus Firm5, a dominant gut symbiont of social bees. Our results show that strains isolated from honey bees and bumble bees belong to separate, highly divergent sublineages of the Firm5 phylotype, which parallels phylogenetic analysis of other bee gut symbionts (Kwong *et al.* 2014; Zheng *et al.* 2016; Kwong *et al.* 2017; Steele *et al.* 2017).

Interestingly, all tested Firm5 strains from bumble bees were able to colonize the gut of microbiota-depleted honey bees, indicating that the divergent evolution of Firm5 strains from different bee species has not resulted in strict host specialization. However, the percentage of successfully colonized bees as well as the number of bacterial cells per gut were both lower for bumble bee strains compared to honey bee strains. Only by increasing the number of bacterial cells in the inoculum by 100-fold, were we able to achieve reliable colonization, which suggests strong negative selection of bumble bee strains during passage through the honey bee gut, possibly due to the lack of host-specific adaptations.

However, we currently do not know whether, inversely, bumble bee strains would perform better, and honey bee strains worse, in microbiota-depleted bumble bees, which would provide further evidence for host-specific adaptations. Nevertheless, our findings show that the fundamental niche of Firm5 strains is larger than the realized niche and includes host species from other bee genera. This is in agreement with a previous study showing that selected bacteria from diverse environments,

423 including zebrafish or termite gut, can establish in the gut of germ-free mice
424 (Seedorf *et al.* 2014). Moreover, the gut symbionts *S. alvi* (social bee gut) and *L.*
425 *reuteri* (vertebrate gut) - for which host specialization has been experimentally
426 demonstrated – are both able to colonize non-native hosts, although at much lower
427 levels than native hosts (Frese *et al.* 2011; Kwong *et al.* 2014).
428
429 If related species have overlapping fundamental niches, such as in the case of the
430 Firm5 isolates from honey bees and bumble bees, classical ecological theory
431 predicts that differences in the realized niche are due to interactions leading to
432 competitive exclusion of one of the species (Macarthur & Levins 2015). This is
433 exactly what we observed when we colonized microbiota-depleted honey bees with
434 a community of five Firm5 strains, including one bumble bee strain and four honey
435 bee strains. The bumble bee strain did not establish in any of the tested bees,
436 although we inoculated the microbiota-depleted bees with up to 100x more
437 bacterial cells of the bumble bee strain than the four honey bee strains. This clearly
438 shows that the tested bumble bee strain has a competitive disadvantage in the
439 honey bee gut compared to honey bee strains. Similar results were obtained for *S.*
440 *alvi*, when a non-native strain was challenged with a native competitor for gut
441 colonization (Kwong *et al.* 2014).
442 Notably, the phylogenetic similarity between the bumble bee and the honey bee
443 strains of the Firm5 phylotype cannot be the underlying reason for the competitive
444 exclusion of the bumble bee strain, because the four more closely related honey bee
445 strains were able to coexist. Therefore, it is more likely that coevolution of the

446 honey bee strains in the honey bee gut has resulted in reciprocal adaptations
447 facilitating coexistence.
448 Honey bees live in large colonies and engage in frequent social interactions. This
449 results in constant exposure to bacteria from nestmates, thereby providing few
450 opportunities for bacteria from non-native hosts to establish in the gut of young
451 worker bees during community assembly. However, even when given the ecological
452 opportunity for gut colonization (as in our colonization experiments), bumble bee
453 strains cannot reliably colonize the gut. Hence we conclude that the competitive
454 disadvantage relative to honey bee strains as well as the suboptimal host adaptation
455 are two important factors contributing to the exclusion of bumble bee Firm5 strains
456 from the honey bee gut in natural populations.
457
458 The bacteria-mediated exclusion of the bumble bee Firm5 strains from the honey
459 bee gut could arise via direct antagonistic interactions between bacteria (e.g. via
460 bacterial toxins), or from resource competition. We identified a number of genes
461 encoding bacteriocins, which are known to mediate interbacterial killing
462 (Kommineni *et al.* 2015). These genes were either shared by strains isolated from
463 both hosts, or they were specific to the bumble bee strains. However, we did not
464 identify any toxin genes specific to the honey bee strains, which could mediate
465 possible antagonistic effects towards bumble bee strains and hence hinder their
466 colonization in the gut. Therefore, we conclude that based on the current
467 knowledge, it is unlikely that direct antagonistic interactions explain the
468 competitive exclusion of bumble bee strains from the honey bee gut.
469

Biofilm formation at the host epithelium has been shown to be a crucial factor for the colonization success and competitiveness of the murine gut symbiont *L. reuteri* (Frese *et al.* 2011; Duar *et al.* 2017). The honey bee gut symbiont *S. alvi* also colonizes the epithelial surface and forms biofilm-like structures, making it conceivable that competition for adherence is also a critical factor for colonization in the bee gut (Kwong & Moran 2016). However, bacteria of the Firm5 phylotype do not attach to the host epithelium, as shown by fluorescence in situ hybridization experiments, but rather colonize the gut lumen in the rectum, where competition for space seems less likely to be a predominant limiting factor (Martinson *et al.* 2012). Moreover, our genomic analysis did not identify genes involved in host interaction or adherence to be specific to strains from one of the two host groups.

Instead our analyses suggest competition for nutrients to be an important underlying factor for competitiveness. We found several metabolic genes that were conserved across all analyzed honey bee strains but not present in any of the bumble bee strains. Among them were several genes encoding amino acid transporters, which may facilitate the acquisition of organic nitrogen from the host diet. The predominant energy metabolism of the Firm5 phylotype is predicted to be fermentation of dietary carbohydrates, which is not surprising given that the diet of social bees (pollen and nectar) is rich in simple sugars, polysaccharides (pectin, hemicellulose and cellulose), and other glycosylated compounds (e.g. flavonoids) (Engel *et al.* 2012; Ellegaard *et al.* 2015; Kešnerová *et al.* 2017). Strikingly, we found that most honey bee strains harbored a much larger arsenal of gene functions for carbohydrate metabolism and transport compared to bumble bee and outgroup

494 strains. This suggests that honey bee strains have a greater capacity to utilize diet-
495 derived carbohydrates, which may give them a growth advantage over bumble bee
496 strains in the bee gut. A similar trend has also been observed for strains of the gut
497 symbiont *G. apicola* (Kwong *et al.* 2014).
498 Bumble bees and honey bees have a similar dietary regime, as both eat nectar and
499 pollen. Hence, the reason why bumble bee strains harbor significantly fewer
500 carbohydrate-related functions is currently unclear. One possibility could be that
501 these strains colonize a different niche in the gut. However in this case, we would
502 not expect to see competition for colonization among honey bee and bumble bee
503 strains. Another possible factor could be differences in the life cycle of bumble bees
504 and honey bees. Honey bees maintain perennial colonies of large population sizes,
505 while bumble bees build smaller colonies from a single overwintering queen every
506 year. This presents a population bottleneck for the bacterial community in the gut of
507 bumble bees, which reduces genetic variation among gut symbionts by genetic drift,
508 which in turn might lead to gene loss and decrease the selective pressure imposed
509 by related bacteria. Consequently, it would slow the acquisition of genes that would
510 allow bacteria to use diverse carbohydrates. The genomes of bumble bee strains
511 tended to be slightly smaller than those of honey bees, which seems to be consistent
512 with this hypothesis. Strikingly, also in the host-specialized vertebrate gut symbiont
513 *L. reuteri*, it was speculated that genomic differences in genome size and pan
514 genome diversity may be due to differences in population bottlenecks across hosts
515 (Frese *et al.* 2011).
516

517 Interestingly, almost none of the carbohydrate-related gene families specific to
518 honey bee strains of the Firm5 phylotype were conserved across all analyzed
519 genomes, suggesting that the genetic basis of host adaptation in regard of
520 carbohydrate metabolism differs between strains. A large proportion of the
521 carbohydrate-related gene content was specifically associated with one of the four
522 sublineages of honey bee Firm5 strains. However, only a few of these functions were
523 conserved (e.g rhamnogalacturonan and rhamnose utilization in Firm5-2 and
524 Firm5-3, respectively) and the number of carbohydrate-related functions varied
525 markedly among strains of some sublineages. Together these findings suggest that
526 metabolic functions with possible roles for adaptation are frequently gained and/or
527 lost, raising the possibility that metabolic flexibility in itself represents an important
528 adaptation to honey bees compared to bumble bees. This may also explain why the
529 four tested honey bee strains were able to coexist in individual bees. In this context,
530 it is important to highlight that the total number of CFUs per bee was similar
531 between colonizations with the individual honey bee strains and the community,
532 suggesting that the four honey bee strains in the community have overlapping
533 niches, but segregate (spatially or metabolically) when present together. Moreover,
534 these results provide additional evidence that interbacterial competition within
535 phylotypes plays an important role in the bee gut environment.
536
537 In conclusion, our study advances the understanding of host specialization of gut
538 symbionts. While previous studies on *L. reuteri* have shown that host interaction,
539 and specifically colonization of the gut surface, determine host specificity, we
540 provide evidence for metabolic flexibility that may facilitate adaptation to the host

541 diet and hence the competitive exclusion of non-adapted strains. As specific dietary
542 preferences are common among animals, similar processes may also be a
543 determining factor of host specialization among other gut symbionts.

Materials and methods

Bee sampling, bacterial culturing and DNA isolation.

Bumble bees were collected from flowers in different locations in Western Switzerland as indicated in **Table S1**. Honey bees were sampled from two healthy looking colonies in the same region located at the University of Lausanne. Within 6h after sampling, bees were immobilized on ice and the entire gut was dissected with sterile scissors and forceps. Each gut tissue was individually placed into a screw cap tube containing 1ml 1x PBS and glass beads (0.75-1mm, SIGMA) and homogenized with a bead-beater (FastPrep-24 5G, MP Biomedicals) for 30s at speed 6.0. Serial dilutions of the gut homogenates were plated on MRS agar and incubated in an incubator at 34°C in an anaerobic chamber (Coy laboratories, MI, USA) containing a gas mix of 8% H_2, 20% CO_2 and 72% N_2. After 3-5 days of incubation, single colonies were picked, restreaked on fresh MRS agar and incubated for another 2-3 days. Then, a small fraction of each restreaked bacterial colony was resuspended in lysis buffer (10 mM Tris-HCl, 1 mM EDTA, 0.1% Triton, pH 8, 2 mg/ml lysozyme and 1mg/ml proteinase K) and incubated in a thermocycler (10 min 37°C, 20 min 55°C, 10 min 95°C). Subsequently, a standard PCR with universal bacterial primers (5'-AGR GTT YGA TYM TGG CTC AG-3', 5'-CCG TCA ATT CMT TTR AGT TT-3') was performed on 1 µl of the bacterial lysate and the resulting PCR products were sent for Sanger sequencing. Sequencing reads were inspected with Geneious v6 (Biomatters Limited) and compared to the NCBI nr database (NCBI Resource Coordinators 2018) using BLASTN. Isolates identified to have high similarity (i.e. >95% sequence identity) to honey bee strains of the Firm-5 phylotype were stocked

in MRS broth containing 25% glycerol at -80°C. Genomic DNA was isolated from fresh bacterial cultures of the strains of interest using the GenElute Bacterial Genomic DNA Kit (SIGMA) according to manufacturers instructions. Bumble bees were genotyped based on the COI gene by performing a PCR on DNA extracted from the carcass, sending the PCR product for Sanger sequencing, and searching the resulting sequence read by BLASTN against the NCBI nr database.

Genome sequencing, assembly and annotation.

Genome sequencing libraries were prepared with the TruSeq DNA kit and sequenced on the MiSeq platform (Illumina) using the paired-end 2x250-bp protocol at the Genomic Technology facility (GTF) of the University of Lausanne. The preliminary genome sequence analysis was carried out in the framework of the student course 'Sequence-a-genome (SAGE)' at the University of Lausanne in 2016-2017. In short, the resulting sequence reads were quality-trimmed with trimmomatic v0.33 (Bolger *et al.* 2014) to remove adapter sequences and low quality reads using the following parameters: ILLUMINACLIP:TruSeq3-PE.fa:3:25:6 LEADING:9 TRAILING:9 SLIDINGWINDOW:4:15 MINLEN:60. The quality-trimmed reads were assembled with SPAdes v.3.7.1 (Bankevich *et al.* 2012), using the "--careful" flag and multiple k-mer sizes (-k 21,33,55,77,99,127). Small contigs (less than 500 bp) and contigs with low kmer coverage (less than 5) were removed from the assemblies, resulting in 11-22 contigs per assembly. The contigs of each assembly were re-ordered according to the complete genome of the honey bee strain ESL0183 using MAUVE v2.4 (Rissman *et al.* 2009). The origin of replication was set to the first base of the *dnaA* gene, which coincided with the sign change of

591 the GC skew. The ordered assemblies were checked by re-mapping the quality-
592 trimmed reads (**Figure S2**). Except for a few prophage regions that showed
593 increased read coverage, no inconsistencies in terms of read coverage or GC skew
594 were revealed suggesting that the overall order of the contigs was correct. The
595 median read coverage of the sequenced genomes ranged between 135x-223x
596 (**Figure S2**). The genomes were annotated using the 'Integrated Microbial Genomes
597 and Microbiomes' (IMG/mer) system (Markowitz *et al.* 2014).
598
599 **Inference of a genome-wide phylogeny.**
600 Gene families, i.e. groups of homologous genes, were determined using OrthoMCL
601 (Li *et al.* 2003) between all publicly available and newly sequenced genomes of the
602 Firm5 phylotype as well as a set of outgroup genomes of other lactobacilli strains.
603 The outgroup strains were selected based on their phylogenetic relatedness with
604 the Firm5 phylotype using a previously published phylogeny of the entire genus
605 *Lactobacillus* (Zheng *et al.* 2015). Based on this analysis, we included the genomes of
606 15 closely related outgroup strains that belong to the same *Lactobacillus* clade as
607 Firm5 ('delbrueckii group') and three more distantly related strains for rooting the
608 phylogeny. All-against-all BLASTP searches were conducted with the proteomes of
609 the selected genomes, and hits with an e-value of $\leq 10^{-5}$ and a relative alignment
610 length of >50% of the query and the hit protein lengths were kept for OrthoMCL
611 analysis. All steps of the OrthoMCL pipeline were executed as recommended in the
612 manual and the mcl program was run with the parameters '--abc -I 1.5'.
613

The core genome phylogeny was inferred from 408 single copy orthologs extracted from the OrthoMCL output (i.e. gene families having exactly one representative in every genome in the analysis). The protein sequences of each of these core gene families were aligned with mafft (Katoh *et al.* 2017). Alignment columns represented by less than 50% of all sequences were removed and then the alignments were concatenated. Core genome phylogenies were inferred on the concatenated trimmed alignments using RAxML (Stamatakis 2014) with the PROTCATWAG model and 100 bootstrap replicates.

Comparison of genome structure, genome divergence, and gene content.

To compare and visualize whole genomes we used the R-package genoPlotR (Guy *et al.* 2010). BLASTN comparison files were generated with DoubleACT (www. hpa-bioinfotools.org.uk) using a bit score cutoff of 100. To estimate sequence divergence between genomes, we calculated pairwise average nucleotide identity (ANI) with OrthoANI (Lee *et al.* 2016) using the exectutable 'OAT_cmd.jar' with the parameter '-method ani'.

For analyzing the distribution of gene families across Firm5 sublineages and closely related outgroup strains, we carried out a second OrthoMCL analysis, in which we excluded the three distantly related outgroup strains. To remove redundancy in our database, we also excluded the genomes of five Firm5 isolates that were identical, or almost identical, to other Firm5 strains in the analysis, based on ANI values of >99.9%. This resulted in a total 37 genomes (22 Firm5 genomes and 15 outgroup genomes) that were included in the analysis. BLASTP and OrthoMCL were run with

638 the same parameters as before. Gene family subsets of interest (e.g. families specific
639 to honey bee, bumble bee or outgroup strains) were extracted from the OrthoMCL
640 output file using custom-made Perl scripts. COGs (Cluster of Orthologous Groups)
641 were retrieved from IMG/mer genome annotations (Markowitz *et al.* 2014).
642
643 For the detection and visualization of the genomic regions encoding bacteriocin
644 genes, Bagel3 (van Heel *et al.*) and MultiGeneBlast (Medema) were used. For the
645 MultiGeneBlast analysis, bacteriocins-encoding genomic regions of strains ESL0233
646 and ESL0247 served as query sequences for searching a custom-made database
647 composed of all non-redundant Firm5 genomes.
648
649 **Identification of glycoside hydrolase gene families.**
650 Glycoside hydrolase gene families were identified in all analyzed genomes
651 (excluding the redundant Firm5 strains and the three distant outgroup strains)
652 using the command-line version of dbCAN (**D**ata**b**ase for automated **C**arbohydrate-
653 active enzyme **An**notation) (Yin *et al.*). In short, we searched each genome against
654 dbCAN using hmmscan implemented in HMMER v3 (Eddy 2009). The output was
655 processed with the parser script 'hmmscan-parser.sh', and genes with hits to Hidden
656 Markov Models of glycoside hydrolase families were extracted.
657
658 For determining the taxonomic distribution of related genes, we searched one
659 homolog of each glycoside hydrolase gene family against the NCBI *nr* database
660 (NCBI Resource Coordinators 2018) using BLASTP. The taxonomy of the first 50
661 BLASTP hits (e-value <10-5) was extracted at the family level using the Perl script

662 'Tax_trace.pl' and the database files nodes.dmp and names.dmp. The latter two files
663 contain the NCBI taxonomy nodes and names.
664
665 **Bee colonization experiments.**
666 Newly emerged, microbiota-depleted bees were generated as described in (Emery
667 *et al.* 2017) and colonized within 24-36h after pupal eclosion. To this end, bacterial
668 strains were grown on MRS agar containing 2% fructose and 0.2% L-cysteine-HCl
669 from glycerol stocks for two days in an anaerobic chamber at 34°C. Then, 1-10
670 colonies were inoculated into 5ml of carbohydrate-free MRS supplemented with 4%
671 fructose, 4% glucose and 1% L-cysteine-HCl and incubated for another 16-18h
672 without shaking. Bacteria were spun down and resuspended in 1xPBS/sugar water
673 (1:1). The optical density (600nm) was adjusted according to the experimental
674 condition (OD=0.0001, 0.001, 0.01, or 0.1) and 5 µl of the final bacterial suspension
675 was fed to each newly emerged bee. Before feeding, the bees were starved for 2-3h.
676 After colonization, bees were given 1 ml of sterilized polyfloral pollen and sugar
677 water ad libitum. Bees were co-housed in groups of 20-40 bees. For the competition
678 experiment, each of the four honey bee strains was adjusted to an optical density
679 (600nm) of 0.001. The bumble bee strain ESL0228 was adjusted to an optical
680 density (600nm) of either 0.001, 0.01, or 0.1. Then equal volumes of the five strains
681 were mixed together and fed to newly emerged bees as described before. As
682 negative control, bees were fed with 5 µl of 1xPBS/sugar water. Dilutions of the
683 bacterial inocula were plated on MRS agar containing 2% fructose and 0.2% L-
684 cysteine-HCl and incubated as described before to determine how many CFUs
685 correspond to a given optical density (see Figure S5).

686 Ten bees per condition were dissected on day 5 after colonization. The hindgut was
687 separated from the midgut with a sterile scalpel and tweezers, and added to 1 ml or
688 500 ul of 1x PBS (depending on the experiment). The tissues were homogenized by
689 bead-beating as described before, dilutions plated on MRS agar containing 2%
690 fructose and 0.2% L-cysteine-HCl and the number of CFUs counted two to three
691 days after incubation. For the negative control, bacterial colonies were detected for
692 only one out of 30 bees with a relatively low abundance (10^3 CFUs per gut).
693 Moreover, the colonies looked different from the colonies of the Firm5 strains and
694 were identified as being *E. coli* and *Staphylococcus aureus* by 16S rRNA gene
695 sequencing.
696 The relative abundance of the five strains in the competition experiment was
697 analyzed using amplicon sequencing of a 199-bp fragment of a conserved
698 housekeeping gene (COG0266). To this end, a two-step PCR protocol was
699 established. In the first PCR, the 199-bp fragment of COG0266 was amplified from
700 crude cell lysates of gut homogenates with primers 1133 (5' -
701 CGTACGTAGACGGCCAGTATGCCNGAAATGCCRGARGTTGA – 3') and 1134 (5' -
702 GACTGACTGCCTATGACGACTAARCGATAYTTRCCYTCCATRCG) (3' – 95˚C; 25x: 30'' -
703 95˚C, 30'' – 64˚C, 30'' – 72˚C; 5' – 72˚C). After removing primers with exonuclease
704 and shrimp alkaline phosphatase, barcoded Illumina adapters were added in the
705 second PCR. The resulting PCR products were pooled at equal volumes, gel purified
706 (MinElute Gel Extraction Kit, Qiagen) and loaded on an Illumina MiniSeq instrument
707 in mid-output mode. Reads were demultiplexed and filtered on quality using
708 trimmomatic (LEADING:28 TRAILING:28 SLIDINGWINDOW:4:15 MINLEN:90)
709 (Bolger *et al.* 2014). Then, each forward and reverse read pair was assembled using

PEAR (-m 290 -n 284 -j 4 -q 26 -v 10 -b 33) (Zhang *et al.* 2014). The resulting contigs were assigned to the five strains based on base positions with discriminatory SNP variants with the help of a custom-made Perl script.

Acknowledgments

We thank the School of Biology of the University of Lausanne for financial support of this project. We would also like to thank Ambrin Farizah Babu, Melvin Berard, Sarah Bergerm, Laurent Casini, Joaquim Claivaz, Yassine El Chazli, Jonas Garessus, Nastassia Gobet, Charlotte Griessen, Olivier Gustarini, Karim Hamidi, Dominique Jacques-Vuarambon, Titouan Laessle, Mirijam Mattei, Cyril Mattheiy-Doret., Jennifer Mayor, Sandrine Pinheiro, Claire Pralong, Virginie Ricci, Shalonie Sheppard, Tatiana Sokoloff, Anthony Sonrel, and Gaëlle Spack who participated as students in the SAGE class 2016/2017 and were involved in a preliminary analysis of the genomic data.

Data accessibility

Genome sequences and short read datasets are available under NCBI Bioproject accession PRJNA392822. Annotations of the Firm5 strains used for this study can be found in IMG/mer. Data analyses including custom scripts and intermediate output files are available under the following link: https://drive.switch.ch/index.php/s/kJiY2tjndjrBwJ5.

Figures

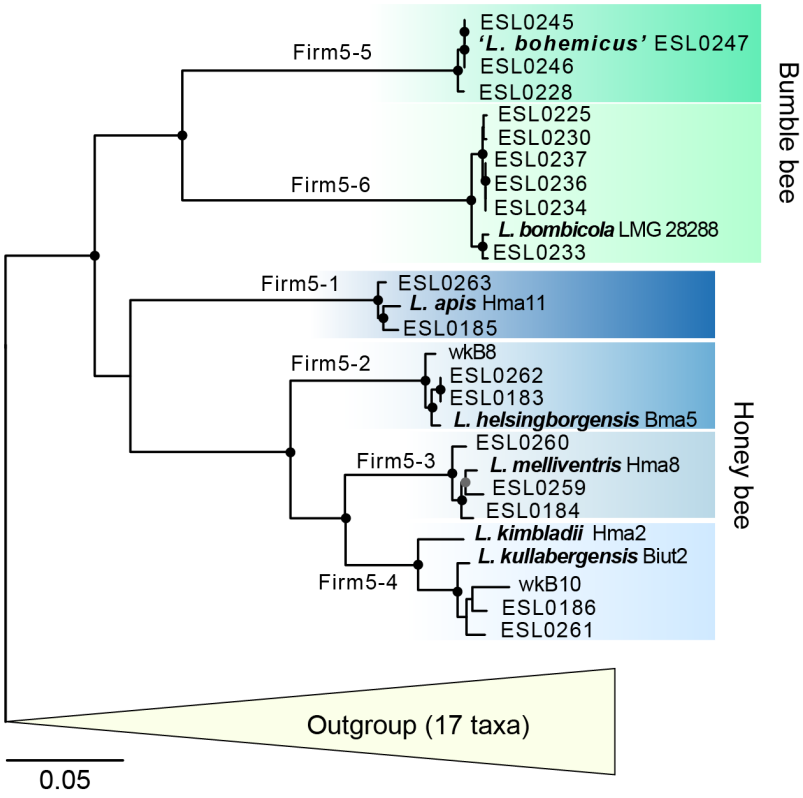
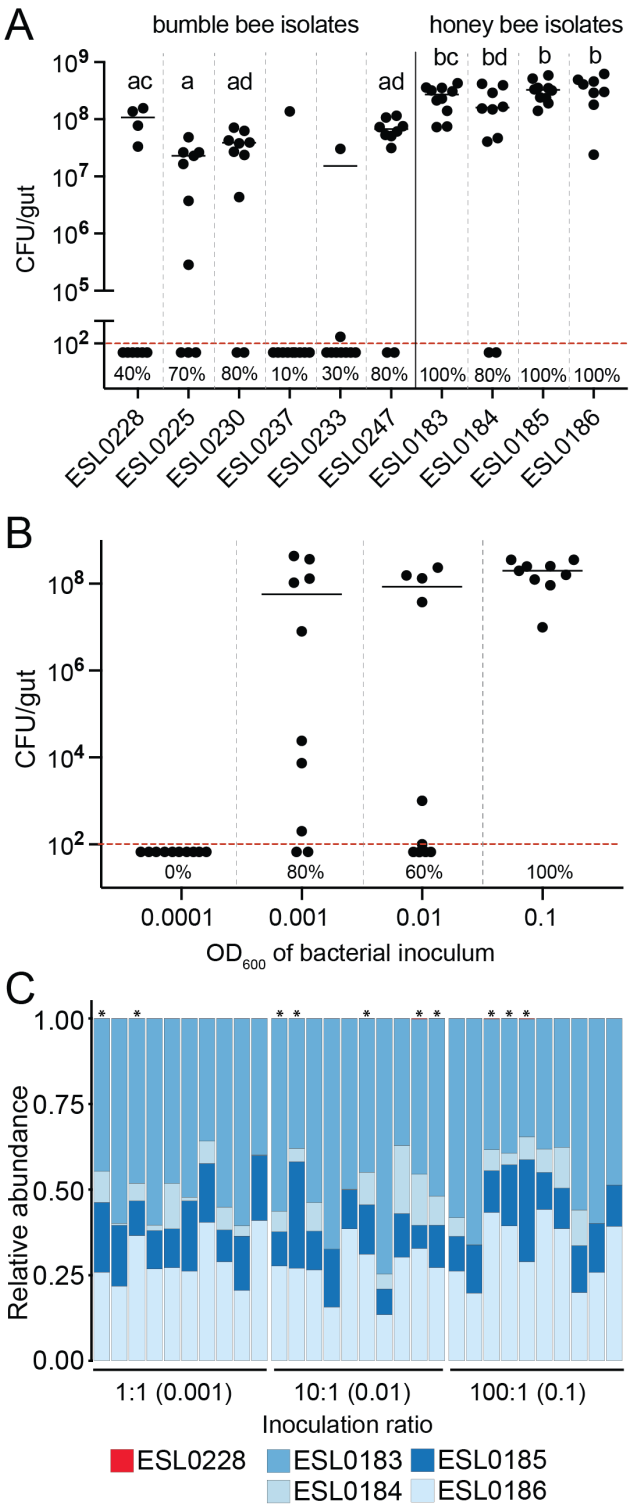


Figure 1. Core genome phylogeny. The tree was inferred using maximum likelihood on the concatenated protein alignments of 408 single-copy core genes (i.e. present in all Firm5 strains and in the outgroup strains). The collapsed outgroup consisted of 18 strains that were used to root the tree (see **Figure S4** for the complete tree). The two lineages of bumble bee strains and the four lineages of honey bee isolates are shown in green and blue color shades, respectively. Black and grey circles indicate bootstrap support values of 100 and ≥80, respectively, out of 100 replicates. The strain designation of each isolate is given and the species names of the typing strains indicated. The candidatus species name *Lactobacillus bohemicus* is depicted by hyphens. The length of the bar indicates 0.05 amino acid substitutions/site.



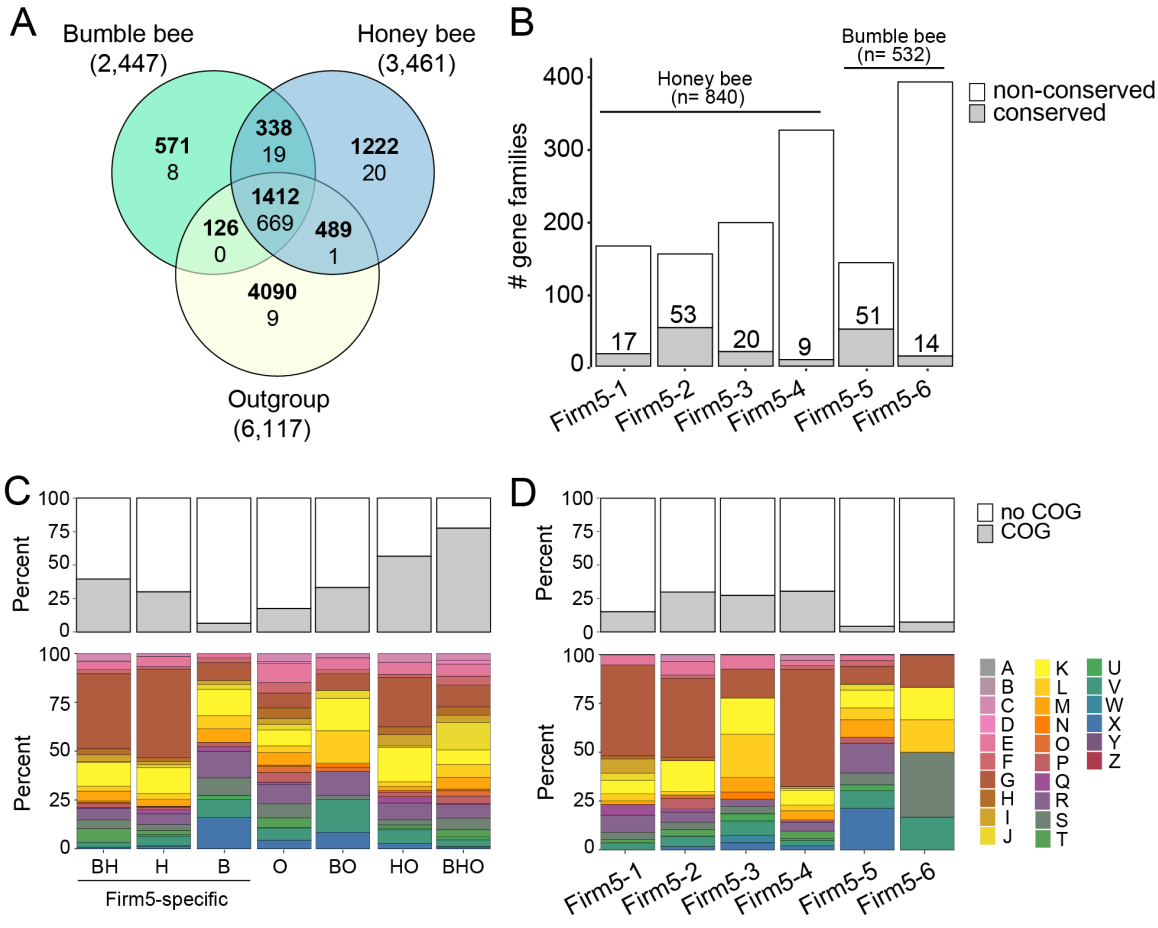
742

743 **Figure 2. Colonization of microbiota-depleted honey bees with Firm5 strains**

744 **from bumble bees and honey bees. (A)** Mono-colonizations of microbiota-

745 depleted honey bees (n=10 per treatment) with six bumble bee strains and four

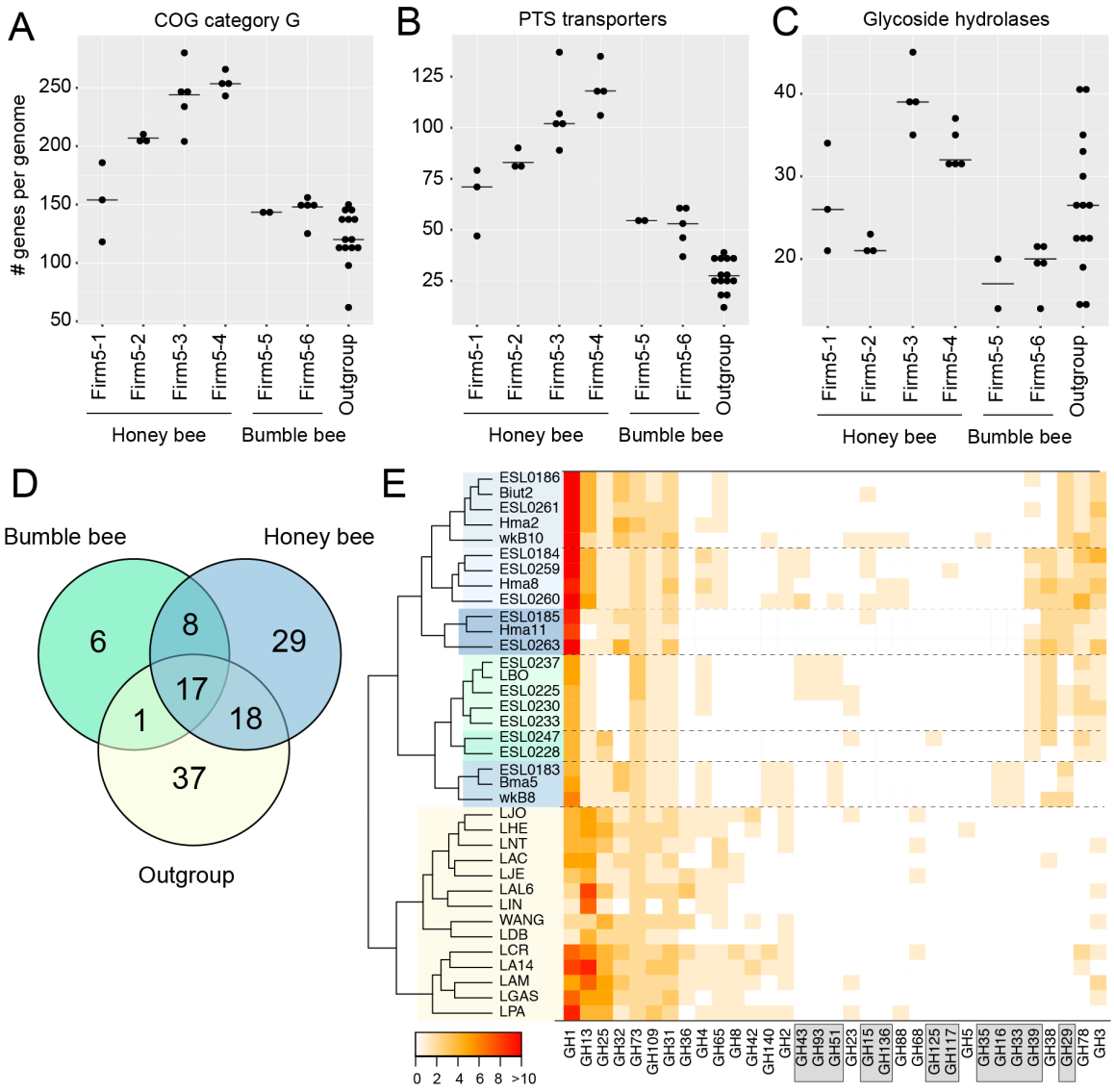
honey bee strains. Each bee was inoculated with 5 µL of an optical density of 0.001. Different letters indicate statistically significant differences between groups, according to two-way ANOVA and Tukey HSD post hoc test (adjusted p-value <0.05). The dashed red line indicates the detection threshold. Data points below the detection limit show bees that had no detectable colonization levels. The percentage of successfully colonized bees is shown. Horizontal lines indicate median. **(B)** Mono-colonizations of microbiota-depleted honey bees with increasing inocula of the bumble bee strain ESL0228. Colony forming units (CFUs) per gut were determined at day 5 post colonization. The graph has the same layout as in panel A. (**C**) Community profiles of microbiota-depleted bees colonized with a community consisting of the bumble bee strain ESL0228 and four honey bee strains (ESL0183, ESL0184, ESL0185, and ESL0186). Three different inoculation ratios of bumble bee strain versus honey bee strains were used. The optical density of the bumble bee strain in the inoculum is given in brackets. Due to the absence or the very low abundance of ESL0228, the red fraction of the graph is not visible. Asterisks indicate samples for which at least a few reads of strain ELS0228 were detected.



763

764 **Figure 3. Pan genome analysis of Firm5 strains from honey bees and bumble**
765 **bees and comparison to outgroup strains (i.e. closely related lactobacilli). (A)**
766 Venn diagram showing gene family distribution into the three major groups: Firm5
767 strains from bumble bees, Firm5 strains from honey bees, and outgroup strains.
768 Numbers in bold indicate pan genome gene families (i.e. present in at least one
769 genome of a given group). Numbers in regular font indicate core genome gene
770 families (i.e. present in all genomes of a given group). **(B)** Number of Firm5-specific
771 gene families exclusively present in strains of one sublineage. The fraction of core
772 genome (present in every genome of a given sublineage) and pan genome (present
773 in at least one genome of a given sublineage) gene families is indicated by grey and
774 white color, respectively. Numbers above the graph indicate total number of

775 lineage-specific gene families for each host group. **(C)** Upper plot shows number of
776 gene families with COG annotation, and lower plot shows COG category distribution
777 of the annotated gene families for each subset of the Venn diagram in panel A. B,
778 specific to bumble bee strains; H, specific to honey bee strains; O, specific to
779 outgroup strains; BH, shared between honey bee and bumble bee strains; BO,
780 shared between bumble bee and outgroup strains; HO, shared between honey bee
781 and outgroup strains; BHO, shared between all three groups. **(D)** Same as in panel C,
782 but for the sublineage-specific gene families shown in panel B. Complete lists of all
783 gene families and their annotations can be found in **Datasets S1 and S2**. The
784 dominant COG category 'G' is shown in dark red and corresponds to 'Carbohydrate
785 transport and metabolism'. Other COG category abbreviations are given in **Dataset**
786 **S2**.



788

789 **Figure 4. Distribution of carbohydrate-related gene families across strains of**
790 **the Firm5 phylotype. (A)** Total number of COG category 'G' gene families per
791 genome per sublineage. **(B)** Total number of PTS (Phosphotransferase system) gene
792 families per genome per sublineage. **(C)** Total number of glycoside hydrolase gene
793 families per genome per sublineage. In all three panels, the genomes of the outgroup
794 strains were included as a reference. **(D)** Venn diagram of glycoside hydrolase gene
795 family distribution into the three major phylogenetic groups: Firm5 strains isolated

796 from bumble bees, Firm5 strains isolated from honey bees, and outgroup strains
797 belonging to related lactobacillus species. **(E)** Heatmap showing the distribution of
798 the identified glycoside hydrolase (GH) families across the analyzed genomes. The
799 dendrogram on the left shows a hierarchical clustering based on glycoside
800 hydrolase distribution. Strains are colored according to the major groups (green,
801 bumble bee strains; blue, honey bee strains; yellow, outgroup) and sublineage (color
802 tones). GH families specific to the Firm5 phylotype are indicated by grey boxes.

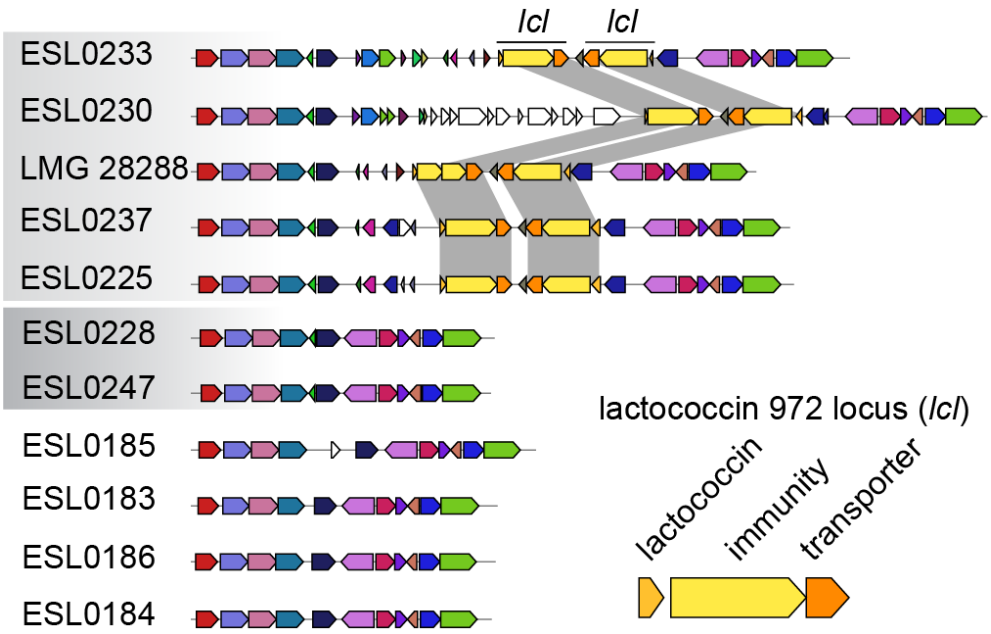


Figure 5. Genomic region encoding class II bacteriocins in Firm5 strains of bumble bee strains. Genomic regions encoding bacteriocin genes were identified and visualized with MultiGeneBlast v1.1.14 (Medema). Arrows present genes and same color indicates homology. A black line indicates the lactococcin 972 locus (*lcl*) and vertical grey blocks connect the homologous genes in other strains. An enlarged version of the three genes of the *lcl* locus with annotation is shown in the lower right. Grey shading over strain names indicates two sublineages of bumble bee strains; the four honey bees strains are representatives of the four sublineages. Other genomic regions encoding bacteriocins genes are given in Figure S8.

Supplementary Figures

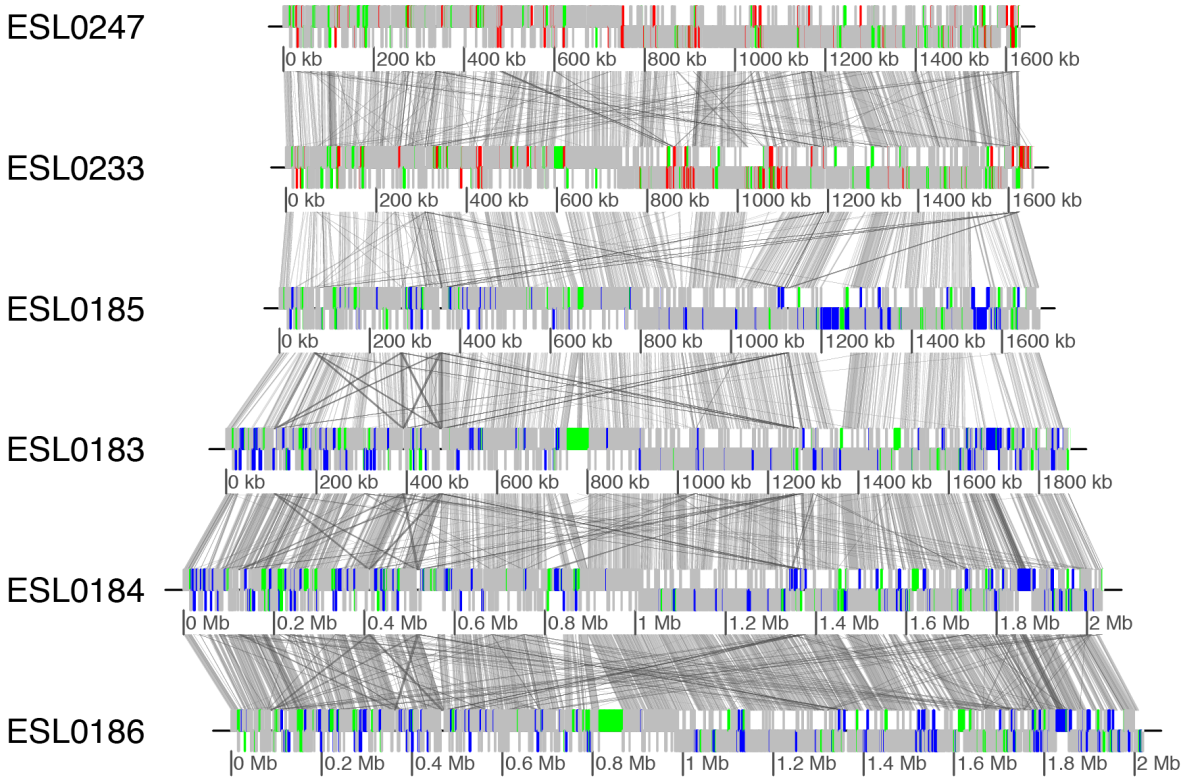
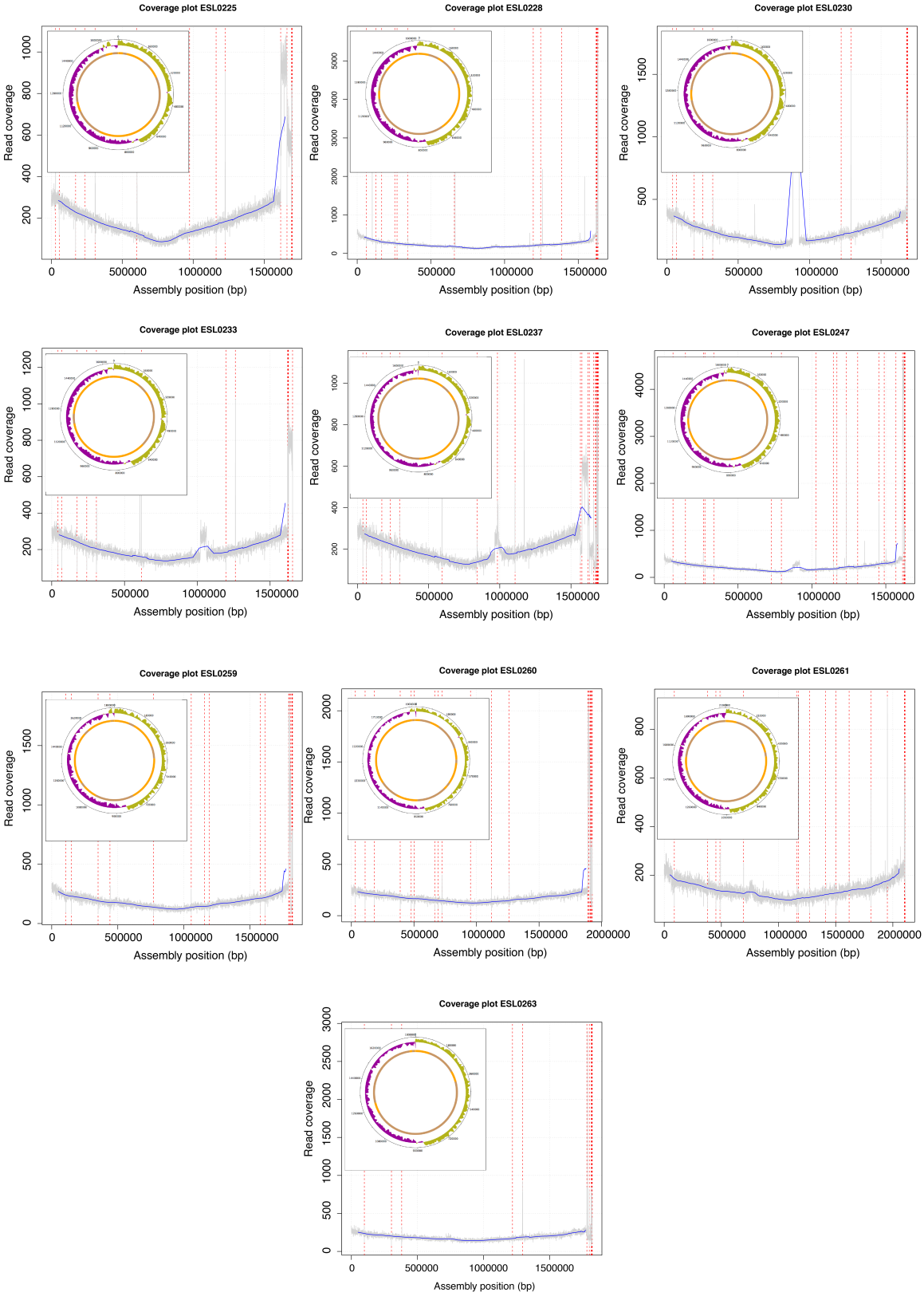


Figure S1. Whole genome alignments of divergent Firm5 strains. ESL0247 and ESL0233 are bumble bee strains. ESL0183-186 are honey bee strains. Vertical grey lines indicate blocks of nucleotide sequence similarity. Different color intensities correspond to different degree of similarity based on BLASTN hits with a bit score of at least 100. Genes in color correspond to Firm5-specific genes relative to the outgroup (blue, genes specific to honey bee strains; red, genes specific to bumble bee strains; green, genes shared between honey bee and bumble bee strains). Other genes are shown in grey.



825

826 **Figure S2. Read coverage and GC skew of the final genome assemblies.**
827 Assembly positions are shown on the x-axis for each sequenced strain. y-Axis shows
828 Illumina read coverage for a sliding window of 100 bp. Red dashed lines indicate
829 contig breaks. Contigs were ordered according to the fully sequenced reference
830 strains ESL0183. Small contigs were left at the end of the assembly. Inset shows the
831 circular form of the assembly with the GC skew indicated. We found a higher read
832 coverage at the origin of replication, which is characteristic for replicating bacteria.
833 Moreover, our assemblies showed the typical GC skew of bacterial genomes. Both
834 characteristics indicate that the contigs of the assemblies were correctly assembled
835 and ordered. Notably, regions of extremely high coverage correspond to prophages
836 that apparently were amplified during culturing of some of the strains.

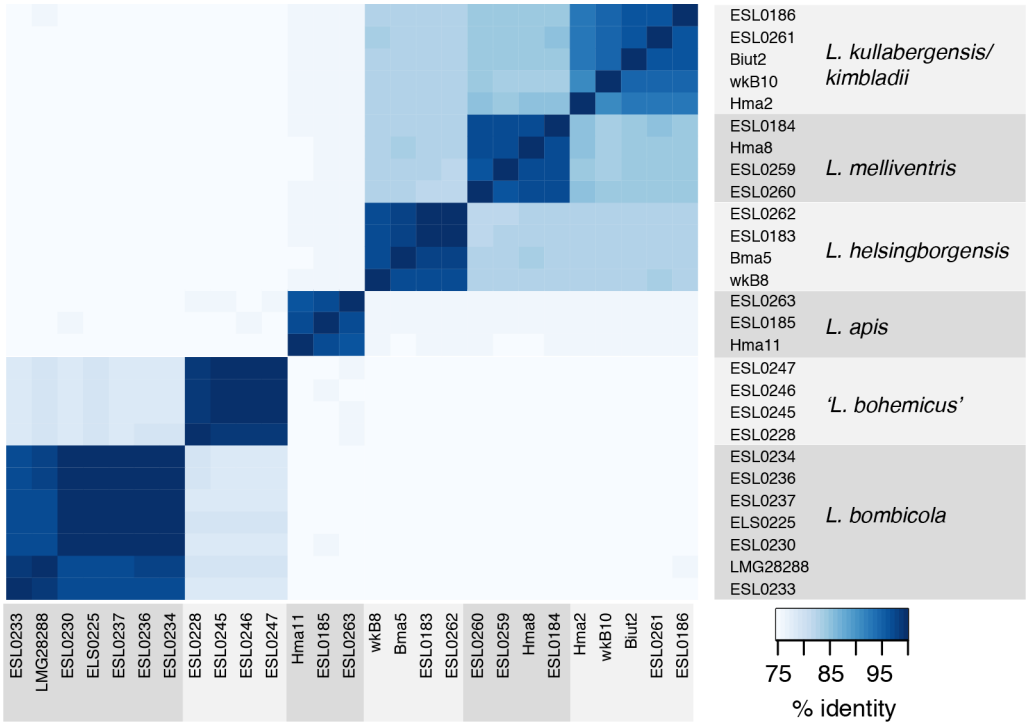


Figure S3. Average nucleotide identity between all analyzed Firm5 genomes. Intensity of heatmap indicates pairwise ANI. White areas correspond to genomes, which were too divergent for ANI calculation. The names of each strain included in the analysis are given next to the plot area (see also **Table S1**). Grey shading indicates the six different sublineages of Firm5. ANI values are given in **Table S3**.

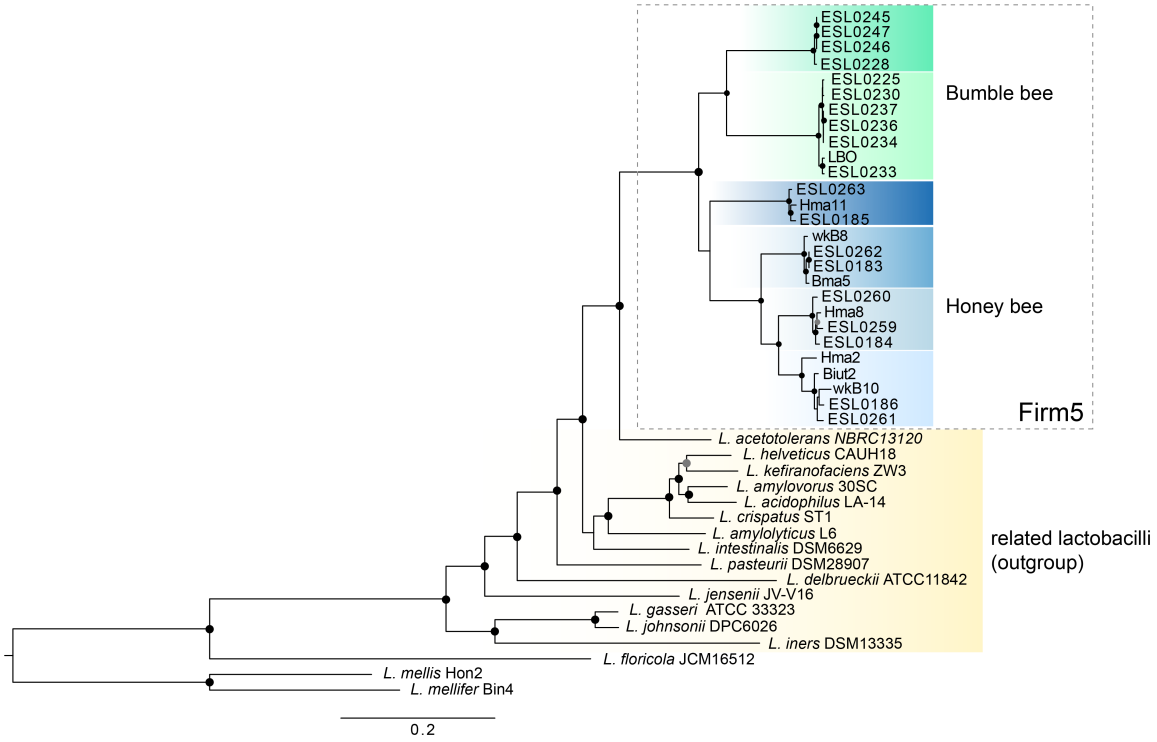


Figure S4. Complete core genome phylogeny. The tree was inferred using maximum likelihood on the concatenated protein alignments of 408 single-copy core gene families (i.e. present in all Firm5 strains and the outgroup strains). The two lineages of bumble bee strains and the four lineages of honey bee strains are shown in green and blue color shades, respectively. As outgroup, 15 representative strains of the *L. delbrueckii* group (to which Firm5 belongs to) were included in the analysis (shown in yellow) based on a previously published phylogeny of the entire genus *Lactobacillus* (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4579461/). In addition, we included three more distantly related strains to root the tree. Noteworthy, these three distantly related lactobacilli were excluded for all subsequent comparative analysis of the Firm5 strains. Filled circles indicate 100 bootstrap support values. The strain designation of each isolate is given. The length of the bar indicates 0.05 amino acid substitutions/sites.

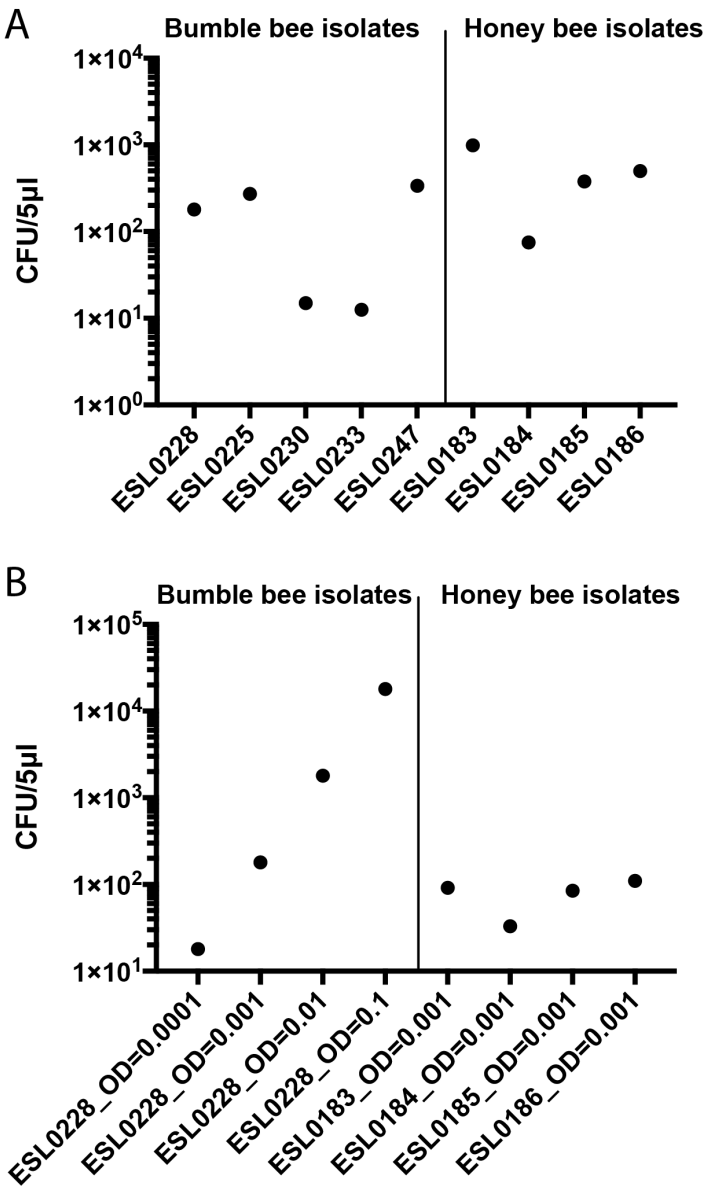
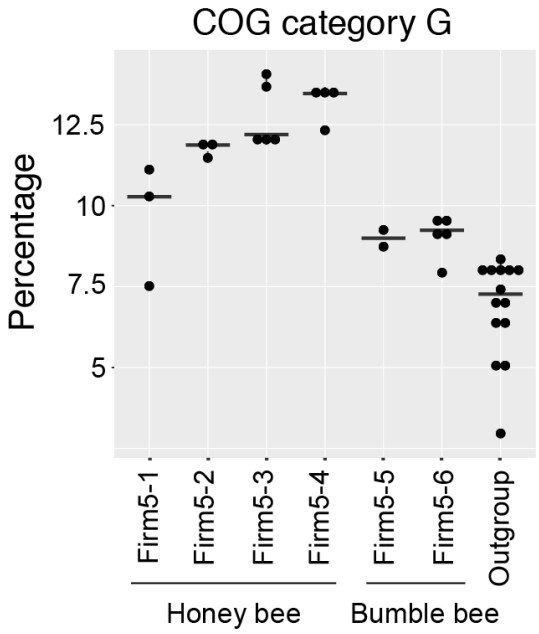


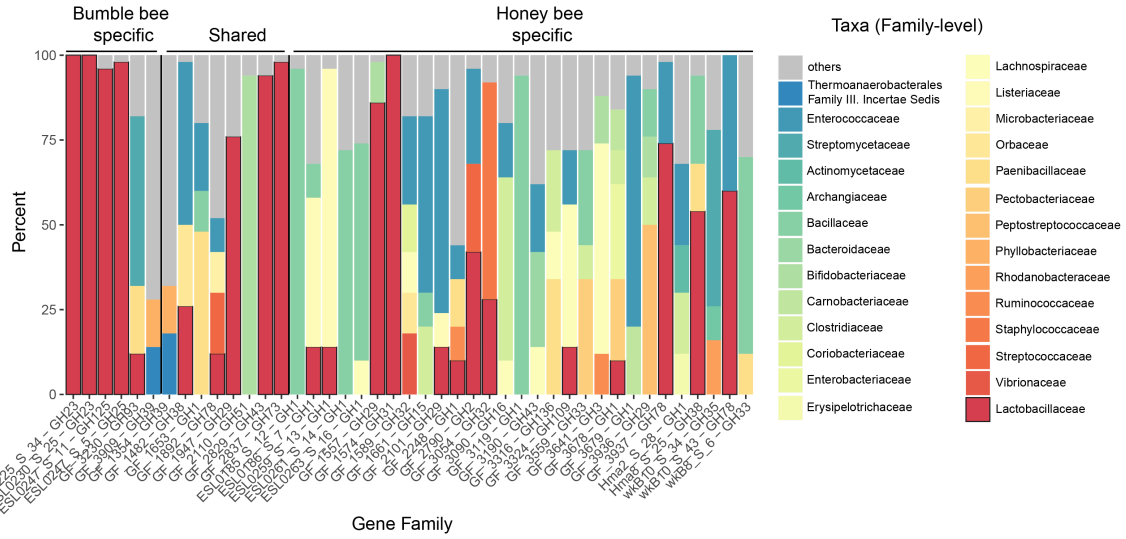
Figure S5. Number of bacterial cell in the inocula used to colonize microbiota-depleted honey bees with Firm5 strains. (A) CFUs in the inocula used for the monocolonization experiments with individual strains. CFUs are given per 5µl, as each bee was inoculated with 5 µl of an OD600 of 0.0001. Despite the adjustment to the same OD600, the amount of live bacteria in each inoculum varied across strains. The inoculum of strain ESL0237 could not be assessed due to a handling mistake during dilution plating. **(B)** CFUs in the incocula used for the colonization

865 experiment with the bumble bee strain ESL0228 (left part) and for the colonization
866 experiment with the five-member community consisting of the bumble bee strain
867 ES0228 (left panel, OD=0.001, 0.01, and 0.1) and the four different honey bee strains
868 (right panel).



869

870 **Figure S6. Percentage of gene families annotated as COG category 'G' per**
871 **genome per sublineage.** Same data as in Figure 4A, but expressed in relative
872 numbers (percentage of all gene families per genome).



873

874 **Figure S7. BlastP hit distribution of glycoside hydrolase (GH) gene families**

875 **specific to Firm5.** A representative protein sequence of each gene family was

876 blasted against the NCBI nr database (NCBI Resource Coordinators 2018). The

877 distribution of the taxonomic classification of the first 50 Blast hits is shown at the

878 family level. The family of lactobacilliales is shown in red with black outlines. For

879 each gene family, the gene family identifier and the glycoside hydrolase enzyme

880 family (GHxx) are given.

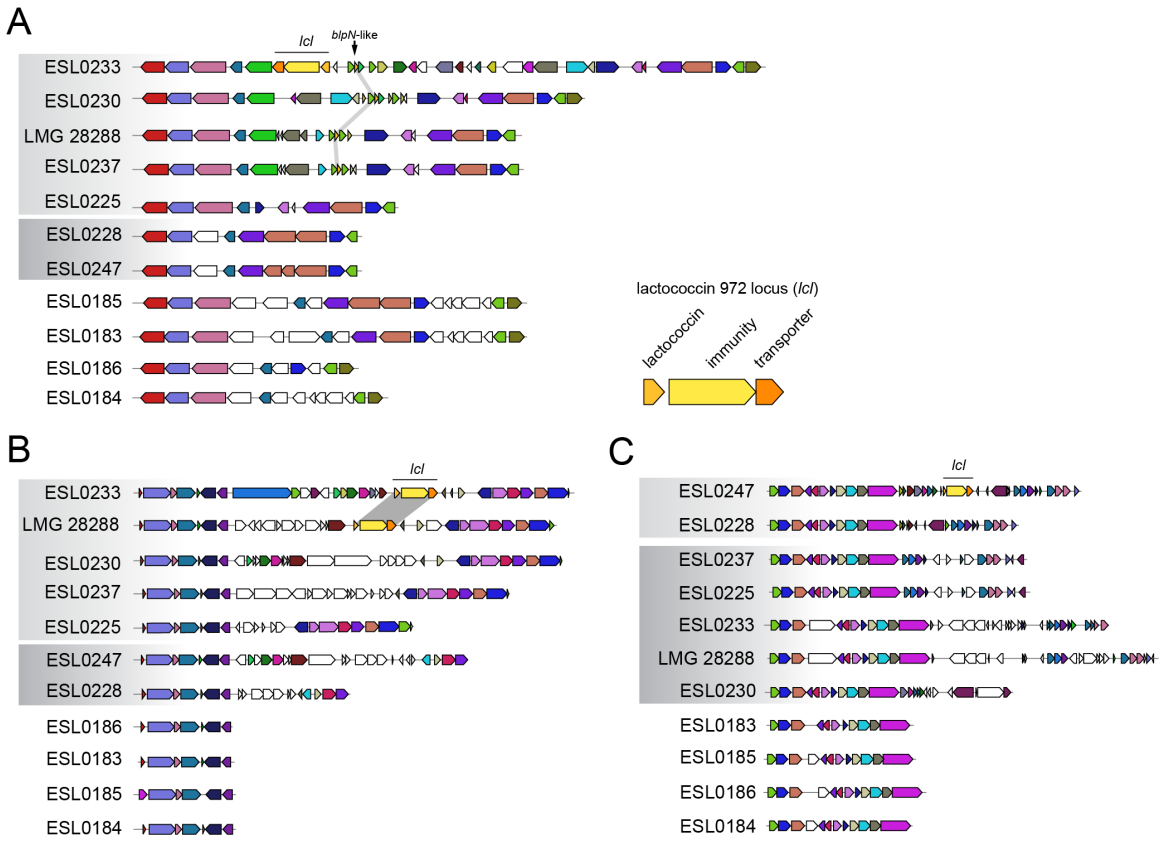
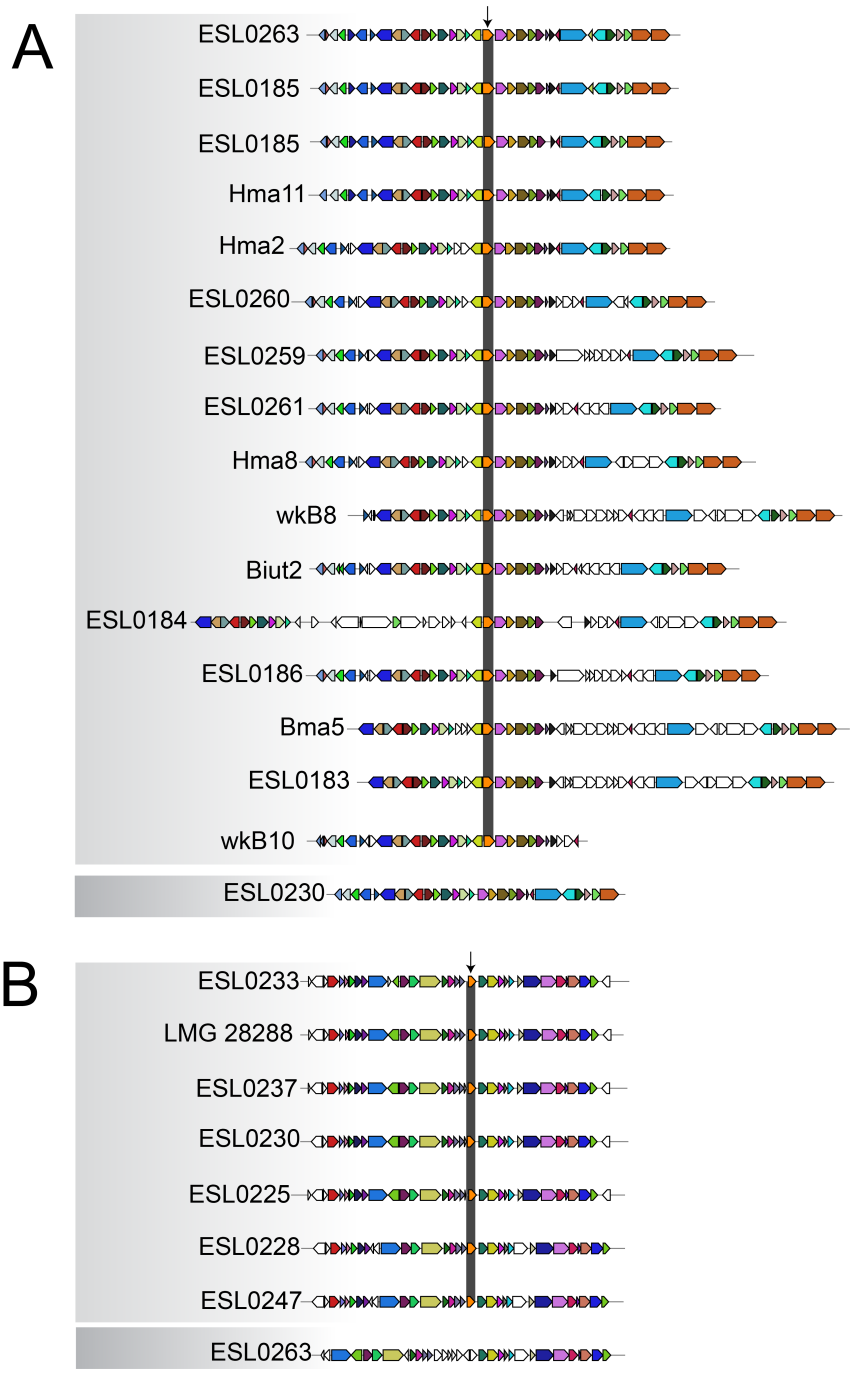


Figure S8. Additional genomic regions encoding class II bacteriocins in Firm5 strains of bumble bee strains. Genomic regions encoding bacteriocin genes were identified and visualized with MultiGeneBlast v1.1.14 (Medema). Arrows represent genes, and same color indicates homology. A black line indicates the lactococcin 972 locus (*lcl*) and vertical grey blocks connect the homologous genes in other strains. An enlarged version of the three genes of the *lcl* locus with annotation is shown in the lower right of panel A. Grey shading over strain names indicates two sublineages of bumble bee strains; the four honey bees strains are representatives of the four sublineages.



892

Figure S9. Genomic regions encoding helviticin-J, a class III bacteriocin. Genomic regions encoding bacteriocin genes were identified and visualized with MultiGeneBlast v1.1.14 (Medema). Arrows represent genes, and same color indicates homology. An arrow points at the helveticin-J gene homolog and vertical grey blocks connect the homologous genes in other strains. Strains with the two different types of grey shadings indicate strains from bumble bees and honey bees. (A) Genomic region encoding helveticin-J in honey bee strains, and (B) genomic region encoding helveticin-J in bumble bee strains.

Supplementary Tables and Datasets

Table S1. Strain list and genome features.

Table S2. Pairwise 16S rRNA gene sequence identities.

Table S3. ANI values.

Dataset S1. List of all pan genome gene families and their distribution according the three major groups: honey bee strains, bumble bee strains, outgroup strains.

Dataset S2. List of sublineage-specific gene families and COG category abbreviations.

Dataset S3. List of genes per genome with hits to the Carbohydrate-active enzyme (CAZY) database.

References

Bankevich A, Nurk S, Antipov D *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, **19**, 455–477.

Bobay L-M, Ochman H (2017) The Evolution of Bacterial Genome Architecture. *Frontiers in genetics*, **8**, 829.

Bolger AM, Lohse M, Usadel B (2014) *Trimmomatic: a flexible trimmer for*.

NCBI Resource Coordinators (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **46**, D8–D13.

Corby-Harris V, Maes P, Anderson KE (2014) The bacterial communities associated with honey bee (*Apis mellifera*) foragers. *PloS one*, **9**, e95056.

Cotter PD, Ross RP, Hill C (2013) Bacteriocins - a viable alternative to antibiotics? Nature reviews. Microbiology, **11**, 95–105.

Cox-Foster DL, Conlan S, Holmes EC *et al.* (2007) A metagenomic survey of microbes in honey bee colony collapse disorder., **318**, 283–287.

Duar RM, Frese SA, Lin XB *et al.* (2017) Experimental Evaluation of Host Adaptation of Lactobacillus reuteri to Different Vertebrate Species. *Applied and environmental microbiology*, **83**, e00132–17.

Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics*, **23**, 205–211.

Ellegaard KM, Tamarit D, Javelind E *et al.* (2015) Extensive intra-phylotype diversity in lactobacilli and bifidobacteria from the honeybee gut. *BMC genomics*, **16**, 284.

934 Emery O, Schmidt K, Engel P (2017) Immune system stimulation by the gut
935 symbiont *Frischella perrara* in the honey bee (*Apis mellifera*). *Molecular Ecology*,
936 **50**, 735.
937 Engel P, Martinson VG, Moran NA (2012) Functional diversity within the simple gut
938 microbiota of the honey bee. *Proceedings of the National Academy of Sciences of*
939 *the United States of America*, **109**, 11002–11007.
940 Eren AM, Sogin ML, Morrison HG *et al.* (2015) A single genus in the gut microbiome
941 reflects host preference and specificity. *The ISME journal*, **9**, 90–100.
942 Frese SA, Benson AK, Tannock GW *et al.* (2011) The Evolution of Host Specialization
943 in the Vertebrate Gut Symbiont Lactobacillus reuteri (DS Guttman, Ed,). *PLoS*
944 *genetics*, **7**, e1001314.
945 Guy L, Kultima JR, Andersson SGE (2010) genoPlotR: comparative gene and genome
946 visualization in R. *Bioinformatics*, **26**, 2334–2335.
947 Hutchinson GE (1957) Concluding remarks. Cold Spring Harbor Symposia on
948 Quantitative Biology*, 22: 415–427*.
949 Katoh K, Rozewicki J, Yamada KD (2017) MAFFT online service: multiple sequence
950 alignment, interactive sequence choice and visualization. *Briefings in*
951 *bioinformatics*, **30**, 3059.
952 Kešnerová L, Mars RAT, Ellegaard KM *et al.* (2017) Disentangling metabolic
953 functions of bacteria in the honey bee gut. *PLoS biology*, **15**, e2003467.
954 Koch H, Abrol DP, Li J, Schmid-Hempel P (2013) Diversity and evolutionary patterns
955 of bacterial gut associates of corbiculate bees. *Molecular Ecology*, **22**, 2028–
956 2044.
957 Kommineni S, Bretl DJ, Lam V *et al.* (2015) Bacteriocin production augments niche

competition by enterococci in the mammalian gastrointestinal tract. *Nature*, **526**, 719–722.

Kostic AD, Howitt MR, Garrett WS (2013) Exploring host-microbiota interactions in animal models and humans. *Genes & development*, **27**, 701–718.

Kwong WK, Moran NA (2015) Evolution of host specialization in gut microbes: the bee gut as a model. *Gut microbes*, **6**, 214–220.

Kwong WK, Moran NA (2016) Gut microbial communities of social bees. *Nature reviews. Microbiology*, **14**, 374–384.

Kwong WK, Engel P, Koch H, Moran NA (2014) Genomics and host specialization of honey bee and bumble bee gut symbionts. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 11509–11514.

Kwong WK, Medina LA, Koch H *et al.* (2017) Dynamic microbiome evolution in social bees. *Science Advances*, **3**, e1600513.

Lee I, Kim YO, Park S-C, Chun J (2016) OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International journal of systematic and evolutionary microbiology*, **66**, 1100–1103.

Letzel A-C, Pidot SJ, Hertweck C (2014) Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC genomics*, **15**, 983.

Ley RE, Hamady M, Lozupone C *et al.* (2008) Evolution of mammals and their gut microbes. *Science*, **320**, 1647–1651.

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**,2178-89.

Ludvigsen J, Porcellato D, L'Abée-Lund TM, Amdam GV, Rudi K (2017)

982 Geographically widespread honeybee-gut symbiont subgroups show locally
983 distinct antibiotic-resistant patterns. *Molecular Ecology*, **26**, 6590–6607.
984 Macarthur R, Levins R (2015) The Limiting Similarity, Convergence, and Divergence
985 of Coexisting Species. *The American Naturalist*, **101**, 377–385.
986 Markowitz VM, Chen I-MA, Chu K *et al.* (2014) IMG/M 4 version of the integrated
987 metagenome comparative analysis system. *Nucleic Acids Research*, **42**, D568–73.
988 Martinson VG, Moy J, Moran NA (2012) Establishment of characteristic gut bacteria
989 during development of the honeybee worker. *Applied and environmental*
990 *microbiology*, **78**, 2830–2840.
991 Martínez B, Rodríguez A, Suárez JE (2000) Lactococcin 972, a bacteriocin that
992 inhibits septum formation in lactococci. *Microbiology*, **146**, 949–955.
993 McCutcheon JP, Moran NA (2012) Extreme genome reduction in symbiotic bacteria.
994 *Nature reviews. Microbiology*, **10**, 13–26.
995 McFall-Ngai M, Hadfield MG, Bosch TCG *et al.* (2013) Animals in a bacterial world, a
996 new imperative for the life sciences. *Proceedings of the National Academy of*
997 *Sciences*, **110**, 3229–3236.
998 Medema MH, Takano E, Breitling R (2013) Detecting Sequence Homology at the
999 Gene Cluster Level with MultiGeneBlast. *Molecular Biolology Evolution*, **30**,
1000 1218-23.
1001 Moeller AH, Caro-Quintero A, Mjungu D *et al.* (2016) Cospeciation of gut microbiota
1002 with hominids. *Science*, **353**, 380–382.
1003 Moran NA, Hansen AK, Powell JE, Sabree ZL (2012) Distinctive gut microbiota of
1004 honey bees assessed using deep sampling from individual worker bees. *PloS one*,
1005 **7**, e36393.

1006 Ochman H, Worobey M, Kuo C-H *et al.* (2010) Evolutionary Relationships of Wild
1007 Hominids Recapitulated by Gut Microbial Communities. *PLoS biology*, **8**,
1008 e1000546.
1009 Oh PL, Benson AK, Peterson DA *et al.* (2010) Diversification of the gut symbiont
1010 Lactobacillus reuteri as a result of host-driven evolution. *The ISME journal*, **4**,
1011 377–387.
1012 Olofsson TC, Alsterfjord M, Nilson B, Butler E, Vásquez A (2014) Lactobacillus
1013 apinorum sp. nov., Lactobacillus mellifer sp. nov., Lactobacillus mellis sp. nov.,
1014 Lactobacillus melliventris sp. nov., Lactobacillus kimbladii sp. nov., Lactobacillus
1015 helsingborgensis sp. nov. and Lactobacillus kullabergensis sp. nov., isolated from
1016 the honey stomach of the honeybee Apis mellifera. *International journal of*
1017 *systematic and evolutionary microbiology*, **64**, 3109–3119.
1018 Praet J, Meeus I, Cnockaert M *et al.* (2015) Novel lactic acid bacteria isolated from
1019 the bumble bee gut: Convivina intestini gen. nov., sp. nov., Lactobacillus
1020 bombicola sp. nov., and Weissella bombi sp. nov. *Antonie van Leeuwenhoek*, **107**,
1021 1337–1349.
1022 Rissman AI, Mau B, Biehl BS *et al.* (2009) Reordering contigs of draft genomes using
1023 the Mauve aligner. *Bioinformatics*, **25**, 2071–2073.
1024 Seedorf H, Griffin NW, Ridaura VK *et al.* (2014) Bacteria from diverse habitats
1025 colonize and compete in the mouse gut. *Cell*, **159**, 253–266.
1026 Sriswasdi S, Yang C-C, Iwasaki W (2017) Generalist species drive microbial
1027 dispersion and evolution. *Nature communications*, **8**, 1162.
1028 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-
1029 analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

1030 Steele MI, Kwong WK, Whiteley M, Moran NA (2017) Diversification of Type VI
1031 Secretion System Toxins Reveals Ancient Antagonism among Bee Gut Microbes.
1032 *mBio*, **8**, e01630–17.
1033 Toft C, Andersson SGE (2010) Evolutionary microbial genomics: insights into
1034 bacterial host adaptation. *Nature Reviews Genetics*, **11**, 465–475.
1035 van Heel AJ, de Jong A, Montalban-Lopez M (2013). BAGEL3: automated
1036 identification of genes encoding bacteriocins and (non-)bactericidal
1037 posttranslationally modified peptides. *Nucleic Acid Research*, **41**, W448-53.
1038 Yin Y, Mao X, Yang J *et al.* dbCAN: a web resource for automated carbohydrate-active
1039 enzyme annotation. *Nucleic Acid Research*, **40**, W445-51.
1040 Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina
1041 Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.
1042 Zheng H, Nishida A, Kwong WK *et al.* (2016) Metabolism of Toxic Sugars by Strains
1043 of the Bee Gut Symbiont Gilliamella apicola. *mBio*, **7**, e01326–16.
1044 Zheng J, Ruan L, Sun M, Gänzle M (2015) A Genomic View of Lactobacilli and
1045 Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology.
1046 *Applied and environmental microbiology*, **81**, 7233–7243.
1047