Genomic signatures of honey bee association in an acetic acid symbiont

Eric A. Smith[a] and Irene L. G. Newton[a]#

[a]Department of Biology, Indiana University, Bloomington, Indiana, USA

Running head: Evolutionary signatures in an acetic acid symbiont

#Address correspondence to Irene L. G. Newton: irnewton@indiana.edu

ABSTRACT

Honey bee queens are central to the success and productivity of their colonies; queens are the only reproductive members of the colony, and therefore queen longevity and fecundity can directly impact overall colony health. Recent declines in the health of the honey bee have startled researchers and lay people alike as honey bees are the most important pollinators in agriculture. Honey bees are important pollinators of many major crops and add billions of dollars annually to the US economy through their services. One factor that may influence queen and colony health is the microbial community. Although honey bee worker guts have a characteristic community of bee-specific microbes, the honey bee queen digestive tracts are colonized predominantly by a single acetic acid bacterium: *Parasaccharibacter apium.* This bacterium is related to flower-associated microbes such as *Saccharibacter floricola,* and initial phylogenetic analyses placed it as sister to these environmental bacteria. We used comparative genomics of multiple bee-associated strains and the flower-associated *Saccharibacter* to identify genomic changes associated with the ecological transition to bee association. We identified several genomic differences in the bee-associated strains, including a complete CRISPR/Cas system. Many of the changes we note here are predicted to confer upon them the ability to survive in royal jelly and defend themselves against mobile elements, including phages. Our results are a first step towards identifying potential benefits provided by the honey bee queen microbiota to the matriarch of the colony.

45 **IMPORTANCE**

46 The health of the world's most important agricultural pollinator, the honey bee, depends on the
47 health of the matriarchs: honey bee queens. These queens are colonized by a specific microbe
48 not found in associated workers but instead, present in the diet given to queen bees. Here we
49 identified genomic signatures of the transition to bee association in this microbe, showing that
50 the queen-associated microbe has adapted to the acids present in honey and royal jelly through
51 the acquisition of acid resistance genes and an enzyme important in de-acidification. Our
52 results highlight potential benefits provided to the queen by her microbial associates.

53

54 **INTRODUCTION**

55 The honey bee (*Apis mellifera*) is extremely important economically because of the
56 pollination services it provides to numerous agricultural crops. As a result, there is increasing
57 interest in determining how the microbiome supports and influences bee function. While a
58 honey bee colony is made up of bees with diverse roles, or castes, the majority of studies on
59 bee microbiomes have focused on workers specifically. The microbial community of worker
60 bees consists of eight to ten core bacterial species (1-6). The characterization of these groups
61 led to speculation about their role in honey bee health and whether or not they provision
62 nutrients (5) or assist in the breakdown of plant-derived carbohydrates, as is the case in other
63 insect-microbe interactions (7-9). There has also been speculation as to the role of the
64 microbiome in resistance to pathogens, as microbial communities have been shown to protect
65 the bumble bee (*Bombus terristris*) from the parasite *Crithidia bombi* (10). Honey bee-
66 associated microbes interact with each other in diverse ways both *in vitro* and *in vivo*,

67 suggesting that they may interact syntrophically within workers (2, 11). While these studies
68 focused on honey bee workers are intriguing, it is surprising that only recently was the
69 microbiome of queen bees throughout development recently characterized (12).

70 Interestingly, the microbial community associated with queen bees is vastly different
71 than associated workers and is comprised of a large percentage of acetic acid bacteria, a group
72 of bacteria present only at very small percentages in workers. One of the primary bacteria that
73 differentiate queens from workers is the recently described *Parasaccharibacter apium* (13). *P.*
74 *apium* is in the family *Acetobacteraceae* and occupies defined niches within the hive, including:
75 queen guts, nurse hypopharyngeal glands, nurse crops, and royal jelly, and is only rarely found
76 outside of these areas (14, 15). Evidence suggests that it might play a role in protecting
77 developing larvae and queens from pathogens such as *Nosema* (13, 16). Given that *P. apium*
78 makes up a large proportion of the queen gut microbiome, it is possible that it plays an
79 important role in queen nutrition, protection from pathogens, and possibly modulating queen
80 fertility, fecundity, and longevity (17).

81 *P. apium* is part of a clade of acetic acid bacteria (AAB, a group within the family
82 *Acetobacteraceae*) that contains both free-living and bee-associated members. Comparative
83 genomics, then, can give us insights into the changes associated with the transition to bee-
84 association in this clade. This comparison can also help elucidate what sets *P. apium* apart from
85 closely related species and the role(s) it might be playing in the hive environment. To that end,
86 we used the genomes of four recently announced *P. apium* strains (18), an unpublished *P.*
87 *apium* genome assembly, as well as four genomes of the closely-related genus *Saccharibacter*
88 (19-21), and a genome of the bumblebee symbiont, *Bombella intestini* (22), to begin to tease

89 apart the unique capabilities of *P. apium*. Insights gained here could prove critical in
90 determining the factors responsible for maintaining queen health in colonies and could
91 ultimately lead to the development of interventions to improve queen health and mitigate the
92 detrimental impacts of queen failure on this economically critical species.
93
94 **RESULTS**
95 *Acetic acid bacteria phylogeny*
96 To robustly determine where the *Parasaccharibacter apium* strains and *Saccharibacter*
97 spp. are placed among the AAB, we constructed a maximum likelihood phylogeny using 16S
98 rRNA sequences derived from each of the 17 AAB genera. Our final tree largely agrees with
99 previously published ABB phylogenies (Figure 1) (23-26). Bootstrap support along the backbone
100 of the tree is low, likely owing to the inclusion of sequences from multiple strains of the same
101 species with little sequence divergence. Sequences were largely grouped into monophyletic
102 clades by genus, with one exception; the genus *Gluconacetobacter* is split into a paraphyletic
103 grade, with each branch dominated by a different *Gluconacetobacter* species (clades 1-5, Figure
104 1A). *Parasaccharibacter* plus *Saccharibacter* comes out as sister to the *Gluconobacter* clade,
105 and the clade encompassing all three of these genera (clade 6, Figure 1A) is sister to
106 *Acetobacter*. Within the *Parasaccharibacter*/*Saccharibacter* clade, the flower-associated *S.*
107 *floricola* is sister to the remaining clade (clade 7, Figure 1B), comprised of a mix of taxonomic
108 classifications from *Parasaccharibacter*, *Saccharibacter*, and *Bombella* species. The inconsistent
109 taxonomic nomenclature of the microbes within this clade is typified by the *Parasaccharibacter*
110 classification, which is applied to one *P. apium* strain (AS1), which clusters with a handful of

111 uncharacterized *Acetobacteraceae* bacteria (clade 8, Figure 1B), to the exclusion of all the other
112 *P. apium* and *Saccharibacter* strains. It should be noted that, while the genome for *P. apium*
113 AS1 is on GenBank, little else is known about this strain. Further taxonomic ambiguity is
114 highlighted by the presence of a member of a completely separate genus, *Bombella apis,*
115 grouped squarely among *P. apium* and *Saccharibacter* strains, indicating that it is likely a
116 member of one of those two genera. However, branch lengths within this clade are generally
117 short, especially among the main *Parasaccharibacter*/*Saccharibacter* clade (clade 9, Figure 1B),
118 making it difficult to resolve exact relationships.
119
120 *Core ortholog phylogeny of* Parasaccharibacter apium *and* Saccharibacter *strains*
121 We used OrthoMCL (v2.0.9 (27)) to define clusters of orthologous genes (COGs) using
122 the *Parasaccharibacter*, *Saccharibacter*, and *Bombella* genomes listed in Table 1; *Gluconobacter*
123 *oxydans* H24 was used as an outgroup (see Supplementary Information for more detail; Table
124 S1). To better resolve the phylogenetic relationships between *P. apium*, *Saccharibacter* spp.,
125 and *Bombella intestini*, we constructed a second maximum likelihood phylogeny using aligned
126 and concatenated amino acid sequences of the 1,259 single-copy COGs. This robustly supported
127 amino acid phylogeny broadly agrees with our previously constructed 16S phylogeny (Figure 2).
128 In the core ortholog tree, *B. intestini* interrupts the monophyly of honey bee-associated
129 *Saccharibacter* spp plus *P. apium*. Notably, this tree groups *B. intestini* more closely to the
130 majority of the *P. apium* strains, while *P. apium* AS1 is more distantly related. Similar to the 16S
131 tree, we again see quite short branch lengths within the bee-associated AAB species,

132 particularly among those in clade D, which includes all *P. apium* and *Saccharibacter* genomes
133 except *P. apium* AS1 and *S. floricola* (Figure 2).
134 Given the discrepancy between nomenclature and phylogeny and the short branch
135 lengths, we calculated genome-wide Average Nucleotide Identity (gANI) and aligned fraction
136 (AF) to clarify species relationships. All pairwise comparisons between genomes in clade D met
137 both gANI and AF thresholds for being considered the same species (namely, gANI > 96.5 and
138 AF > 0.6 (28)), while no other genome pairs reach both thresholds (Figures S1 and S2). Because
139 of the genetic similarity within clade D and phylogenetic distinction from the rest of the
140 genomes in this tree, the remainder of our analyses focus largely on this clade; for clarity we
141 will refer to this clade as the '*Parasaccharibacter* clade' from here forward.
142
143 *Signatures of bee association in the P. apium genomes*
144 To identify genes associated with the transition to bee association, we identified COGs
145 that contained at least one gene from each bee-associated AAB genome (*Parasaccharibacter*
146 clade, *B. intestini*, and *P. apium* AS1) and were also missing in *S. floricola*. There were a total of
147 1,286 COGs containing at least one gene from each of the aforementioned genomes, but only
148 89 were also missing in *S. floricola*. We determined the putative functions of these genes using
149 the *P. apium* reference genome representative for each COG (Table S2). It should be noted that
150 all annotations discussed from here forward are putative and require further functional
151 characterization.
152 Several bee-associated unique genes stood out as particularly interesting, the first being
153 gluconolactonase. Lactonases, such as gluconolactonase, reversibly catalyze the hydrolysis of

154 lactones (such as gluconolactone) to the hydroxyl acid form (such as gluconic acid).
155 Gluconolactone is found in both honey and royal jelly and is thought to be partially responsible
156 for the antibacterial properties of both compounds (29). In water, this compound can be
157 hydrolyzed into gluconic acid, acidifying the environment and preventing bacterial growth (30-
158 32). The presence of this gene capable of reversing this acidification – at least locally – may
159 explain how *Parasaccharibacter* is able to thrive in the presence of royal jelly (13). BLAST
160 searches of the metatranscriptomes and metagenomes of bacteria in the "core" honey bee
161 microbiome (33) resulted in zero hits, indicating that none of the "core" microbiome members
162 possess a homolog of this gene. The presence of gluconolactonase may help explain the unique
163 distribution of *Parasaccharibacter* within the hive. Another bee-associated unique gene is an
164 HdeD family acid-resistance protein, which in *E. coli* participates in resistance to acids at high
165 cell densities (34). The presence of this gene in *P. apium* may indicate an adaptation to living in
166 low pH environments – such as the queen bee digestive tract or royal jelly (35).
167 An AI-2 E family transporter was identified as unique to the bee-associated AAB. AI-2 is
168 an auto-inducer responsible for activating cascades associated with quorum sensing. While *P.*
169 *apium* does not contain any AI-2 synthesis genes, the presence of an AI-2 E family transporter
170 indicates that it may be responding to exogenous AI-2 produced by other bacteria, possibly in a
171 competitive interaction. Bolstering the competition hypothesis is the presence of fusaric acid
172 resistance (FUSC) genes in the *P. apium* genomes. Fusaric acid and its analogs can be quorum
173 sensing inhibitors (36), so the presence of FUSC genes might be an adaptation that allows *P.*
174 *apium* to evade quorum sensing inhibition attempts by other microbes. Alternatively, these
175 FUSC genes may play a role in competition with fungal species. Fusaric acid is produced by

176 several species of fungus and is antibacterial (37). Therefore, the FUSC genes may play a role in
177 *P. apium*'s protection of honey bee larvae and queens from infection with *Nosema* pathogens
178 (16).

179 An invasion-associated locus B (Ialb) protein was identified as present in bee-associated
180 AAB, but absent in *S. floricola*. In *Bartonella bacilliformis*, *ialb* mutants are impaired in their
181 ability to colonize human erythrocytes, suggesting a role for this protein in eukaryotic cell
182 invasion (38). While it is not clear whether *P. apium* is ever intracellular, the presence of *ialb*
183 suggests that it may have this capability.

184 The final set of genes of particular interest in this analysis is a complete Type I-E
185 CRISPR/Cas cassette. To determine if this CRISPR/Cas cassette was active, we annotated the
186 genomes for the presence of CRISPR arrays, and found that all of the genomes that have this
187 CRISPR/Cas cassette contain multiple CRISPR arrays. It is possible that these CRISPR arrays were
188 present in the most recent common ancestor of the *Parasaccharibacter* clade and have simply
189 been maintained in these current genomes; if that were the case, we would expect the spacers
190 in these CRISPR arrays to be highly similar between all strains. However, if these arrays are part
191 of an active CRISPR/Cas system, we would expect the spacers to differ from strain to strain,
192 reflecting unique challenges encountered by each strain. To rule out the possibility that these
193 arrays are ancestral, we aligned each spacer sequence from a given genome to all other spacer
194 sequences from the other genomes and calculated the percent identity. The minimum best
195 intergenomic match for any spacer was 40%, while the maximum was just 65% identical over
196 the length of the spacer, indicating that the spacer sequences are unique from genome to
197 genome and the CRISPR/Cas systems identified here are likely active.

198 *B. intestini* was isolated from a bumble bee gut, so we also looked at genes that were
199 unique to this bacterium to determine whether there are any obvious signatures of bumble bee
200 association in its genome. There were a total of 65 genes that were unique to *B. intestini*,
201 including a complete type IV secretion system (T4SS) and several genes involved in antibiotic
202 production or resistance. Putative annotations of these 65 genes are in Table S3.
203
204 *Identification of horizontally transferred gene regions*
205 Horizontal transfer of DNA between unrelated bacteria is a commonly known
206 mechanism by which bacteria can acquire new traits and adapt to novel environments (39-42).
207 We identified two regions of phage origin, one in *S. floricola* and one in the *P. apium* reference
208 genome (Figure 3; see Supplementary Information for more detail). Movement and insertion of
209 bacteriophage sequences in a genome can have profound effects on the evolution of that
210 genome (43-45) and future work will determine whether these phage are lytic and what their
211 host ranges may be. To determine whether the bacteria in the *Parasaccharibacter* clade have
212 undergone other potential horizontal gene transfer (HGT) events, we determined the spatial
213 distribution of genes of particular interest (e.g. clade-specific, species-specific, or strain-specific
214 genes) across the bacterial genomes (Figure 3). Some of the genes specific to different clades
215 occur in clusters, an indication that they may have originated elsewhere and been horizontally
216 inherited as a chunk of contiguous DNA. We then looked for anomalies in sequence
217 composition (%GC) and phylogeny to determine whether they were putatively horizontally
218 transferred. Using this combination of methods, we identified a total of five HGT regions in the

219 *Parasaccharibacter* clade, which we have numbered 1-5 (see Table S4 for %GC and lineage
220 probability index (LPI)-difference deviations for each gene in each HGT).
221 HGT1 (Figure 4A) is present in all genomes in the *Parasaccharibacter* clade, and contains
222 10 genes, although *P. apium* C6 is missing one of the genes (the second-to-last gene at the 3'
223 end of the HGT, annotated as an ABC transport auxiliary component). The three most 5' genes
224 show homology to YfaP (an uncharacterized conserved protein), SrfB (part of the surfactin
225 antibiotic synthesis machinery), and an uncharacterized bacterial virulence factor. The genes in
226 the 3' half of this HGT contain a number of domains involved in membrane transport. We
227 hypothesize that the two halves of this HGT work together to synthesize and export antibiotics
228 as a form of defense or regulation of competing bacteria. Lending support to the hypothesis
229 that this HGT is involved in defense or immunity is the fact that a CRISPR array lies immediately
230 5' of this HGT in each genome (Table S5). Bacterial defense mechanisms tend to occur in
231 clusters of "defense islands" (46), so the presence of this CRISPR array is perhaps a further
232 indication of this HGTs role in bacterial immunity.
233 HGTs 2 and 3 (Figure 4B and 4C) are restricted solely to the *P. apium* reference genome
234 and are both bacterial restriction-modification (R-M) systems. Bacterial R-M systems are a
235 defense against invading DNA (i.e. bacteriophage). They act by methylating host DNA at specific
236 sites; invading DNA with the same recognition site will be un-methylated, recognized as foreign,
237 and targeted for degradation (47). HGT2 contains 6 genes, which make up the core components
238 of a bacterial (R-M) system. Interestingly, the domain architecture in this R-M system has been
239 recognized as a precursor to eukaryotic defenses against transposable elements (48). HGT3
240 (Figure 4C) consists of 3 genes comprising 5 domains; the 5'-most gene consists of a predicted

241 restriction-modification DNA methylase coupled to a specificity domain, the middle gene is
242 predicted to be an XhoI restriction enzyme, and the 3'-most gene is a PHP phosphoesterase
243 coupled to a RecN DNA repair ATPase. Taken together, it appears that HGTs 2 and 3 are
244 responsible for recognition of and defense against foreign DNA.

245 HGT4 (Figure 4D) is present in all *P. apium* genomes in the *Parasaccharibacter* clade and
246 contains 3 genes: two GDP-D-mannose dehydratases (GMD) and an O-linked N-
247 acetylglucosamine transferase (OGT). GMD plays a role in the metabolism of mannose and
248 fructose, sugars commonly found in nectar (49). The presence of GMD in *P. apium* genomes
249 might allow for the consumption of nectar or nectar components by these bacteria. OGT, on
250 the other hand, plays a role in post-translational modification of thousands of identified
251 proteins (50). However, while OGT-mediated post-translational modification is common in
252 eukaryotes, it is far more rare in bacteria (51). To date, only a handful of prokaryotic OGTs have
253 been identified, and the targets of these OGTs remain unclear (52, 53). Given the role OGTs
254 play in eukaryotic post-translational modification and the fact that many bacterial effector
255 proteins show homology to eukaryotic proteins (54), it is possible that the presence of OGT in *P.*
256 *apium* represents a pathway for host-microbe interaction and symbiont-mediated protein
257 modification.

258 HGT5 (Figure 4E) is unique to the *P. apium* strains A29, B8, and C6, all strains that had
259 been isolated from honey bee larvae. Like HGTs 1-3, HGT5 contains genes that may play a role
260 in protection against foreign DNA. There are four genes in the 5' section of HGT5, three of
261 which are kinases, and the fourth contains a SAD/SRA domain in its 5' end, and an HNH
262 endonuclease domain in its 3' end. In bacteria, the SAD/SRA domain is often found associated

263 with an HNH domain (55) and it is thought that the two domains act together to recognize and
264 cleave foreign DNA (56). The 3' section of HGT5 consists of a conjugative relaxase, a TraG/TraD
265 family ATPase (a coupling protein involved in bacterial conjugation and/or T4SS), a homolog of
266 the pyocin activator protein PrtN, a homolog of a yeast RNA polymerase I subunit, and two
267 additional genes with no annotations. The presence of a PrtN homolog is particularly
268 interesting, as in *Pseudomonas aeruginosa* pyocins are antibacterial agents, often acting to
269 depolarize the membrane of target cells (57),(58). Interestingly, one of the two unannotated
270 genes in the 3' region of HGT5 shows weak homology to a phage shock protein, which are
271 proteins involved in the response to stress that may weaken the energy status of the cell (59).
272 This protein, then, may play a part in immunity to membrane depolarization. Given the
273 presence in HGT5 of: an HNH endonuclease coupled to a SAD/SRA domain, a conjugative
274 relaxase, a TraG/TraD family ATPase, a pyocin activator protein, and a protein with at least
275 some homology to a phage shock protein, we hypothesize that it may play a role in
276 pathogenesis or defense.
277
278 **DISCUSSION**
279 Here, we used the genomes of five *P. apium* strains, four *Saccharibacter* strains, and the
280 closely related *B. intestini* to gain insight into the genomic changes associated with the
281 transition to honey bee symbiosis in this group. We note several genomic differences – some of
282 which were horizontally acquired – between bee-associated bacteria and the flower-associated
283 *S. floricola* that may have allowed for the expansion of *P. apium* into previously unoccupied
284 niches within the honey bee colony. These differences can be classified as changes that

285 introduce: 1) novel metabolic capabilities, 2) defense and/or virulence mechanisms, and 3)
286 mechanisms for interaction with other microbes and/or the host.

287 Metabolic genes identified here include gluconolactonase, which may allow for the de-
288 acidification of royal jelly (29-32), and two copies of GMD, a gene that plays a role in the
289 metabolism of mannose and fructose, components of nectar and honey (49). Distinct defense
290 and/or virulence mechanisms were identified, including: a functional CRISPR/Cas system, two R-
291 M systems, and an HGT with some homology to known virulence mechanisms. Interestingly,
292 the R-M systems were identified in the only genome in the clade that also contains a phage
293 sequence (the *P. apium* reference sequence). Restriction modification systems, like phages, can
294 act as selfish genetic elements (60), so their presence in this genome may indicate that it was
295 historically more permissive to invading DNA. These R-M systems may also have been coopted
296 by the prophage to prevent super-infection with additional phages (61).

297 Genes involved in the interaction with other microbes and/or the host that we identified
298 include: an AI-2 family transporter, fusaric acid resistance genes, *ialb*, and *ogt*. Given that *P.*
299 *apium* does not encode any of the canonical genes for the production of quorum-sensing
300 molecules, it seems likely that *P. apium* is responding to exogenous AI-2 (and/or fusaric acid
301 and its analogs) produced by other members of the bee microbiome (62). The *ialb* and *ogt*
302 genes provide routes for interaction with the host, as *ialb* may play a role in eukaryotic cell
303 invasion (38) and *ogt* is known to modify thousands of eukaryotic proteins (50). Taken together,
304 we hypothesize that the novel combination of metabolic, quorum-sensing, defense/virulence,
305 and eukaryotic interaction genes in the *Parasaccharibacter* clade genomes allowed for the

306 utilization of a unique food source and protection from an onslaught of previously un-
307 encountered challenges and facilitated the transition to honey bee association in this clade.
308 *P. apium* has been shown to benefit honey bee larval development and provide
309 protection against *Nosema* (16). Some of the genes identified here, while allowing *P. apium* to
310 transition to honey bee symbiosis, may also be related to its ability to protect the bee host from
311 infection with *Nosema* or other pathogens. If indeed these genes are responsible for the
312 transition to, and maintenance of, honey bee symbiosis, we would expect to see a modified
313 evolutionary trajectory relative to those genes not involved in the symbiosis. We currently lack
314 sufficient sampling of non-bee-associated bacteria in this clade to do such analyses; however,
315 future studies addressing this question should allow for further elucidation of the genes
316 involved in the transition to honey bee association. Those analyses, coupled with functional
317 characterization of the genes of interest identified here, should lay the foundation for the
318 development of beneficial intervention strategies in this economically critical insect.
319
320 **METHODS**
321 *Acetic acid bacteria phylogeny*
322 To determine the placement of *Parasaccharibacter* and *Saccharibacter* among the AAB,
323 we downloaded all 16S rRNA sequences from the Silva database (63-65) that met the following
324 criteria: 1) from a species belonging to one of the seventeen genera of AAB (26), 2) length at
325 least 1200 bases, and 3) sequence quality >90. Additionally, the 16S sequence for *Rhodopila*
326 *globiformis* – which is in the family *Acetobacteraceae* but is not part of AAB – was included as
327 an outgroup. Given the close relation to *Saccharibacter floricola*, 16S sequences for *Bombella*

328 *intestini* (66) and *Bombella sp.* MRM1 (*Bombella apis*)(67) were included. We BLASTed the
329 *Saccharibacter floricolca* 16S sequence against the *Parasaccharibacter* and *Saccharibacter*
330 genomes (Table 1) to pull out their respective 16S sequences for use in this phylogeny. All
331 sequences were aligned using the SINA aligner (68); parameters used were set using the --auto
332 option. A maximum likelihood phylogeny was constructed using RAxML with the GTRGAMMA
333 substitution model and 1000 bootstrap replicates (v8.2.11, (69)). The final tree was visualized
334 using FigTree (v1.4.2, http://tree.bio.ed.ac.uk/software/figtree/).
335
336 *Orthology analysis*
337 To facilitate downstream analyses, we clustered genes from all genomes in Table 1 –
338 plus *Gluconobacter oxydans* H24 as an outgroup – into clusters of orthologous genes (COGs)
339 using OrthoMCL (v.2.0.9, (27)). Amino acid sequences were downloaded from NCBI and
340 clustering was performed using default OrthoMCL parameters. These clusters were then
341 classified as single-copy orthologs (defined as containing exactly one representative from each
342 genome), variable (defined as missing a representative from at least one genome and having
343 varying numbers of representatives from each of the other genomes), multi-copy ortholog
344 (containing at least one representative from each genome, but multiple copies from at least
345 one genome), or genome-specific (containing at least two genes that all came from the same
346 genome) using an in-house Perl script.
347
348
349

350 Parasaccharibacter *and* Saccharibacter *core ortholog phylogeny*

351 We constructed a phylogeny using concatenated amino acid alignments of all single-
352 copy COGs. The amino acid sequences were aligned using the MAFFT L-INS-I algorithm (v7.310,
353 (70)), and alignments were then concatenated, and used to construct a maximum likelihood
354 phylogeny using RAxML with substitution model PROTGAMMALGF and 1000 bootstrap
355 replicates (v8.2.11, (69)). The final tree was visualized using FigTree (v1.4.2,
356 http://tree.bio.ed.ac.uk/software/figtree/).

357

358 *Calculation of genomic similarity*

359 To determine relatedness and species assignment, we calculated genome-wide Average
360 Nucleotide Identity (gANI) and aligned fraction (AF) for each pairwise comparison using
361 ANIcalculator (28). Predicted transcript sequences for each pairwise comparison were passed to
362 the software, which output gANI and AF in each direction for the pairwise comparison. As gANI
363 and AF can vary depending on the direction of comparison due to differences in genome length,
364 we report the average of the pairwise calculations in each direction.

365

366 *Synteny analysis*

367 We used Mauve (71, 72) to determine the syntenic regions between the
368 *Parasaccharibacter apium* and *Saccharibacter* spp. genomes. The *Parasaccharibacter apium*
369 reference genome is resolved to a single chromosome, so it was used as the reference
370 sequence in Mauve's "move contigs" tool, and the likely order and orientation of contigs in the
371 other genomes was determined. To facilitate downstream analyses, the output of Mauve's

372 “move contigs” tool was used to order, orient, and concatenate contigs into single pseudo-
373 chromosomes for each genome. Structural rearrangements were then visualized using Mauve’s
374 built-in graphical interface.
375
376 *Annotation of CRISPR arrays and phage sequences*
377 Pseudo-chromosomes for each genome were uploaded to CRISPRFinder to determine
378 location and sequence of CRISPR arrays (73). We used an in-house Perl script to determine the
379 maximum intergenomic percent identity of spacer sequences. We used PHAge Search Tool
380 Enhanced Release (PHASTER) (74, 75) to define phage-like regions. Any region determined to be
381 “questionable” or “intact” by PHASTER that also appeared as an insertion in the host genome in
382 our synteny analysis was labeled as likely to be of phage origin.
383
384 *Determination of bee-associated bacteria-specific orthologs*
385 We identified all COGs that contained at least one gene from each genome of bee-
386 associated bacteria and no genes from *S. floricola.* We then took the *P. apium* reference
387 genome representative for each of these COGs and got KEGG annotations for as many as
388 possible using BlastKOALA (76). For those genes that we were not able to get KEGG
389 annotations, we used NCBI’s BLAST to aid in determining potential function of these bee-
390 associated bacteria-specific genes. This list of genes and their potential functions was then
391 manually curated to hypothesize genes that may have allowed for the transition to bee-
392 association.
393

394 *Analysis of horizontal gene transfers*

395 To determine whether or not genes in any of the *Parasaccharibacter* and *Saccharibacter*
396 genomes arrived via horizontal transfer, we employed a combination of sequence-composition,
397 phylogenetic, and synteny approaches. We mapped genes of particular interest (e.g. genes
398 unique to certain clades, species, or strains) to their locations on the linear pseudo-
399 chromosomes constructed during synteny analysis. Additionally, we calculated the %GC for
400 each gene. We then determined how many standard deviations each gene was from the
401 genome-wide mean %GC. The third prong of this analysis involved identifying genes that were
402 phylogenetically aberrant. To do this, we used Darkhorse (77) to calculate the lineage
403 probability index (LPI) twice for each gene, once including BLAST hits to *Parasaccharibacter* and
404 *Saccharibacter* subject sequences, and once excluding such hits. In doing so, genes that are
405 likely to be horizontally transferred will have a larger discrepancy between LPI values than
406 genes that were vertically inherited (see supplemental methods for details). We then identified
407 regions as likely to be HGTs if they met the following criteria: 1) a block of at least three
408 syntenic genes that show interesting phylogenetic distributions (e.g. unique to clade, species,
409 or strain) where 2) a majority of genes in the region are at least 1 standard deviation from the
410 mean %GC or LPI difference (or both).

411

412 *Domain annotation of genes of interest*

413 We used HHpred (https://toolkit.tuebingen.mpg.de/#/tools/hhpred, (78)) to determine
414 domain architecture and gain an understanding of potential function of the genes in each HGT.
415 For genes of interest that were part of a COG, all members of the COG were first aligned using

416 the MAFFT L-INS-I algorithm (v7.310, (70)). These multiple sequence alignments (or single
417 amino acid sequences in the case of strain-unique genes) were then uploaded to HHpred's
418 online tool and homology was determined using HMMs in the COG_KOG_v1.0, Pfam-A_v31.0,
419 and SMART_v6.0 databases; only domains scoring above 60% probability are discussed here.
420 Gene models for each region of interest were then constructed and visualized using the HHpred
421 results and in-house R scripts. HGT5 occurs at the junction of two contigs in the linear pseudo-
422 chromosomes we constructed. The abutting ends of each contig have annotations for partial
423 pseudogenes, such that when they are joined a complete gene is created. We BLASTed the
424 nucleotide sequence of this gene against the NCBI nr database to determine a putative
425 function.
426

427 **ACKNOWLEDGEMENTS**

428 We thank Amelia R.I. Lindsey for feedback on initial versions of this manuscript. E.A.S. was
429 supported by a Faculty Research Support grant from Indiana University.
430

431 **REFERENCES**

432 1. Martinson VG, Danforth BN, Minckley RL, Rueppell O, Tingek S, Moran NA. 2011. A
433 simple and distinctive microbiota associated with honey bees and bumble bees. Mol Ecol
434 20:619–628.

435 2. Martinson VG, Moy J, Moran NA. 2012. Establishment of Characteristic Gut Bacteria
436 during Development of the Honeybee Worker. Appl Environ Microbiol 78:2830–2840.

437 3. Moran NA, Hansen AK, Powell JE, Sabree ZL. 2012. Distinctive Gut Microbiota of Honey
438 Bees Assessed Using Deep Sampling from Individual Worker Bees. PLoS ONE 7.
439

440 4. Sabree ZL, Hansen AK, Moran NA. 2012. Independent Studies Using Deep Sequencing
441 Resolve the Same Set of Core Bacterial Species Dominating Gut Communities of Honey
442 Bees. PLoS ONE 7.

443 5. Moran NA. 2015. Genomics of the honey bee microbiome. Curr Opin Insect Sci 10:22–28.

444 6. Anderson KE, Rodrigues PAP, Mott BM, Maes P, Corby-Harris V. 2016. Ecological
445 Succession in the Honey Bee Gut: Shift in Lactobacillus Strain Dominance During Early
446 Adult Development. Microb Ecol 71:1008–1019.

447 7. McCutcheon JP, Moran NA. 2007. Parallel genomic evolution and metabolic
448 interdependence in an ancient symbiosis. Proc Natl Acad Sci USA 104:19392–19397.

449 8. Gündüz EA, Douglas AE. 2009. Symbiotic Bacteria Enable Insect to Use a Nutritionally
450 Inadequate Diet. Proceedings: Biological Sciences 276:987–991.

451 9. Douglas AE. 2013. Microbial Brokers of Insect-Plant Interactions Revisited. J Chem Ecol
452 39:952–961.

453 10. Koch H, Schmid-Hempel P. 2011. Socially transmitted gut microbiota protect bumble
454 bees against an intestinal parasite. Proc Natl Acad Sci USA 108:19288–19292.

455 11. Rokop ZP, Horton MA, Newton ILG. 2015. Interactions between Cooccurring Lactic Acid
456 Bacteria in Honey Bee Hives. Appl Environ Microbiol 81:7261–7270.

457 12. Tarpy DR, Mattila HR, Newton ILG. 2015. Development of the Honey Bee Gut
458 Microbiome throughout the Queen-Rearing Process. Appl Environ Microbiol 81:3182–
459 3191.

460 13. Corby-Harris V, Snyder LA, Schwan MR, Maes P, McFrederick QS, Anderson KE. 2014.
461 Origin and Effect of Alpha 2.2 Acetobacteraceae in Honey Bee Larvae and Description of
462 Parasaccharibacter apium gen. nov., sp nov. Appl Environ Microbiol 80:7460–7472.

463 14. Anderson KE, Sheehan TH, Mott BM, Maes P, Snyder L, Schwan MR, Walton A, Jones BM,
464 Corby-Harris V. 2013. Microbial Ecology of the Hive and Pollination Landscape: Bacterial
465 Associates from Floral Nectar, the Alimentary Tract and Stored Food of Honey Bees (Apis
466 mellifera). PLoS ONE 8.

467 15. Vojvodic S, Rehan SM, Anderson KE. 2013. Microbial Gut Diversity of Africanized and
468 European Honey Bee Larval Instars. PLoS ONE 8.

469 16. Corby-Harris V, Snyder L, Meador CAD, Naldo R, Mott B, Anderson KE. 2016.
470 Parasaccharibacter apium, gen. nov., sp nov., Improves Honey Bee (Hymenoptera:
471 Apidae) Resistance to Nosema. J Econ Entomol 109:537–543.

472

473 17. Anderson KE, Ricigliano VA, Mott BM, Copeland DC, Floyd AS, Maes P. 2018. The queen's
474 gut refines with age: longevity phenotypes in a social insect model. Microbiome 6:108.

475 18. Corby-Harris V, Anderson KE. 2018. Draft Genome Sequences of Four Parasaccharibacter
476 apium Strains Isolated from Honey Bees. Genome Announc 6:e00165–18.

477 19. Jojima Y, Mihara Y, Suzuki S, Yokozeki K, Yamanaka S, Fudou R. 2004. Saccharibacter
478 floricola gen. nov., sp nov., a novel osmophilic acetic acid bacterium isolated from pollen.
479 Int J Syst Evol Microbiol 54:2263–2267.

480 20. Chouaia B, Gaiarsa S, Crotti E, Comandatore F, Degli Esposti M, Ricci I, Alma A, Favia G,
481 Bandi C, Daffonchio D. 2014. Acetic Acid Bacteria Genomes Reveal Functional Traits for
482 Adaptation to Life in Insect Guts. Genome Biol Evol, 3rd ed. 6:912–920.

483 21. Veress A, Wilk T, Kiss J, Olasz F, Papp PP. 2017. Draft Genome Sequences of
484 Saccharibacter sp. Strains 3.A.1 and M18 Isolated from Honey and a Honey Bee (Apis
485 mellifera) Stomach. Genome Announc 5:e00744–17.

486 22. Li L, Illeghems K, Van Kerrebroeck S, Borremans W, Cleenwerck I, Smagghe G, De Vuyst L,
487 Vandamme P. 2016. Whole-Genome Sequence Analysis of Bombella intestini LMG
488 28161(T), a Novel Acetic Acid Bacterium Isolated from the Crop of a Red-Tailed Bumble
489 Bee, Bombus lapidarius. PLoS ONE 11.

490 23. Roh SW, Nam Y-D, Chang H-W, Kim K-H, Kim M-S, Ryu J-H, Kim S-H, Lee W-J, Bae J-W.
491 2008. Phylogenetic characterization of two novel commensal bacteria involved with
492 innate immune homeostasis in Drosophila melanogaster. Appl Environ Microbiol
493 74:6171–6177.

494 24. Cleenwerck I, De Vos P, De Vuyst L. 2010. Phylogeny and differentiation of species of the
495 genus Gluconacetobacter and related taxa based on multilocus sequence analyses of
496 housekeeping genes and reclassification of Acetobacter xylinus subsp. sucrofermentans
497 as Gluconacetobacter sucrofermentans (Toyosaki et al. 1996) sp. nov., comb. nov. Int J
498 Syst Evol Microbiol 60:2277–2283.

499 25. Matsutani M, Hirakawa H, Yakushi T, Matsushita K. 2011. Genome-wide phylogenetic
500 analysis of Gluconobacter, Acetobacter, and Gluconacetobacter. FEMS Microbiol Lett
501 315:122–128.

502 26. Yamada Y. 2016. Systematics of Acetic Acid Bacteria, pp. 1–50. *In* Acetic Acid Bacteria,
503 3rd ed. Springer, Tokyo, Tokyo.

504 27. Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for
505 eukaryotic genomes. Genome Res 13:2178–2189.

506

507 28. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC,
508 Pati A. 2015. Microbial species delineation using whole genome sequences. Nucleic Acids
509 Res 43:6761–6771.

510 29. Sagona S, Turchi B, Fratini F, Giusti M, Bull BT. 2015. Preliminary evaluation of glucose
511 oxidase and its products in vitro antimicrobial activities on Paenibacillus larvae ATCC9545
512 vegetative form. Bulletin of Insectology 68:233–237.

513 30. Furusawa T, Rakwal R, Nam HW, Shibato J, Agrawal GK, Kim YS, Ogawa Y, Yoshida Y,
514 Kouzuma Y, Masuo Y, Yonekura M. 2008. Comprehensive Royal Jelly (RJ) Proteomics
515 Using One- and Two-Dimensional Proteomics Platforms Reveals Novel RJ Proteins and
516 Potential Phospho/Glycoproteins. J Proteome Res 7:3194–3229.

517 31. Li J, Wang T, Zhang Z, Pan Y. 2007. Proteomic analysis of royal jelly from three strains of
518 western honeybees (Apis mellifera). J Agric Food Chem 55:8411–8422.

519 32. Schönleben S, Sickmann A, Mueller MJ, Reinders J. 2007. Proteome analysis of Apis
520 mellifera royal jelly. Anal Bioanal Chem 389:1087–1093.

521 33. Lee FJ, Miller KI, McKinlay JB, Newton ILG. 2018. Differential carbohydrate utilization and
522 organic acid production by honey bee symbionts. FEMS Microbiol Ecol.

523 34. Mates AK, Sayed AK, Foster JW. 2007. Products of the Escherichia coli acid fitness island
524 attenuate metabolite stress at extremely low pH and mediate a cell density-dependent
525 acid resistance. J Bacteriol 189:2759–2768.

526 35. Anderson KE, Sheehan TH, Eckholm BJ, Mott BM, DeGrandi-Hoffman G. 2011. An
527 emerging paradigm of colony health: microbial balance of the honey bee and hive (*Apis*
528 *mellifera*). Insect Soc 58:431–444.

529 36. Tung TT, Jakobsen TH, Dao TT, Fuglsang AT, Givskov M, Christensen SB, Nielsen J. 2017.
530 Fusaric acid and analogues as Gram-negative bacterial quorum sensing inhibitors. Eur J
531 Med Chem 126:1011–1020.

532 37. Crutcher FK, Puckhaber LS, Stipanovic RD, Bell AA, Nichols RL, Lawrence KS, Liu J. 2017.
533 Microbial Resistance Mechanisms to the Antibiotic and Phytotoxin Fusaric Acid. J Chem
534 Ecol 43:996–1006.

535 38. Coleman SA, Minnick MF. 2001. Establishing a direct role for the Bartonella bacilliformis
536 invasion-associated locus B (IalB) protein in human erythrocyte parasitism. Infect Immun
537 69:4373–4381.

538 39. Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene
539 transfer. Mol Biol Evol 19:2226–2238.

540

541 40. Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of
542 horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–766.

543 41. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic
544 transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976.

545 42. Roberts AP, Kreth J. 2014. The impact of horizontal gene transfer on the adaptive ability
546 of the human oral microbiome. Front Cell Infect Microbiol 4:1.

547 43. Casas V, Maloy S. 2011. Role of bacteriophage-encoded exotoxins in the evolution of
548 bacterial pathogens. Future Microbiol 6:1461–1473.

549 44. Koskella B, Brockhurst MA. 2014. Bacteria-phage coevolution as a driver of ecological and
550 evolutionary processes in microbial communities. FEMS Microbiol Rev 38:916–931.

551 45. Harrison E, Hall JPJ, Paterson S, Spiers AJ, Brockhurst MA. 2017. Conflicting selection
552 alters the trajectory of molecular evolution in a tripartite bacteria-plasmid-phage
553 interaction. Mol Ecol 26:2757–2764.

554 46. Makarova KS, Wolf YI, Snir S, Koonin EV. 2011. Defense islands in bacterial and archaeal
555 genomes and prediction of novel defense systems. J Bacteriol 193:6039–6056.

556 47. Rodic A, Blagojevic B, Zdobnov E, Djordjevic M, Djordjevic M. 2017. Understanding key
557 features of bacterial restriction-modification systems through quantitative modeling.
558 BMC Systems Biology 2017 11:1 11:1–15.

559 48. Iyer LM, Abhiman S, Aravind L. 2008. MutL homologs in restriction-modification systems
560 and the origin of eukaryotic MORC ATPases. Biol Direct 3.

561 49. Freeman CE, Madroño RW. 1985. Some floral nectar-sugar compositions of species from
562 southeastern Arizona and southwestern New Mexico. Madrono 32:78–86.

563 50. Love DC, Hanover JA. 2005. The hexosamine signaling pathway: deciphering the "O-
564 GlcNAc code". Sci STKE 2005:re13–re13.

565 51. Ostrowski A, Gundogdu M, Ferenbach AT, Lebedev AA, van Aalten DMF. 2015. Evidence
566 for a Functional O-Linked N-Acetylglucosamine (O-GlcNAc) System in the Thermophilic
567 Bacterium Thermobaculum terrenum. J Biol Chem 290:30291–30305.

568 52. Shen A, Kamp HD, Grundling A, Higgins DE. 2006. A bifunctional O-GlcNAc transferase
569 governs flagellar motility through anti-repression. Genes Dev 20:3283–3295.

570 53. Sokol KA, Olszewski NE. 2015. The Putative Eukaryote-Like O-GlcNAc Transferase of the
571 Cyanobacterium Synechococcus elongatus PCC 7942 Hydrolyzes UDP-GlcNAc and Is
572 Involved in Multiple Cellular Processes. J Bacteriol 197:354–361.

573 54. Galán JE. 2009. Common themes in the design and function of bacterial effectors. Cell
574 Host Microbe 5:571–579.

575 55. Makarova KS, Aravind L, Wolf YI, Tatusov RL, Minton KW, Koonin EV, Daly MJ. 2001.
576 Genome of the extremely radiation-resistant bacterium Deinococcus radiodurans viewed
577 from the perspective of comparative genomics. Microbiol Mol Biol Rev 65:44–79.

578 56. Iyer LM, Abhiman S, Aravind L. 2011. Natural History of Eukaryotic DNA Methylation
579 Systems. Progress in Molecular Biology and Translational Science 101:25–104.

580 57. Jacob F. 1954. Induced biosynthesis and mode of action of a pyocine, antibiotic produced
581 by Pseudomonas aeruginosa. Ann Inst Pasteur (Paris) 86:149–160.

582 58. Michel-Briand Y, Baysse C. 2002. The pyocins of Pseudomonas aeruginosa. Biochimie
583 84:499–510.

584 59. Flores-Kim J, Darwin AJ. 2016. The Phage Shock Protein Response. Annu Rev Microbiol
585 70:83–101.

586 60. Kobayashi I. 2001. Behavior of restriction-modification systems as selfish mobile
587 elements and their impact on genome evolution. Nucleic Acids Res 29:3742–3756.

588 61. Van Etten JL, Xia YN, Burbank DE, Narva KE. 1988. Chlorella Viruses Code for Restriction
589 and Modification Enzymes. Gene 74:113–115.

590 62. Newton I, Miller KI, Franklin CD, Mattila HR. 2018. Social communication between
591 microbes colonizing the social honey bee Apis mellifera. bioRxiv 287995.

592 63. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Gloeckner FO. 2013.
593 The SILVA ribosomal RNA gene database project: improved data processing and web-
594 based tools. Nucleic Acids Res 41:D590–D596.

595 64. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig
596 W, Gloeckner FO. 2014. The SILVA and "All-species Living Tree Project (LTP)" taxonomic
597 frameworks. Nucleic Acids Res 42:D643–D648.

598 65. Gloeckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, Bruns G, Yarza P, Peplies
599 J, Westram R, Ludwig W. 2017. 25 years of serving the community with ribosomal RNA
600 gene reference databases and tools. J Biotechnol 261:169–176.

601 66. Li L, Praet J, Borremans W, Nunes OC, Manaia CM, Cleenwerck I, Meeus I, Smagghe G, De
602 Vuyst L, Vandammel P. 2015. Bombella intestini gen. nov., sp nov., an acetic acid
603 bacterium isolated, from bumble bee crop. Int J Syst Evol Microbiol 65:267–273.

604

605 67. Yun J-H, Lee J-Y, Hyun D-W, Jung M-J, Bae J-W. 2017. Bombella apis sp nov., an acetic
606 acid bacterium isolated from the midgut of a honey bee. Int J Syst Evol Microbiol
607 67:2184–2188.

608 68. Pruesse E, Peplies J, Gloeckner FO. 2012. SINA: Accurate high-throughput multiple
609 sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829.

610 69. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
611 with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

612 70. Katoh K, Misawa K, Kuma KI, Miyata T. 2002. MAFFT: a novel method for rapid multiple
613 sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.

614 71. Darling A, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved
615 genomic sequence with rearrangements. Genome Res 14:1394–1403.

616 72. Darling AE, Mau B, Perna NT. 2010. progressiveMauve: Multiple Genome Alignment with
617 Gene Gain, Loss and Rearrangement. PLoS ONE 5.

618 73. Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered
619 regularly interspaced short palindromic repeats. Nucleic Acids Res 35:W52–W57.

620 74. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. 2011. PHAST: A Fast Phage Search Tool.
621 Nucleic Acids Res 39:W347–W352.

622 75. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better,
623 faster version of the PHAST phage search tool. Nucleic Acids Res 44:W16–W21.

624 76. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for
625 Functional Characterization of Genome and Metagenome Sequences. Journal of
626 Molecular Biology 428:726–731.

627 77. Podell S, Gaasterland T. 2007. DarkHorse: a method for genome-wide prediction of
628 horizontal gene transfer. Genome Biol 8.

629 78. Soding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics
630 21:951–960.

631 79. Ge X, Zhao Y, Hou W, Zhang W, Chen W, Wang J, Zhao N, Lin J, Wang W, Chen M, Wang
632 Q, Jiao Y, Yuan Z, Xiong X. 2013. Complete Genome Sequence of the Industrial Strain
633 Gluconobacter oxydans H24. Genome Announc 1:e00003–13–e00003–13.

634

635

Table 1. Genome names, accession number, and isolation sources for complete genomes used in these analyses. *Parasaccharibacter apium* G7_7_3c is the *P. apium* reference genome.

Genome	GenBank accession number	Isolation source
Parasaccharibacter apium G7_7_3c	GCA_002079945.1	*Apis mellifera* hindgut(18)
Parasaccharibacter apium A29	GCA_002917995.1	*Apis mellifera* larva(18)
Parasaccharibacter apium B8	GCA_002917945.1	*Apis mellifera* larva(18)
Parasaccharibacter apium C6	GCA_002917985.1	*Apis mellifera* larva(18)
Parasaccharibacter apium AS1	GCA_002592045.1	*Apis melliera* larva
Saccharibacter sp. AM169	GCA_000723565.1	*Apis mellifera* stomach(20)
Saccharibacter sp. M18	GCA_002150105.1	*Apis mellifera* stomach(21)
Saccharibacter sp. 3.A.1	GCA_002150125.1	Honey(21)
Saccharibacter floricola	GCA_000378165.1	Flower(19)
Bombella intestini	GCA_002003665.1	*Bombus lapidarius* crop(22)
Gluconobacter oxydans H24	GCA_000311765.1	Industrial sample(79)

FIGURES AND LEGENDS

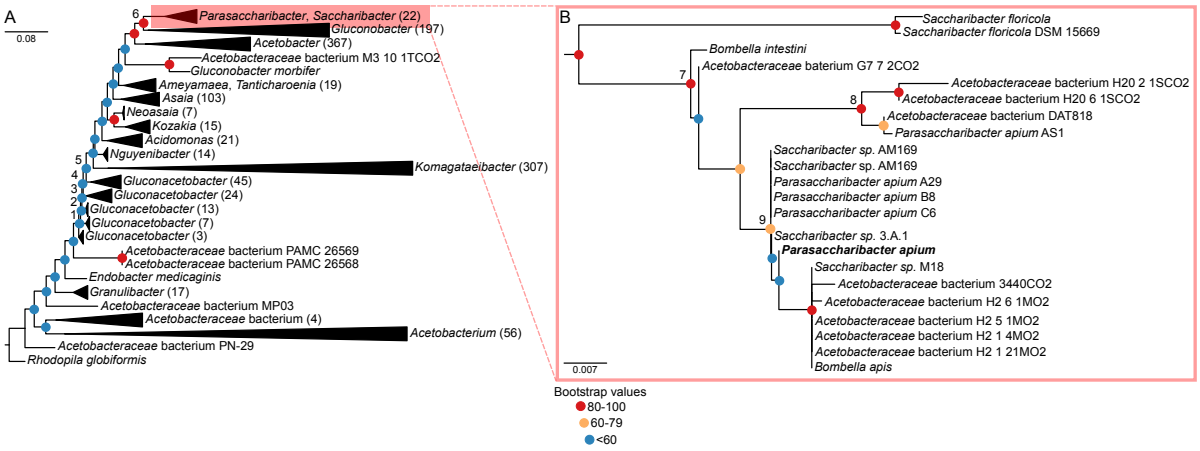

Figure 1. Maximum likelihood phylogenetic tree of Acetic Acid Bacteria constructed from full-length 16S rRNA sequences. Bootstrap scores are indicated at each node, and numbers at certain nodes are for reference in the main text. A) Full tree of all 17 genera of Acetic Acid Bacteria. Clades were collapsed based on genus. Genera that make up at least 40% of the sequences in a clade are listed, except for the highlighted clade. Numbers in parentheses represent the total number of sequences in each collapsed clade (Supplemental File S3 for full tree). B) Zoomed-in look at the *Parasaccharibacter*/*Saccharibacter* clade. The bolded *Parasaccharibacter apium* represents the reference *P. apium* sequence.

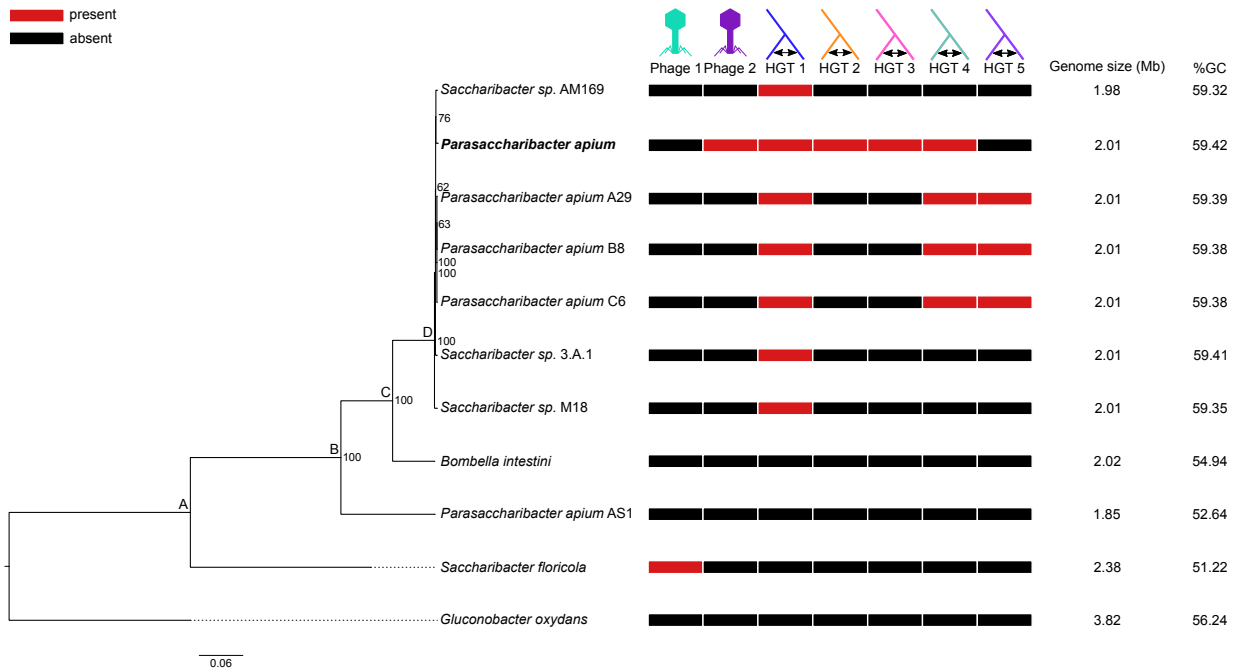

Figure 2. Maximum likelihood phylogenetic tree of the *Parasaccharibacter*/*Saccharibacter* clade constructed from concatenated amino acid alignment of 1259 single copy orthologous genes. Bootstrap scores are indicated at each node. Letters at certain nodes are for reference in the main text. Colored boxes represent the presence (red) or absence (black) of each of 7 genomic regions of interest. Genome size and %GC are also displayed. Bolded *Parasaccharibacter apium* represents the reference *P. apium* sequence.

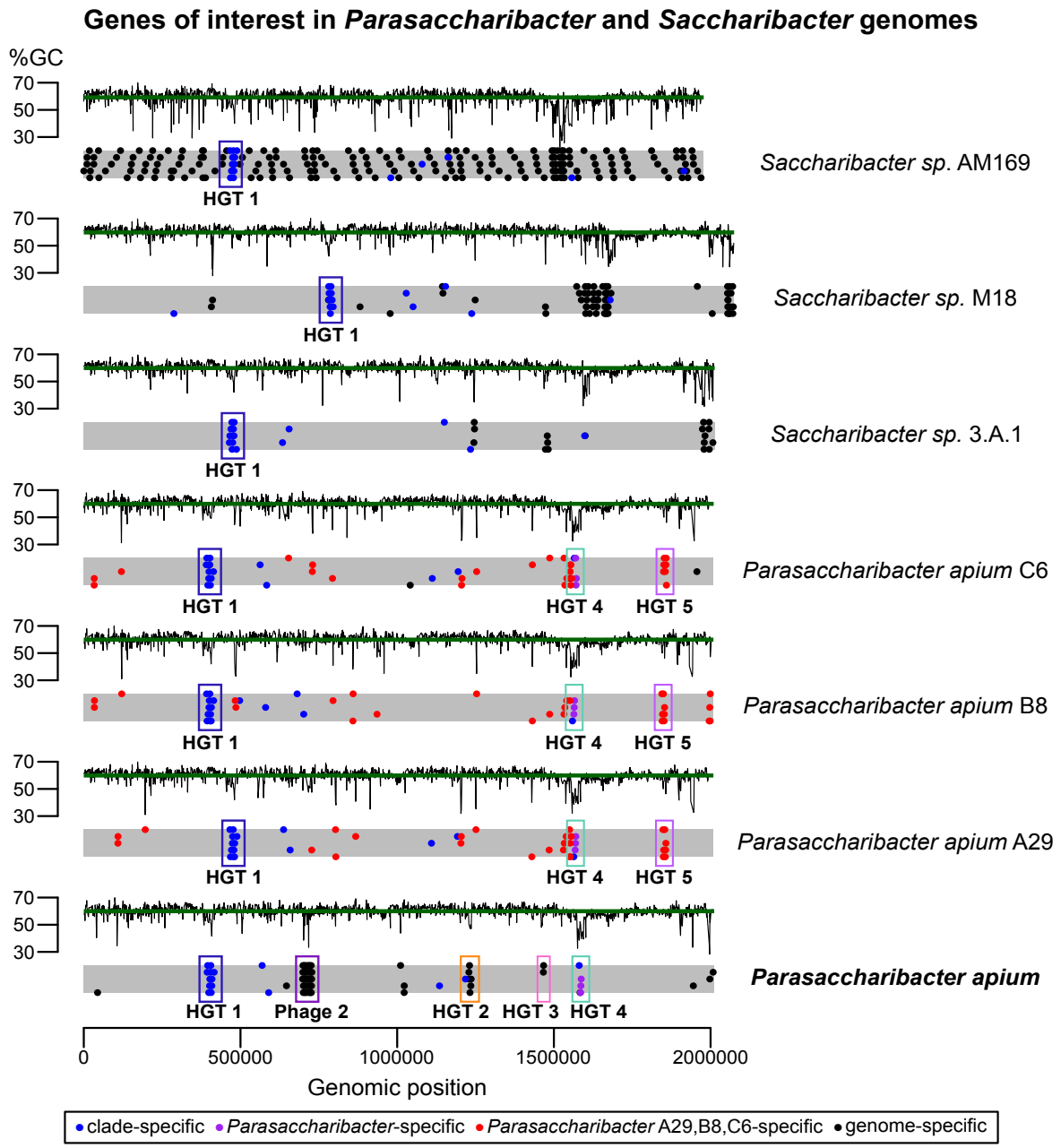

Figure 3. Genomic locations of genes of interest in *Parasaccharibacter* and *Saccharibacter* genomes. Each gray bar is a representation of the genome, with each dot representing the location of a gene in each of four categories (see legend). Regions of interest mentioned in the text are highlighted and labeled. %GC for every gene is plotted above each genome representation, with the green line indicating the genome-wide average %GC. Bolded *Parasaccharibacter apium* represents the *P. apium* reference sequence.

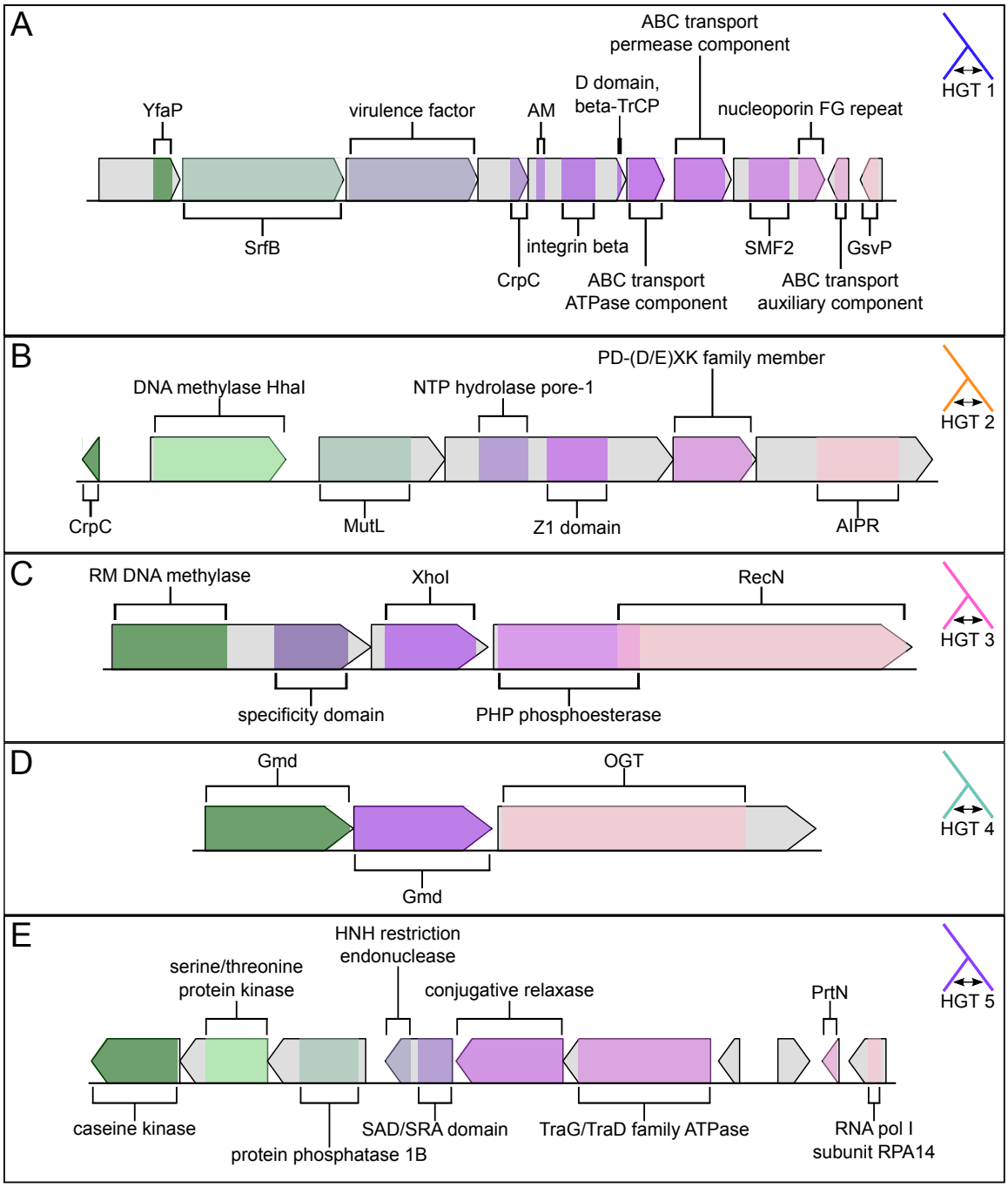

Figure 4. Gene models for each of 5 genomic regions of interest. Gene models are drawn to scale within each panel, but not across panels. A) HGT1. Abbreviations are: CrpC: cysteine rich protein C, AM: automated matches, SMF2: sulfatase modifying factor 2, GsvP: gas vesicle protein C. B) HGT2. Abbreviations are: CrpC: cysteine rich protein C, AIPR: abortive infection phage resistance protein. C) HGT3. Abbreviations are: RM: restriction-modification. D) HGT 4. Abbreviations are: Gmd: GDP-D-mannose dehydratase, OGT: O-linked N-acetylglucosamine transferase OGT. E) HGT5. Abbreviations are: SAD/SRA: SET and Ring finger Associated, PrtN: pyocin activator protein.

673 **SUPPLEMENTAL FIGURES AND TABLE LEGENDS**

674 **Table S1.** OrthoMCL clusters and gene counts for each type of COG in each genome.

675 **Table S2.** Accession number, annotation source, annotation score, and putative annotation for

676 each of the bee-specific genes identified.

677 **Table S3.** Accession number, annotation source, annotation score, and putative annotation for

678 each of the *Bombella intestini* genes identified.

679 **Table S4.** %GC and LPI-difference standard deviations for each gene in each genome harbing

680 each HGT.

681 **Table S5.** Positions and spacer counts for each CRISPR array identified in the *Parasaccharibacter*

682 clade genomes.

683 **Figure S1.** Pairwise genome-wide average nucleotide identity (gANI) for all genomes analyzed.

684 **Figure S2.** Pairwise aligned fraction (AF) for all genomes analyzed.

685 **Supplemental file S1.** Supplemental methods and results.

686 **Supplemental file S2.** BLAST results for genes present in phage 1, present in *Saccharibacter*

687 *floricola.*

688 **Supplemental file S3.** BLAST results for genes present in phage 2, present in the

689 *Parasaccharibacter apium* reference sequence.

690 **Supplemental file S4.** Newick format file for complete 16S rRNA gene tree (Figure 1).