Genomic signatures of honey bee association in an acetic acid symbiont

Eric A. Smith*a and Irene L. G. Newton*a#

aDepartment 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.
IMPORTANCE

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

INTRODUCTION

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

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

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

P. apium is part of a clade of acetic acid bacteria (AAB, a group within the family Acetobacteraceae) that contains both free-living and bee-associated members. Comparative genomics, then, can give us insights into the changes associated with the transition to bee-association in this clade. This comparison can also help elucidate what sets P. apium apart from closely related species and the role(s) it might be playing in the hive environment. To that end, we used the genomes of four recently announced P. apium strains (18), an unpublished P. apium genome assembly, as well as four genomes of the closely-related genus Saccharibacter (19-21), and a genome of the bumblebee symbiont, Bombella intestini (22), to begin to tease
apart the unique capabilities of *P. apium*. Insights gained here could prove critical in determining the factors responsible for maintaining queen health in colonies and could ultimately lead to the development of interventions to improve queen health and mitigate the detrimental impacts of queen failure on this economically critical species.

**RESULTS**

*Acetic acid bacteria phylogeny*

To robustly determine where the *Parasaccharibacter apium* strains and *Saccharibacter* spp. are placed among the AAB, we constructed a maximum likelihood phylogeny using 16S rRNA sequences derived from each of the 17 AAB genera. Our final tree largely agrees with previously published ABB phylogenies (Figure 1) (23-26). Bootstrap support along the backbone of the tree is low, likely owing to the inclusion of sequences from multiple strains of the same species with little sequence divergence. Sequences were largely grouped into monophyletic clades by genus, with one exception; the genus *Gluconacetobacter* is split into a paraphyletic grade, with each branch dominated by a different *Gluconacetobacter* species (clades 1-5, Figure 1A). *Parasaccharibacter* plus *Saccharibacter* comes out as sister to the *Gluconobacter* clade, and the clade encompassing all three of these genera (clade 6, Figure 1A) is sister to *Acetobacter*. Within the *Parasaccharibacter/Saccharibacter* clade, the flower-associated *S. floricola* is sister to the remaining clade (clade 7, Figure 1B), comprised of a mix of taxonomic classifications from *Parasaccharibacter*, *Saccharibacter*, and *Bombella* species. The inconsistent taxonomic nomenclature of the microbes within this clade is typified by the *Parasaccharibacter* classification, which is applied to one *P. apium* strain (AS1), which clusters with a handful of
uncharacterized *Acetobacteraceae* bacteria (clade 8, Figure 1B), to the exclusion of all the other

*P. apium* and *Saccharibacter* strains. It should be noted that, while the genome for *P. apium* AS1 is on GenBank, little else is known about this strain. Further taxonomic ambiguity is highlighted by the presence of a member of a completely separate genus, *Bombella apis*, grouped squarely among *P. apium* and *Saccharibacter* strains, indicating that it is likely a member of one of those two genera. However, branch lengths within this clade are generally short, especially among the main *Parasaccharibacter/Saccharibacter* clade (clade 9, Figure 1B), making it difficult to resolve exact relationships.

**Core ortholog phylogeny of Parasaccharibacter apium and Saccharibacter strains**

We used OrthoMCL (v2.0.9 (27)) to define clusters of orthologous genes (COGs) using the *Parasaccharibacter*, *Saccharibacter*, and *Bombella* genomes listed in Table 1; *Gluconobacter oxydans* H24 was used as an outgroup (see Supplementary Information for more detail; Table S1). To better resolve the phylogenetic relationships between *P. apium*, *Saccharibacter* spp., and *Bombella intestini*, we constructed a second maximum likelihood phylogeny using aligned and concatenated amino acid sequences of the 1,259 single-copy COGs. This robustly supported amino acid phylogeny broadly agrees with our previously constructed 16S phylogeny (Figure 2).

In the core ortholog tree, *B. intestini* interrupts the monophyly of honey bee-associated *Saccharibacter* spp plus *P. apium*. Notably, this tree groups *B. intestini* more closely to the majority of the *P. apium* strains, while *P. apium* AS1 is more distantly related. Similar to the 16S tree, we again see quite short branch lengths within the bee-associated AAB species,
particularly among those in clade D, which includes all _P. apium_ and _Saccharibacter_ genomes except _P. apium_ AS1 and _S. floricola_ (Figure 2).

Given the discrepancy between nomenclature and phylogeny and the short branch lengths, we calculated genome-wide Average Nucleotide Identity (gANI) and aligned fraction (AF) to clarify species relationships. All pairwise comparisons between genomes in clade D met both gANI and AF thresholds for being considered the same species (namely, gANI > 96.5 and AF > 0.6), while no other genome pairs reach both thresholds (Figures S1 and S2). Because of the genetic similarity within clade D and phylogenetic distinction from the rest of the genomes in this tree, the remainder of our analyses focus largely on this clade; for clarity we will refer to this clade as the ‘Parasaccharibacter clade’ from here forward.

Signatures of bee association in the _P. apium_ genomes

To identify genes associated with the transition to bee association, we identified COGs that contained at least one gene from each bee-associated AAB genome (Parasaccharibacter clade, _B. intestini_, and _P. apium_ AS1) and were also missing in _S. floricola_. There were a total of 1,286 COGs containing at least one gene from each of the aforementioned genomes, but only 89 were also missing in _S. floricola_. We determined the putative functions of these genes using the _P. apium_ reference genome representative for each COG (Table S2). It should be noted that all annotations discussed from here forward are putative and require further functional characterization.

Several bee-associated unique genes stood out as particularly interesting, the first being gluconolactonase. Lactonases, such as gluconolactonase, reversibly catalyze the hydrolysis of
lactones (such as gluconolactone) to the hydroxyl acid form (such as gluconic acid).

Gluconolactone is found in both honey and royal jelly and is thought to be partially responsible for the antibacterial properties of both compounds (29). In water, this compound can be hydrolyzed into gluconic acid, acidifying the environment and preventing bacterial growth (30-32). The presence of this gene capable of reversing this acidification – at least locally – may explain how *Parasaccharibacter* is able to thrive in the presence of royal jelly (13). BLAST searches of the metatranscriptomes and metagenomes of bacteria in the “core” honey bee microbiome (33) resulted in zero hits, indicating that none of the “core” microbiome members possess a homolog of this gene. The presence of gluconolactonase may help explain the unique distribution of *Parasaccharibacter* within the hive. Another bee-associated unique gene is an HdeD family acid-resistance protein, which in *E. coli* participates in resistance to acids at high cell densities (34). The presence of this gene in *P. apium* may indicate an adaptation to living in low pH environments – such as the queen bee digestive tract or royal jelly (35).

An AI-2 E family transporter was identified as unique to the bee-associated AAB. AI-2 is an auto-inducer responsible for activating cascades associated with quorum sensing. While *P. apium* does not contain any AI-2 synthesis genes, the presence of an AI-2 E family transporter indicates that it may be responding to exogenous AI-2 produced by other bacteria, possibly in a competitive interaction. Bolstering the competition hypothesis is the presence of fusaric acid resistance (FUSC) genes in the *P. apium* genomes. Fusaric acid and its analogs can be quorum sensing inhibitors (36), so the presence of FUSC genes might be an adaptation that allows *P. apium* to evade quorum sensing inhibition attempts by other microbes. Alternatively, these FUSC genes may play a role in competition with fungal species. Fusaric acid is produced by
several species of fungus and is antibacterial (37). Therefore, the FUSC genes may play a role in
P. apium’s protection of honey bee larvae and queens from infection with Nosema pathogens
(16).

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

The final set of genes of particular interest in this analysis is a complete Type I-E CRISPR/Cas cassette. To determine if this CRISPR/Cas cassette was active, we annotated the genomes for the presence of CRISPR arrays, and found that all of the genomes that have this CRISPR/Cas cassette contain multiple CRISPR arrays. It is possible that these CRISPR arrays were present in the most recent common ancestor of the Parasaccharibacter clade and have simply been maintained in these current genomes; if that were the case, we would expect the spacers in these CRISPR arrays to be highly similar between all strains. However, if these arrays are part of an active CRISPR/Cas system, we would expect the spacers to differ from strain to strain, reflecting unique challenges encountered by each strain. To rule out the possibility that these arrays are ancestral, we aligned each spacer sequence from a given genome to all other spacer sequences from the other genomes and calculated the percent identity. The minimum best intergenomic match for any spacer was 40%, while the maximum was just 65% identical over the length of the spacer, indicating that the spacer sequences are unique from genome to genome and the CRISPR/Cas systems identified here are likely active.
B. intestini was isolated from a bumble bee gut, so we also looked at genes that were unique to this bacterium to determine whether there are any obvious signatures of bumble bee association in its genome. There were a total of 65 genes that were unique to B. intestini, including a complete type IV secretion system (T4SS) and several genes involved in antibiotic production or resistance. Putative annotations of these 65 genes are in Table S3.

Identification of horizontally transferred gene regions

Horizontal transfer of DNA between unrelated bacteria is a commonly known mechanism by which bacteria can acquire new traits and adapt to novel environments (39-42). We identified two regions of phage origin, one in S. floricola and one in the P. apium reference genome (Figure 3; see Supplementary Information for more detail). Movement and insertion of bacteriophage sequences in a genome can have profound effects on the evolution of that genome (43-45) and future work will determine whether these phage are lytic and what their host ranges may be. To determine whether the bacteria in the Parasaccharibacter clade have undergone other potential horizontal gene transfer (HGT) events, we determined the spatial distribution of genes of particular interest (e.g. clade-specific, species-specific, or strain-specific genes) across the bacterial genomes (Figure 3). Some of the genes specific to different clades occur in clusters, an indication that they may have originated elsewhere and been horizontally inherited as a chunk of contiguous DNA. We then looked for anomalies in sequence composition (%GC) and phylogeny to determine whether they were putatively horizontally transferred. Using this combination of methods, we identified a total of five HGT regions in the
Parasaccharibacter clade, which we have numbered 1-5 (see Table S4 for %GC and lineage probability index (LPI)-difference deviations for each gene in each HGT).

HGT1 (Figure 4A) is present in all genomes in the Parasaccharibacter clade, and contains 10 genes, although *P. apium* C6 is missing one of the genes (the second-to-last gene at the 3’ end of the HGT, annotated as an ABC transport auxiliary component). The three most 5’ genes show homology to YfaP (an uncharacterized conserved protein), SrfB (part of the surfactin antibiotic synthesis machinery), and an uncharacterized bacterial virulence factor. The genes in the 3’ half of this HGT contain a number of domains involved in membrane transport. We hypothesize that the two halves of this HGT work together to synthesize and export antibiotics as a form of defense or regulation of competing bacteria. Lending support to the hypothesis that this HGT is involved in defense or immunity is the fact that a CRISPR array lies immediately 5’ of this HGT in each genome (Table S5). Bacterial defense mechanisms tend to occur in clusters of “defense islands” (46), so the presence of this CRISPR array is perhaps a further indication of this HGTs role in bacterial immunity.

HGTs 2 and 3 (Figure 4B and 4C) are restricted solely to the *P. apium* reference genome and are both bacterial restriction/modification (R-M) systems. Bacterial R-M systems are a defense against invading DNA (i.e. bacteriophage). They act by methylating host DNA at specific sites; invading DNA with the same recognition site will be un-methylated, recognized as foreign, and targeted for degradation (47). HGT2 contains 6 genes, which make up the core components of a bacterial (R-M) system. Interestingly, the domain architecture in this R-M system has been recognized as a precursor to eukaryotic defenses against transposable elements (48). HGT3 (Figure 4C) consists of 3 genes comprising 5 domains; the 5’-most gene consists of a predicted
restriction-modification DNA methylase coupled to a specificity domain, the middle gene is predicted to be an XhoI restriction enzyme, and the 3’-most gene is a PHP phosphoesterase coupled to a RecN DNA repair ATPase. Taken together, it appears that HGTs 2 and 3 are responsible for recognition of and defense against foreign DNA.

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

HGT5 (Figure 4E) is unique to the P. apium strains A29, B8, and C6, all strains that had been isolated from honey bee larvae. Like HGTs 1-3, HGT5 contains genes that may play a role in protection against foreign DNA. There are four genes in the 5’ section of HGT5, three of which are kinases, and the fourth contains a SAD/SRA domain in its 5’ end, and an HNH endonuclease domain in its 3’ end. In bacteria, the SAD/SRA domain is often found associated
with an HNH domain (55) and it is thought that the two domains act together to recognize and cleave foreign DNA (56). The 3’ section of HGT5 consists of a conjugative relaxase, a TraG/TraD family ATPase (a coupling protein involved in bacterial conjugation and/or T4SS), a homolog of the pyocin activator protein PrtN, a homolog of a yeast RNA polymerase I subunit, and two additional genes with no annotations. The presence of a PrtN homolog is particularly interesting, as in Pseudomonas aeruginosa pyocins are antibacterial agents, often acting to depolarize the membrane of target cells (57),(58). Interestingly, one of the two unannotated genes in the 3’ region of HGT5 shows weak homology to a phage shock protein, which are proteins involved in the response to stress that may weaken the energy status of the cell (59).

This protein, then, may play a part in immunity to membrane depolarization. Given the presence in HGT5 of: an HNH endonuclease coupled to a SAD/SRA domain, a conjugative relaxase, a TraG/TraD family ATPase, a pyocin activator protein, and a protein with at least some homology to a phage shock protein, we hypothesize that it may play a role in pathogenesis or defense.

DISCUSSION

Here, we used the genomes of five P. apium strains, four Saccharibacter strains, and the closely related B. intestini to gain insight into the genomic changes associated with the transition to honey bee symbiosis in this group. We note several genomic differences – some of which were horizontally acquired – between bee-associated bacteria and the flower-associated S. floricola that may have allowed for the expansion of P. apium into previously unoccupied niches within the honey bee colony. These differences can be classified as changes that
introduce: 1) novel metabolic capabilities, 2) defense and/or virulence mechanisms, and 3) mechanisms for interaction with other microbes and/or the host.

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

Genes involved in the interaction with other microbes and/or the host that we identified include: an AI-2 family transporter, fusaric acid resistance genes, ialb, and ogt. Given that P. apium does not encode any of the canonical genes for the production of quorum-sensing molecules, it seems likely that P. apium is responding to exogenous AI-2 (and/or fusaric acid and its analogs) produced by other members of the bee microbiome (62). The ialb and ogt genes provide routes for interaction with the host, as ialb may play a role in eukaryotic cell invasion (38) and ogt is known to modify thousands of eukaryotic proteins (50). Taken together, we hypothesize that the novel combination of metabolic, quorum-sensing, defense/virulence, and eukaryotic interaction genes in the Parasaccharibacter clade genomes allowed for the
utilization of a unique food source and protection from an onslaught of previously un-
encountered challenges and facilitated the transition to honey bee association in this clade.

*P. apium* has been shown to benefit honey bee larval development and provide
protection against *Nosema* (16). Some of the genes identified here, while allowing *P. apium* to
transition to honey bee symbiosis, may also be related to its ability to protect the bee host from
infection with *Nosema* or other pathogens. If indeed these genes are responsible for the
transition to, and maintenance of, honey bee symbiosis, we would expect to see a modified
evolutionary trajectory relative to those genes not involved in the symbiosis. We currently lack
sufficient sampling of non-bee-associated bacteria in this clade to do such analyses; however,
future studies addressing this question should allow for further elucidation of the genes
involved in the transition to honey bee association. Those analyses, coupled with functional
characterization of the genes of interest identified here, should lay the foundation for the
development of beneficial intervention strategies in this economically critical insect.

**METHODS**

*Acetic acid bacteria phylogeny*

To determine the placement of *Parasaccharibacter* and *Saccharibacter* among the AAB,
we downloaded all 16S rRNA sequences from the Silva database (63-65) that met the following
criteria: 1) from a species belonging to one of the seventeen genera of AAB (26), 2) length at
least 1200 bases, and 3) sequence quality >90. Additionally, the 16S sequence for *Rhodopila
globiformis* – which is in the family *Acetobacteraceae* but is not part of AAB – was included as
an outgroup. Given the close relation to *Saccharibacter floricola*, 16S sequences for *Bombella*
*intestini* (66) and *Bombella sp.* MRM1 (*Bombella apis*) (67) were included. We BLASTed the
*Saccharibacter floricolca* 16S sequence against the *Parasaccharibacter* and *Saccharibacter*
genomes (Table 1) to pull out their respective 16S sequences for use in this phylogeny. All
sequences were aligned using the SINA aligner (68); parameters used were set using the --auto
option. A maximum likelihood phylogeny was constructed using RAxML with the GTRGAMMA
substitution model and 1000 bootstrap replicates (v8.2.11, (69)). The final tree was visualized
using FigTree (v1.4.2, [http://tree.bio.ed.ac.uk/software/figtree/](http://tree.bio.ed.ac.uk/software/figtree/)).

**Orthology analysis**

To facilitate downstream analyses, we clustered genes from all genomes in Table 1 –
plus *Gluconobacter oxydans* H24 as an outgroup – into clusters of orthologous genes (COGs)
using OrthoMCL (v.2.0.9, (27)). Amino acid sequences were downloaded from NCBI and
clustering was performed using default OrthoMCL parameters. These clusters were then
classified as single-copy orthologs (defined as containing exactly one representative from each
genome), variable (defined as missing a representative from at least one genome and having
varying numbers of representatives from each of the other genomes), multi-copy ortholog
(containing at least one representative from each genome, but multiple copies from at least
one genome), or genome-specific (containing at least two genes that all came from the same
genome) using an in-house Perl script.
Parasaccharibacter and Saccharibacter core ortholog phylogeny

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

Calculation of genomic similarity

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

Synteny analysis

We used Mauve (71, 72) to determine the syntenic regions between the Parasaccharibacter apium and Saccharibacter spp. genomes. The Parasaccharibacter apium reference genome is resolved to a single chromosome, so it was used as the reference sequence in Mauve’s “move contigs” tool, and the likely order and orientation of contigs in the other genomes was determined. To facilitate downstream analyses, the output of Mauve’s
“move contigs” tool was used to order, orient, and concatenate contigs into single pseudo-chromosomes for each genome. Structural rearrangements were then visualized using Mauve’s built-in graphical interface.

Annotation of CRISPR arrays and phage sequences

Pseudo-chromosomes for each genome were uploaded to CRISPRFinder to determine location and sequence of CRISPR arrays (73). We used an in-house Perl script to determine the maximum intergenomic percent identity of spacer sequences. We used PHAge Search Tool Enhanced Release (PHASTER) (74, 75) to define phage-like regions. Any region determined to be “questionable” or “intact” by PHASTER that also appeared as an insertion in the host genome in our synteny analysis was labeled as likely to be of phage origin.

Determination of bee-associated bacteria-specific orthologs

We identified all COGs that contained at least one gene from each genome of bee-associated bacteria and no genes from S. floricola. We then took the P. apium reference genome representative for each of these COGs and got KEGG annotations for as many as possible using BlastKOALA (76). For those genes that we were not able to get KEGG annotations, we used NCBI’s BLAST to aid in determining potential function of these bee-associated bacteria-specific genes. This list of genes and their potential functions was then manually curated to hypothesize genes that may have allowed for the transition to bee-association.
Analysis of horizontal gene transfers

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

Domain annotation of genes of interest

We used HHpred (https://toolkit.tuebingen.mpg.de/#/tools/hhpred, (78)) to determine domain architecture and gain an understanding of potential function of the genes in each HGT. For genes of interest that were part of a COG, all members of the COG were first aligned using
the MAFFT L-INS-I algorithm (v7.310, (70)). These multiple sequence alignments (or single amino acid sequences in the case of strain-unique genes) were then uploaded to HHpred’s online tool and homology was determined using HMMs in the COG_KOG_v1.0, Pfam-A_v31.0, and SMART_v6.0 databases; only domains scoring above 60% probability are discussed here.

Gene models for each region of interest were then constructed and visualized using the HHpred results and in-house R scripts. HGT5 occurs at the junction of two contigs in the linear pseudo-chromosomes we constructed. The abutting ends of each contig have annotations for partial pseudogenes, such that when they are joined a complete gene is created. We BLASTed the nucleotide sequence of this gene against the NCBI nr database to determine a putative function.

ACKNOWLEDGEMENTS

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

REFERENCES


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.

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

**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 Paracchcaribacter/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.
SUPPLEMENTAL FIGURES AND TABLE LEGENDS

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

Table S2. Accession number, annotation source, annotation score, and putative annotation for each of the bee-specific genes identified.

Table S3. Accession number, annotation source, annotation score, and putative annotation for each of the Bombella intestini genes identified.

Table S4. %GC and LPI-difference standard deviations for each gene in each genome harboring each HGT.

Table S5. Positions and spacer counts for each CRISPR array identified in the Parasaccharibacter clade genomes.

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

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

Supplemental file S1. Supplemental methods and results.

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

Supplemental file S3. BLAST results for genes present in phage 2, present in the Parasaccharibacter apium reference sequence.

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