Short reads from honey bee (*Apis* sp.) sequencing projects reveal microbial associate diversity

Michael Gerth Corresp., [1] , **Gregory DD Hurst** [1]

[1] Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom

Corresponding Author: Michael Gerth
Email address: gerth@liverpool.ac.uk

High throughput (or ‘next generation’) sequencing has transformed most areas of biological research and is now a standard method that underpins empirical study of organismal biology, and (through comparison of genomes), reveals patterns of evolution. For projects focused on animals, these sequencing methods do not discriminate between the primary target of sequencing (the animal genome) and ‘contaminating’ material, such as associated microbes. A common first step is to filter out these contaminants to allow better assembly of the animal genome. Here, we aimed to assess if these ‘contaminations’ provide information with regard to biologically important microorganisms associated with the individual as part of the ‘hologenome’. To achieve this, we examined whether the short read data from *Apis* retrieved elements of its well established microbiome. To this end, we screened almost 1,000 short read libraries of honey bee (*Apis* sp.) sequencing project for the presence of microbial sequences, and find sequences from known honey bee microbial associates in at least 9% of them. Further to this, we used the data to reconstruct draft genomes of three *Apis* associated bacteria *de novo*. We conclude that ‘contamination’ in short read sequencing libraries can provide useful genomic information on microbial taxa known to be associated with the target organisms, and may even lead to the discovery of novel associations. However, we also find that sequences deriving from microbes outside of the natural microbiome may present a challenge to our approach.

Short reads from honey bee (*Apis* sp.) sequencing projects reveal microbial associate diversity

Michael Gerth* & Gregory D. D. Hurst

University of Liverpool, Institute of Integrative Biology, Biosciences Building, Crown Street
Liverpool L69 7ZB, United Kingdom

*Correspondence & material requests
gerth@liv.ac.uk

Abstract

High throughput (or ‘next generation’) sequencing has transformed most areas of biological research and is now a standard method that underpins empirical study of organismal biology, and (through comparison of genomes), reveals patterns of evolution. For projects focused on animals, these sequencing methods do not discriminate between the primary target of sequencing (the animal genome) and ‘contaminating’ material, such as associated microbes. A common first step is to filter out these contaminants to allow better assembly of the animal genome. Here, we aimed to assess if these ‘contaminations’ provide information with regard to biologically important microorganisms associated with the individual as part of the ‘hologenome’. To achieve this, we examined whether the short read data from *Apis* retrieved elements of its well established microbiome. To this end, we screened almost 1,000 short read libraries of honey bee (*Apis* sp.) sequencing project for the presence of microbial sequences, and find sequences from known honey bee microbial associates in at least 9% of them. Further to this, we used the data to reconstruct draft genomes of three *Apis* associated bacteria *de novo*. We conclude that ‘contamination’ in short read sequencing libraries can provide useful genomic information on microbial taxa known to be associated with the target organisms, and may even lead to the discovery of novel associations. However, we also find that sequences deriving from microbes outside of the natural microbiome may present a challenge to our approach.

Introduction

Novel DNA sequencing methods have revolutionized biological and medical research in the last two decades (Goodwin et al. 2016). High throughput sequencing (or ‘massively parallelized sequencing’, ‘next generation sequencing’,‘NGS’) facilitated the creation of enormous amounts of data for a fraction of the costs associated with traditional Sanger sequencing (Kircher & Kelso 2010; Sboner et al. 2011). This ‘genomics revolution’, has not only enhanced our understanding of molecular and genome evolution (Wolfe & Li 2003), but also contributed to the recognition that eukaryotes are commonly associated with a plethora of microbial taxa.

In eukaryote genome sequencing projects, sequences deriving from these microbes may obstruct genome assembly efforts, and measures directed at removing microbial associates are routinely performed. This is achieved either by antibiotic treatment of the target organism prior to sequencing (Colbourne et al. 2011), or by removing microbial sequences bioinformatically after sequencing (Schmieder & Edwards 2011). While eliminating microbes may facilitate eukaryotic genome reconstruction, it neglects the recently emerging appreciation of microbes as a biologically important component of all multicellular organisms. Numerous examples illustrate the impact of microbes on animal and plant biology, including physiology, behavior, and evolution (McFall-Ngai et al. 2013). These findings have led to a concept that defines an individual eukaryote with all its associated microbes (microbiome) as an entity (holobiont-hologenome) (Bordenstein & Theis 2015). Although this concept is contentious (Moran & Sloan 2015; Douglas & Werren 2016), it is undisputed that some aspects of organismal biology can only be understood by deciphering interactions with microbial symbionts.

To characterize microbiome composition, three approaches are commonly used. First, microbes may be isolated from the host and cultured axenically. Their properties can then be determined through traditional microbiological methods or by sequencing (Browne et al. 2016).

This approach has the benefit of providing both biological and genomic information, but limits discovery to culturable taxa. Second, microbiome taxa may be identified by amplicon sequencing. Specific primers are used to amplify a short informative region from all bacterial taxa in a sample (usually a part of the 16S rRNA gene), and then sequenced (today typically via NGS methods) (Caporaso et al. 2012). This mechanism discovers broad patterns of community diversity, but at a coarse scale, and with weaker functional information. Finally, microbiome composition can be determined via metagenomics, i.e., collective genome sequencing of all bacteria present in a sample (Riesenfeld et al. 2004). This is unbiased, fine scaled, and provides an assessment of biological potential at a community scale, but resolution of genome sequences is more complex

In this study, we examined if the data generated in eukaryotic sequencing projects can be used to identify microbiome taxa, and thus to inform about the composition of the wider ‘holobiont’. Previously, this approach was used to recover genomes of heritable microbes that occur in high densities in many arthropod species, and are therefore prone to be retrieved in arthropod sequencing projects. For example, the genomes of multiple *Wolbachia* strains were discovered in *Drosophila* sequencing data, revealing novel *Wolbachia* diversity and patterns of *Wolbachia* evolution (Salzberg et al. 2005; Richardson et al. 2012).

Here, we examine short reads of honey bee (*Apis* sp.) sequencing projects to investigate whether this archived data can be used to retrieve a wider set of microbial associates, including pathogens and gut symbionts. We focus on honey bees because 1) there is a large number of short read sequencing projects targeting *Apis*; 2) the components of healthy and unhealthy *Apis* microbiomes are well established (Evans & Schwarz 2011; Kwong & Moran 2016); 3) managed populations of the economically important honey bees have been in decline worldwide (Neumann & Carreck 2010), and it was hypothesized that certain bacteria and viruses are key players in this decline (Cox-Foster et al. 2007). Thus, any novel genomic data on honey bee symbionts may directly contribute to our understanding of bee disease.

To identify ‘contaminants’, we here use short signature ‘bait’ sequences of symbionts and pathogens to screen a large number of short read libraries from *Apis* sequencing projects. We demonstrate that the libraries contain non-target sequences from many sources, some of which reflect the natural honey bee microbiome. We further show that highly covered, and possibly novel symbiont genomes can be retrieved from this contamination. Our study highlights the value of database sequences for exploratory symbiont screens and argues against neglecting the filtered ‘contaminants’ in sequencing projects.

Materials & methods

Reference sequences of 18 common *Apis* associated symbionts and pathogens were compiled to be used as baits to detect presence of the microbe (Table S1). In order to reduce the computational expense of all following steps, only short signature sequences were used instead of complete genomes; where possible these were of slowly evolved housekeeping genes to allow a range of diversity to be recovered through sequence similarity to the bait. For previously identified bacterial symbionts for example, we included a 16S rRNA sequence for each known associate. Next, we searched for honey bee sequencing projects in NCBI’s short read archive, using the search term *‘Apis*’, and excluding transcriptome and microbiome (e.g., metagenome or amplicon sequencing) projects. At the time of the search, 306 experiments matched these criteria, including 32 using museum specimens. We downloaded all short read libraries associated with these experiments (993 in total, Table S2) and mapped all reads of each of the libraries to the reference sequences using NextGenMap version 0.4.12 (Sedlazeck et al. 2013). If at least 1,000 reads of a library were aligned to one or more sequence baits, we extracted the matching reads and assembled them using SPAdes version 3.7 (Bankevich et al. 2012). Contigs resulting from this assembly were then subject to taxonomic annotation via BLAST+ (Camacho et al. 2009) searches against a local copy of the NCBI 'nt' database, and the Blobtools package (Kumar et al. 2013). Detailed description of all steps outlined above can be found under https://github.com/gerthmicha/symbiont-sra.

Since this approach yielded a high number of hits to various *Lactobacillus* species, we repeated the entire procedure using 620 16S bait sequences from *Lactobacillus* only. These sequences were taken from a previously compiled dataset of Lactobacilli associated with *Apis*, other Hymenoptera, and other *Lactobacillus* sequences retrieved from public databases (McFrederick et al. 2013). All hits short than 250bp were discarded, and remaining contigs were combined with the reference sequences. We used SSU-ALIGN version 0.1 (Nawrocki 2009) to align and mask this dataset based on conserved secondary structure. Original and masked alignments are available from https://github.com/gerthmicha/symbiont-sra. A maximum likelihood phylogeny was reconstructed from the complete 16S alignment (740 sequences in total) using IQTREE version 1.3.10 (Nguyen et al. 2015) with automated model selection and 1,000 ultrafast bootstraps (Minh et al. 2013) to assess node support. The resulting tree was visualized using the online tool Evolview (He et al. 2016). Furthermore, as an approximate measure for the number of *Lactobacillus* OTUs recovered with our approach, we used the average neighbor clustering algorithm as implemented in mothur version 1.34.4 (Schloss et al. 2009).

Although our aim was not to recover all, but only the highly covered symbiont data from honey bee short reads, we wanted to test if our screening approach yields comparable results to more commonly used metagenomic approaches. To this end, we screened the reads of a metagenomic dataset created from the pooled DNA of 150 honeybee worker hindguts (Engel et al. 2012; ~43M 150bp paired-end reads, SRA accession: SRR5237156) for *Lactobacillus* in the same way as described above. We found 6 different *Lactobacillus* 16S sequences, all within the

Firm-4 and Firm-5 *Lactobacillus* groups (Fig. S1). This was in agreement to the results obtained from taxonomic profiling approaches performed by Engel et al. (2012) and thus confirmed the general effectiveness of our approach (Fig. S1).

Next, we aimed to validate that whole symbiont genomes can in principle be recovered from *Apis* sequencing projects. To this end, we chose one sequencing library (SRR1046114, ~85.5M 100bp paired-end reads) that contained 'contamination' from two *Lactobacillus* strains (*Lactobacillus kunkeei* & *Fructobacillus* sp.). We performed a *de novo* assembly using all reads with MEGAHIT version 1.0.4-beta (Li et al. 2015). All resulting contigs of this assembly were taxonomically assigned to either *L. kunkeei*, *Fructobacillus* sp. or 'other' based on BLAST searches, GC distributions, and read coverage. Reads matching to contigs from either *Lactobacillus* strain were then separately re-assembled using SPAdes, and all contigs smaller than 500bp discarded. Completeness and contamination of the novel draft genomes were assessed using CheckM version 1.0.6 (Parks et al. 2015), and annotation performed with PROKKA version 1.12 (Seemann 2014). The annotated draft genomes are available under under https://github.com/gerthmicha/symbiont-sra and via NCBI accession numbers XXXX00000000 (*L. kunkeei*) and YYYY00000000 (*Fructobacillus* sp.). To evaluate the evolutionary relationships of newly assembled genomes in a broader taxonomic context, we assessed their phylogenetic placement. Whole-genome datasets were compiled for both strains (13 *L. kunkeei* genomes, 9 *Fructobacillus* & *Leuconostoc* genomes altogether, Table S3). For each of the datasets, single copy orthologs were identified using OrthoFinder version 0.2.8 (Emms & Kelly 2015). Recombining loci were identified by using the pairwise homoplasy index test (Bruen et al. 2006), and removed from subsequent analyses (window size = 20 amino acid positions, significance cutoff at 0.05). Using IQTREE, we performed maximum likelihood analysis of two final supermatrices (947 loci and 290,774 aa for the *L. kunkeii* dataset, 435 loci and 145,069 positions for the *Fructobacillus*/*Leuconostoc* dataset). Prior to this, best-fitting partitioning schemes and models were selected using the 'greedy' scheme implemented in IQTREE (Lanfear et al. 2012).

Using the same approach, we assembled and annotated a *Spiroplasma melliferum* genome (NCBI accession ZZZZ00000000) from library SRR957082, (~224.5M 50bp single end reads). Phylogenetic analysis was performed based on a dataset of 206 concatenated single copy genes (58,950 amino acid positions) shared among 17 *Spiroplasma* strains (Table S3). Furthermore, to assess synteny, the newly assembled draft genome was ordered against and aligned with other *Spiroplasma melliferum* genomes (one genome each of strains IPMB4A and KC3) using the progressiveMauve algorithm of Mauve development snapshot version 2015-02-13 (Darling et al. 2010).

Results

Using bait sequences of 18 common *Apis*- associated microbes, we found non-target symbiont reads in 89 of the 993 investigated libraries (~9%). Taxonomic annotation revealed that the detected sequences belong to one of three categories (Fig. 1): 1) *Apis*- associated symbionts that were targeted with our bait sequences, 2) *Apis*- associated taxa that we did not target with our approach, 3) microbial sequences from other sources for which there is no current evidence of *Apis* association. Category 1 included sequences from 3 of the 18 targeted *Apis*- associated taxa (*Crithidia*, *Nosema*, and *Spiroplasma*, Fig. 1, see also Table S4). The second category included mostly honey bee gut bacteria, such as *Lactobacillus*, *Gilliamella* and *Bartonella* (Fig. 1, Table S4). The third category included sequences from fungi (Ascomycota), plants, and the bacterium *Thermus*, that were likely not part of the native microbiome of the sequenced samples. All of these contaminations were crossed-checked via manual online BLAST searches and were confirmed to represent 'true' hits with high and continuous identities with the respective database sequences.

Because the majority of hits in this first screening process were Lactobacilli, we repeated the screening, this time using only *Lactobacillus* 16S sequences as baits. We found 121 *Lactobacillus* sequences in 40 of the 993 investigated libraries, corresponding to 25 OTUs (estimated with mothur using a 5% cutoff). In our phylogenetic analysis based on 16S rRNA sequences, most of the detected strains clustered within *Lactobacillus* groups known to be associated with honey bees (Fig. 2a). Of the recovered sequences not clustering within these lineages, three were found to group with other *Apis*- associated Lactobacilli as sister group to the *Lactobacillus coryniformis* group (Fig. 2a). Online BLAST searches revealed *Fructobacillus* species as closest matches based on 16S rRNA sequence.

Next, we aimed at recovering draft genome sequences of bee-associated Lactobacilli. We chose a sequencing library from which 16S sequences of both *L. kunkeei* and *Fructobacillus* isolates were detected in our screen. The contigs of a meta-assembly were taxonomically annotated, and reads matching to the respective target taxa were then assembled and annotated separately. For each assembly, we performed a phylogenetic analysis based on all single copy orthologs shared with related genomes (Fig. 2b, c), thus confirming the identity of the strains as *L. kunkeei* (Fig. 2b) and *Fructobacillus* (Fig. 2c). Both genomes were highly covered and mostly complete based on the presence of conserved markers (Fig. 2d). Finally, we recovered the genome of a *Spiroplasma melliferum* strain from another *Apis* sequencing library (Fig. 3). In the meta-assembly, *Spiroplasma* and *Apis* contigs could be clearly separated by coverage and taxonomic annotations (Fig. 3b). The refined assembly resulted in a highly covered draft genome of *Spiroplasma melliferum*, which is very similar to the two previously sequenced *Spiroplasma melliferum* strains (Alexeev et al. 2012; Lo et al. 2013), based on shared ortholog clusters, genome organisation, and phylogeny (Fig. 3a, c, d).

Discussion

We used two screens to determine if microbial symbiont data can be retrieved from sequencing projects targeting *Apis* (honey bees). First, by using bait sequences of *Apis* symbionts and pathogens, we found evidence for the presence of these taxa in 9% of 993 *Apis* short read libraries. This measure of non-target 'contamination' can be considered as conservative, since our approach only reports relatively high levels of contamination (at least 1000 reads per bait sequence). Three common honey bee pathogens were detected with this approach: *Nosema*, *Crithidia*, and *Spiroplasma. Nosema* are microsporidian gut parasites of various honey bee species, and while the sampling of our screen is not representative, this finding corroborates the recognition of *Nosema* as widespread pathogen of honey bee colonies worldwide (Nixon 1982; Klee et al. 2007). *Crithidia* (Trypanosomatidae), another gut pathogen of *Apis* and related bee species (Schwarz et al. 2015) was detected at an even higher frequency (Fig. 1, Table S4). Finally, we found *Spiroplasma melliferum* in one of the investigated sequencing libraries. *Spiroplasma* are common symbiotic bacteria of many invertebrates (Duron et al. 2008) and have been connected to pathogenicity in honey bees (Clark 1977). The bait sequences of all of these pathogens showed a high coverage in our screen, suggesting that novel genetic variants can be recovered from already available data, or from data that will become available as by-product of future honey bee sequencing projects. We did not find any viral sequences in our screen (Table S1), probably because most honey bee viruses are RNA viruses (Chen et al. 2004), that are in retrospect unlikely to be picked up with WGS approaches (but could potentially be retrieved from RNAseq data).

This first screen also revealed the presence of many reads originating from *Apis* gut microbes. These reads were the most common 'contamination' detected in the libraries, despite these taxa not being specifically targeted. The microbiome of healthy honey bees is dominated by Lactobacilli (Kwong & Moran 2016), and this is also reflected in our results (Fig1, Table S1). Furthermore, a number of taxa that are likely not part of the natural *Apis* microbiome were detected. For example, we detected *Aspergillus* in several sequencing libraries that originated from museum material, which likely represents post mortem saprophytic growth. We also retrieved hits to plant sequences which might originate from co-amplified and sequenced pollen DNA (Fig. 1). We further detected *Thermus*, which is best explained by contaminated laboratory reagents or sequencing kits (Salter et al. 2014). This 'false discovery' illustrates an important caveat in our approach: the differentiation between host-associated microbes and microbes from other sources may not always be possible, and will be particularly difficult for museum specimens. Though not problematic in the examples we present, the situation is likely more complicated in hosts with a less well-investigated microbiome, or for symbionts that are very similar to environmental taxa. In these cases, the approach will establish candidates that will then require direct validation.

In the second screen, targeted only at *Lactobacillus*, our protocol detected 25 taxonomically different *Lactobacillus* strains. Our phylogenetic reconstruction of *Lactobacillus* relationships based on 16S rRNA generally reflected the current understanding of this genus' taxonomy (Felis & Dellaglio 2007; Salvetti et al. 2012), and revealed that most Lactobacilli known to be associated with honey bees are also present in *Apis* short read libraries. This includes Firm-4 and Firm-5 Lactobacilli, both of which are honey bee hindgut colonizers, and *L. kunkeei*, which is common in nectar and other hive material, and sometimes found in honey bee crops (Kwong & Moran 2016). Furthermore, we found *Fructobacillus,* which share an ecological niche with *L. kunkeei,* i.e., they are found in flowers, nectar, and in honey bee guts (Endo et al. 2009; Endo & Salminen 2013) . Although not classified as such, recent phylogenomic evidence suggests that *Fructobacillus* (and the closely related *Leuconostoc*) are part of the *Lactobacillus* radiation (Sun et al. 2015). Here, we also infer *Fructobacillus* grouping within, rather than outside of *Lactobacillus* (Fig. 2a). These results show that a reasonably accurate understanding of

Lactobacillus community composition in honey bees can be gained from non-target sequences produced as a by-product of honey bee sequencing projects.

Finally, we demonstrate that draft genomes of microbial symbionts can be recovered from *Apis* short reads. For example, inspecting the non-target components of just a single *Apis* sequencing library produced novel, highly covered, and near complete draft genomes of *Lactobacillus kunkeii* and a *Fructobacillus* strain (Fig. 2 b,c,d). Although the 16S sequence of the *Fructobacillus* strain best matched *F. fructosus*, our analysis suggests it belongs to a species so far not represented by genomic sequences in public databases, or even a novel species (Fig. 2c). Conceivably, many additional *Lactobacillus* variants could be retrieved from the libraries investigated here, potentially providing a more complete picture of the *Apis* microbiome composition and function. It should be noted that draft genomes reconstructed this way must be regarded as 'population consensus' genomes, as opposed to genomes sequenced from cultured bacterial clones. While these genomes cannot be linked to a bacterial clone, they still provide information of metabolic capacities within the *Apis* microbiome.

Although our study was focused on *Apis*, it is conceivable that the amount of non-target 'contamination' is similar for other sequencing projects. As a best practice in any sequencing project, we therefore suggest that all non-target taxa should be identified, and their genomes assembled, annotated, and published alongside the target genome. This requires less effort than it may seem, as de-contamination is already a standard post-processing step. Instead of discarding the contaminated reads, they can be processed with one of many available software solutions that automate the process of identifying and assembling genomes from metagenomes (Oulas et al. 2015), thus minimizing the additional workload. Not only would this provide the community with valuable genomic data of microbial symbionts from known host taxa, but it can additionally be argued that this is the most sensible thing to do from a biological point of view. Evidence is mounting that symbiotic microbes influence almost all aspects of their host's biology (Douglas

2014; Bordenstein & Theis 2015). Taking into account the total genomic information recovered in sequencing projects may therefore provide a more complete picture of the target organism's biology.

Conclusion

The biological properties of an individual are a composite of the functions encoded in their genome and that of microbial associates, the 'hologenome'. We here revisited published short read data from *Apis* spp. sequencing projects to investigate if these give insight into the wider set of associates that are commonly disgarded as 'contaminants'. We found that a large variety of distinct *Apis*-associated microbial symbionts and pathogens can be detected as 'contamination' in these data. Further, due to the large depths of today's sequencing projects, the genomes of some microbial associates (which are typically much smaller than the target genomes) can often be recovered in high quality. Honey bees have a comparatively simple microbiota (Kwong & Moran 2016) and are thus considered suitable models for microbiome-animal interactions and evolution (Engel et al. 2016). Their enormous economic importance (Calderone 2012), has driven the large (and still increasing) number of honey bee sequencing projects. Our examination of the output of these projects suggests that large amounts of genomic information on bee-associated microbes are included in these data. While genomes gained from contaminated bee samples cannot and should not replace focused microbiological and metagenomic investigations, they might still improve our understanding of honey bee microbiome composition and functioning.

Acknowledgements

The authors would like to thank Dr Seth Barribeau for comments on an earlier version of this manuscript. We further thank Dr Philipp Engel and Olivier Emery for providing access to short read data. This work was supported by the European Commission through H2020 funding in the form of an EMBO long term fellowship (ALTF 48-2015, LTFCOFUND2013, GA-2013-609409) and a Marie Curie Fellowship (H2020-MSCA-IF-2015, 703379) to MG.

References

Alexeev D, Kostrjukova E, Aliper A, Popenko A, Bazaleev N, Tyakht A, Selezneva O, Akopian T, Prichodko E, Kondratov I, et al (2012) Application of *Spiroplasma melliferum* proteogenomic profiling for the discovery of virulence factors and pathogenicity mechanisms in host-associated spiroplasmas. *Journal of Proteome Research* **11**, 224–236.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477.

Bordenstein SR, Theis KR (2015) Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS Biology* **13**, e1002226.

Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, Goulding D, Lawley TD (2016) Culturing of `unculturable´ human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543–546.

Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* **172**, 2665–2681.

Calderone NW (2012) Insect pollinated crops, insect pollinators and US agriculture: trend analysis of aggregate data for the period 1992–2009. *PLoS ONE* **7**, e37235.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**, 1621–1624.

Chen Y, Zhao Y, Hammond J, Hsu H-t, Evans J, Feldlaufer M (2004) Multiple virus infections in the honey bee and genome divergence of honey bee viruses. *Journal of Invertebrate Pathology* **87**, 84–93.

Clark T (1977) *Spiroplasma* sp, a new pathogen in honey bees. *Journal of Invertebrate Pathology* **113**, 112–113.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi J-H, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011) The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555–561.

Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan P-L, Briese T, Hornig M, Geiser DM, Martinson V, vanEngelsdorp D, Kalkstein AL, Drysdale A, Hui J, Zhai J, Cui L, Hutchison SK, Simons JF, Egholm M, Pettis JS, Lipkin WI (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* **318**, 283–287.

Darling AE, Mau B, Perna NT (2010) Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147.

Douglas AE (2014) Multiorganismal insects: diversity and function of resident microorganisms. *Annual Review of Entomology* **60**, 1–18.

Douglas AE, Werren JH (2016) Holes in the hologenome: why host-microbe symbioses are not holobionts. *mBio* **7**, e02099–15.

Duron O, Bouchon D, Boutin S, Bellamy L, Zhou L, Engelstädter J, Hurst GDD (2008) The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. *BMC Biology* **6**, 27.

Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157.

Endo A, Futagawa-Endo Y, Dicks LM (2009) Isolation and characterization of fructophilic lactic acid bacteria from fructose-rich niches. *Systematic and Applied Microbiology* **32**, 593–600.

Endo A, Salminen S (2013) Honeybees and beehives are rich sources for fructophilic lactic acid bacteria. *Systematic and Applied Microbiology* **36**, 444–448.

Engel P, Kwong WK, McFrederick Q, Anderson KE, Barribeau SM, Chandler JA, Cornman RS, Dainat J, de Miranda JR, Doublet V, Emery O, Evans JD, Farinelli L, Flenniken ML, Granberg F, Grasis JA, Gauthier L, Hayer J, Koch H, Kocher S, Martinson VG, Moran N, Munoz-Torres M, Newton I, Paxton RJ, Powell E, Sadd BM, Schmid-Hempel P, Schmid-

Hempel R, Song SJ, Schwarz RS, vanEngelsdorp D, Dainat B (2016) The bee microbiome: impact on bee health and model for evolution and ecology of host-microbe interactions. *mBio* **7**, e02164–15.

Engel P, Martinson VG, Moran NA (2012) Functional diversity within the simple gut microbiota of the honey bee. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 11002–11007.

Evans JD, Schwarz RS (2011) Bees brought to their knees: microbes affecting honey bee health. *Trends in Microbiology* **19**, 614–620.

Felis GE, Dellaglio F (2007) Taxonomy of lactobacilli and bifidobacteria. *Current Issues in Intestinal Microbiology* **8**, 44.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351.

He Z, Zhang H, Gao S, Lercher MJ, Chen W-H, Hu S (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. *Nucleic Acids Research* **44**, W236–W241.

Kircher M, Kelso J (2010) High-throughput DNA sequencing - concepts and limitations. *BioEssays* **32**, 524–536.

Klee J, Besana AM, Genersch E, Gisder S, Nanetti A, Tam DQ, Chinh TX, Puerta F, Ruz JM, Kryger P, Message D, Hatjina F, Korpela S, Fries I, Paxton RJ (2007) Widespread dispersal of the microsporidian *Nosema ceranae*, an emergent pathogen of the western honey bee, *Apis mellifera*. *Journal of Invertebrate Pathology* **96**, 1–10.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter ML (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics* **4**, 237.

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

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **29**, 1695–701.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676.

Lo W-S, Chen L-L, Chung W-C, Gasparich GE, Kuo C-H (2013) Comparative genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. *BMC Genomics* **14**, 22.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, Hentschel U, King N, Kjelleberg S, Knoll AH, Kremer N, Mazmanian SK, Metcalf JL, Nealson K, Pierce NE, Rawls JF, Reid A, Ruby EG, Rumpho M, Sanders JG, Tautz D, Wernegreen JJ (2013) Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 3229–3236.

McFrederick QS, Cannone JJ, Gutell RR, Kellner K, Plowes RM, Mueller UG (2013) Specificity between lactobacilli and hymenopteran hosts is the exception rather than the rule. *Applied and Environmental Microbiology* **79**, 1803–12.

Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution* **30**, 1188–1195.

Moran NA, Sloan DB (2015) The Hologenome Concept: Helpful or Hollow?. *PLOS Biology* **13**, e1002311.

Nawrocki EP (2009) *Structural RNA homology search and alignment using covariance models*. PhD Thesis, Washington University in Saint Louis.

Neumann P, Carreck NL (2010) Honey bee colony losses. *Journal of Apicultural Research* **49**, 1–6.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268–274.

Nixon M (1982) Preliminary world maps of honeybee diseases and parasites. *Bee World* **63**, 23–42.

Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos4 I (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights* , 75.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043–1055.

Richardson MF, Weinert LA, Welch JJ, Linheiro RS, Magwire MM, Jiggins FM, Bergman CM (2012) Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. *PLoS Genetics* **8**, e1003129.

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* **38**, 525–552.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, .

Salvetti E, Torriani S, Felis GE (2012) The genus *Lactobacillus*: a taxonomic update. *Probiotics & Antimicrobial Proteins* **4**, 217–226.

Salzberg SL, Hotopp J, Delcher A, Pop M, Smith D, Eisen M, Nelson W (2005) Serendipitous discovery of *Wolbachia* genomes in multiple *Drosophila* species. *Genome Biology* **6**, R23.

Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think!. *Genome Biology* **12**, 125.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537–7541.

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* **6**, e17288.

Schwarz RS, Bauchan GR, Murphy CA, Ravoet J, Graaf DC, Evans JD (2015) Characterization of two species of Trypanosomatidae from the honey bee *Apis mellifera*: *Crithidia mellificae* Langridge and McGhee, and *Lotmaria passim* n gen, n sp. *Journal of Eukaryotic Microbiology* **62**, 567–583.

Sedlazeck FJ, Rescheneder P, von Haeseler A (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791.

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069.

Sun Z, Harris HMB, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Ross RP, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW (2015)

Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nature Communications* **6**, 8322.

Tamarit D, Ellegaard KM, Wikander J, Olofsson T, Vasquez A, Andersson SGE (2015) Functionally structured genomes in *Lactobacillus kunkeei* colonizing the honey crop and food products of honeybees and stingless bees. *Genome Biology and Evolution* **7**, 1455–1473.

Wolfe KH, Li W-H (2003) Molecular evolution meets the genomics revolution. *Nature Genetics* **33**, 255–265.

Figure legends

Figure 1: Taxonomic annotation of contigs assembled from 'contaminated' *Apis* short read libraries. Bar chart shows the frequency of each taxonomic category assigned by best BLAST matches against NCBI's 'nt' database, as the number of libraries in which that taxon was detected (in the sample of 993 SRA libraries). Bold categories are 'phyla', as defined in https://www.ncbi.nlm.nih.gov/taxonomy, taxa in italics represent typical genera that were recovered within each phylum. See Table S4 for a complete list.

Figure 2: 'Contamination' from Lactobacilli in *Apis* short read libraries. a) Maximum likelihood tree of 720 16S rRNA sequences from Lactobacilli. Branch colors and the color of the outer annotation circle correspond to *Lactobacillus* species groups according to Felis & Dellaglio (2007). Inner circle demarks taxa found Hymenoptera (grey squares) and in corbiculate apids (honey bees and relatives, black squares). *Lactobacillus* sequences recovered in this study from contaminated *Apis* libraries are labeled with blue triangles. The Lactobacilli typically associated with honey bees (Firm-4, Firm-5, *L. kunkeei*) are further highlighted with a blue background color. Two dotted blue lines denote the taxa of which whole draft genomes were recovered. See text for details. An interactive version of the tree containing all node labels is available under http://www.evolgenius.info/evolview/#shared/wZcKHbwJuT. Abbreviations: al-far- alimentarius-farciminis, bre- brevis, buch- buchneri, cas- casei, cor- coryniformis, del- delbrueckii, fru-fructivorans, per- perolens, plan- plantarum, reu- reuteri, sak- sakei, sal- salivarius, OUT-outgroup. b) Phylogeny of *Lactobacillus kunkeei* strains based on maximum likelihood analyses of 947 concatenated single copy orthologs (290,774 amino acid positions). Tree is rooted with *Lactobacillus apinorum* Fhon13 (taxon not shown). Strain names correspond to the names used in Tamarit et al. (2015; see Table S3). Blue taxon label corresponds to the *L. kunkeei* strain recovered from 'contaminants' in library SRR1046114. Bootstrap values are given on nodes. See

Table S3 for sources of genomes. c) Maximum likelihood tree of *Fructobacillus* (F.) and *Leuconostoc* (L.) species based on 435 concatenated single copy orthologs (145,069 amino acid positions). Tree is rooted with *Lactobacillus delbruecki*. Numbers on nodes correspond to bootstrap values. Again, blue taxon label denotes the *Fructobacillus* genome recovered from the 'contaminated' library SRR1046114. Note that the phylogenetic distance between *Fructobacillus fructosus* and the novel genome is similar to other between-species distances in this tree. See Table S3 for accession numbers of all genomes used for phylogenetic analysis. d) Assembly statistics for the two novel draft genomes recovered from library SRR1046114. Abbreviations: CDS- coding sequences predicted with PROKKA, Comp. & Cont.- completeness and contamination as estimated with CheckM version 1.0.6 (Parks et al. 2015) based on the number of conserved marker loci. Phylogenetic affiliations of the two strains are depicted in Fig. 3b and 3c, respectively.

Figure 3: Characteristics of *Spiroplasma melliferum* isolated from a 'contaminated' *Apis* sequencing library (SRR957082). a) Venn diagram illustrating the number of orthologs shared between the novel strain and its closest sequenced relatives IBMB4A (Lo et al. 2013) and KC3 (Alexeev et al. 2012). b) Taxon-annotated GC-coverage plot of SRR951082 metaassembly created with Blobology. *Spiroplasma* and *Apis* contigs can be differentiated by coverage. c) Synteny across *Spiroplasma melliferum* genomes. Contigs from assemblies SRR957082 and IPMB4A were ordered against KC3, the most complete of the three *S. melliferum* genomes. d) Phylogenetic relationships within the genus *Spiroplasma*. Maximum likelihood tree is based on 206 concatenated loci (58,950 amino acid positions), numbers on branches correspond to bootstrap values. *Spiroplasma* groups are highlighted with colors. The taxon label of the novel genome is highlighted in bold. Accession numbers for all taxa are listed in Table S4.

Supplementary files

Fig S1: Verification of screening approach employed here using the dataset of Engel et al. (2012). All short reads from this dataset were mapped against *Lactobacillus* 16S reference sequences as detailed in the materials & methods section. Thus retrieved 16S sequences are highlighted with thick, dark blue lines. All other taxa in this tree are identical to the ones in Fig. 2A, as is the color scheme. Although the topology differs between these two *Lactobacillus* trees, it is evident that the strains recovered from the Engel et al. (2012) dataset cluster within the Firm-4 and Firm-5 *Lactobacillus* groups. Engel et al. (2012) essentially find the same ("These distinct clusters reflect the eight dominant species with the two closely related Firmicutes (Firm-4 and Firm-5) [...]"; see also their Fig. 1c) using the programs MetaPhyler (http://metaphyler.cbcb.umd.edu/) and IMG/M (https://img.jgi.doe.gov/) for taxonomic profiling.

Table S1: Accession numbers for all signature reference sequences used in the initial screen. The sequence for *Arsenophonus* 16S was recovered from honey bee short read data (unpublished).

Table S2: A list of NCBI accession numbers for all short reads downloaded and screened in this work.

Table S3: NCBI accession numbers for all genomes employed for comparative/phylogenetic analyses of *Lactobacillus kunkeei*, *Fructobacillus* sp., and *Spiroplasma* sp.

Table S4: Taxonomic summary of BLAST hits for contigs created in the first round of screening. File created with a Blobtools script (see Materials & methods).

Figure 1(on next page)

Taxonomic annotation of contigs assembled from ‘contaminated’ *Apis* short read libraries.

Bar chart shows the frequency of each taxonomic category assigned by best BLAST matches against NCBI’s ‘nt’ database, as the number of libraries in which that taxon was detected (in the sample of 993 SRA libraries). Bold categories are ‘phyla’, as defined in https://www.ncbi.nlm.nih.gov/taxonomy , taxa in italics represent typical genera that were recovered within each phylum. See Table S4 for a complete list.

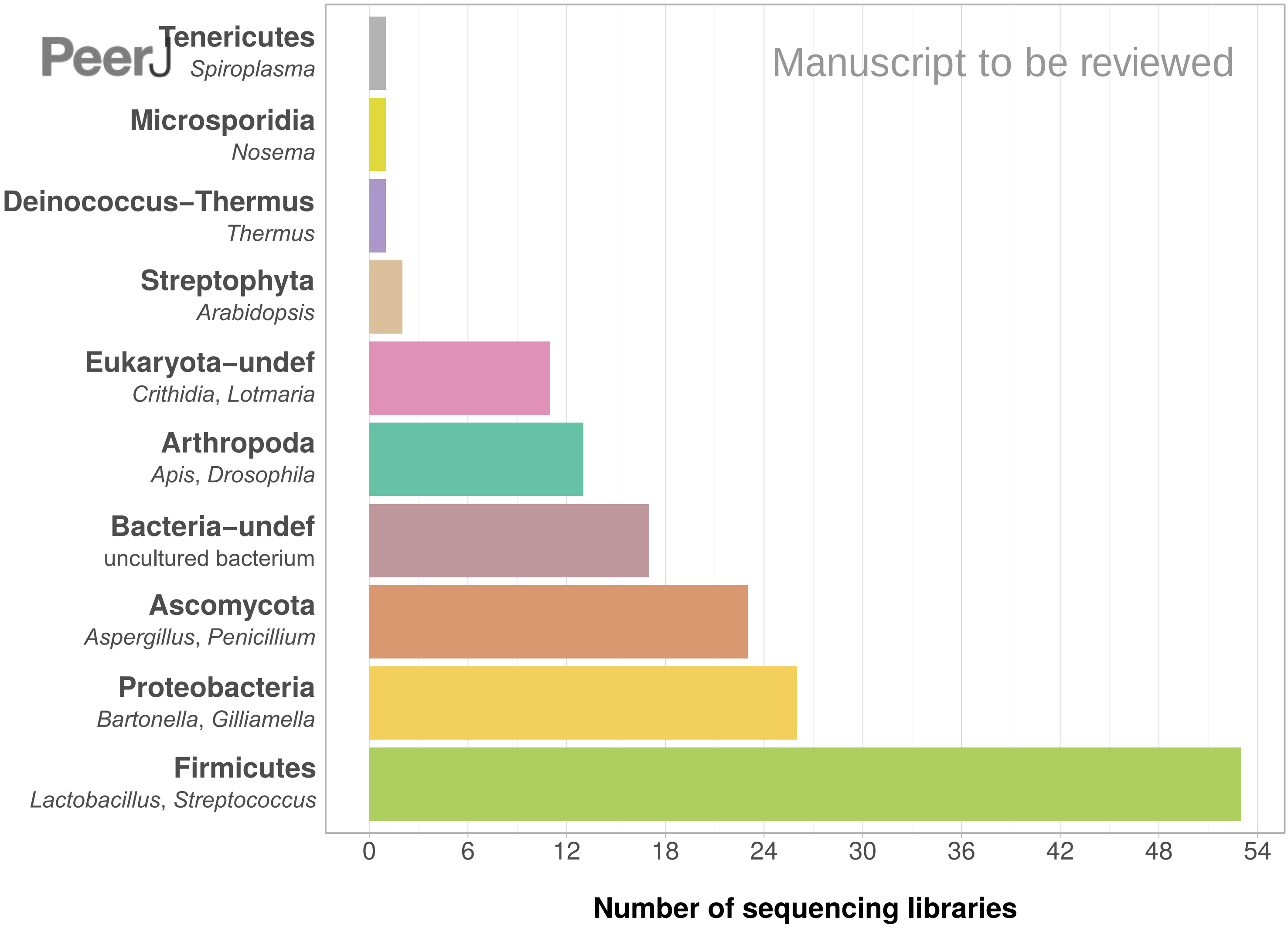
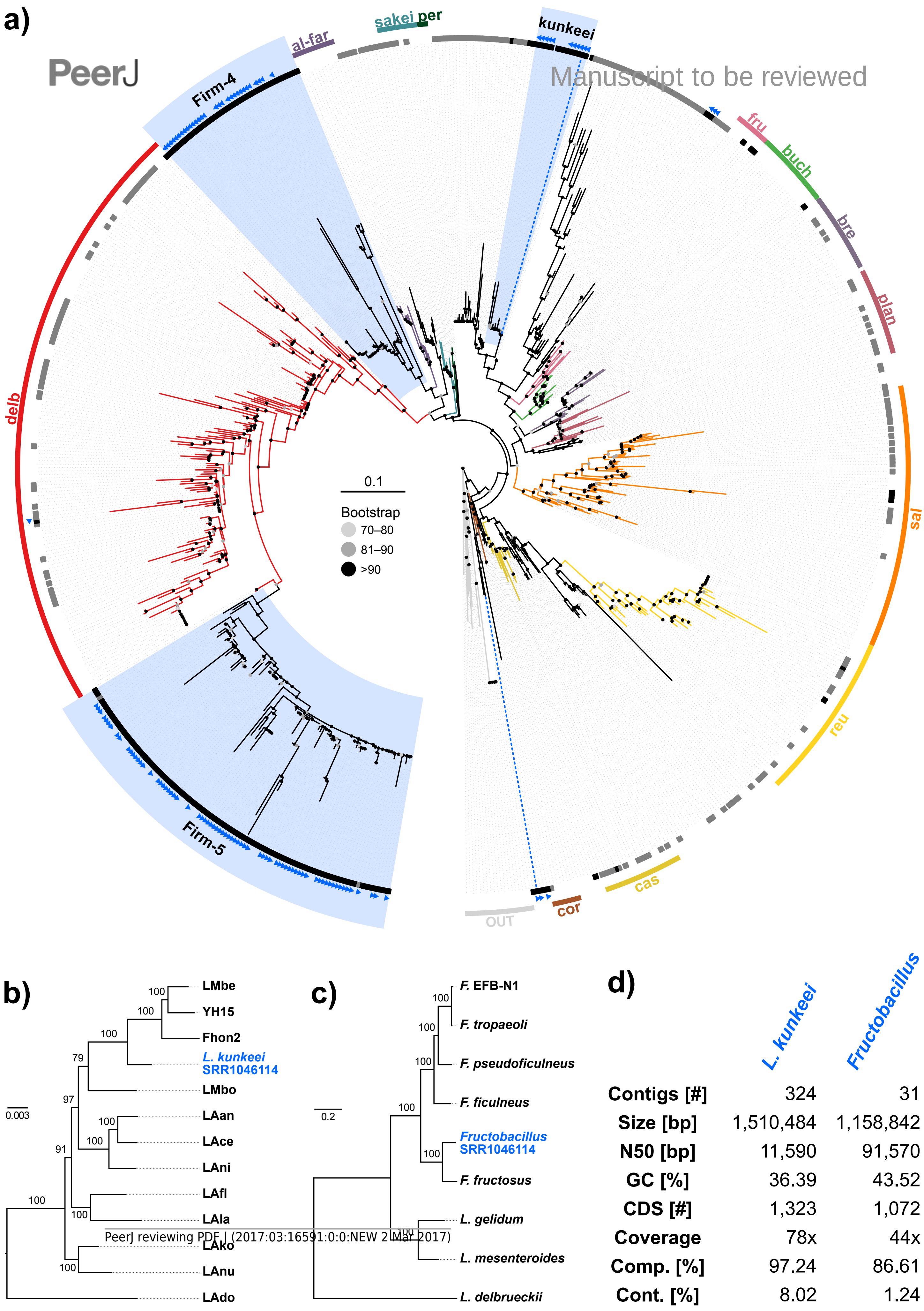

Tenericutes
Spiroplasma
Microsporidia
Nosema
Deinococcus–Thermus
Thermus
Streptophyta
Arabidopsis
Eukaryota–undef
Crithidia, *Lotmaria*
Arthropoda
Apis, *Drosophila*
Bacteria–undef
uncultured bacterium
Ascomycota
Aspergillus, *Penicillium*
Proteobacteria
Bartonella, *Gilliamella*
Firmicutes
Lactobacillus, *Streptococcus*
0
6
12
18
24
30
36
42
48
54
Number of sequencing libraries

Figure 2(on next page)

‘Contamination’ from Lactobacilli in *Apis* short read libraries.

a) Maximum likelihood tree of 720 16S rRNA sequences from Lactobacilli. Branch colors and the color of the outer annotation circle correspond to *Lactobacillus* species groups according to Felis & Dellaglio (2007). Inner circle demarks taxa found Hymenoptera (grey squares) and in corbiculate apids (honey bees and relatives, black squares). *Lactobacillus* sequences recovered in this study from contaminated *Apis* libraries are labeled with blue triangles. The Lactobacilli typically associated with honey bees (Firm-4, Firm-5, *L. kunkeei*) are further highlighted with a blue background color. Two dotted blue lines denote the taxa of which whole draft genomes were recovered. See text for details. An interactive version of the tree containing all node labels is available under http://www.evolgenius.info/evolview/#shared/wZcKHbwJuT . Abbreviations: al-far- alimentarius-farciminis, bre- brevis, buch- buchneri, cas- casei, cor- coryniformis, del- delbrueckii, fru- fructivorans, per- perolens, plan- plantarum, reu- reuteri, sak- sakei, sal- salivarius, OUT- outgroup. b) Phylogeny of *Lactobacillus kunkeei* strains based on maximum likelihood analyses of 947 concatenated single copy orthologs (290,774 amino acid positions). Tree is rooted with *Lactobacillus apinorum* Fhon13 (taxon not shown). Strain names correspond to the names used in Tamarit et al. (2015; see Table S3). Blue taxon label corresponds to the *L. kunkeei* strain recovered from ‘contaminants’ in library SRR1046114. Bootstrap values are given on nodes. See Table S3 for sources of genomes. c) Maximum likelihood tree of *Fructobacillus* (F.) and *Leuconostoc* (L.) species based on 435 concatenated single copy orthologs (145,069 amino acid positions). Tree is rooted with *Lactobacillus delbruecki*. Numbers on nodes correspond to bootstrap values. Again, blue taxon label denotes the *Fructobacillus* genome recovered from the ‘contaminated’ library SRR1046114. Note that the phylogenetic distance between *Fructobacillus fructosus* and the novel genome is similar to other between-species distances in this tree. See Table S3 for accession numbers of all genomes used for phylogenetic analysis. d) Assembly statistics for the two novel draft genomes recovered from library SRR1046114. Abbreviations: CDS- coding sequences predicted with PROKKA, Comp. & Cont.- completeness and contamination as estimated with CheckM version 1.0.6 (Parks et al. 2015) based on the number of conserved marker loci. Phylogenetic affiliations of the two strains are depicted in Fig. 3b and 3c, respectively.

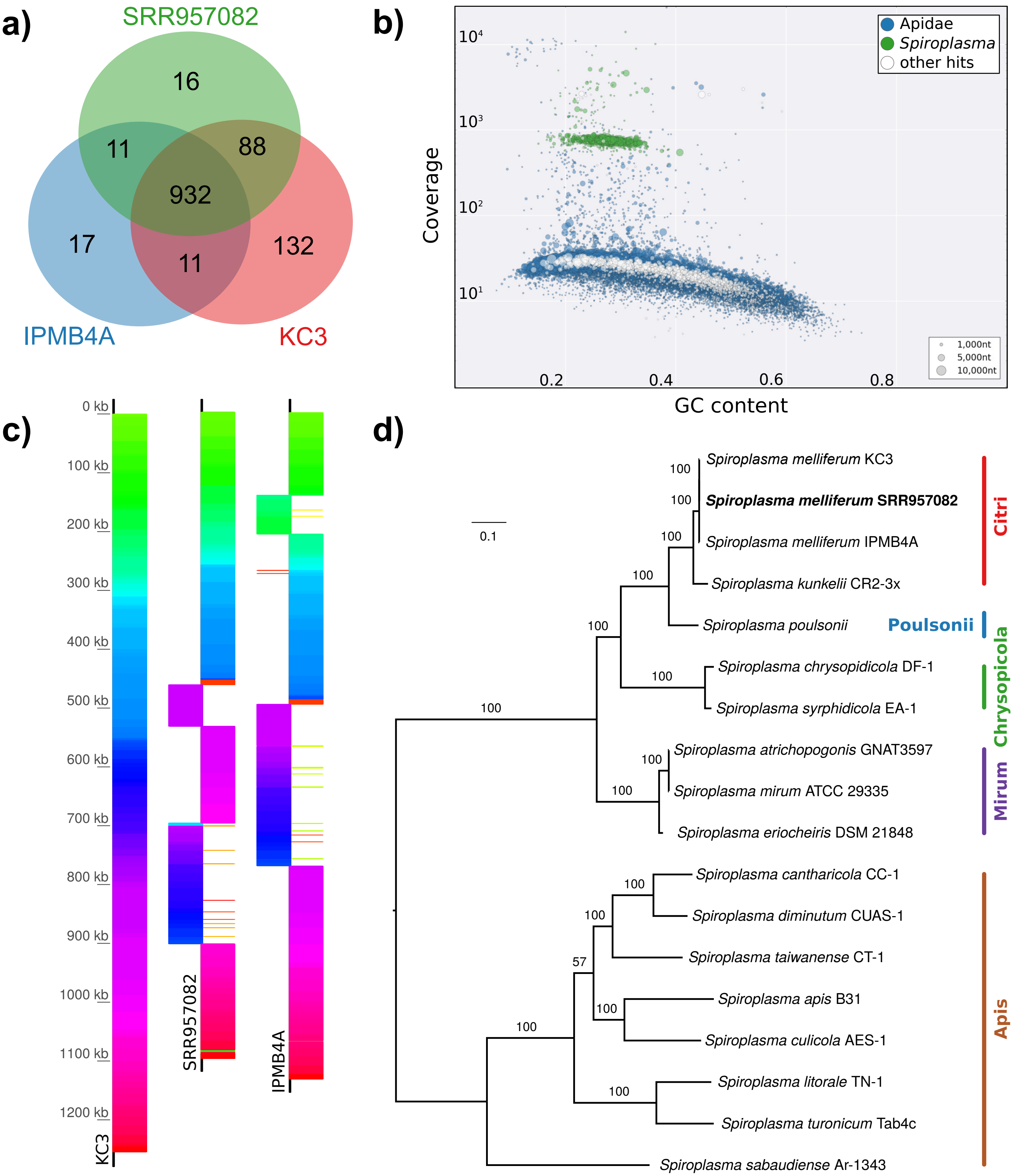


	L. kunkeei	Fructobacillus
Contigs [#]	324	31
Size [bp]	1,510,484	1,158,842
N50 [bp]	11,590	91,570
GC [%]	36.39	43.52
CDS [#]	1,323	1,072
Coverage	78x	44x
Comp. [%]	97.24	86.61
Cont. [%]	8.02	1.24

Figure 3(on next page)

Characteristics of *Spiroplasma melliferum* isolated from a ‘contaminated’ *Apis* sequencing library (SRR957082).

a) Venn diagram illustrating the number of orthologs shared between the novel strain and its closest sequenced relatives IBMB4A (Lo et al. 2013) and KC3 (Alexeev et al. 2012). b) Taxon-annotated GC-coverage plot of SRR951082 metaassembly created with Blobology. *Spiroplasma* and *Apis* contigs can be differentiated by coverage. c) Synteny across *Spiroplasma melliferum* genomes. Contigs from assemblies SRR957082 and IPMB4A were ordered against KC3, the most complete of the three *S. melliferum* genomes. d) Phylogenetic relationships within the genus *Spiroplasma*. Maximum likelihood tree is based on 206 concatenated loci (58,950 amino acid positions), numbers on branches correspond to bootstrap values. *Spiroplasma* groups are highlighted with colors. The taxon label of the novel genome is highlighted in bold. Accession numbers for all taxa are listed in Table S4.


a)
SRR957082
16
11
88
932
17
132
11
IPMB4A
KC3
b)
Coverage
10⁴
10³
10²
10¹
GC content
0.2
0.4
0.6
0.8
Apidae
Spiroplasma
other hits
1,000nt
5,000nt
10,000nt
c)
0 kb
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
700 kb
800 kb
900 kb
1000 kb
1100 kb
1200 kb
KC3
SRR957082
IPMB4A
d)
0.1
Spiroplasma melliferum KC3
Spiroplasma melliferum SRR957082
Spiroplasma melliferum IPMB4A
Spiroplasma kunkelii CR2-3x
Spiroplasma poulsonii
Spiroplasma chrysopidicola DF-1
Spiroplasma syrphidicola EA-1
Spiroplasma atrichopogonis GNAT3597
Spiroplasma mirum ATCC 29335
Spiroplasma eriocheiris DSM 21848
Spiroplasma cantharicola CC-1
Spiroplasma diminutum CUAS-1
Spiroplasma taiwanense CT-1
Spiroplasma apis B31
Spiroplasma culicola AES-1
Spiroplasma litorale TN-1
Spiroplasma turonicum Tab4c
Spiroplasma sabaudiense Ar-1343
100
100
100
100
100
100
100
100
100
100
100
100
57
100
100
100
Citri
Poulsonii
Chrysopicola
Mirum
Apis