Phylogeny and genomic characterization of bee-associated *Lactobacillus* species and their closest relatives

Eric Gordon, David Haisten, Kaleigh Russell, Hoang Vuong

Introduction

While widely used in bacteria, it is difficult to fully understand the evolutionary relationships between bacterial species only using limited "housekeeping" genes such as 16S or RecA in a phylogenetic analysis. For instance, there are strains of *Brevundimonas alba* with identical or extremely similar 16S rDNA sequences, yet they have dramatically different genomic sequences (Jaspers and Overmann 2004). With similar findings likely to be found in other genera of microbes, phylogenetics utilizing only rDNA and other limited data from universal DNA sequences are likely to be misleading. Meanwhile with the accessibility of whole-genome sequencing, divergence by species via analysis of genome-wide functional diversity and analysis for selection across multiple loci is made possible. These analyses often rely upon the detection and analysis of orthologous sequences. Large levels of ortholog sequence divergence would be expected across a genus inhabiting many niches, such as *Lactobacillus*, a group of gram-positive bacteria (Marakova 2006).

From breweries, to animal guts including human and bees, and the vaginal microbiome, *Lactobacillus* spp. have piqued the interest for their possible roles as symbionts for animals, for instance in bees (Petrova, 2015; Marakova 2006). Over the last decade bee population, both native and wild bees have been in a decline. An area of interest for researchers investigating this decline are the microbes associated with bees. Of these microbes, many *Lactobacillus* spp. are associated with bees and their environment (McFrederick et al. 2012). Species of the Firm4, Firm5, and *kunkeei* clade in *Lactobacillus* are important for nutrition, and are members of the core gut microbiome for honey bees. Additionally, *Lactobacillus kunkeei* is a major symbiont of honey bees and wild bees worldwide, present in their hive substrate and provisions, guts, and their associated flowers.

Previous studies have already characterized and analyzed genomes, one of which investigated the ortholog enrichment in the *Lactobacillus kunkeei* clade (Tamarit 2015), while another compared the bee-associated Firm4 and Firm5 clades (Ellegaard 2014).
The paper that focused on L. kunkeei used ~700 orthologs, DNA sequence for making their phylogeny. Their search executed 1000 rapid bootstrap replicates followed by 200 searches for the best Ln LH tree. The other study investigated the Firm-4 and Firm-5 clades used less genes and from glancing at their methods their tree searches were inadequate (only 100 rapid replicates followed by 20 searches). The authors for both papers included analyses of the ratio between the rate of non-synonymous nucleotide changes and synonymous nucleotide changes to detect the presence of selection occurring in protein coding DNA sequences. Finally the authors used their full collection of orthologs to characterize the phylogenetic relationships of their taxa in their respective clades.

For our project, we incorporated taxonomic sampling for sister groups of bee associated *Lactobacillus* spp. including the Firm4/5 and kunkeei clades. We seek to pool together the bee associated *Lactobacillus* and complete an orthologous sequence based analysis with non-bee associated relatives. We employed both concatenation, and for the first time in the group, coalescent analyses.

Results & Methods

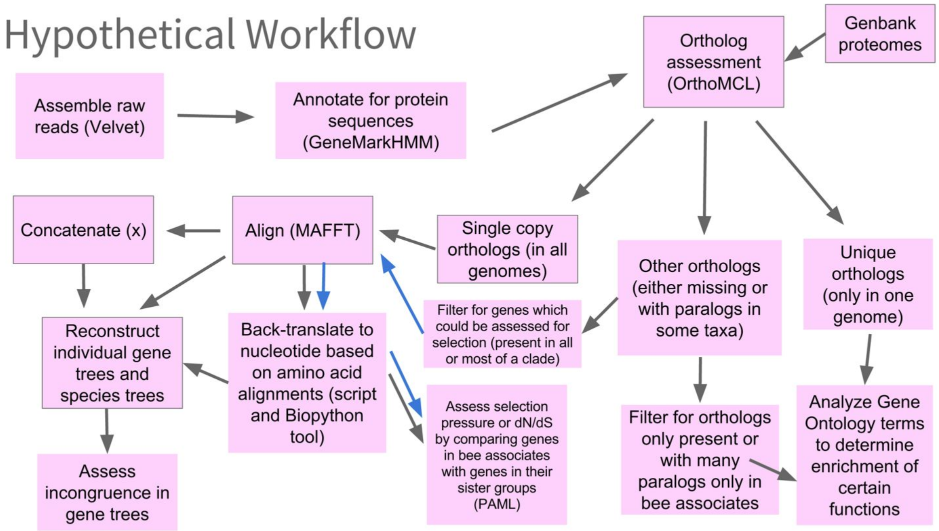
Hypothetical Workflow
Assemble raw reads (Velvet)
Annotate for protein sequences (GeneMarkHMM)
Ortholog assessment (OrthoMCL)
Genbank proteomes
Concatenate (x)
Align (MAFFT)
Single copy orthologs (in all genomes)
Other orthologs (either missing or with paralogs in some taxa)
Unique orthologs (only in one genome)
Filter for genes which could be assessed for selection (present in all or most of a clade)
Reconstruct individual gene trees and species trees
Back-translate to nucleotide based on amino acid alignments (script and Biopython tool)
Assess selection pressure or dN/dS by comparing genes in bee associates with genes in their sister groups (PAML)
Filter for orthologs only present or with many paralogs only in bee associates
Analyze Gene Ontology terms to determine enrichment of certain functions
Assess incongruence in gene trees

Figure 1. Flowchart of our workflow with programs utilized in parentheses.

An overview of our methodology is summarized in Fig. 1. More extensive descriptions of the individual components and results follow below. Since the goal of our pipeline was to be able to incorporate raw sequence data in from future experiments, we used the tool Velvet (Zerbino and Birney, 2008) to assemble raw Illumina sequence data downloaded from the SRA (Leinonen et al., 2011) of one *Lactobacillus* strain (*L. kimbladii*). The simple script used for velvet assembly (velvet assembly.txt) used a k-mer size of 29 to assemble each interleaved paired-end Illumina .fastq data in given folder. Our final assembly had an estimated coverage of 54.67x with an n50 of 30,992 but was not nearly as completely assembled (618 contigs) as the data on Genbank (130x coverage; n50: 170,214; number of contigs: 40). This is likely because additional long read next-gen data (2 runs of 454 GS FLX Titanium) was used to complete the genome of this strain as evidenced in the SRA record and we did not attempt to use this data with the Velvet assembler. Next, we annotated the 618 contigs of our genome for feeding into OrthoMCL (Fischer et al. 2011). We planned to use the program Prokka (Seemann, 2014) on the cluster, but one of it's dependencies, aragorn, was not installed in the cluster and it did not seem that we could skip the step of tRNA scanning that required this module. Thus, we tried using the program GeneMarkHMM (Lukashin and Borodovsky, 1998) on the cluster but encountered problems and weren't certain that the version installed was not only intended for eukaryotic DNA. We settled on using a web service of GeneMarkHMM specifically for prokaryotes for our assembled genome using the strain *L. kefiranofaciens* as a reference. We returned 2176 predicted protein genes for this taxon in comparison to the 1891 original annotated proteins.

We incorporated the newly annotated genome along with the original genbank annotation and proteome of that strain for comparison making a total of 38 proteomes. OrthoMCL is a tool which takes a set of protein sequences per taxon and used a reciprocal best blast approach to assign orthologs to groups of other orthologs. We acquired the proteome for each strain of *Lactobacillus* we wanted to include in our analysis and put the fasta file in a folder named by taxon (but abbreviated to four characters as required for OrthoMCL). Eric wrote a script (nonarrayjob_orthomcl_adjust_fasta.sh) for taking each folder in the working directory and adjusting the fasta files based on the folder name for input into OrthoMCL. The next steps mostly use OrthoMCL specific commands and are described in the file, Orthomcl first part.txt, but in brief, Blast (Camacho et al., 2009) is used to search all proteins against each other, these results are loaded into a MySQL database (Widenius et al., 2002), several other intermediate steps are conducted by OrthoMCL before the program mcl (Dongen, 2002) is used to cluster ortholog groups. We output all singletons (any protein which OrthoMCL did not pair with any other protein in our dataset) as a fasta file. We also used a couple of already written scripts (CopyNumberGen.sh, ExtractSCOs.sh, ExtractSeq.sh) to pull out the sequences of all of our single copy ortholog group which totalled up to 400 genes in our final dataset. This was down from 420 in our first run through that didn't include our newly assembled genome and had some mistakes in the fasta files for a couple of other genomes. Unfortunately, there was not enough time to repeat tree construction on all the final files although it is expected that they would be similar if not identical.

As a way to look at the relative enrichment in different ortholog groups in bee versus non-bee associates, we calculated the average number of orthologs present in each ortholog group for bee-associated species and in environmentally-associated strains. We then took the ratio of those two values to find the top 100 ortholog groups most and least enriched in bee associates. This data is seen in the file xx and with the top and bottom 100 in files xx xx. This measure of the most frequently duplicated orthologs in bee associates in comparison to environmental strains may relate to the bee niche of these species and these genes can later be assessed for function to see if certain categories of proteins are abundant in these genes. However, because we have multiple representative strains of the same species in one case (for *L. kunkeei*) some of the genes may suffer from being more highly represented in just this one particular species and not necessarily enriched in all bee associates.

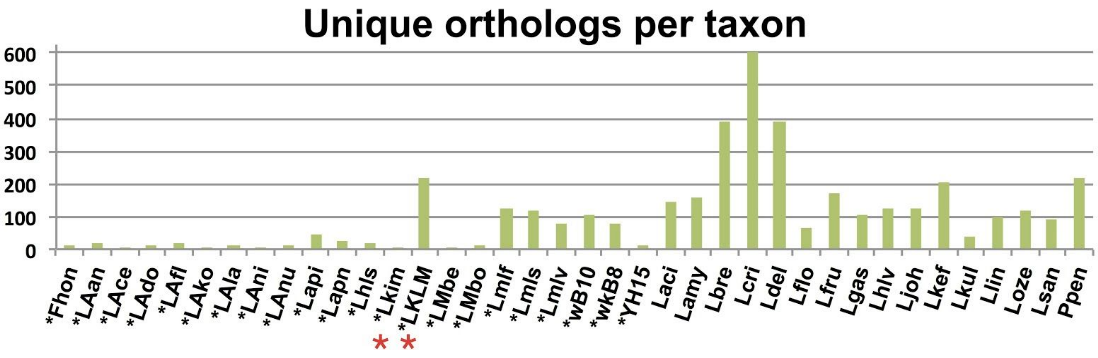

Figure 2. Unique orthologs per taxon. Bee associates on left, marked with *. Red asterisks indicate two versions of one genome

A graph of the number of unique ortholog graphs show that some taxa have essentially no unique orthologs (<10) while a few have more than 100. Certain taxa (in shell terms: Fhon or LA* or LM*) represent the same species (*L. kunkeii*) and correspondingly have very low levels of unique orthologs. Looking at the number of unique orthologs in our newly assembled genome (LKLM: 216) versus the original annotation (Lkim: 1), it is apparent there are many more unique proteins in the newly assembled version of the genome. This raises the question of whether all genomes we incorporated were perfectly assembled and annotated and whether taxa with a large amount of unique orthologs could be an artifact of imperfect annotation. The very low number for the Lkim genome also probably reflects the inclusion of much of the same data in the newly assembled genome. Three environmental *Lactobacillus* strains, *L. brevis, L. crispatus* and *L. delbrueckii,* have an exceptional complement of unique orthologs.

The output file of OrthoMCL allowed us to extract gene identification numbers for each bacterial species within each orthologous gene. We were able to write a script (batch_retreival_uniprot.pl) that which uses an identification number to retrieve gene ontology data from annotated genomes online. This script works specifically for uniprot.org and therefore needed uniprot specific identification numbers. Since, OrthoMCL output only gives gene identification numbers (GI), automating this process was difficult, and although works for uniprot IDs, did not work for GIs. Visualization of the most common gene ontology terms was done by text analysis to create weighed lists in R.3.2.2 (2015). We used parameters of a word frequency of at least ten. We were able to visualize the top gene ontology terms for enriched ortholog groups in bee associated *Lactobacillus* (Figure 3A), non-bee associated *Lactobacillus* species(Figure 3B), and for all "singleton" genes (found in only a single species) (Figure 4).

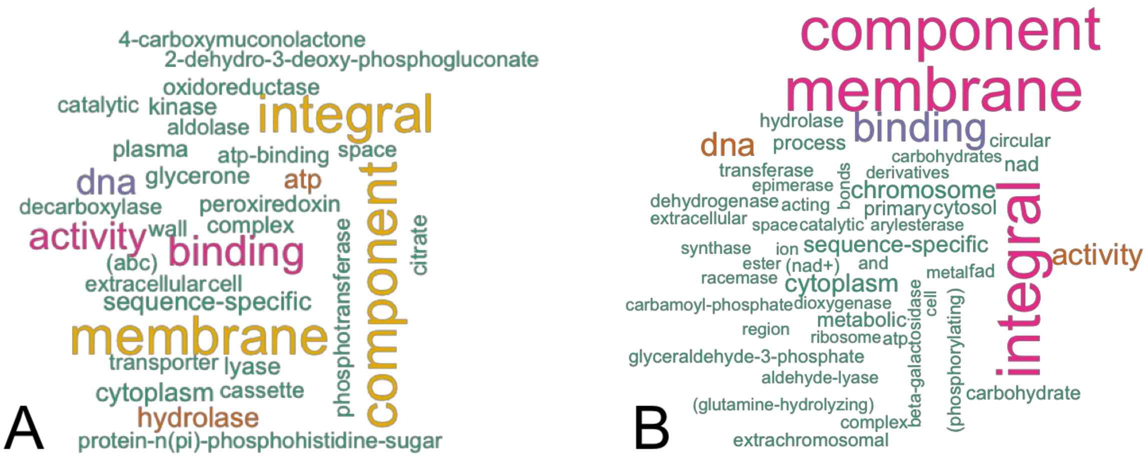

Figure 3. Gene ontology terms of top 100 enriched orthologs groups for bee-associated *Lactobacillus* species (A) and non-bee-associated *Lactobacillus* species (B) ranked by frequency term is used, according to scale size of words.

Bee associated *Lactobacillus* species had 58 top gene ontology terms according to our parameters, the most used terms being "integral membrane component", "activity", and "binding" (Figure 3A). Of the 1,258 GIs identified by OrthoMCL only 843 had information on gene ontology. Non-bee associated *Lactobacillus* species had a total of 1,200 GIs with 819 returning gene ontology specialization. Enriched orthologs for non-bee associated species had similar gene ontology designations as those for bee-associates (Figure 3B). More interesting data may be in those terms less frequently assigned. Singleton orthologs had the same top hit gene ontology terms, however, had many more terms overall (Figure 4). These results are most likely due to the abundance of unidentified gene ontologies or housekeeping genes.

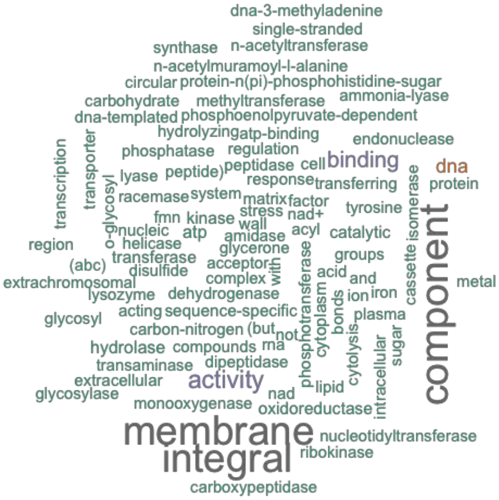

Figure 4. Gene ontology terms of single copy ortholog genes ranked by frequency term is used, according to scale size of words.

In total, exactly 420 orthologs sampled for 35 taxa were represented in our first data sets. Four taxa, however, had duplicate orthologs. We first employed MAFFT v7.266 (Katoh and Standley, 2013) with the fast "-- retree 1" option for all amino acid (AA) ortholog alignments. Ambiguous regions were then trimmed from alignments via trimAl 1.2rev59 (Capella-Gutierrez et al., 2009) using the automated1 setting. Because all ortholog OrthoMCL output files contained multiple copies of the same four taxa, we used the "mogrify --deduplicate-tax" command of seqmagick v0.6.1 (https://github.com/fhcrc/seqmagick/) to remove duplicate taxa and associated sequences. The first occurrence of a taxon in an alignment was subsequently retained. Trimmed aligned ortholog fasta files were then converted via FASconCAT-G_v1.02 (Kück and Longo, 2014) into individual phylip files, as well as a concatenated supermatrix, for individual gene tree and supermatrix analyses. Simple Unix loops incorporating the appropriate commands for each of the above utilities were integrated into a single shell script, suitable for pipeline analysis, which is freely available here. The relaxed hierarchical clustering algorithm was employed via PartitionFinder v.1.1.1 (Lanfear et al., 2012) for selection of best fit AA substitution models for the partitioned supermatrix according to the corrected Akaike information theoretic criterion. Execution of data partitioning scheme selection with PartitionFinder was, however, time prohibitive for this data set. The initial PartitionFinder selection of best-fitting models for all 420 subsets (partitions) in the supermatrix was implemented for partitioned maximum likelihood (ML) analysis via RAxML hybrid-MPI on CIPRES using the "-f a" option: 1000 rapid bootstrap replicates (RBS) followed by 200 tree searches for the best scoring ML topology. A second unpartitioned supermatrix analysis was similarly executed with the RAxML PROTGAMMAAUTO setting on CIPRES. Analysis of the 420 individual AA alignments was executed via RAxML 8.2.4 (Stamatakis, 2014) on the UCR biocluster with the aforementioned 1000 RBS unpartitioned tree searches and model settings. A custom script for executing RAxML batch analyses on the UCR biocluster was provided by J. Stajich. The script is available here.

The 420 resulting gene topologies were then input into ASTRAL v. 4.7.12 (Mirarab and Warnow, 2015) for coalescent-based species tree reconstruction. RAxML RBS were utilized for ASTRAL bootstrapping, and 1000 ASTRAL bootstraps were executed for assessing nodal support of the optimal ASTRAL topology. It is worthwhile to note that coalescent-based species tree inference does not account for gene duplication or genetic exchange, and assumes the following conditions apply to the taxa being sampled: panmictic population structure, no recombination within loci, and no natural selection acting upon genetic loci. Furthermore, shortcut coalescent analyses, such as ASTRAL, also assume that the input gene trees are reconstructed without error. Here ASTRAL was tested for suitability of phylogenetic reconstructions for a group in which horizontal genetic exchange is likely, and to compare to hypotheses of vertical descent (i.e., bifurcating trees) obtained with concatenated supermatrix analyses.

Resulting topologies for partitioned and unpartitioned ML analyses were identical (2 x RF=0 calculated via PAUP* v.4.0a146), differing slightly only for nodal support within *L. kunkeii* (Fig. 5A). Clades recovered via ASTRAL II were largely congruent with the supermatrix topology, and differed only for marginally supported nodes within *L. kunkeii* and at one other weakly supported node in the ASTRAL topology (2 x RF = 8; Fig. 5B). However, resolution within *L. kunkeii* was poor for ML supermatrix topologies obtained with the AA dataset due to short branch lengths. The above pipeline was modified and FASconCAT-G_v1.02 was employed for backtranslation and exclusion of ambiguous third codon positions prior to trimming alignments with trimAl. This modification of the analysis pipeline was necessary to preserve reading frames. Exclusion of ambiguous alignment regions for the 1st and 2nd codon position dataset eradicated five of the 420 orthologs, resulting in a supermatrix of 415 data partitions. Furthermore, the original AA supermatrix consisted of 127546 total characters, of which 10758 were parsimony informative. The backtranslated supermatrix consisted of 85053 characters, of which 40891 were parsimony informative. Because of multiple short gene partitions within the DNA supermatrix (<<100 bp), partitioned analysis was excluded from analyses. Likewise, phylogenetic reconstruction of individual genes would have resulted in a substantially reduced dataset after filtering for short alignments. Thus, a single unpartitioned supermatrix run was executed on CIPRES using the GTR+Γ model of sequence evolution and the above tree search settings. Nodal support for the resultant ML DNA based topology did not change for higher order relationships included in our taxonomic sampling, with the exception of the basal position of *L crispatus*□ sister to *L kefiranofaciens* + *L amylovorus*□+ *L acidophilus*. The same clade was subject to incongruence for ASTRAL vs. RAxML AA based topologies (Fig. 5). An alternative hypothesis for relationships within *L. kunkeii*, with greater nodal support for clades in comparison to the AA based topologies was recovered with the backtranslated dataset (Fig. 6).

For detecting selection among ortholog groups, a nucleotide data set must be organized by ortholog class. OrthoMCL outputs all orthologs present in sampled taxa and categorizes and pools protein sequences by ortholog groups. Combined with the subsequent protein alignment, the data is well organized and it was a priority to retain this organization. The headers in every protein sequence contains a GI(genInfo identifier) that links to a genpept record on NCBI. Most genpept records report the origin nucleotide sequence and its range in its origin sequence. Taking advantage of this we parsed out all GI numbers by ortholog group, then used NCBI's Edirect through Biopython to access Entrez and pulled out the genpept record. A precoded biopython script extracting corresponding nucleotide sequences from genpept records were available from biopython pipermail, only modifications to input variables and adjusting output variables were changed. Next were to adjust the headers of the aligned protein sequences as these sequences in each ortholog group had to have identical headers with their corresponding nucleotide sequences. Once adjusted the sequences can be processed by a biopython script available on the galaxy project called "align_back_trans.py".

Discussion

To improve similar analyses in the future, we would change a couple of things that we learned during this process. We would only analyze very well assembled genomes to limit the amount of artificial singleton orthologs. We would also make sure to download data from individual genomes in the first place instead of all proteins from one species which could include multiple genomes. Downloading data of multiple genomes created problems with back translation because the protein records were "non-redundant" meaning they corresponded to multiple DNA sequences. Another thing we would implement is individual analyses of Gene

Ontology terms for each genome instead of artificially pooling these terms all together which was done for the sake of time.

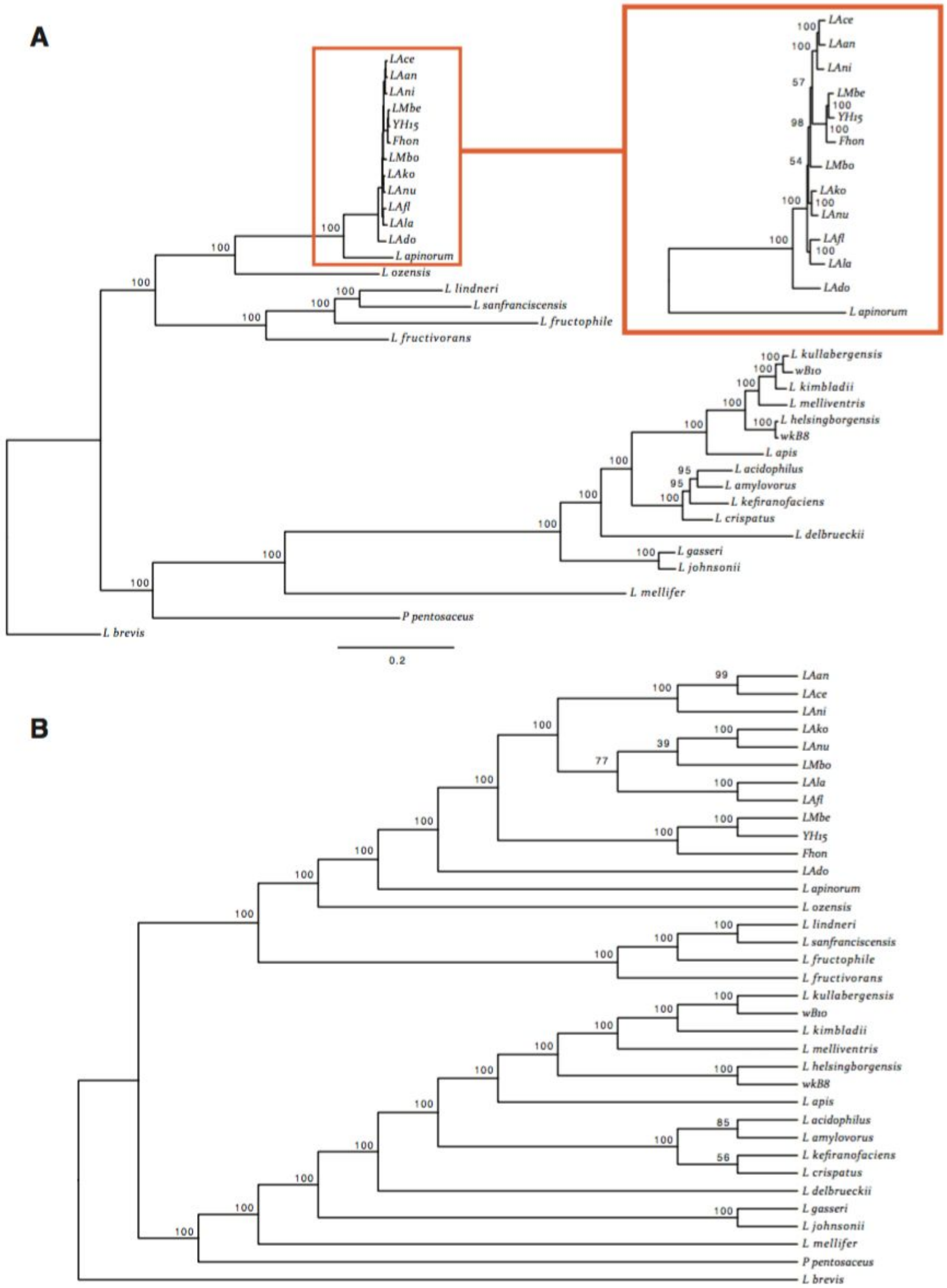

Figure 5. (Above).. A) RAxML AA based hypothesis for 12 *L. kunkeei* strains (inset in orange box) and outgroup taxa. B) ASTRAL AA based hypothesis.

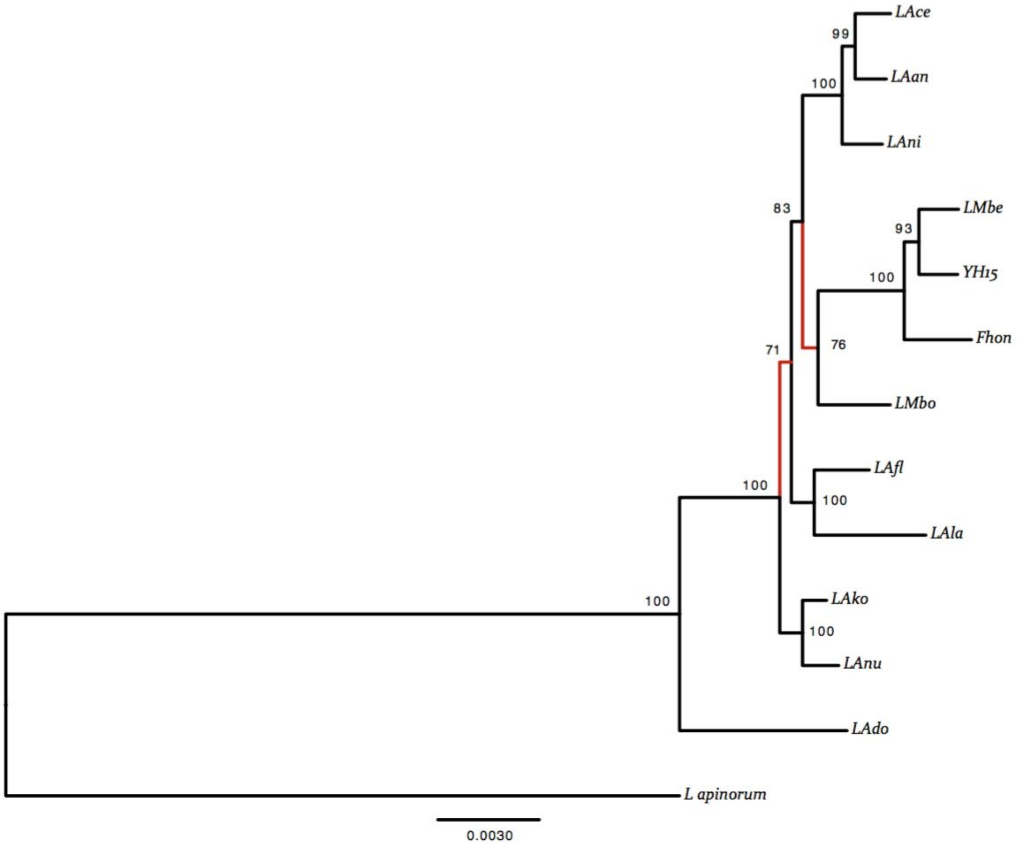

Figure 6. RAxML DNA based hypothesis for 12 *L. kunkeei* strains. Clades that differ from the RAxML AA based hypothesis are marked in red.

OrthoMCL outputted a lot of information that could be utilized in additional ways. A way to filter for orthologs groups which are well represented in enough genomes (but not necessarily only single copy orthologs) that we could then use to detect selection pressure would be useful. Looking at ortholog groups that are enriched or impoverished in all bee associates pooled together is how we chose to look at gene function, but ideally this would be done on a clade-by-clade basis instead of grouping unrelated bee associates together. This comparison also is restricted to orthologs that were present in both bee and non-bee associates thus leaving ortholog groups that are only present in bee-associates and not in other strains unanalyzed.

Most single copy orthologs found in all species and strains are likely housekeeping genes, which are usually conserved. With the abundance of genome data available, it is indeed worth dividing up the strains and species into subset by niche or clade location. This approach will benefit other analyses as well as there are a diversity of orthologs present at an intermediate amount of strains and species that will provide interesting results to analyze.

Our DNA-based RAxML *L. kunkeei* topology was largely congruent with the previous analyses of Tamarit et al. (2015), with the exception of one node within LMbe+YH15+Fhon. The recovery of a virtually identical topology with a fraction of the orthologs of Tamarit et al.'s previous analysis bolsters confidence that our phylogenetic hypotheses are plausible. Likewise, our higher order relationships were identical to Ellegaard et al. (2015), with no visible incongruence. Less apparent is the comparison between ML based supermatrix analyses and ASTRAL coalescent analysis. Without an identical dataset of nucleotides for all three tree reconstructions, it is not possible to optimize topologies under a single model of sequence evolution to discern which topology is the most likely.

Unfortunately, no nucleotide based ortholog analysis by back-translation from a protein alignment was completed due to the obstacles posed by the data format required for analysis. Phylogenetic reconstructions utilizing protein and first and second codon position ambiguous nucleotide back-translation data already show limited incongruence, only limited in the *L. kunkeei* clade. This suggests that with the full nucleotide sequence comparison that the incongruence may be amplified in that clade as well as arise in other clades. The third codon position provides more depth in the data for tree construction as the inclusion of only first and second codon position nucleotides in ambiguous back-translation provides less data to analyze. Mutations in the third codon position have a more conservative effect on an organism, as many may not change the protein sequence. This is where analysis of non-synonymous and synonymous mutations in protein sequences can be inspected for selection-based insight. The analysis of this ratio would provide insight on how selection is acting on a given gene. Though worth inspecting, given the single copy orthologs (likely to be common housekeeping genes) in our phylogenetic analysis, it is very unlikely to find orthologs under heavy positive selection.

Our GitHub repository for the project is located here and our presentation for the class is located here.

Author contribution statement:

Eric did assembly and annotation of the raw sequence reads. Eric and Kaleigh output single copy orthologs for phylogenetic analysis along with unique per genome, bee-enriched and environmentally enriched ortholog groups for GO term analysis. Kaleigh analyzed gene ontology terms and made graphs and word clouds in R for our figures. David strung together multiple programs into an analysis pipeline for the conversion of OrthoMCL output fasta files into individual gene tree and supermatrix alignments. Modification of the pipeline was implemented for ambiguous backtranslation of protein alignments. David also executed phylogenetic analyses with RAxML and ASTRAL II. Hoang processed the OrthoMCL and alignment results from Eric and David respectively into format for the nucleotide back-translation alignment. This involved parsing all GI numbers, piping them to request their genpept reports from Edirect, then piped to a pre-coded nucleotide parser. Hoang also modified the script of the parser for unix loop input and for appropriate header format for the nucleotide back-translation. Likewise with the nucleotide sequences headers, the protein sequence headers needed to be parsed before making the nucleotide back-translation. We all gathered the Genbank data on GitHub, met about progress, and helped each other with problems we encountered if we could.

Citations:

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. doi:10.1093/bioinformatics/btp348

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. doi:10.1186/1471-2105-10-421

Cock et al (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11) 1422-3. http://dx.doi.org/10.1093/bioinformatics/btp163 pmid:19304878.

Dongen, S., 2000. A cluster algorithm for graphs. Report-Information Systems 10, 1–40.

Ellegaard, K.M., Tamarit, D., Javelind, E., Olofsson, T.C., Andersson, S.G.E., Vásquez, A., 2015. Extensive intra-phylotype diversity in lactobacilli and bifidobacteria from the honeybee gut. BMC Genomics 16, 284.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., Stoeckert, C.J., 2011. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr. Protoc. Bioinformatics Chapter 6, Unit 6.12.1–19. doi:10.1002/0471250953.bi0612s35

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30, 772–780. doi:10.1093/molbev/mst010

Kück, P., Longo, G.C., 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. 11, 81. doi:10.1186/s12983-014-0081-x

Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701. doi:10.1093/molbev/mss020

Leinonen, R., Sugawara, H., Shumway, M., 2011. The sequence read archive. Nucleic Acids Res. 39, D19–21. doi:10.1093/nar/gkq1019

Lukashin, A. V, Borodovsky, M., 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26, 1107–15.

McFrederick, Q.S., Wcislo, W.T., Taylor, D.R., Ishak, H.D., Dowd, S.E., Mueller, U.G., 2012. Environment or kin: whence do bees obtain acidophilic bacteria? Mol. Ecol. 21, 1754–1768.

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31, i44–i52. doi:10.1093/bioinformatics/btv234

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–9. doi:10.1093/bioinformatics/btu153

Tamarit, D., Ellegaard, K.M., Wikander, J., Olofsson, T., Vásquez, A., Andersson, S.G.E., 2015. Functionally Structured Genomes in Lactobacillus kunkeei Colonizing the Honey Crop and Food Products of Honeybees and Stingless Bees. Genome Biol. Evol. 7, 1455–1473.

Widenius, M., Axmark, D., AB, M., 2002. MySQL Reference Manual: Documentation from the Source. “O’Reilly Media, Inc.”

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–9. doi:10.1101/gr.074492.107