Laurent Casini1 , Cyril Matthey-Doret2

**1** Affiliation not available
**2** Affiliation not available

# Abstract

This is the abstract where we write stuff about bee gut bacteria

Host-Specific Gene Content of Bee Gut Lactobacilli

Sequence a Genome II : 2017, Group 4

Laurent Casini, Cyril Matthey-Doret

Investigate which gene functions are specific to Firm-5 in either Bumble Bees (B) or Honey Bees (H).

Introduction

This project is a direct follow up of the Sequence a genome I (SAGE) course in which we assembled and annotated the genomes of several Lactobacillus strains colonizing honeybee and bumblebee gut. Our goal is to analyze the genome of the Lactobacillus strains with a focus on the host-related differences in gene content. Since the Lactobacillus strains are established in different hosts, their genomes likely show signs of adaptation reflected on their gene content. In addition and to compare the level of gene conservation with more distant isolates, we will also use other Lactobacillus strains that are not associated with a host.

The project is divided into three parts:

The first objective is to determine the host-specific gene families present in the different groups. A gene family is a set of homologous genes, both orthologs and paralogs (between and within a given strains, respectively). The second goal is to investigate the level of gene families conservation across-strains. Finally, we wanted to determine what function are enriched in the gene families of the different groups of strains. The combination of these approaches allowed us to explore which functions are important for the Lactobacillus to establish itself in a given host.

Fig 1:  Phylogenetic tree based on 16srRNA displaying the relative position of various Latobacillus spp. The green, red and blue bar represent the differents group considered in this project (Outgroup, Bumblebee and Honeybee, respectively). The strains in bold are strains that have been sequenced during the first part of SAGE. Adapted from Olofsson., et al 2014.

Methods

We used data from 35 Lactobacillus strains in this project, among which 10 were bumblebee associated (B), 17 were honeybee associated (H). The outgroup (O) consists of the 8 remaining strains which have no host. We were provided the following data by Kirsten M Ellegaard: the FASTA files containing all genes of each strain, the orthoMCL tables containing homolog gene families (orthologs and paralogs) for all 35 strains and the strain list. An overview of the pipeline is presented in Fig. 2.

Fig 2:  Pipeline overview of the project. All the objectives start with the same original data, which include FASTA files, OrthoMCL table and the strains list. Data of gene families are then split for each group (Bumblebee, Honeybee and Outgroup) and in every cases, a set of all and core gene families is generated. Data that are needed for the different objectives are indicated with arrow. In addition, for the gene function analysis, eggNOG mapper was used to annotate gene families. A more detailed version of the pipeline overview is available on Github (<https://github.com/cmdoret/SAGE2>.).

Host-specific gene families

We split the orthoMCL table into 7 separate ortholog sub-tables. Each sub-table contain gene families that are contained in a given “host-group”,  that is any combination of Bumblebee, Honeybee or Outgroup Lactobacilli. For instance, the B sub-table contains gene families contained exclusively in bumblebee strains, whereas the BH sub-table contains those shared between bumblebee and honeybee strains but absent from the outgroup.

When counting the number of gene families, we also included singletons (i.e. genes without any homolog) to include all unique genes. Each singleton is considered as a gene family on its own.

For each host group, we distinguish a subset of gene families we call “core” gene families. These are gene families present in all strains of their host group.

Distribution of gene families across strains

To investigate the conservation of gene families across strains, we quantified the number of strains in which each gene family is found. This was done using the ortholog sub-tables of each group.

Functional role

We concatenated the FASTA files containing the protein sequences of all strains and submitted the resulting file (53,023 genes) for annotation to the eggNOG-mapper server (version 4.5.1). When setting the parameters for the eggNOG-mapper tool, we used the DIAMOND mapping mode and Firmicutes taxonomic group. To prioritize coverage, we included all orthologs and non-electronic terms. EggNOG-mapper annotated 14,900 genes (28%) from the concatenated file and these annotations were used for downstream analyses.

We then performed GO term enrichment tests, comparing the relative abundance of all GO terms in each host-group to other strains. We measured enrichment for every GO term using Fisher exact test and corrected p-values for multiple testing using the Benjamini-Hochberg method. We then retrieved the biological annotations corresponding to GO terms by sending SQL requests to the Gene Ontology database (<http://geneontology.org/>). We filtered annotations to include only those that were significantly enriched or depleted (corrected p-value <= 0.05) and odd ratio was used to measure enrichment strength. A schematic of the functional role pipeline is displayed in Fig.3.

Fig3:  Flowchart of GO term enrichment. FASTA files were submitted for annotation to eggNOG mapper and we extracted biological annotations using the GO identifiers. For each GO term, we used Fisher exact test to test for enrichment. The contingency table on the right shows the configuration of the test: The number of occurrences of GO terms in the different groups are used to quantify enrichment. Odd ratio is used to measure enrichment strength and q-values (corrected p-values) to assess significance.

Data availability

The whole pipeline was implemented in Makefile and is available on github, along with original data and the eggNOG-mapper annotation file, at: <https://github.com/cmdoret/SAGE2>. Implementation details and instructions are documented in the README file of the repository.

Results

Host-specific gene families

Results of the host-specific gene families are visible in Fig. 2. We found very few core gene families in all categories, except for the BHO group (i.e. gene families found in bumblebees, honeybees and outgroup) where there were 595 core gene families. This is likely due to the presence of gene families responsible for maintenance of the cell and essential functions in this group. We found that the outgroup (O) shares almost 10X more gene families with bumblebee strains (B) than honeybee strains (H), however this does not necessarily imply bumblebee strains underwent more genomic adaptations. Indeed, this result is susceptible to differences in group sizes and should be interpreted with caution as there are 17 bumblebee strains, but only 10 honeybee strains.

Fig 2: Venn diagram representing the number of gene families per group. Each color correspond to a different group. Black numbers correspond to gene families that are present in at least one strain in a given host. Blue numbers correspond to “core” gene families that are present in all strains in a given host.

Distribution of gene families across strains

Results of the across strains distribution is shown in Fig.3. Since the notion of core gene family is very stringent and to assess the conservation of gene families, we quantified their distribution across strains. We observed that gene families shared between honeybee and bumblebee (BH) strains are more conserved than those specific to honeybee (H) or bumblebee (B). The number of strains differs across groups, but this only affects the scales of the distributions and the proportion of highly conserved gene families is different between host shared than in host specific gene families.

Fig 3: Distribution of gene families across strains. Y-axis corresponds to the density of gene families. X-axis corresponds to the number of strains possessing a given gene family. B stands for Bumblebee group, H for Honeybee group and BH for the Bumblebee-Honeybee group.

Functional role

The gene families annotation results are displayed in Fig.4.

Fig 4: Gene Ontology (GO) terms enrichment per host.  (A) For Honeybee group and (B) for Bumblebee group. In each case, the categories of the GO terms are represented on the right. Enrichment are measured using odd ratio value superior. An odd ration superior to 0 indicates an enrichment and an odd ratio value inferior to 0 correspond to a depletion. The color code corresponds to the corrected p-value (qval). Annotations that are enriched are represented with a green (+) and annotations that are depleted are displayed with a red (-).