UPPSALA
UNIVERSITET

Selfish, mobile genes in honeybee gut bacteria

Aleksei Põlajev

Degree project in bioinformatics, 2018
Examensarbete i bioinformatik 30 hp till masterexamen, 2018
Biology Education Centre and Department of Cell and Molecular Biology, Uppsala University
Supervisors: Siv Andersson and Andrea Garcia Montaner

Abstract

Transposons are selfish, mobile genetic elements, moving within the genome. The transposase gene makes this possible, as it codes for the enzyme that catalyzes the movement. In the case of bacteria, they can also move horizontally between individual bacteria, and sometimes even between species. By default, they are a burden for the host organism, coding for a protein that the host does not need. They also pose the risk of disabling the host's crucial genes by inserting themselves into it. Transposons are under some pressure to benefit the host, to help propagate themselves more effectively. And some transposons have indeed evolved to benefit the host.

Lactobacillus kunkeei is a bacterial species known to reside in honeybee guts. It is known for its role in honey preservation and wine spoilage. The genome of *L. kunkeei* is reduced because it is a symbiont, however it contains an unusually high amount of transposons in its genome.

In this study, the transposase genes (transposon enzymes) found in *L. kunkeei* are studied and categorized. The *L. kunkeei* have been extracted from honeybees (*Apis mellifera*). The honeybees themselves have been collected from the islands Åland and Gotland.

This study focuses on the transposase genes that come in pairs, one after another in the genome. Transposase genes were identified using annotation software and orthology-based methods. The annotation software provides numbering for the genes, which allows finding paired genes. The paired genes were categorized based on alignments and phylogenetic software. Pseudogenized transposons were identified based on length and/or clustering into triplets.

A total of 766 paired transposase genes were found. The transposase genes were found to take up 1.9% of the genome, on average. A low level of diversity has been found when performing alignments and generating phylogenetic trees. The positions of the transposase genes are generally conserved within phylogenetic groups. Pseudogenization has been detected for some transposase genes – 4.5 per genome, on average. All of the studied transposons belong to the IS3 family, which is a family of Class I transposons.

Selfish, mobile genes in honeybee gut bacteria

Popular science summary

Aleksei Põlajev

Most genes have a constant location in an organism's genome. At least, the order of the genes does not generally change with every generation. Much like an instruction manual that may be updated, but the order of the chapters rarely changes.

However, there is an exception: the Mobile Genetic Elements (MGE), and more specifically in this study, the transposons. They are genetic systems that move around the genome. Sometimes the method that they use is copy-paste (Class I), sometimes cut-and-paste (Class II). Transposons appear to be present in all lifeforms. With bacteria, things get even more messy – the transposons can travel from individual to individual, and even between species! This is because bacteria are known for absorbing DNA from their environment (known as transformation, one of the Horizontal Gene Transfer (HGT) mechanisms) and keeping it, rather than always rejecting or destroying it.

By default, these mobile genetic elements are a burden for the host organism. A waste of resources for their copies. And what if they land on an important gene? That is like copying a sentence from one book and pasting it into the middle of a sentence in another book – the sentence will not make any sense anymore. So the transposase can disrupt an important gene and damage the host organism.

However, some transposases have also evolved to carry beneficial genes as passengers, therefore copying good genes and/or spreading them around between individual bacteria.

This study is about transposase genes (the central element of transposons) in a honeybee gut bacteria – the *Lactobacillus kunkeei*. This species is otherwise known as the "ferocious lactobacilli" and a spoiler of wine. Later on it was found to be crucial for healthy honeybee colonies. Their genomes are rather reduced (just 1.5Mb), and hence it would be expected that they would tend to evolve smaller genomes due to their specialized, sheltered existence. They would also be more isolated from "transposase invasions" from the outside. However, surprisingly many transposase genes have been found in the genomes of *Lactobacillus kunkeei*.

Patterns of paired transposase genes have been detected. The transposase genes were divided into groups according to different traits, but mostly according to their similarity in sequence alignments and phylogenetic clustering. It appears that they have been active recently in the honeybee gut bacteria, and their positions appear to be quite similar between related groups of this bacterium. Moreover, based on sequence length it was possible to identify transposase genes that most likely have undergone a pseudogenization process, inactivating the transposon. With these results, the current study tries to shed light on the biological implications of the trasnposons abundance by studying the diversity, the chromosomal distribution and sequence conservation.

Table of Contents

1 Introduction

Transposons are selfish, mobile genetic elements. They are capable of moving within a genome, and in the case of bacteria, even between genomes of different species. By default, they are a burden for the host organism. In worst cases, they can also disable important genes by inserting themselves into them, severely impacting the fitness of the host. This study focuses on the transposase genes. That is the enzyme which catalyses the movement of the transposon to another part of the genome. As transposons can travel within the genome and even between individual bacteria, they offer an opportunity to study (recent) evolution. Considering their default harmful effects, one could ask how prevalent they are in reduced genomes.

Lactobacillus kunkeei is a bacterial species known to reside in honeybee guts and protecting the bees' sugar-rich honey supplies by inhibiting yeast growth. It is also known for "spoiling" wine by, once again, inhibiting yeast growth (Bisson *et al.* 2016). The genome of *L. kunkeei* is reduced because it is a symbiont, however it contains an unusually high amount of transposase genes in its genome. Previous studies show that symbiotic bacterial genomes tend to contain few selfish genetic elements, as these bacteria live in relative isolation, which stops the horizontal spread of selfish genetic elements (Frank *et al.* 2002). The reduction of the genome also tends to delete existing transposons along with other genes.

Honeybees (*Apis mellifera*) were previously collected from the islands Åland and Gotland (see Figure 1) as part of an ongoing project. From these bees' guts, *L. kunkeei* was extracted and analysed. Therefore, all the genomic data in this project has been provided by the research group, and was used to study the transposase genes in these genomes.

One goal of this study is to check if the percentage of transposons matches the expectation: reduced genome – few transposons. Another is analysing the transposase genes to determine their conservation patterns and functionality. The resulting data and findings of this project will help the research group continue studying the genome dynamics of *Lactobacillus kunkeei*.

Paired transposase genes have been found and categorized based on phylogenetic analyses and comparisons with databases. In addition, gene maps have been drawn and more general stats have been provided.

Figure 1: Origins of the samples. Source: Google Maps.

2 Background

Mobile genetic elements are genetic materials that can change place in the genome, or can be transferred from one species to another. Among them are transposable elements, plasmids (because of bacterial conjugation) and bacteriophage elements. Transposable elements (transposons) are defined as DNA sequences that are able to move from one location to another in the genome (Munoz-Lopez & Garcia-Perez 2010). Transposases are the enzymes that are responsible for catalyzing the movement of the transposable element to a different location in the genome. This study focuses on transposase genes. More precisely, it focuses on paired transposase genes, their categorization and their positions in the genomes.

In bacteria, transposons can also jump to different specimens, and even between species. They are a burden to the host, by default, as they are essentially an example of synthesizing a protein that the host organism does not need. For the host, there is an evolutionary pressure to delete transposons from its genome. For example, a deletion in the transposon would benefit the host.

For the transposon, there is evolutionary pressure to benefit the host, to counteract the aforementioned burden that it gives. This is done to promote the genomes containing the

transposon, and thus promoting the transposon. The reduction in fitness suffered by the host due to transposition ultimately affects the transposon, since host survival is critical to perpetuation of the transposon (Munoz-Lopez & Garcia-Perez 2010). Therefore, a given transposon can evolve to carry beneficial genes on the edges of its sequence, as it relocates. The genes could benefit the host individually, or the entire population of this species in a given area (Rankin *et al.* 2011). For example, secreting mucous or toxins into the environment that hinder other species of bacteria, but not the hosts of the transposons. So for the host organism, it would be beneficial to reduce its genome by deleting any non-beneficial transposons. Severe genome reduction is a common trait in symbionts, i.e. organisms that started living inside other organisms (Moran & Bennett 2014). *Lactobacillus kunkeei* has experienced a reduction in its genome compared to other closely related species, as it became a symbiont of the honeybee. Furthermore, the isolation granted by the symbiotic lifestyle generally hinders the horizontal spread of transposons (Frank *et al.* 2002). However, in this study, a surprisingly high number of transposase genes have been detected.

Phylogenetic studies are used to infer evolutionary relationships between species and strains. The applications include studies of the history of evolution, tracing the spread of an epidemic (Yang & Rannala 2012), tracing a family lineage and comparing rates and patterns of sequence evolution in a genome. Phylogenetic software is used to infer phylogenetic trees based on sequence alignments (Yang & Rannala 2012), whether they are nucleotide or protein sequences. Although many different types of methods are available for these studies, such as for example distance-based, maximum likelihood (Stamatakis 2014) and Bayesian methods (Ronquist *et al.* 2012), they are all based on the assumption that sequences change through a gradual accumulation of mutations. Of all the methods used to categorize transposase genes in this study, the phylogenetic method is the quality standard, accounting even for minor differences in the sequences.

3 Materials and methods

The transposase genes need to be detected in the genome and verified with a different approach/program. This will already allow us to see how prevalent these genes are, provide their positions and provide the sequences themselves for further analysis. Afterwards, the genes need to be categorized based on their similarity and preferably in an automatable way. The genes also need to be checked for pseudogenization. Finally, the genes bordering with the transposases need to be looked into as well, possibly providing explanations for their prevalence. For these purposes, a variety of programs and scripts were used. A summary of the software used is provided in Table 1.

3.1 Genome data

A total of 26 genomes of *L. kunkeei* (8 from Åland and 18 from Gotland) had been previously sequenced, assembled and annotated. All genomes are complete, and have an average size of 1604562 nucleotides (1.6 Mb). Both the nucleotide and proteome data are in fasta files. Additional information is available in Genbank files.

3.2 Identification of transposase genes

The genomes of *L. kunkeei* have been searched for transposases based on the annotation provided by different software. The genomes had been annotated with Prokka (Seemann 2014), which uses sequence homology via BLAST to identify the predicted genes. Besides, a functional annotation against the database of Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.* 2000) had also been performed. The aforementioned analyses had already been performed by this research group. Therefore, all genes annotated as "transposase" that were identified by any of these two methods were extracted and subject to further analyses. The COG-detected transposase genes are trusted more in this study, as COG is based on the principle of orthology (Tatusov *et al.* 2000), which allows us to assume not just an adequate sequence similarity, but a reasonable expectation of functional similarity.

The transposases were categorized based on two broad categories: whether they are detected by COG or only by Prokka. The latter ones were verified later via alignment to COG-analysis based transposases. Interproscan (Jones *et al.* 2014) has also been used to analyse the proteome data. This program includes several analyses under it, such as Pfam and SUPERFAMILY. The Pfam database groups proteins based on the presence of different functional regions (domains) and their combinations. The SUPERFAMILY analysis groups proteins together based on evolutionary relationship, with consideration for domains.

3.3 Transposon classification

The analysis focuses on the transposases that are "paired" - they are adjacent to each-other in the genome. According to a previous analysis in this research group, the overwhelming majority of "single" transposase genes appear to be bacteriophages. Attempts were made to categorize the paired transposases further – by length, length ratio (left-side length divided by right-side length) and by grouping them based on alignments/phylogeny.

For sequence alignment, MAFFT (Katoh *et al.* 2002) was used. Trimal (Capella-Gutiérrez *et al.* 2009) was used for trimming the alignments (remove those positions where more than 50% of the positions were a gap). RAxML (Stamatakis 2014) was used for the generation of phylogenetic trees. FigTree was used to visualize the aforementioned trees. As an extra classification method, the protein families defined by OrthoMCL (Li *et al.* 2003) were used to study orthology between the studied transposases.

Table 1: A summary of the software used for this study.

Software used	Version	Extra parameters, if applicable
MAFFT	7.305	--auto --reorder
RaxML	8.1.17.	RaxmlHPC-PTHREADS-SSE3 -m PROTGAMMALG -# 100 -T 4 -p 12345 -f a -x 123456
FigTree	1.4.2.	
Blastp	2.2.30.	-outfmt 7 -evalue 1e-6
Trimal	1.4.rev15	-gt 0.5
Prokka	1.12-beta	
COG	4.19.2012	
InterProScan	5.20-59.0	

After identifying all putative transposase genes based on the previously mentioned methods, patterns of transposases in the genomes have been studied to categorize the transposons. The locus tags (sequence Ids) extracted from the proteomes essentially number the genes in succession. Out of two adjacent genes, the second gene has the same locus tag number, plus 10. This is because of how the individual genes received their locus tags (ID) from the annotation program. For example, the genes G00401_00200 and G00401_00210 come after one another, no other genes in between. Using this, pairs of transposases have been identified and studied in further detail. Looking at the genome maps, these transposases were observed to come in pairs too.

Blastp (Altschul *et al.* 1990) has been used in this study, for the categorization of transposases in the Excel datatable of this work as well as for verification of dubious (not supported by COG) transposases. For extra verification, transposases unique to Prokka have been aligned against other transposases, separately against ”left-side” transposases and ”right-side” transposases.

3.4 Identification of pseudogenized transposases

A pseudogenized transposase must be considered differently from a functioning one. The pseudogenized ones are less significant as evidence for a recent expansion of transposons. On the other hand, conserved pseudogenized transposase positions (same position, different genomes) hint at a common origin for different genomes.

Pseudogenized transposases were determined based on lengths within homologous groups of transposases. Within each group, all transposases shorter than 80% of the longest, are considered pseudogenized. It is difficult to accurately predict pseudogenization without studying each sequence individually. However, the longest sequence in a homologous alignment can be considered as (closest to) the original transposase, with no/fewer deletion events. If a sequence is more than 20% shorter than that “original”, it is a deletion(s) long enough to disable an important functional unit (domain) in the protein.

In addition, three transposase genes in succession (triplets) are considered a pseudogenized unit here. When searching the sequences in the NCBI database, concatenating the second and third genes in the triplet gave more clear results than searching all the genes individually. Otherwise, there are more top matches that are putative.

3.5 Programming languages

The languages Bash and Python were used to write scripts. The purposes include data analysis, visualization and the automated generation of an Excel table with lists of transposase pairs and some of their attributes. The R language was used mainly for additional visualization.

3.6 Bordering genes

As previously mentioned, transposons might benefit the host organism by carrying useful genes. In other words, when the transposon moves (or is copied) to a new place in the genome, genes bordering with the transposase gene might be moved also. The genes and their function (according to the Prokka annotation) were observed.

4 Results

The ISfinder website (Siguier *et al.* 2006) was used to determine transposase families. However, all of them are from family IS3, with no further distinction between the transposases in this study. The IS3 family is a family of Class I transposons, meaning that it spreads via a "copy-paste" process, keeping the original transposon in its position (Chandler *et al*. 2015).

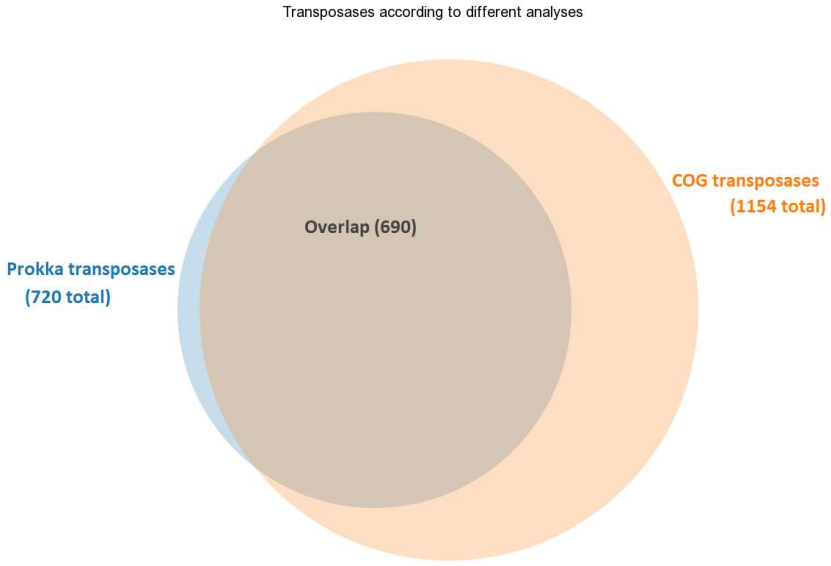

Figure 2: The number of transposase genes, according to the Prokka program (blue) and the COG database (pink). Some of the genes overlap (grey) - they were marked as transposases by both analyses.

A difference between the transposase detection of Prokka and Interproscan has been observed. On our genomes, Prokka and Interproscan have no overlap for transposase categorization. All genes marked as transposases by InterProScan are called "hypothetical proteins" by Prokka.

For the genomes, the detected transposase statistics are: 720 transposases detected by Prokka. 1154 transposases detected by COG. From those, 690 transposases overlap between the two analyses. 30 are unique to the Prokka analysis. 464 are unique to COG. So, in total, 1184 transposases have been detected. This can be seen in Figure 2. The numbers of transposase genes per genome can be seen in Figure 3.

Transposase genes take up roughly 1.9% of the whole genome on average, although the number of detected transposase genes is not constant across all the genomes.

We observed that 766 of all transposases detected came in “pairs” - they are located next to each-other in the genome, with no other genes between them. This is illustrated in Figure 4. This pattern repeats itself for several genomes. In most of these cases, the first transposase gene is shorter than the second one. In other words, transposase pair "Type A" is more widespread.

Further alignment analyses show that all 30 of the Prokka-specific transposases are pseudogenized. That would be 1,1 pseudogenized transposase genes per genome on average. These sequences have good matches against more trustworthy verified transposase sequences, but the Prokka-specific ones are significantly shorter – 50% of the length or less.

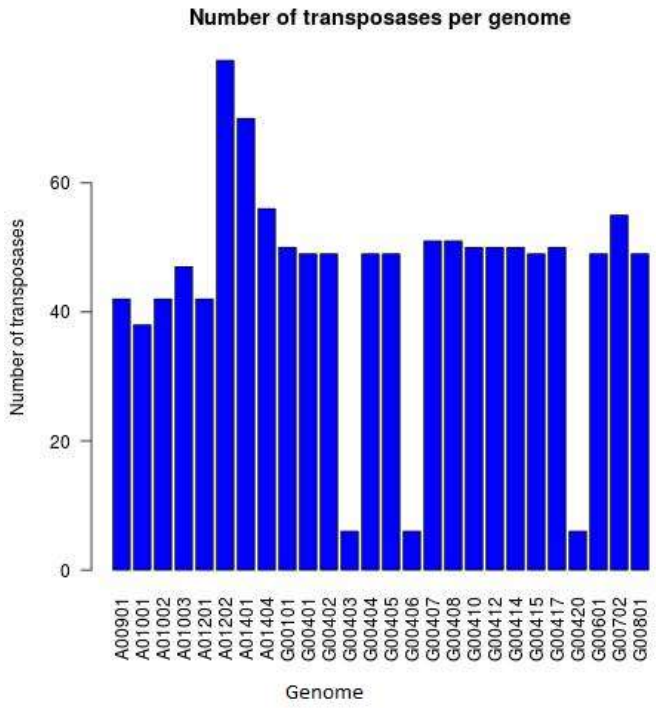

Figure 3: Number of transposases (y axis) per genome (x axis).

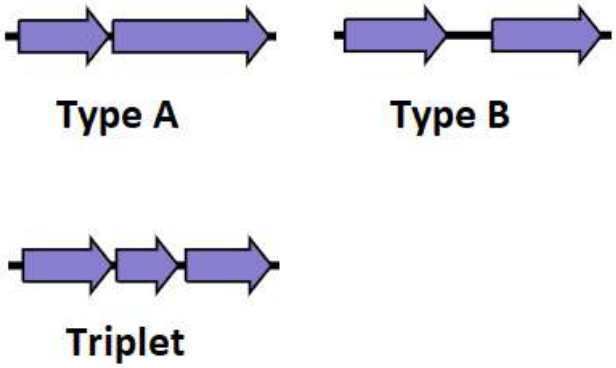

Figure 4: Patterns of paired transposase genes that have been observed in the genomes. Types A and B based on phylogenetic similarity. Triplets have also been found and considered as pseudogenized.

The purpose of the alignments and trees is to find groups of related transposases, potentially tracking their origin. When performing sequence alignments of transposases, and generating phylogenetic trees, distinct groups of transposases can be found. In every genome, there is always a dominant, relatively homogenous group, which are called Type A in this study, followed by branches with fewer transposases. When aligning all paired transposases of all genomes (left and right separately), there is also a dominant, relatively homogenous group.

There are multiple ways to group the transposases, such as lengths, length ratio (left-to-right) and using phylogenetic methods. In the end, the latter was used most extensively, as it is arguably the most precise one, although it is very difficult to automate. It is precise, as it accounts for all differences in sequences, proportionately to their severity, and it can be seen in the phylogenetic trees. Meanwhile, other methods to group the transposases (such as top Blast match) can divide them into some groups of similar sequences, but fail to see that two of these groups of sequences are more similar than different.

Genomes with no transposase-based trees (due to lack of paired transposases) have no phylogenetic classification automatically. Therefore, these transposases were aligned manually to a related genome's transposases to find the type.

When aligning all left-side transposases and generating the tree, a low level of diversity was detected between the transposases, as seen in Figure 5. In other words, most of the transposases are quite similar to each-other, if we compare the sequences. There are two distant branches dominated by a specific island's transposase genes, except for one interloper. But those two branches are a minority, greatly outnumbered by the main branch with the majority of the transposase sequences represented there, with little sequence diversity between Åland and Gotland transposases. That is, Åland and Gotland transposases are difficult to distinguish from one-another.

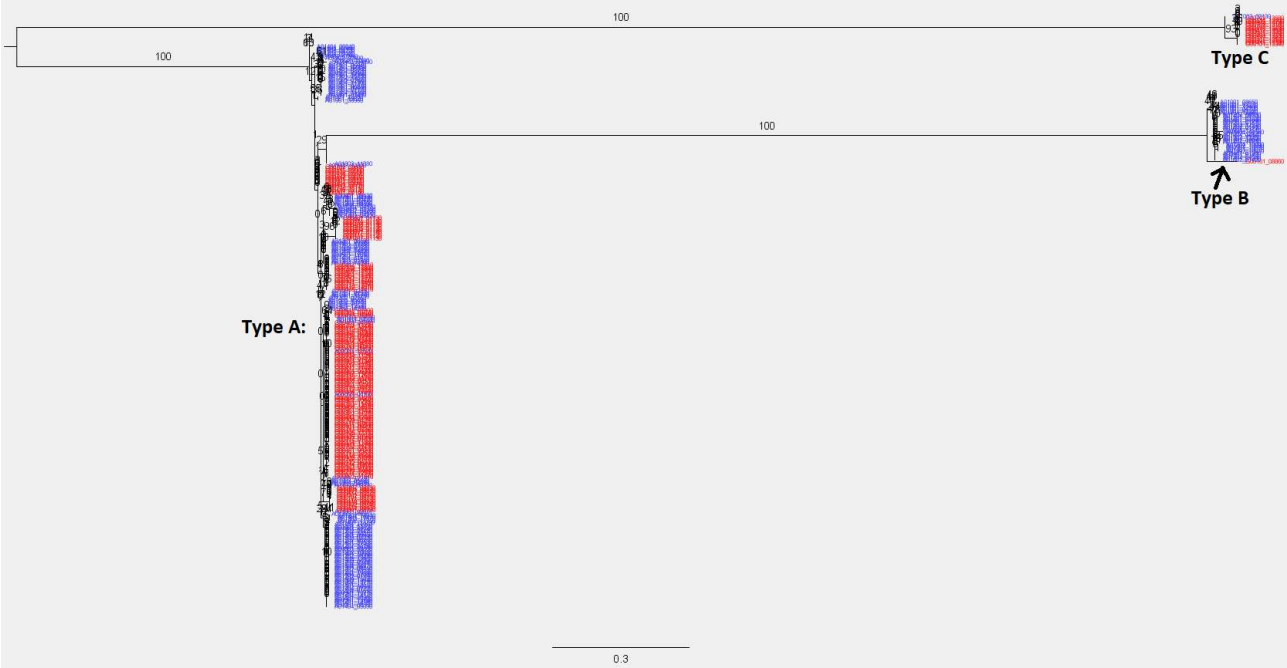

Figure 5: Phylogenetic tree of left-side transposases. Transposases from Åland specimens colored blue. Gotland specimens colored red. Numbers on branches are bootstrap values. Branches are drawn to scale according to substitutions per site.

In the end, 118 transposase genes were found to be pseudogenized. That is 4.5 genes per genome on average.

As a way of summarizing all the retrieved attributes on the detected paired transposase genes, a script was written to generate an Excel table (see supplements). It lists the transposases, left-side and right-side separately, and attributes such as detected Pfam families, detected SUPERFAMILY families, best blastp match when blasted agains our *L. kunkeei* proteomes collection, start position in genome, length of transposase, left/right length ratio, type of transposase judging by length ratio, type according to alignments/trees based on this group, OrthoMCL analysis results (the resulting cluster name) and notes concerning pseudogenization. Out of 12 columns, only 2 are filled manually, rather than generated by the script.

An overview of other genes bordering on transposase genes proved to be inconclusive. An overwhelming majority of these genes was given a vague name by the annotation software, such as “hypothetical protein”, “unique protein” or “putative protein”. Such genes might have had the potential to be carried along when the transposon relocates.

Finally, we present all the transposase genes, showing their type and position in the genome, in Figure 6. There are visible patterns in paired transposases and their positions in the genome on each individual phylogroup. Notice the positions of the transposases and the phylogenetic tree on the left side of the image. The transposases appear to be conserved within previously described phylogenetic groups. The major exceptions to the patterns observed were genomes A01003, A01202 and A01401, which show completely different transposon distribution compared to the rest

of the members of the same group. The positions of the transposases (and occasionally types) in the genome differ significantly.

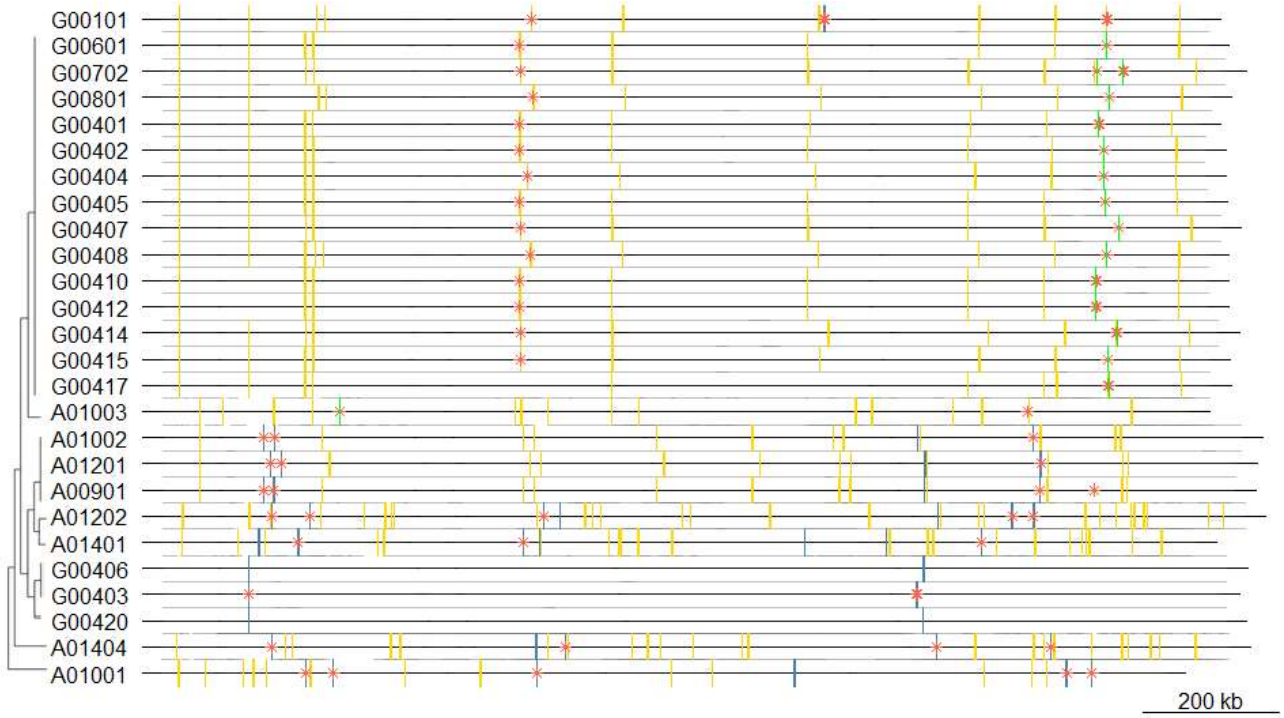

Figure 6: Paired transposase genes and their positions in the genomes. Genomes whose names start with A are Åland genomes, while G stands for Gotland. Yellow vertical lines represent Type A, the dominant group of transposase genes. Blue lines are Type B. Green is Type C. Red stars mean that the transposase has been pseudogenized. Some of the features are spaced very closely and therefore appear as one feature.

5 Discussion

The level of abundance for the transposons has been found. The transposons were categorized and their family found. Pseudogenized transposons were detected. Studying the bordering genes has proven inconclusive.

Prokka and InterProScan are fundamentally different tools, as Prokka relies on Blast searches to a reference genome(s) to find genes, including transposases. However, InterProScan includes multiple tools that search for functional units in sequences and try to group proteins into families and domains. Thus, it is not particularly surprising that these two tools diverge in terms of detected transposase genes.

The expectation is that symbiont bacteria have not only small genomes, but also few transposons (Frank *et al.* 2002). Here we see an average of 45 transposase genes per genome, or 1.9% of the whole genome's length. For the sake of comparison, another study found that the Baltic

Sea bacteria have an average of 1.7% of transposases per genome, which is already relatively high (Vigil-Stenman *et al.* 2017). It should be emphasized that the Baltic Sea study is based on samples of free-living bacteria.

The positions of transposases is remarkably well conserved within a given phylogroup. This supports previous findings about how the strains are related to each-other. Not only are there divisions in transposase patterns between the islands Åland and Gotland, but also further subdivisions within each of the islands, and also within phylogroups. The pattern of paired transposases is remarkable and seems to indicate that the two transposases function as a unit, thus copying/displacing in unison.

The genomes G00403, G00406 and G00420 form their own branch and have very few transposase genes compared to other genomes. It would appear that they have deleted (via mutation) the main type of transposase (Type A) from their genomes early on, preventing it from spreading throughout the genome. Another explanation is that they were never infected with it in the first place, because of some kind of barrier between *L. kunkeei* strains. In both of these cases, as a strain reproduces, it transmits transposases, or lack thereof, to its daughter cells. After that, they are subject to random deletions and transpositions, of course.

Due to large groups of homologous transposases, we suspect that each of the groups has the same origin, i.e. started from a common ancestor transposase. Building on that, the genomes with the most non-pseudogenized homologous transposases are genomes with most recent transposase activity.

6 Future

There is still room for improvements. More manual work can provide higher accuracy.

Despite the abilities of annotation software, there are cases when checking transposons individually, in their context, can provide different conclusions. This manual verification might include looking at the genome map and checking concatenated adjacent genes in an online database. Manual verification could potentially identify more genes as pseudogenized as well.

In addition, the Prokka annotation can still be unclear at times, marking genes as “hypothetical protein”. While predicting potential genes is a start, it is not the final answer. Genome annotation can be a large project in and of itself.

Finally, experimental verification is still a “gold standard” for verifying genes and the resulting proteins, should specific genes be of high interest.

7 Acknowledgements

I would like to thank Siv Andersson, my supervisor, for providing the opportunity to do the thesis and for the feedback during my work. I would like to thank my mentor in the project, Andrea Garcia Montaner, for all the support and regular guidance. Thank you to Björn Nystedt, the subject reader, for reviewing the progress and adding his perspective. Last but not least, I would like to thank my family. Without their support, I think I would not have gotten this far in my studies.

8 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Bisson L, Walker G, Ramakrishnan V, Luo Y, Fan Q, Wiemer E, Luong P, Ogawa M, Joseph CM. 2016. The Two Faces of Lactobacillus kunkeei: Wine Spoilage Agent and Bee Probiotic. Catalyst: Discovery Into Practice, doi 0.5344/catalyst.2016.16002.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England) 25: 1972–1973.

Chandler M, Fayet O, Rousseau P, Ton Hoang B, Duval-Valentin G. 2015. Copy-out-Paste-in Transposition of IS911: A Major Transposition Pathway. Microbiology Spectrum, doi 10.1128/microbiolspec.MDNA3-0031-2014.

Frank AC, Amiri H, Andersson SGE. 2002. Genome deterioration: loss of repeated sequences and accumulation of junk DNA. Genetica 115: 1–12.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30: 1236–1240.

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

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research 13: 2178–2189.

Moran NA, Bennett GM. 2014. The tiniest tiny genomes. Annual Review of Microbiology 68: 195–215.

Muñoz-López M, García-Pérez JL. 2010. DNA transposons: nature and applications in genomics. Current Genomics 11: 115–128.

Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements, and why? Heredity 106: 1–10.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61: 539–542.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Research 34: D32-36.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30: 1312–1313.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33–36.

Vigil-Stenman T, Ininbergs K, Bergman B, Ekman M. 2017. High abundance and expression of transposases in bacteria from the Baltic Sea. The ISME Journal 11: 2611–2623.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. Nature Reviews Genetics 13: 303–314.