University of South Bohemia
Faculty of Science

Comparison of metabolic pathways in insect symbiotic bacteria

Bachelor Thesis

Robin Schürz

Supervisor: Prof. RNDr. Václav Hypša, CSc.
Co-Supervisor: Prof. RNDr. František Vácha, Ph.D.

České Budějovice, 2021

Schürz R. (2021): Comparision of metabolic pathways in insect symbiotic bacteria. Bc Thesis, in English. -47 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation:

Genomes of lice endosymbionts, *Legionella* from two different host species (*Polyplax serrata* and *Polyplax spinulosa*), were compared in their metabolic pathways. Preserved genes were annotated, mapped out for comparison and checked for their functionality. The annotated genes also were assessed for their Ka/Ks values and the COG categories they were included in.

Declaration:

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature displayed in the list of used sources only.

Linz, 14.12.2021

Acknowledgments

I would like to thank Prof. RNDr. Václav Hypša, CSc. for his help throughout this whole thesis even though it took more time than intended through various difficulties and changes of priorities in the thesis.

Moreover, it has to be acknowledged that his help in the writing process was immensely and his patience with me was immense.

Next I want to thank Prof. RNDr. František Vácha, Ph.D. which helped me with the writing process and interpretation of the metabolic pathways, so I wouldn´t miss their functionality.

Also I wanted to thank the members of the Department of Parasitology, even though the time working together was cut short through the pandemic it was a pleasure to meet and work with them and their help was instrumental in finishing this thesis.

Table of Content

1. Introduction

1.1. Endosymbiosis

Endosymbiosis describes an interaction between different species in which one of the organisms, the so called symbiont, lives within tissue or cells of another organism called host. Usually if you talk about intracellular endosymbionts, bacteria are meant but in lesser frequency also fungi can live inside eukaryotic cells (Taylor *et al.* 2012). Endosymbiosis is defined by their interactions (which can be beneficial, neutral or harmful) between organisms from various domains, mainly viruses, archaea, bacteria or eukaryotes (Kubiak *et al.* 2018).

Endosymbionts and their host organisms represent a distinct domain of life in which they generate new biochemical capabilities allow these two to survive and live in otherwise inhospitable environments. Typically, these relationships involve a bacterial endosymbiont living within a eukaryotic host and these are also the ones most often studied (Wernegreen 2012).

1.2. Endosymbiotic bacteria

Symbiotic bacteria is a huge part of nature and influences the eukaryotic evolution and diversity immensely. They can be harmful as parasites or play important roles in their host's survival as mutualists. One of the most important forms found is the endosymbiosis form in which the symbiotic bacteria live inside the host body which allows interaction between them. They are mostly located in the gut of the insects but also can lie within specialized cells. (Kikuchi 2009).

One bacterium which is found in around 40% of insects is the *Wolbachia pipientis* (or more generally the genus *Wolbachia*), which is maternally inherited, an obligate intracellular bacterium and is best known for invading invertebrate populations by modifying the host's reproductive system. Recently, there have been discoveries that a group of flies survives longer when infected by *Wolbachia* as well as producing less viruses than flies which were cured of *Wolbachia*. Subsequently, research also showed that mosquitos infected had a lower transmission rate of the human arboviruses DENV and CHIKV (Johnson 2015).

Apart from insects, *Wolbachia* was also found in an endosymbiotic relationship with the Onchocercidae family which is a part of the filarial nematodes. Here surveys suggest that *Wolbachia* has been integrated rather recently as it is absent from ancestral groups but found in more recent species. Here *Wolbachia* is an important part in the larval and embryonic growth and development and fertility of the host (Taylor *et al.* 2012).

Endosymbiotic bacteria have varying interdependence with their host, ranging from facultative to obligate. In facultative symbiosis the host mostly suffers low or no consequences if the symbiont is absent. On the other hand in the obligate symbiosis neither host nor symbiont would survive in the absence of their partner (Kikuchi 2009).

Mutualistic symbiosis between bacteria and insects is one of the best described symbiotic interactions in nature. In general insects have been successful through mutualistic primary endosymbiotic bacteria, which helps insects survive niche lifestyles like nutrient-poor diets, such as wood (termites), plant sap (aphids) and blood (sucking lice) (Allen *et al.* 2007). This primary endosymbiotic bacteria are typically maintained inside specialized host cells and exhibit a nucleotide A+T bias greater than 50%. Most primary endosymbionts then leave their specialized host cell to migrate into the ovaries and incorporate themselves into developing eggs to be passed onto the next host generation (Allen *et al.* 2007).

For example, the symbiosis between the pea aphid and its bacterial endosymbiont *Buchnera aphidicola* shows a good example of mutualistic and beneficial symbiosis. These genomes of the aphid and the bacterium have been together for millions of generations, influencing each other throughout it. For examples the bacterium provides nutrients for its host to balance the otherwise insufficient diet of pea saps (Sabater-Muñoz *et al.* 2017).

However, for aphids around 40%-60% of their population are in an endosymbiotic relationship with another symbiont, which indicates that there has to be a way in which aphids lose or gain symbionts. As aphid symbionts are normally transmitted vertically during reproduction a transmission failure rate of up to 40% can occur which can explain the numbers shown above (Zytynska 2019).

Also recent studies have shown that vector borne disease are greatly influenced by tripartite interactions of viruses, endosymbiotic bacteria and their host. As mentioned before *Wolbachia* protects mosquitos from viral infections which has major implications for naturally infected insects and in disease control (Johnson 2015).

1.3. Endosymbiotic bacteria in lice

Various insect species with nutritionally incomplete diets have mutualistic bacteria which synthesize missing nutrients. These endosymbiotic bacteria are intracellular and occupy specialized host cells, so called bacteriocytes, and are transmitted vertically (Zytynska 2019).

Sucking lice (Anoplura) are one big group of insects which have endosymbiotic bacteria and form relationships with them. The suborder consists of 532 described species, and every species is parasitizing one or some closely related species of mammals. Many of these species have been recorded to live in endosymbiotic relationships. For human lice, *Pediculus humanus* the implication is that the endosymbionts major role is providing B-vitamins. However, the notion has been made that the symbiosis of lice and bacteria have arisen multiple times in different louse species and not from one common ancestor (Hypsa & Krízek, 2007; Boyd *et al.* 2017).

On the other side the symbiotic bacteria gains host-derived amino acids which supports the growth and lifespan of said bacteria (Burkhart & Burkhart 2006). As explained above the endosymbiotic bacteria are transferred vertically which can also be called transovarial transmission. This has been shown by removing the bacteriomes from young female lice. This procedure kills the louse shortly after and the eggs are deformed. Also if the bacteria are removed from the eggs directly, it also dies shortly after. However studies have shown that lice without endosymbiotic bacteria were able to survive if their diet was supplemented with nicotinic acid, pantothenic acid and beta-biotin (Allen *et al.* 2007).

Endosymbionts in different lice species originate from different bacterial groups and are classified differently like the endosymbionts of lice that parasitize hominids (humans, chimpanzees and gorillas) are classified into the genus *Candidatus* Riesia. On the other hand lice that parasitize old world monkeys have the genus *Candidatus* Puchtella to name some of them. This shows that a replacement of an endosymbiont in common ancestors had to take place (Boyd *et al.* 2017).

1.4. Modifications of the endosymbiotic bacteria genome

Most symbionts have one feature and that is that the endosymbiotic organism possesses a reduced genome size compared to their free-living counterparts. There are varieties of size change from some mild reduction in *Sodalis glossinidius* to high size reduction of the genome in *Buchnera* (Manzano-Marin & Latorre 2016).

The reduction of the genome size is a product of isolation within the host followed by a massive pseudogenization and gene loss which often includes DNA repair mechanisms (Nicks & Rahn-Lee 2017).

Generally, studies have shown that older obligate endosymbionts have highly reduced genomes as small as 112 kbp (*Nasuia*-ALF) (Bennett & Moran, 2013). Meanwhile, more recent endosymbionts still have a larger genome with up to 4.5 Mbp (Manzano-Marin & Latorre 2016).

The bacteria experience major genetic and phenotypic changes. These can be detected by comparing them against their free-living relatives. Some changes include the bias towards AT-rich genomes, an accumulation of small deleterious mutations and a loss of mobile elements. These changes show that the bacteria shifts towards an obligate symbiosis. The genes which are lost in the reduction are mainly genes that have become unnecessary in the new nutrient-rich intracellular environment or codons for functions the host carries out. The only genes regained are either essential for maintenance of the bacterial cell or important for the mutualistic association with its host (Latorre & Manzano-Marin 2017).

This genome reduction in endosymbionts is considered to be driven by genetic drift and because of that seen as consequence of chance and mutational bias. The intensity if the genetic drift in every species is determined by its effective population size. Also, as endosymbionts have few opportunities to recombine, they suffer from the accumulation of deleterious mutations by Muller´s ratchet. These facts combined with the above-mentioned bias towards the mutation pattern of GC to AT mutation and the excess of deletion can explain the AT richness and massive gene loss in endosymbionts (Marais *et al.* 2008).

1.5. *Legionella* of *Polyplax* lice

Legionellae are gram-negative bacteria found in freshwater environments, which were first isolated from guinea pigs in 1943 by Tatlock. Legionellosis is known in two different clinical entities. Legionaires disease, which is a multisystem disease involving pneumonia and Pontiac fever, a self-limited flu-like illness (Fields *et al.* 2002). The organisms grow in water systems at temperatures of 20-50 °C. *Legionella* is able to form relationships with freshwater or soil amoebae where the bacteria is provided with support for multiplication and rises the resistance against disadvantageous environmental factors in the amoebae (Wang *et al.* 2019).

As mentioned above *Legionella* normally is known from aquatic environments, however recently studies have shown exception in form of an endosymbiotic lineage described from the rodent lice of the genus *Polyplax,* in which these bacteria are obligate symbionts (Říhová *et al.* 2017).

The 16S rDNA sequence of a *Legionella* like bacteria has be found in two different *Polyplax* species: the *Polyplax serrata* and *P. spinulosa* (Hypsa & Krizek 2007). The full genome of the bacteria has been reconstructed recently (Říhová *et al.* 2017).

2. Aim of the work

This bachelor's thesis aims to compare the genomes of the symbiotic bacterium *Legionella polyplacis* from two different species of host lice, *Polyplax serrata* and *Polyplax spinulosa*. Two different approaches have been used for this. The first approach focused on reconstructing and comparing the metabolic capacities of the two genomes. The other approach addresses the process of natural selection. The aim is to prepare a background for the following comparative study by calculating Ka/Ks values as selection estimator. The louse symbiont *Riesia pediculicola* was used as a reference for this part.

In detail this thesis focuses on

- Annotating genomes of the two symbiotic bacteria.
- Translating the annotated genes.
- Obtaining K-Numbers from the BlastKOALA-platform
- Obtaining the metabolic pathways from the KEGG Mapper-platform
- Reconstruction and analysing of the metabolic pathways
- Identifying orthogroups across the annotated genes.
- Getting the Ka/Ks values for the identified orthologues.
- Matching the orthogroups to their respective gene.

3. Materials and Methods

The genomes of *Legionella-polyplacis*$_{ser}$ (from *P. serrata*), *L. polyplacis*$_{spi}$ (from *P. spinulosa*) and *R. pediculiculicola* , were provided by Professor Václav Hypša RNDr. CSc or obtained from GenBank. The accession numbers are for *L. polyplacis*$_{ser}$ NZ_CP021497 and for *R. pediculicola* CP012849, the genome of *L. polyplacis*$_{spi}$ was provided so there is no accession number.

These genomes were then imported to Geneious Prime 2019.2.3 with a license provided by the Faculty of Science USB.

3.1. Annotation

The annotations were done using a RAST-platform, implementing the standard RAST-method with automatic error correction (Overbeek et. al, 2014; Brettin et. al, 2015; Aziz et. al, 2008). These annotated genomes were then imported into Geneious Prime and translated. From there, annotated genes were obtained which were exported as fasta files. To optimize the fasta file a self-written command was used to shorten the names of the genes and adding numbering to them.

3.2. Metabolic reconstruction

The files then were sent to the BlastKOALA-platform which returned KEGG-numbers for all the genes (Kanehisa et. al, 2016). The K-numbers were implemented into KEGG Mapper, where metabolic pathways were shown (Kanehisa & Sato, 2020; Kanehisa et. al, 2021). These metabolic pathways were then compared and categorized into functional, partly functional or non-functional. This was done by analysing the pathways and check if the necessary enzymes are present to achieve the wanted product in a similar fashion as in Rihova et. al. (2021).

3.3. Orthologs identification

Afterwards the annotated genes from RAST, which we already used for the metabolic reconstruction, were taken and used to find their single copy orthologue sequences by using OrthoFinder v2.5.4 (Emms & Kelly, 2019). OrthoFinder only works with amino acids, so a script was written in Python to get all the single copy orthologue sequences in nucleotides.

3.4. Ka/Ks determination

The resulting files were then uploaded to the CBU Ka/Ks calculation tool to obtain the Ka/Ks values for the all the genes (http://services.cbu.uib.no/tools/kaks). CDS from the three genomes were also uploaded to eggNOG mapper to obtain their COGs (Huerta-Cepas et. al, 2019; Cantalapiedra et. al, 2021).

All this information was then put into one Excel file and compared to find various outliers like the genes with the highest and smallest Ka/Ks values and which orthogroups these belonged to. Also, it was checked to which COG categories these genes belonged, to provide a database and first insight for the future analyses

3.5. Ka/Ks values

The Ka/Ks value is an important parameter for molecular evolutionary analyses and is determined as the ratio of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks). To explain these terms, think of aligning genes from different species and checking the differences. If the nucleotide substitution causes a differences in the amino acids of the protein, they are called nonsynonymous changes. However, if there is no change in the protein, the substitution is called synonymous or silent (Hurst, 2002).

The consensus is that a Ka/Ks value above one is considered as mark of positive selection, a value of one means neutral mutation and a value below one is considered as a negative selection. There are several methods to estimate Ka and Ks, which are based on various substitution models. These are categorized into two main types: approximate methods and maximum likelihood methods (Wang et. al, 2009).

There are several types of algorithm models for the Ka/Ks value to be calculated like the LPB93 algorithm or the PAML, which follow the same three steps. Counting the number of synonymous and nonsynonymous sites, counting the numbers of synonymous and nonsynonymous substitutions, and correcting for multiple substitutions (Li et. al, 2009).

3.6. Cluster of Orthologous groups

Orthologous groups are a useful identification method for genome annotation, comparison and studies on gene/protein evolution. To classify proteins from complete genomes the Cluster of Orthologous groups (COG) of protein database had been established (Li et.al, 2003).

Originally the first set was created in 1997 and included proteins from five bacterial, one archaeal and one eukaryotic genome which concluded in 720 COGs (Tatusov et. al, 2000). Since

then, more and more COGs have been added to the database with the latest release including 1309 genomes and 4877 COGs (Galperin et. al, 2021).

To construct a COG, they have to be identified on a basis of an all-against all sequence comparison of the proteins encoded in complete genomes. This procedure is based on the notion that in a group of at least three proteins from various genomes, which are similar to each other, are more likely to be in the same orthologous family than sharing one with other proteins from the same genome (Tatusov et. al, 2000).

COGs are also classified into their functionality, which can be seen on (https://ecoliwiki.org/colipedia/index.php/Clusters_of_Orthologous_Groups_(COGs)), and allows to easily compare organisms based on their preference for certain types of metabolic, signal transduction, repair or other pathways (Galperin et. al, 2021).

4. Results

On the next pages the metabolic pathway of *L. polyplacis* in the lice species *Polyplax spinulosa* is shown. According to the resulted K-numbers obtained, it is shown which enzyme is present and working and which are not. The enzymes which are present are marked in green, missing ones are colourless.

4.1. *Legionella polyplacis* in *P.spinulosa*

The genome of *L. polyplacis*$_{spi}$ is 532.296 bp long. It´s GC-content is similar to *L. polyplacis*$_{ser}$ at 23.1% which also indicates a sizeable change in the genome with an old endosymbiotic relationship. Through annotations 525 genes were found. There were 3 rRNA and 36 tRNA genes so the leftover 486 genes were used to obtain their K-Numbers. 442 of these genes returned K-numbers which were then put into KEGG Mapper to show the metabolic pathways of the genome.

4.1.1. Amino acids

The bacterium *L. polyplacis*$_{spi}$ is missing most of the pathways to synthesize various amino acids. The missing amino acids are as follows: alanine (A), arginine (R), asparagine (N), aspartic acid (D), cysteine (C), glycine (G), histidine (H), leucine (L), isoleucine (I), methionine (M), phenylalanine (F), proline (P), serine (S), threonine (T), tryptophan (W), tyrosine (Y) and valine (V) as seen in *Figure 1*.

The bacterium *L.polyplacis*$_{spi}$ has also three semi-functional or functional amino acid metabolic pathways. Starting with the pathway for lysine (K) the starting point is aspartic acid and through the enzymes K12526 (aspartate kinase), K00133 (aspartate-semialdehyde dehydrogenase), K01714 (4-hydroxy-tetrahydrodipicolinate synthase), K00215 (4-hydroxy-tetrahydrodipicolinate reductase), K00674 (2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase) N-succinyl-2-L-amino-6-oxoheptanedioate is synthesized. The enzyme K00821 (acetylornithine/N-succinyldiaminopimelate aminotransferase) is missing to synthesize N-succinyl-LL-2,6-diaminoheptanedioate. From there the pathway to lysine (K) is present with the enzymes of K01439 (succinyl-diaminopimelate desuccynlase), K01778 (diaminopimelate epimerase) and K12526 (diaminopimelate decarboxylase).

The same assumption can be made that the degradation process has started for the metabolic pathway of lysine (K)

The next amino acid which is semi-functional for glutamic acid (E). Glutamic acid is synthesized from 2-oxoglutarate, an intermediate of citrate cycle (see chapter 4.1.6). As it is stated in chapter 4.1.6, citrate cycle contains all enzymes of its reactions, and since it can run both directions, it means that glutamate can be synthetized from any intermediate of citrate cycle.

The enzyme for the initial reaction of the citrate cycle, the reaction of oxaloacetate with acetyl-CoA that produces citrate is missing, however *L. polyplacis*$_{spi}$ possesses a functional enzyme K00027 (malate dehydrogenase) that produces malate (an intermediate of citrate cycle) from pyruvate and therefore pyruvate can be considered also as a precursor of glutamate biosynthesis.

Lastly glutamine (Q) can also be synthesized as there is only one step left from glutamic acid (E) and this is done using K01915 (glutamine synthetase) as it can be seen in *Figure 1*.

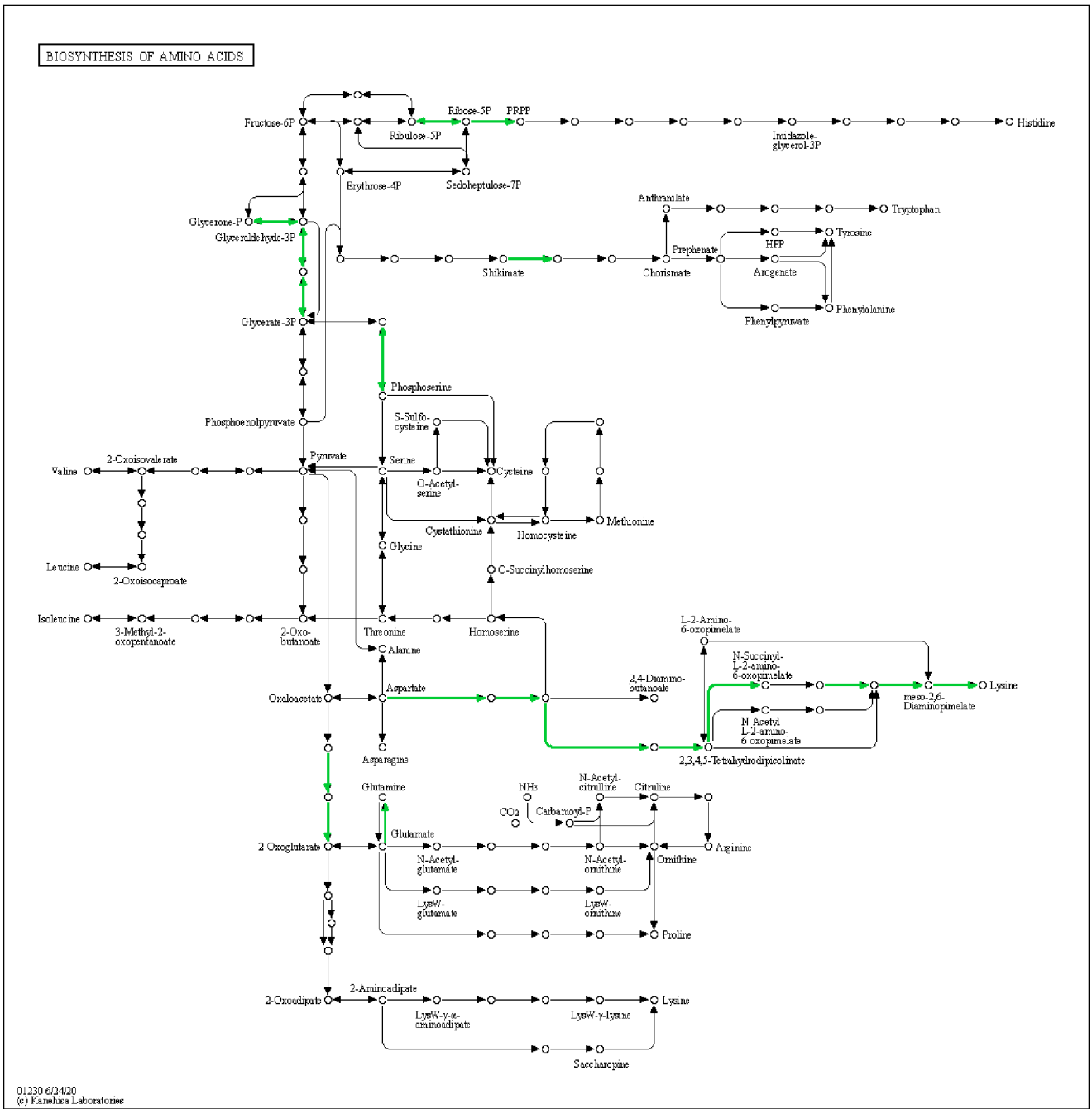

Figure 1: metabolic pathway of amino acids (*L. polyplacis*$_{spi}$)

4.1.2. B-Vitamins

Of the B-vitamins the bacterium *L. polyplacis*$_{spi}$ is able to synthesize three important ones, either completely or with support from the host. The vitamins where either some enzymes or all of them are missing are thiamine (B1), niacin (B3), pantothenic acid (B5), pyridoxine (B6) and cobalamin (B12).

Beginning with riboflavin (B2) the starting point is GTP which synthesizes to 5-amino-6-(5´-phospho-D-ribitylamino)-uracil using the enzymes K14652 (GTP cyclhydrolase II) and K11752 (diaminohydroxyphosphoribosylaminopyrimidine deaminase / 5-amino-6-(5-phosphoribosylamino)-uracil reductase). For the next step the enzyme K22912 (5-amino-6-(5-phospho-D-ribitylamino)-uracil phosphatase) is missing for the synthesis of 5-amino-6-(D-ribitylamino)-uracil. From this compound on the last two steps to obtain riboflavin (B2) are again present, namely the enzymes K00794 (6-7-dimethyl-8-ribityllumazine synthase) and K00793 (riboflavin synthase).

It can be seen in *Figure 2* that there would be a path from D-ribulose 5-phosphate to riboflavin (B2) but the bacteria, as shown in *chapter 4.1.5*, is not able to produce D-ribulose 5-phosphate so it would only work through help from the host.

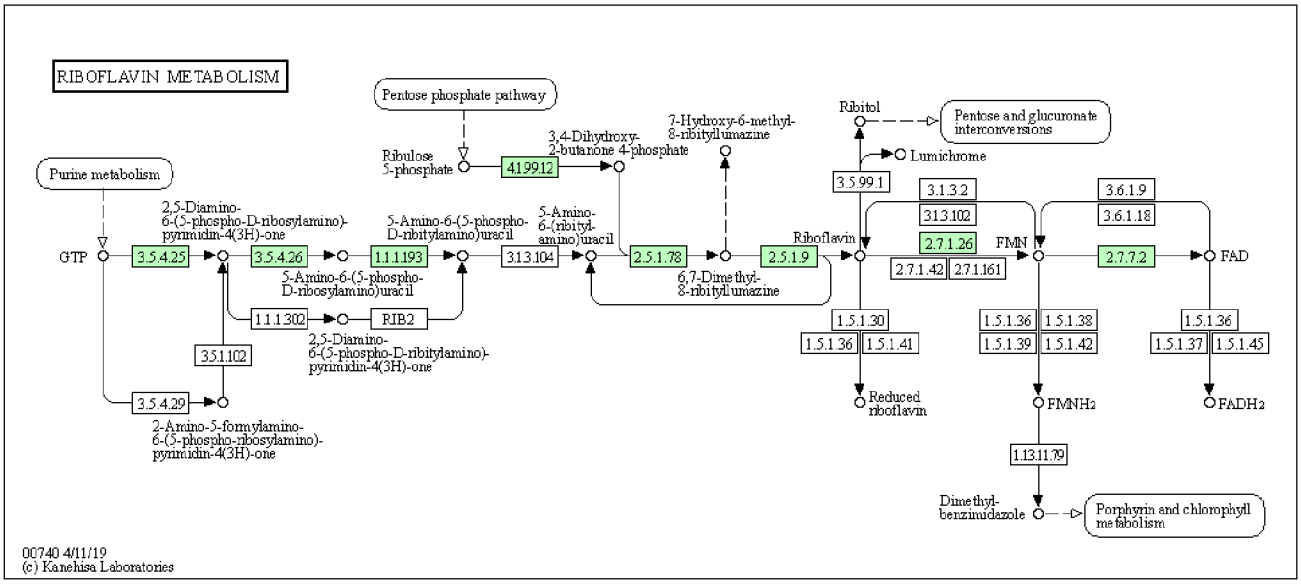

Figure 2: metabolic pathway of riboflavin (*L. polyplacis*$_{spi}$)

The next vitamin which can be synthesized by *L. polyplacis*$_{spi}$ is biotin (B7). For the synthesis to work the host must provide pyruvate which the bacteria can first synthesize malonyl-[acyl-carrier-protein] using K00163 (pyruvate dehydrogenase E1 component), K00627 (pyruvate dehydrogenase E2 component) and K01962 (acetyl-CoA carboxylase carboxyl transferase subunit alpha). From malonyl-[acyl-carrier-protein] the whole path to biotin (B7) is present as shown in *Figure 3*. Using the enzymes K02169 (malonly-CoA O-methyltransferase), K09458 (3-oxoacyl-

[acyl-carrier-protein] synthase II), K00059 (3-oxyl-[acyl-carrier-protein] reductase), K02372 (3-hydoxylacyl-[acyl-carrier-protein] dehydratase, K00208 (enoyl-[acyl-carrier-protein] reductase I), K02170 (pimeloyl-[acyl-carrier-protein] methyl ester esterase), K00652 (8-amino-7-oxononanoate synthase), K00833 (adenosylmethionone-8-amino-7-oxononanoate aminotransferase), K01935 (dethiobiotin synthetase) and K01012 (biotin synthase). This means the whole metabolic pathway for biotin (B7) should be fully functional.

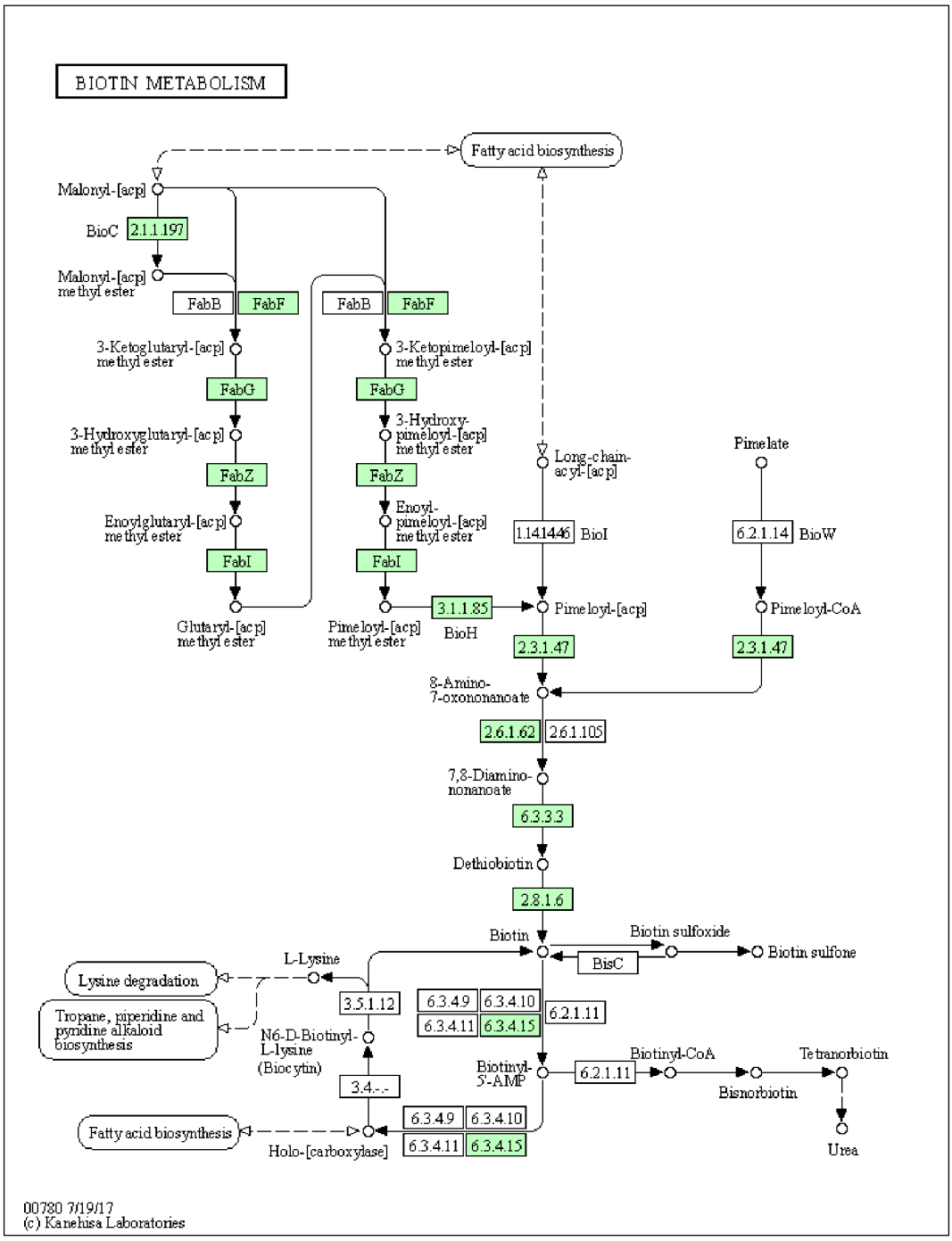

Figure 3: metabolic pathway for Biotin (*L. polyplacis*$_{spi}$)

The last vitamin where many enzymes are preserved in *L. polyplacis*$_{spi}$ is folic acid (B9) which similar to riboflavin (B2) also starts at GTP. The enzyme K01495 (GTP cyclohydrolase IA) synthesizes GTP to 7,8-dihydroneopterin 3´-triphosphate. The enzyme K01077 (alkaline phosphatase) is missing which means there is no synthesis to 7,8-dihydroneopterin. Subsequently, from this compound on all enzymes to synthesize folic acid are present which are K01633 (7,8-dihydroneopterin aldolase/epimerase/oxygenase), K00950 (2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase), K00796 (dihydropteroate synthase), K11754 (dihydrofolate synthase/folylpolyglutamate synthase) and K00287 (dihydrofolate reductase).

However, for the synthesis of 7,8-dihydropteroate the compound 4-aminobenzoate is needed as seen in *Figure 4.* This compound cannot be produced by the bacteria so if they host is not able to provide it the folic acid (B9) pathway is non-functional.

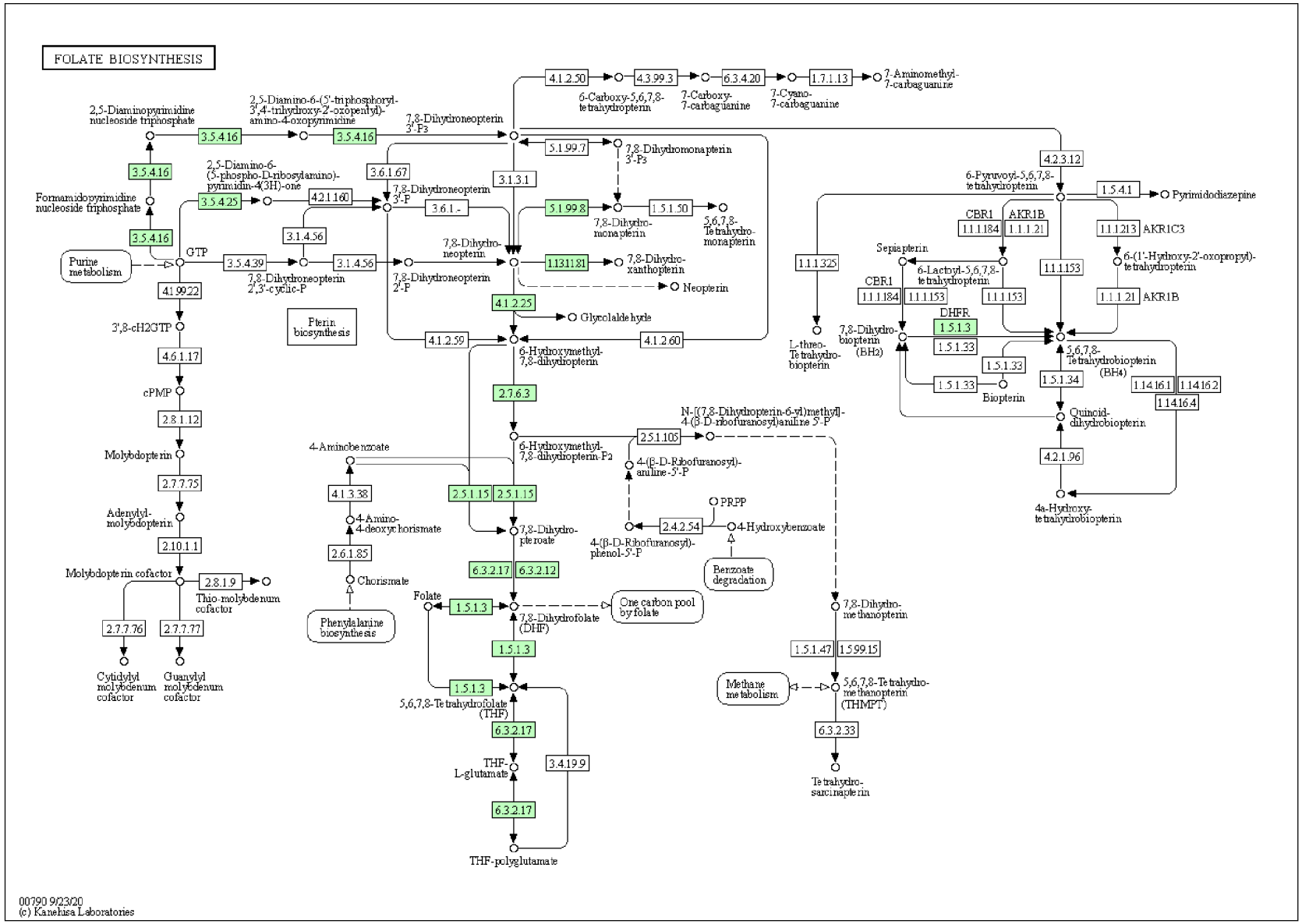

Figure 4: metabolic pathway of folic acid (*L. polyplacis*$_{spi}$)

4.1.3. Fatty acids

The starting point for the synthesis of fatty acids is acetyl-CoA which can be synthesized using either pyruvate or several amino acids (leucine, lysine etc.) by *L. polyplacis*$_{spi}$. From there the enzymes K01962 (acetyl-CoA carboxylase carboxyl transferase subunit alpha) and K00645 ([acyl-carrier-protein] S-malonyltransferase) are used to synthesize malonyl-[acyl-carrier-protein]. This compound is kind of a way point to all the various fatty acids as can be seen in *Figure 5*.

Through the enzymes K00648 (3-oxoacyl-[acyl-carrier-protein] synthase III, K09458 3-oxoacyl-[acyl-carrier-protein] synthase II), K00059 (3-oxoacyl-[acyl-carrier-protein] reductase), K02372 (3-hydroxyacyl-[acyl-carrier-protein] dehydratase) and K00208 enoyl-[acyl-carrier-protein] reductase I) acp-bound fatty acids can be synthesized by the bacteria. However, for the last step to obtain fatty acids *L.polyplacis*$_{spi}$ is missing the enzymes K01071 (medium-chain acyl-[acyl-carrier-protein] hydrolase and K10781 (fatty acyl-ACP thioesterase B). This means subsequently that the bacteria is not able to produce fatty acids.

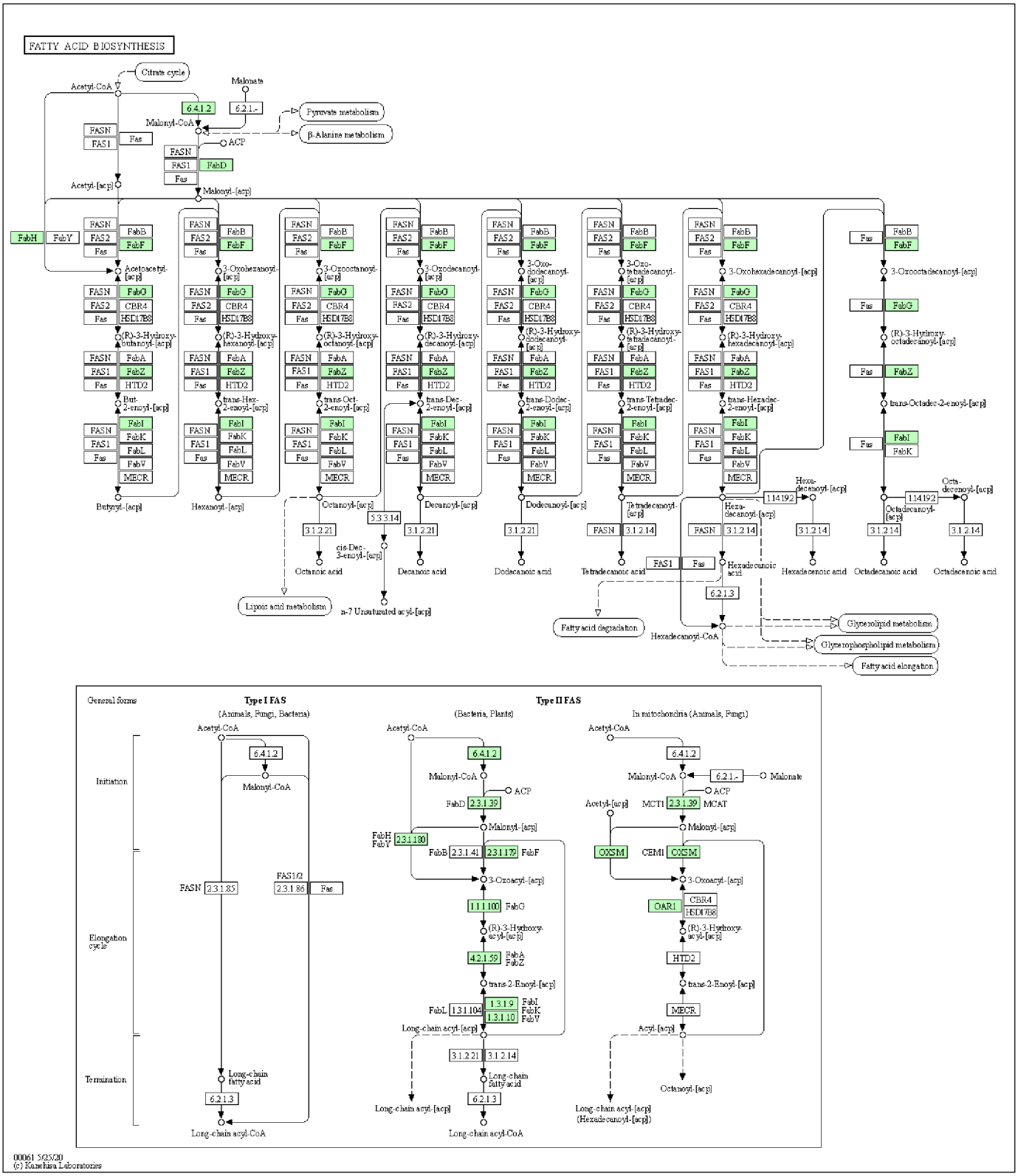

Figure 5: metabolic pathways of fatty acids (*L. polyplacis*$_{spi}$)

4.1.4. Glycolysis

For the first phase (the preparatory phase) of glycolysis there are no enzymes found in *L. polyplacis*$_{spi}$ which could synthesize any compound important for the process shown in *Figure 6*. However, in the pay-off phase there are some enzymes like K00134 (glyceraldehyde 3-phosphate dehydrogenase) and K00927 (phosphoglycerate kinase) which synthesize 3-phospho-D-glycerate from D-glyceraldehyde 3-phosphate. As only these steps are present and all other enzymes are missing it can be said that the bacteria has no functional glycolysis process. On the other hand, these two subsequent reactions (from glyceraldehyde 3-phosphate to 3-phosphoglycerate) are the key bioenergetics reactions of glycolysis that produce NADH and ATP. Retaining of genes for the NADH and ATP producing reaction suggest that the organism most probably utilizes this part of glycolysis for production of energetic compounds.

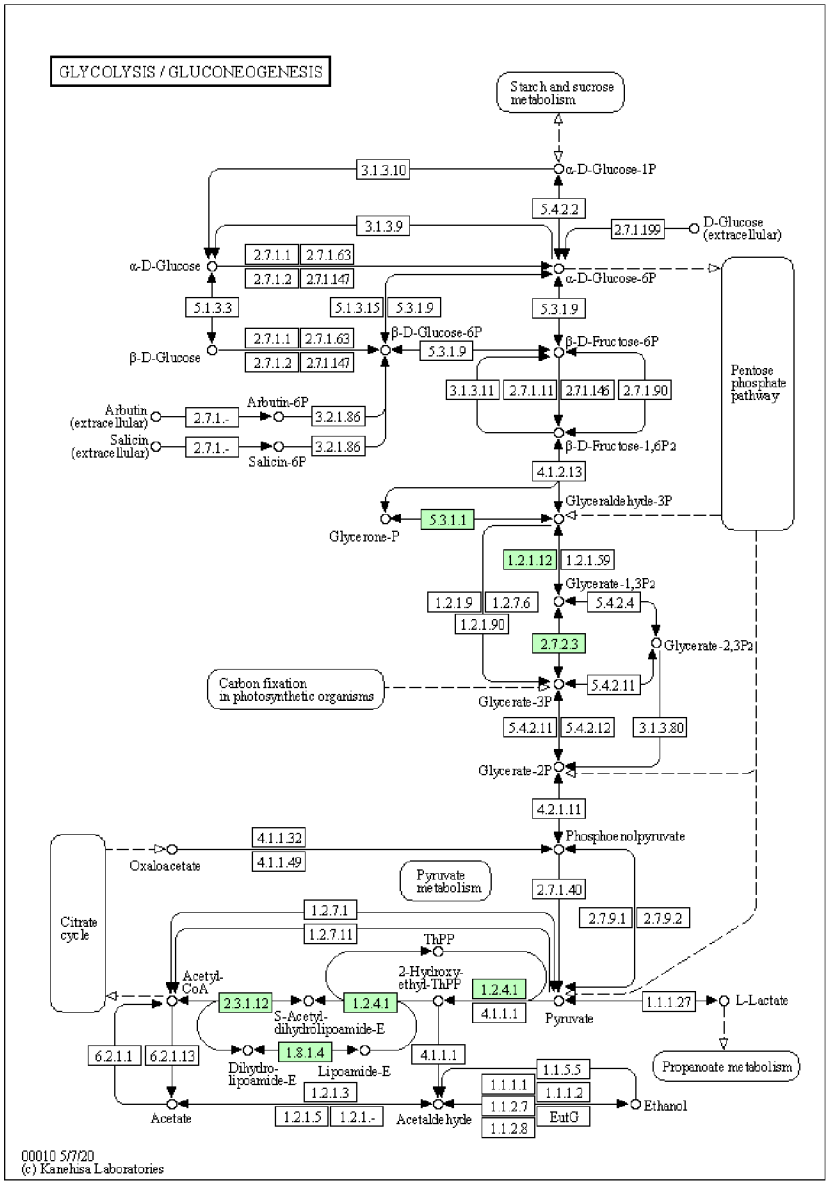

Figure 6: metabolic pathway of glycolysis (*L. polyplacis*$_{spi}$)

4.1.5. Pentose phosphate cycle

L. polyplacis$_{spi}$ shows no enzymes in the oxidative phase at all which means the whole metabolic pathway is non-functional (*Figure 7).*

Furthermore, even though in the non-oxidative phase there are some enzymes present to synthesize D-ribose-5-phosphate from D-ribulose-5-phosphate using K01807 (ribose 5-phosphate isomerase A) and subsequently D-ribose-1-phosphate using K01839 (phosphopentomutase), the process itself is non-functional as too many enzymes are missing to obtain the products.

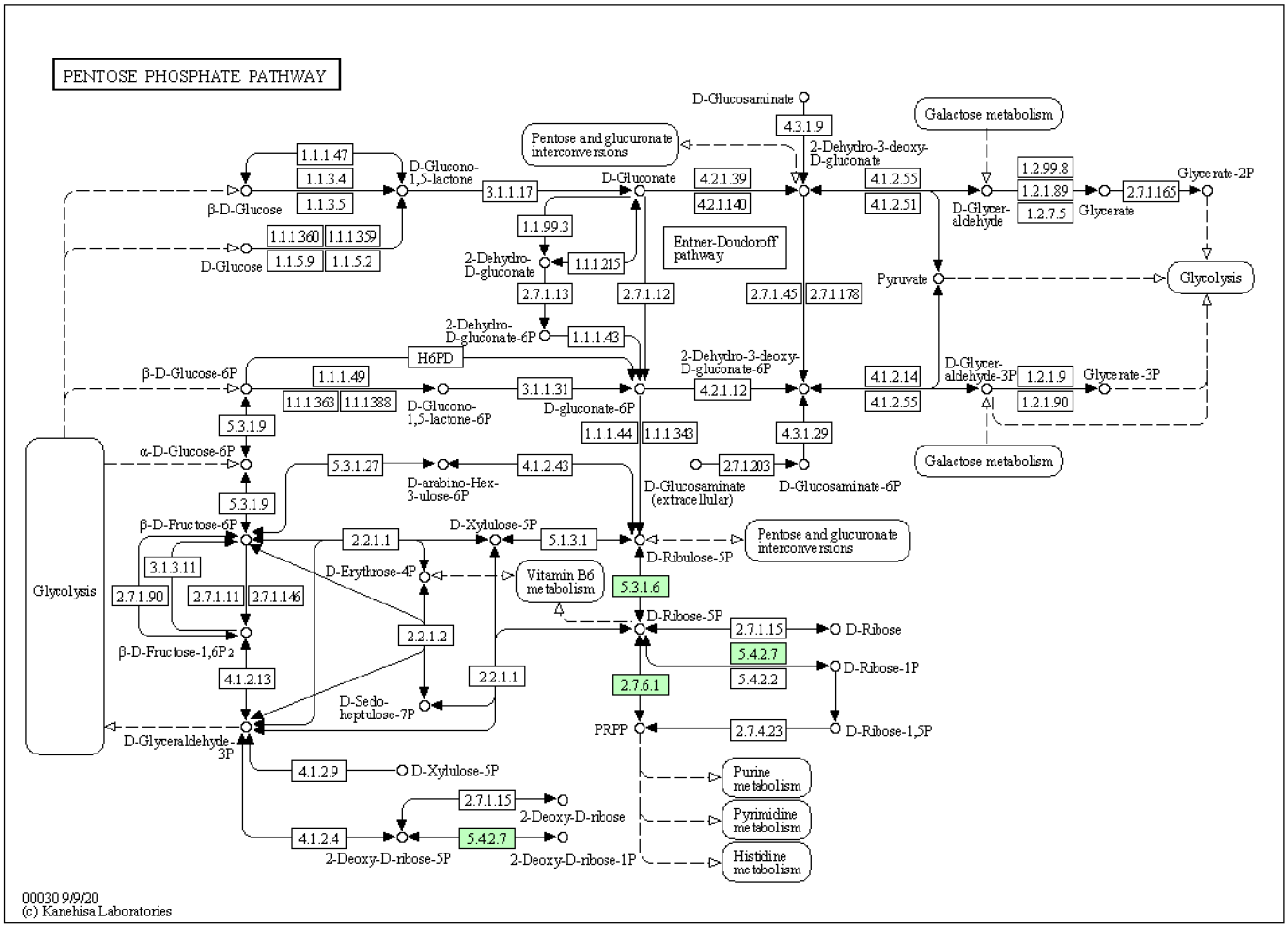

Figure 7: metabolic pathway of pentose phosphate pathway (*L. polyplacis*$_{spi}$)

4.1.6. Citrate cycle

The citric cycle in *L. polyplacis*$_{spi}$ contains almost all necessary enzymes needed to be fully functional (see *Figure 8*). The first compound the bacteria needs is pyruvate which needs to be provided by the host. From there on Acetyl-CoA can be synthesized using K00163 (pyruvate dehydrogenase E1 component) and K00627 (pyruvate dehydrogenase E2 component/dihydrolipoamide acetyltransferase). Only the next step is problematic as the enzyme K01648 (ATP citrate (pro-S)-lyase) to synthesize citrate from acetyl-CoA is missing. However, *L.polyplacis*$_{spi}$ possesses a functional enzyme K00027 (malate dehydrogenase) that produces malate from pyruvate and therefore Citrate cycle can run "counter clockwise" in an anabolic direction.

Namely, are present, the enzymes K01681 (aconitate hydratase), K00031 (isocitrate dehydrogenase), K00164 (2-oxoglutarate dehydrogenase E1 component), K00658 (2-oxoglutarate dehydrogenase E2 component/dihydrolipoamide succinyltransferase), K01902 (succinyl-CoA synthetase alpha subunit), K00239 (succinate dehydrogenase/fumarate reductase/flavoprotein subunit), K01679 (fumarate hydratase class II) and K00024 (malate dehydrogenase) are all present. This means that the bacteria can with some support from the host run the citric cycle and produce compounds such as succinate, fumarate, succinyl-CoA and others.

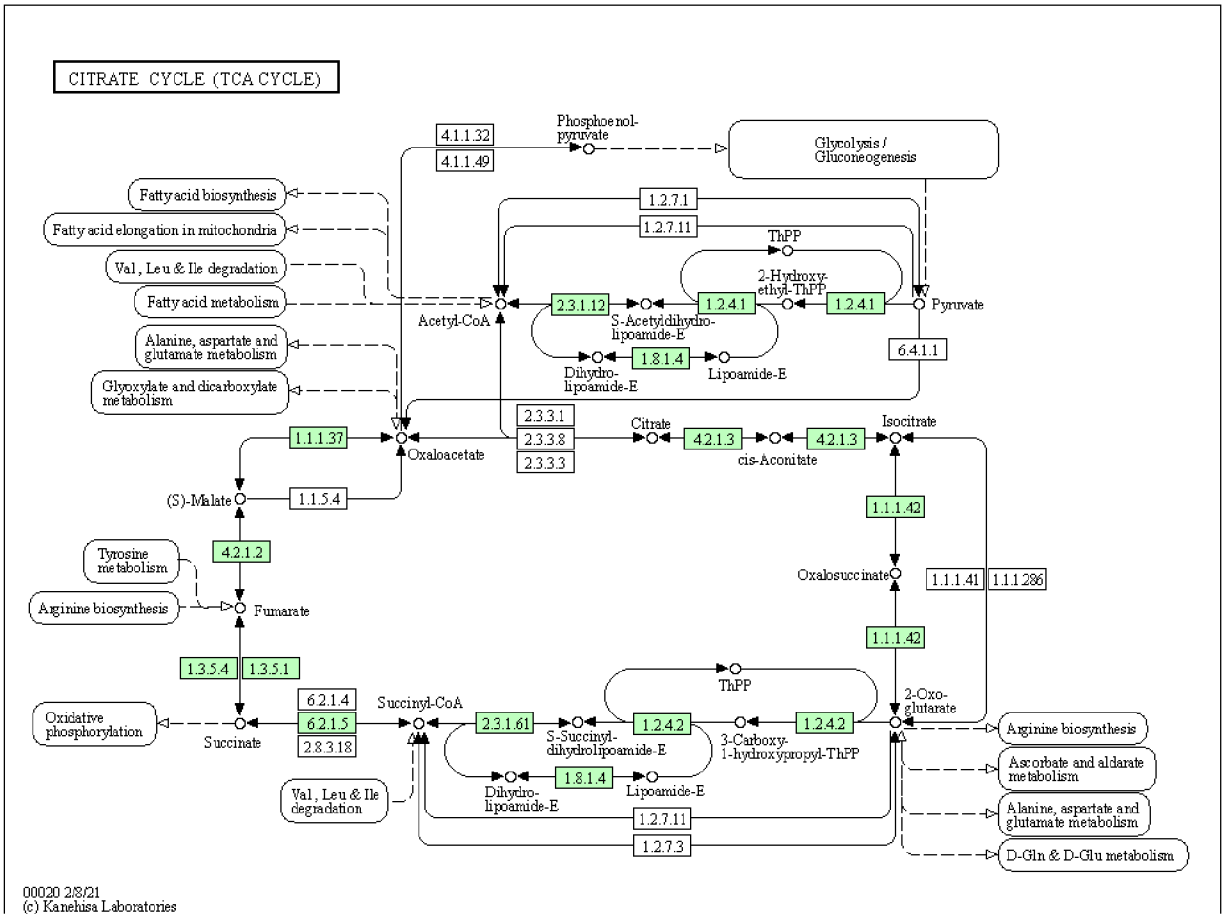

Figure 8: metabolic pathway of citric cycle (*L. polyplacis*$_{spi}$)

4.1.7. Purine

The metabolic pathway of purine is filled with working enzymes present in *L. polyplacis*$_{spi}$ especially around IMP, AMP, XMP and GMP. Starting with IMP (inosine 5´-monophosphate) there are several ways to synthesize different compounds. Using K00088 (IMP dehydrogenase) to synthesize XMP (xanthosine 5´-phosphate), following K01951 (GMP synthase) synthesizes the compound GMP (guanosine 5´-phosphate). Furthermore K00942 (guanylate kinase) synthesizes GDP (guanosine 5´-diphosphate). From this compound K00940 (nucleoside-diphosphate kinase), K00525 (ribonucleoside-diphosphate reductase alpha chain), K00942 (guanylate kinase) and K01081 (5´-nucleotidase) are used to synthesize GTP, dGDP, dGTP, dGMP and deoxyguanosine respectively.

Once again starting at IMP the bacteria also provide enzymes to synthesize AMP (adenosine 5´-monophosphate), namely K01939 (adenylosuccinate synthase) and K01756 (adenylosuccinate layase). Following that the enzyme K00939 (adenylate kinase) synthesizes AMP to ADP (adenosine-5´-diphosphate). From there similar as from GDP, K00940 (nucleoside-diphosphate kinase), K00525 (ribonucleoside-diphosphate reductase alpha chain), K00939 (adenylate kinase) and K01081 (5´-nucleotidase) are used to synthesize ATP, dADP, dATP, dAMP and deoxyadenosine respectively.

Lastly, GMP, XMP and IMP all can use the enzymes K03787 (5´-nucleotidase) and K00760 (hypoxanthine phosphoribosyltransferase) to synthesize respectively, Guanosine or guanine, xanthosine or xanthine and inosine or hypoxanthine. In *Figure 9* also shown is that there are some enzymes found practically without purpose as following enzymes are missing.

However, for the purine metabolic pathway to function the host has to provide IMP to the bacteria as it is not able to produce it itself.

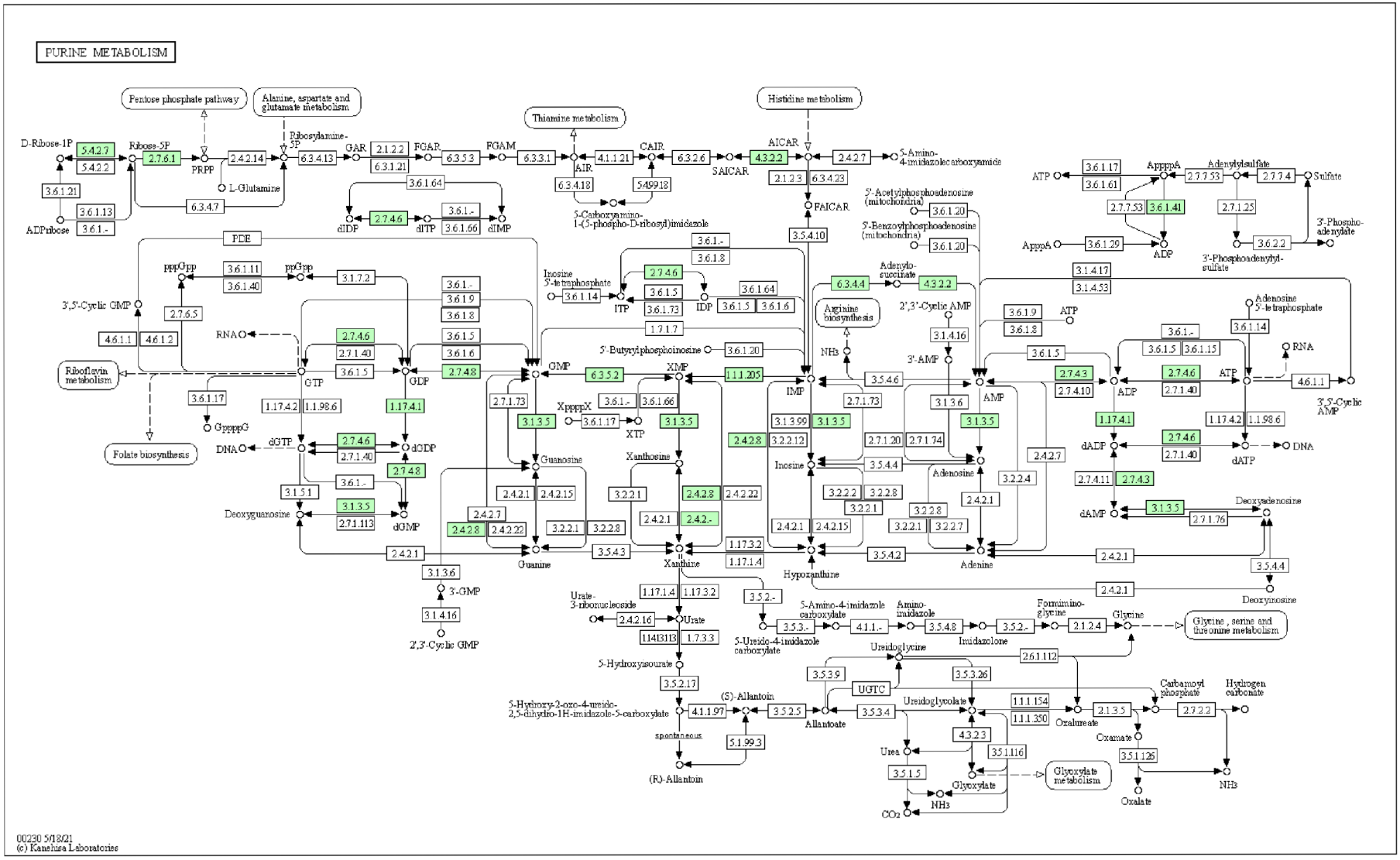

Figure 9: metabolic pathway of purine (*L. polyplacis*$_{spi}$)

4.1.8. Pyrimidine

The starting compound, which must be provided by the host, is UMP (uridine 5´-monophosphate). This compound can be either synthesized by K03787 (5´-nucleotidase) to uridine or by K09903 (uridylate kinase) to UDP (uridine 5´-diphosphate). The path after uridine is missing so uracil cannot be synthesized from uridine, however UDP can be synthesized by the enzyme K00940 (nucleoside-diphosphate kinase) to UTP. This compound then uses K01937 (CTP synthase) to synthesize CTP. First using K00940 CDP is synthesized which then uses K00945 (CMP/dCMP kinase) to obtain CMP. The last step is then using K03787 (5´-nucleotidase) to synthesize cytidine. Also, from CDP another path is presented by using K00525 (ribonucleoside-diphosphate reductase alpha chain) to synthesize dCDP. Here K00945 and K03787 are used again to obtain dCMP and deoxycytidine respectively.

On the other side from dCDP using K00940 gives dCTP which uses K01494 (dCTP deaminase) to synthesize dUTP. Now dUDP can be synthesized either from dUTP using K00940 or from UDP using K00525. Subsequently, using K00943 (dTMP kinase) and K03787, dUMP and deoxyuridine is synthesized.

Lastly from dUMP one last path to dTMP is shown by using K00560 (thymidylate synthase) and from there either synthesizing thymidine with the enzyme K03787 or obtaining dTDP and dTTP by using K00943 and K00940, respectively.

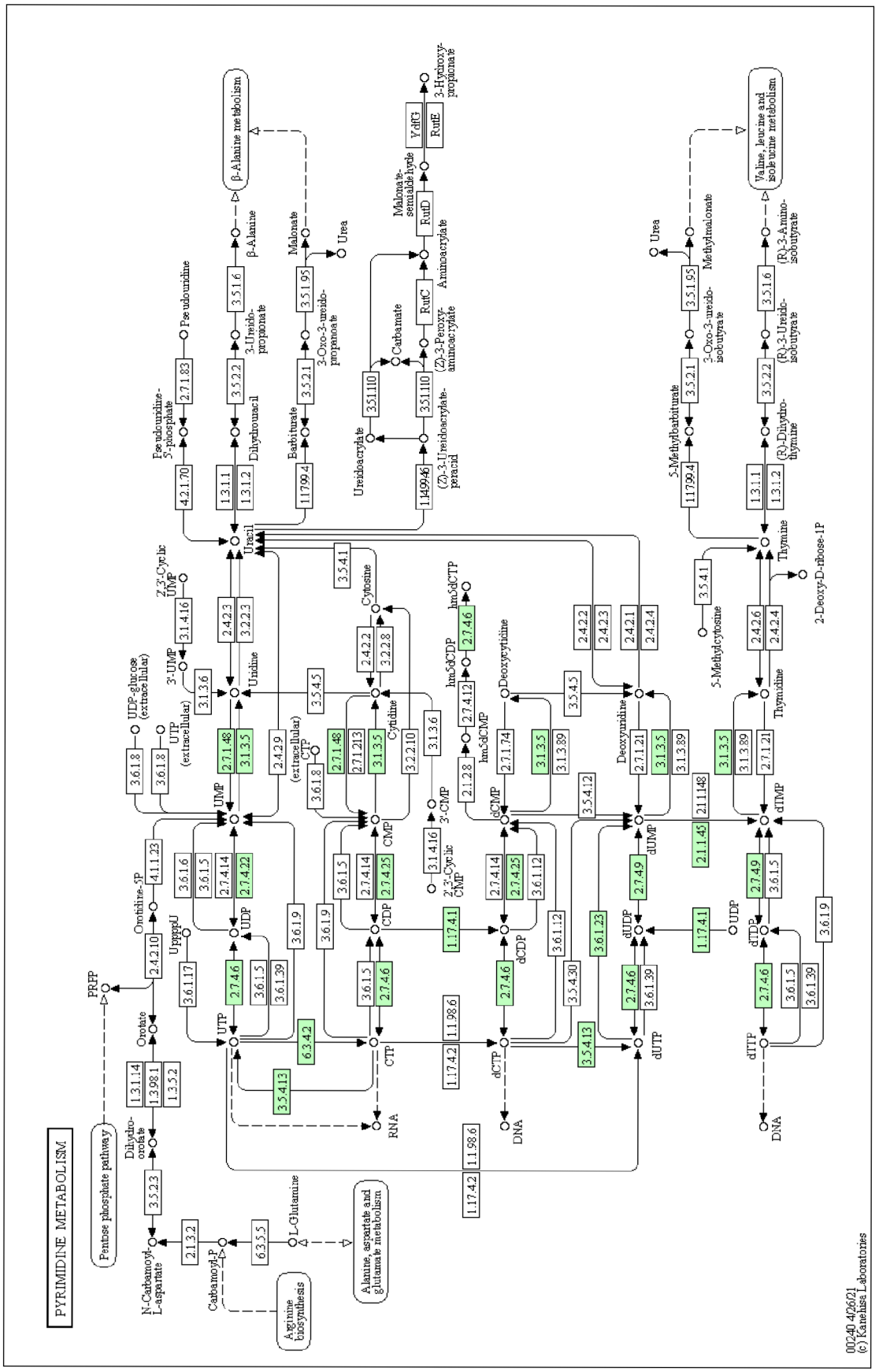

Figure 10: metabolic pathway of pyrimidine (*L. polyplacis*$_{spi}$)

4.1.9. Oxidative phosphorylation

The oxidative phosphorylation as seen in *Figure 11* shows various complexes. Many of them are present in *L. polyplacis*$_{spi}$.

Starting with the NADH dehydrogenase where all enzymes needed are present. The first enzyme is K00330 (NADH-quinone oxidoreductase subunit A) followed by the enzymes K00331 (NADH-quinone oxidoreductase subunit B), K00332 (NADH-quinone oxidoreductase subunit C), K00333 (NADH-quinone oxidoreductase subunit D), K00334 (NADH-quinone oxidoreductase subunit E), K00335 (NADH-quinone oxidoreductase subunit F), K00336 (NADH-quinone oxidoreductase subunit G), K00337 (NADH-quinone oxidoreductase subunit H), K00338 (NADH-quinone oxidoreductase subunit I), K00339 (NADH-quinone oxidoreductase subunit J), K00340 (NADH-quinone oxidoreductase subunit K), K00341 (NADH-quinone oxidoreductase subunit L), K00342 (NADH-quinone oxidoreductase subunit M) and K00343 (NADH-quinone oxidoreductase subunit N). This means that the first complex is fully functional.

For the complex succinate dehydrogenase all necessary enzymes are present and are named K00241 (succinate dehydrogenase / fumarate reductase, cytochrome b subunit), K00242 (succinate dehydrogenase / fumarate reductase, membrane anchor subunit), K00239 (succinate dehydrogenase / fumarate reductase, flavoprotein subunit) and K00240 (succinate dehydrogenase / fumarate reductase, iron-sulfur subunit)

Contrasting the complex cytochrome c reductase shows not one enzyme present in the bacteria and is non-functional

The cytochrome c oxidase complex is missing one enzyme for the process in K02277 (cytochrome c oxidase subunit IV). All the other enzymes are present like K02257 (heme o synthase), K02276 (cytochrome c oxidase subunit III), K02274 (cytochrome c oxidase subunit I) and K02275 (cytochrome c oxidase subunit II). The complex should be functional with some help of the host.

Lastly, the F-type bacterial ATP synthase has all enzymes needed starting with K02111 (F-type H+/Na+-transporting ATPase subunit alpha), following that the enzymes are K02112 (F-type H+/Na+-transporting ATPase subunit beta), K02115 (F-type H+/Na+-transporting ATPase subunit gamma), K02113 (F-type H+/Na+-transporting ATPase subunit delta), K02114 (F-type H+/Na+-transporting ATPase subunit epsilon), K02108 (F-type H+-transporting ATPase subunit a), K02109 (F-type H+-transporting ATPase subunit b) and K02110 (F-type H+-transporting ATPase subunit c). This means this complex is also fully functional.

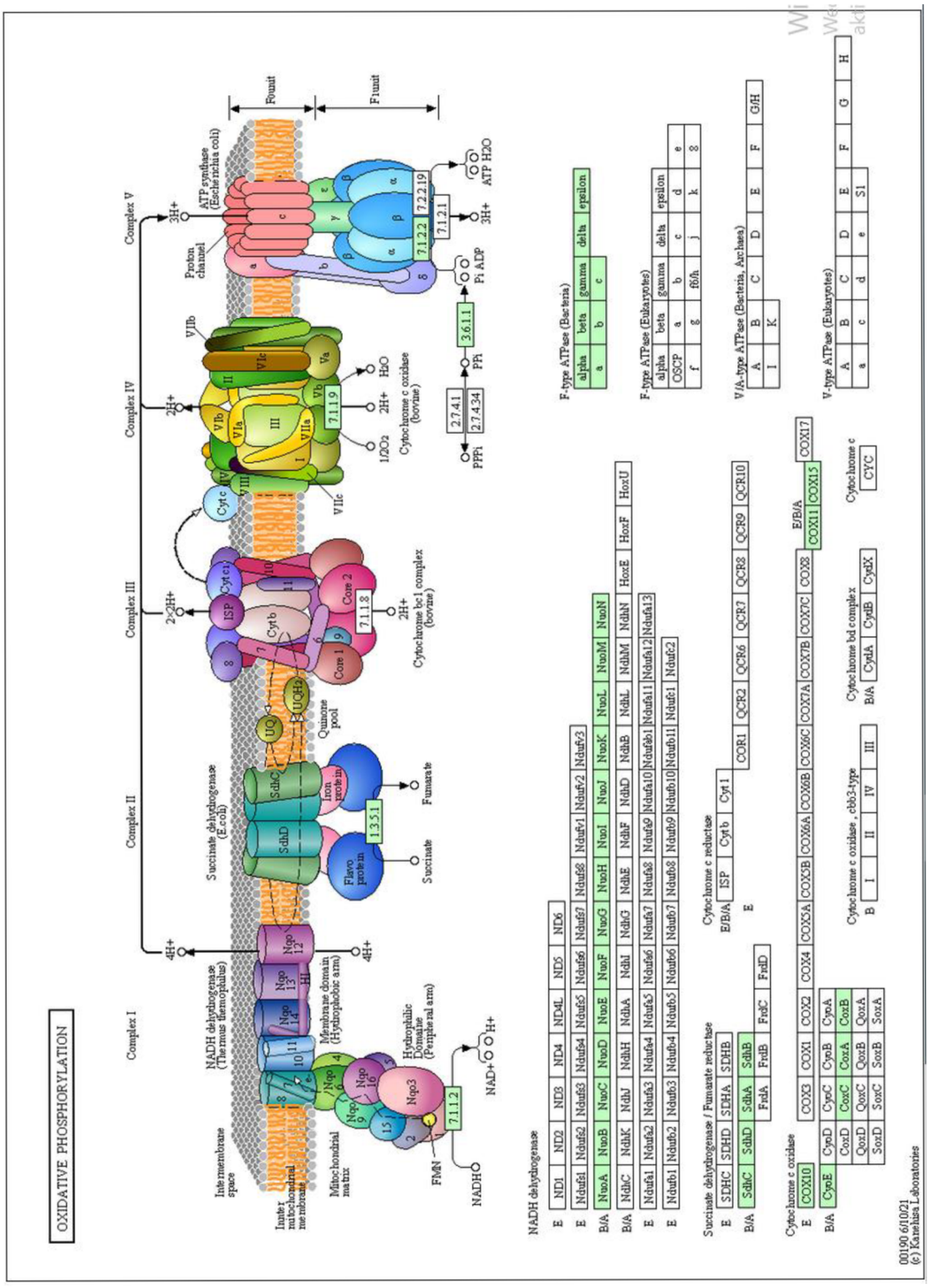

Figure 11: metabolic pathways for oxidative phosphorylation (*L. polyplacis*$_{spi}$)

4.2. Differences between *L.polyplacis in Polyplax serra*ta and *Polyplax spinulosa* and *R. pediculischaefii*

For the differences between *L.polyplacis* and *R. pediculischaefii* the Thesis of Ms.Zadinova (2021) is used as template and only comparison for *L. polyplacis* for *Polyplax serrata* and *Polyplax spinulosa* is done in this Thesis. For these two, there is only slight difference as they are closely related. The results show that there is no difference in the metabolic pathways for biotin, citric cycle, fatty acids, glycolysis, purine, pyrimidine and riboflavin. For the other four metabolic pathways shown before there are several differences, most of them are however minor and do not change the interference functionalities of the pathways.

4.2.1. Amino acids

The differences between the two genomes are not of functionally important kind as even though in *L. polyplacis*$_{ser}$ there are more enzymes found, none of these enzymes help to produce another amino acid. So basically, there are more enzymes present, which are K00615 (transketolase) and K01783 (ribulose-phosphate 3-epimerase) which are shown in *Figure 12*.

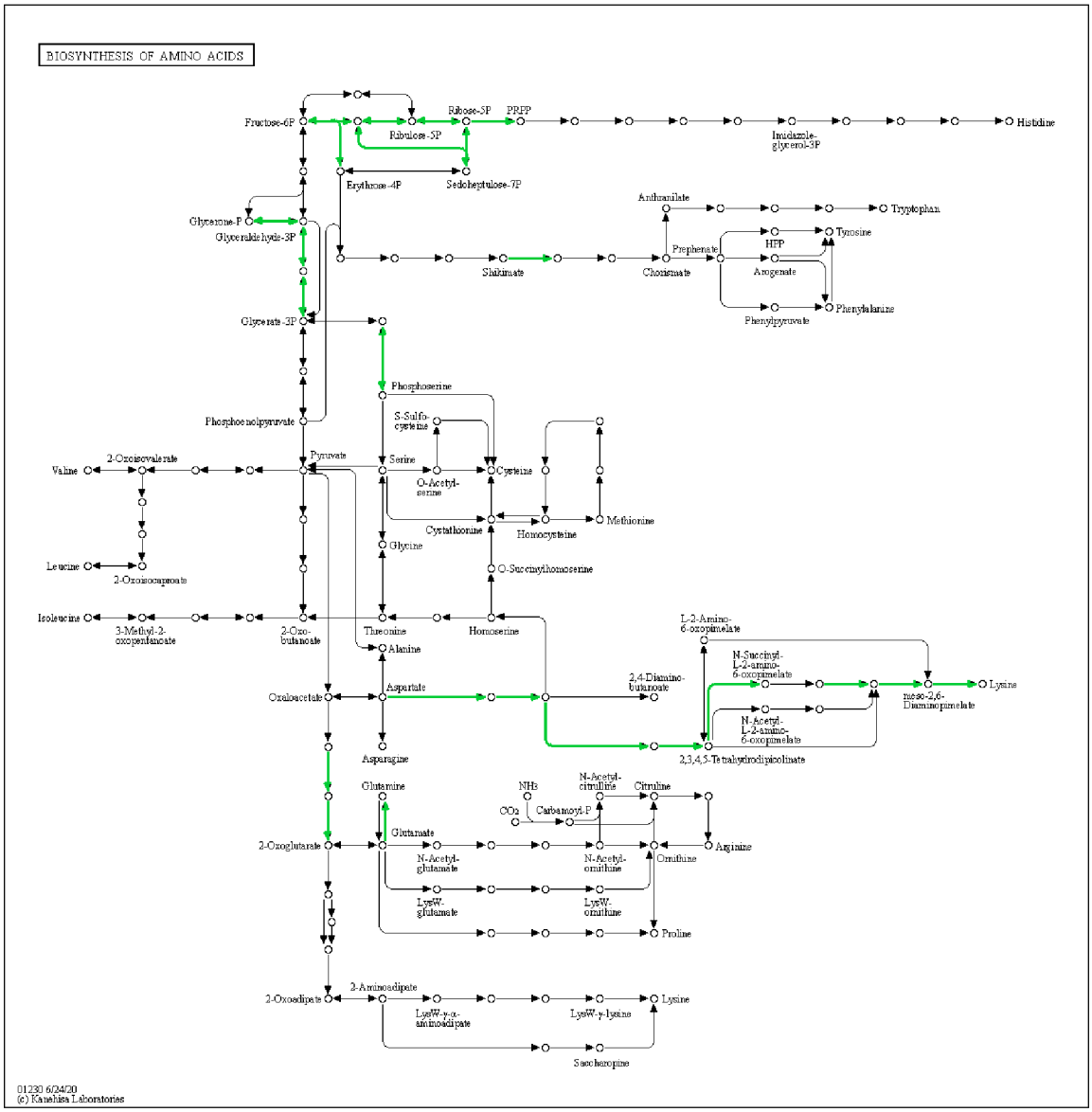

Figure 12: metabolic pathway of amino acids (*L. polyplacis*$_{ser}$)

4.2.2. Folic acids

Folic acid pathway display a single difference, as can be seen in Figure 13 for *L. polyplacis*$_{ser}$ the enzyme K06879 (7-cyano-7-deazaguanine reductase) is present. This enzyme however is not part of the process to produce folic acid, so this difference is not functionally significant.

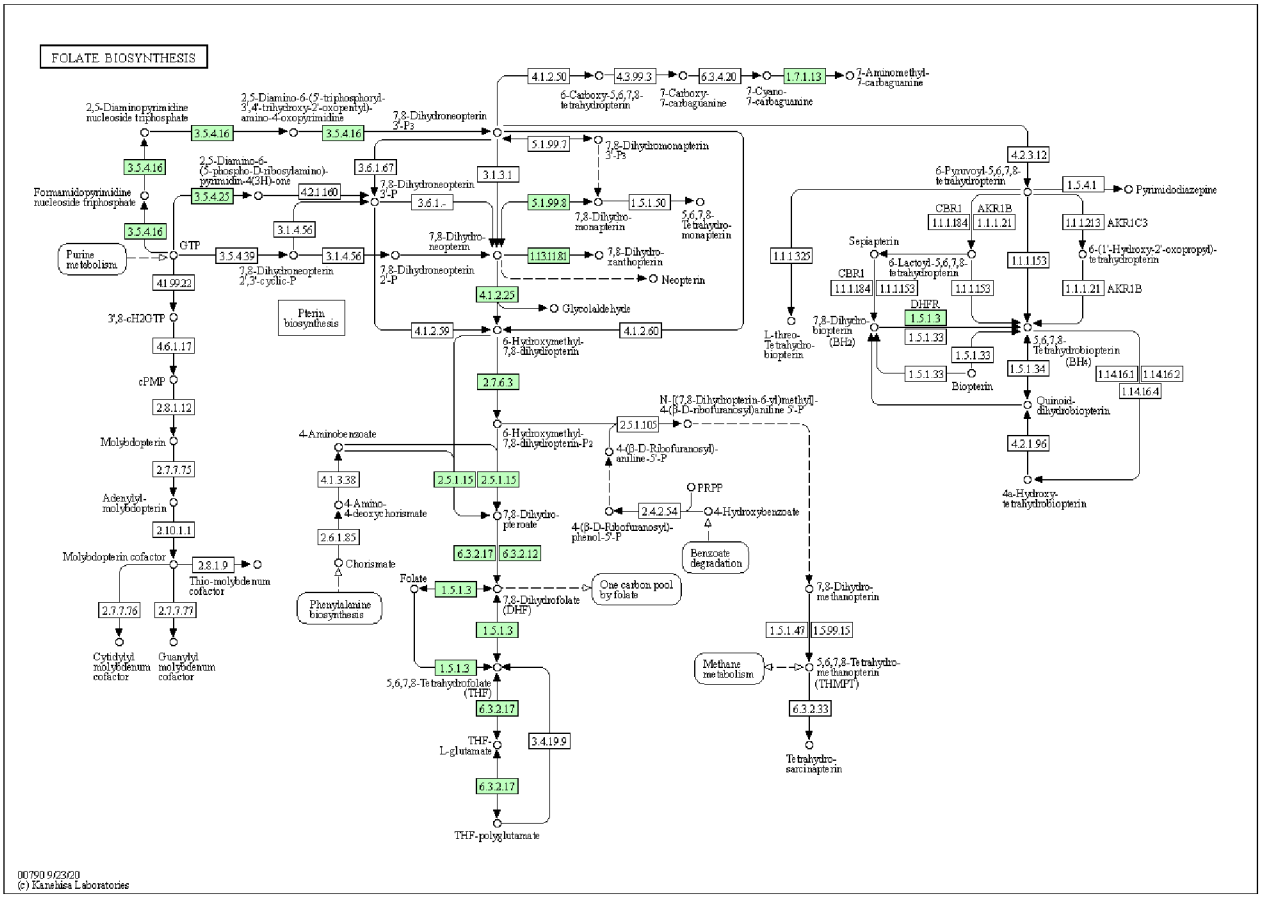

Figure 13: metabolic pathway of folic acid (*L. polyplacis*$_{ser}$)

4.2.3. Oxidative phosphorylation

The only difference here is that *L. polyplacis*$_{ser}$ the missing the enzyme K02257 (heme o synthase) however this has no major implication in the functionality.

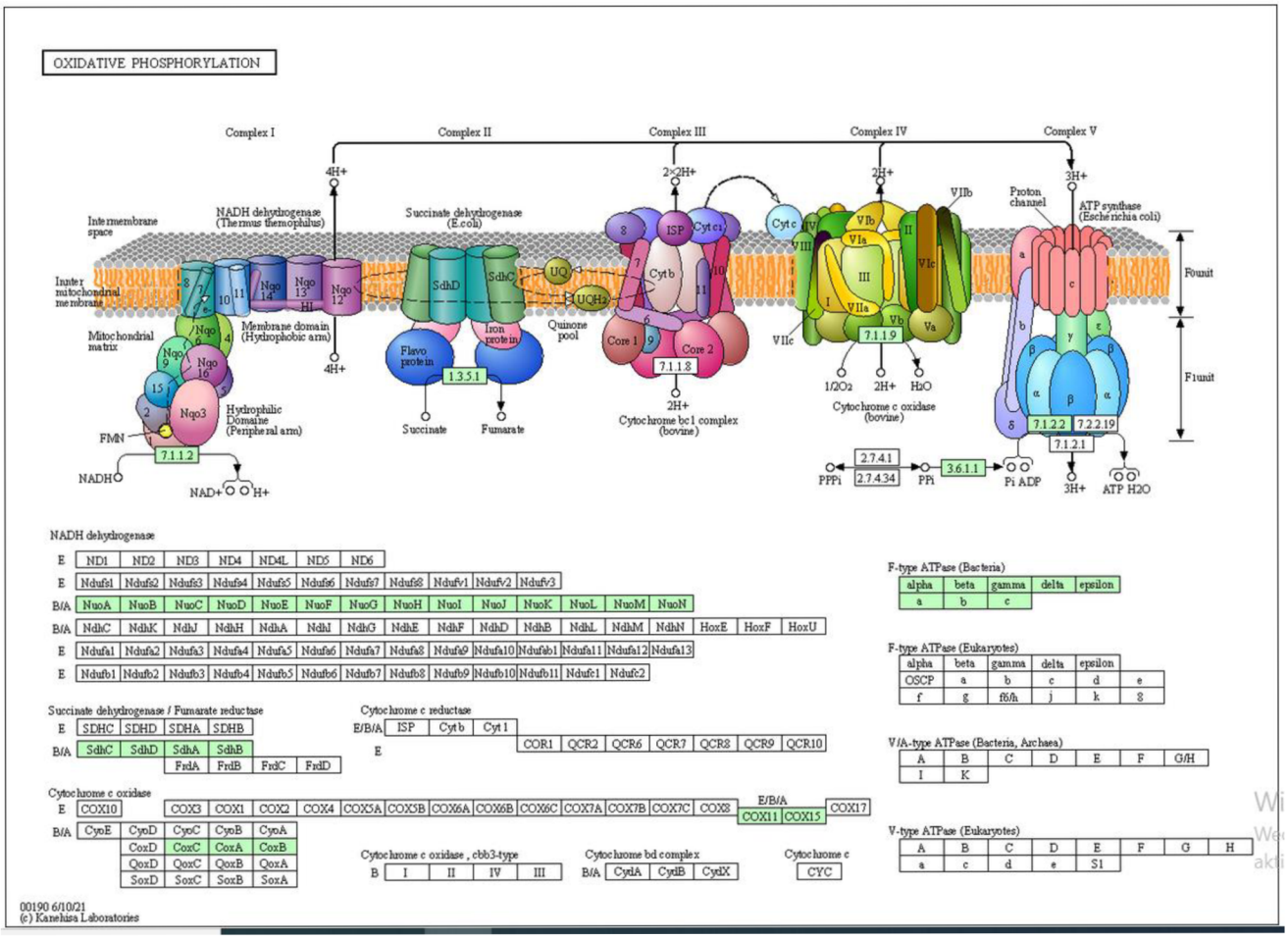

Figure 14: metabolic pathway of oxidative phosphorylation (*L. polyplacis*$_{ser}$)

4.2.4. Pentose Phosphate

As for the metabolic pathway of pentose phosphate the same enzymes as for the amino acids are present for *L. polyplacis*$_{ser}$ which means K00615 (transketolase) and K01783 (ribulose-phosphate 3-epimerase) which could help to produce either D-ribulose 5-phosphate or beta-D-fructose 6-phosphate from each other.

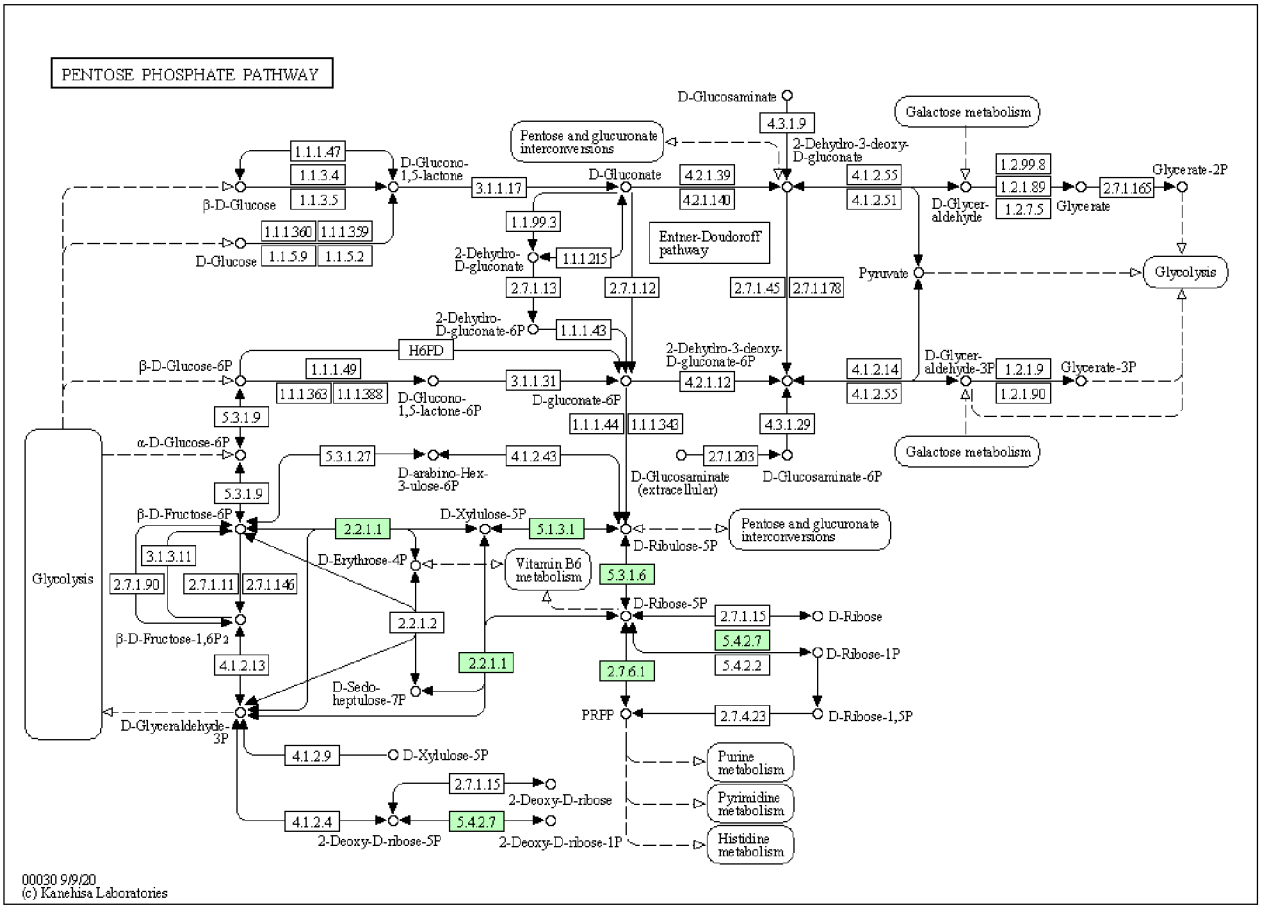

Figure 15: metabolic pathway of pentose phosphate (*L. polyplacis*$_{ser}$)

4.3. OrthoFinder results and their Ka/Ks values

OrthoFinder was able to find 220 orthogroups of single copy orthologue sequences from the three genomes. The alignments of them were then used to find the Ka/Ks values using a specific tree.

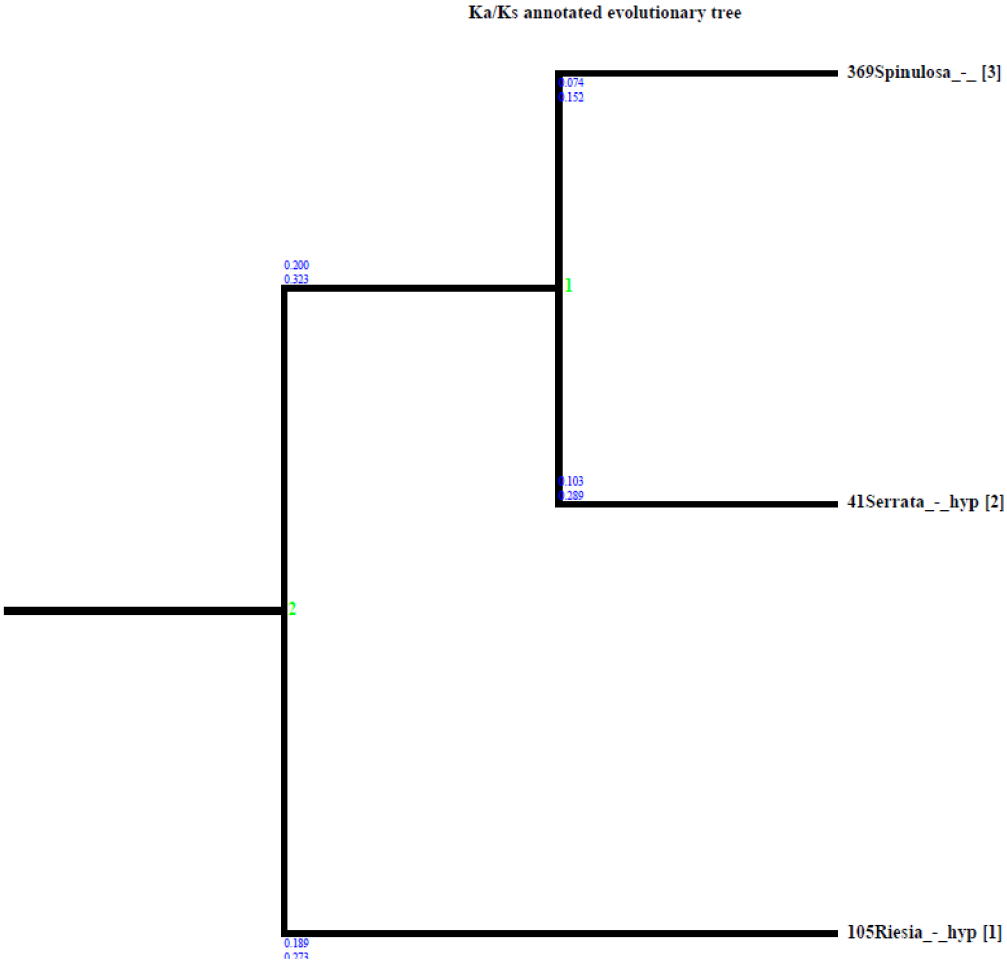

Figure 16: Evolutionary tree for Orthogroup 1

The Ka/Ks values were then ordered by their Ka/Ks values. This resulted in 218 of the orthogroups having a Ka/Ks value under one and two of the orthogroups having a Ka/Ks value over one. This means almost all the orthogroups are more likely to act against change. To further show the ten genes with the highest and lowest Ka/Ks values are shown in *Table 1*.

Table 1: Genes with the ten highest and lowest Ka/Ks values

Gene number	Orthogroup	Ka/Ks value	COG Category	Gene name
119Serrata	OG0000248	0	K	Cold shock protein of CSP family
244Spinulosa	OG0000151	0,01219979	O	Heat shock protein 10 kDa family chaperone GroES
245Spinulosa	OG0000150	0,01677127	O	Heat shock protein 60 kDa family chaperone GroEL
448Spinulosa	OG0000248	0,02820025	K	Cold shock protein of CSP family
401Serrata	OG0000150	0,02824856	O	Heat shock protein 60 kDa family chaperone GroEL
471Spinulosa	OG0000189	0,03622709	K	Transcription termination factor Rho
141Serrata	OG0000189	0,04173525	K	Transcription termination factor Rho
178Spinulosa	OG0000070	0,05173499	J	SSU ribosomal protein S10p (S20e)
174Spinulosa	OG0000066	0,06679395	J	SSU ribosomal protein S12p (S23e)
334Serrata	OG0000069	0,07231404	J	Translation elongation factor Tu
202Serrata	OG0000263	0,7638	J	hypothetical protein
317Serrata	OG0000098	0,7777	J	Peptide deformylase (EC 3.5.1.88)
400Spinulosa	OG0000129	0,7868	J	Translation elongation factor Ts
165Spinulosa	OG0000123	0,8618	H	Dihydrofolate reductase (EC 1.5.1.3)
303Serrata	OG0000174	0,8689	J	SSU ribosomal protein S15p (S13e)
440Spinulosa	OG0000047	0,8767	J	hypothetical protein
435Spinulosa	OG0000116	0,9007	J	LSU ribosomal protein L31p @ LSU ribosomal protein L31p, zinc-dependent
190Spinulosa	OG0000081	0,9486	J	LSU ribosomal protein L24p (L26e)
273Spinulosa	OG0000114	1,4225	J	LSU ribosomal protein L32p @ LSU ribosomal protein L32p, zinc-independent
147Spinulosa	OG0000174	1,5505	J	SSU ribosomal protein S15p (S13e)

As shown in *Table 1* there are some outliers like the only two orthogroups which showed a Ka/Ks value above one are in orthogroup 114, containing the gene LSU ribosomal protein L32p @ LSU ribosomal protein L32p, zinc-independent with the gene being from 273Spinulosa, and in orthogroup 174, containing the gene SSU ribosomal protein S15p (S13e) with the gene being from 147Spinulosa.

Another outlier is in orthogroup 248 as the Ka/Ks value of the gene from 119Serrata, Cold shock protein of CSP family, is zero.

Also shown in *Table 1* are the COG categories of theses outliers which show three major categories, O (Post-translational modification, protein turnover, chaperone functions), K (Transcription) and J (Translation), with one outlier in orthogroup 123 the gene from 165Spinulosa being in category H (Coenzyme metabolism).

4.4. COG categories of the orthogroups

The first part was to determine if all three genes in each orthogroup belonged to the same COG category. This was the case for 201 of the orthogroups. In 13 orthogroups at least one of the genes belonged to a different COG category. Lastly, six orthogroups had either one gene or all three genes with no attribution to a COG category.

The next step was to determine in which COG category the genes were present, how they were divided in these categories and the range of the Ka/Ks values for each category. This can be seen in *Table 2*.

Table 2: Amount of Genes found in COG categories and their Ka/Ks value range, X stands for no genes in this category

Short name	Description	Amount of Genes	Ka/Ks-values range
A	RNA processing and modification	2	0.2458-0.424
B	Chromatin Structure and dynamics	X	X
C	Energy production and conversion	30	0,09158611-0,6248
D	Cell cycle control and mitosis	4	0,3029-0,3739
E	Amino Acid metabolis and transport	16	0.1919-0.4628
F	Nucleotide metabolism and transport	58	0.09012921-0.7345
G	Carbohydrate metabolism and transport	6	0.2117-0.6884
H	Coenzyme metabolis	38	0.2011-0.8618
I	Lipid metabolism	28	0.1278-0.5186
J	Tranlsation	265	0.05173499-1.5505
K	Transcription	27	0-0.4509
L	Replication and repair	42	0.1511-0.6247
M	Cell wall/membrane/envelop biogenesis	21	0.2592-0.6637
N	Cell motility	X	X
O	Post-translational modification, protein turnover, chaperone functions	52	0.012199794-0.6018
P	Inorganic ion transport and metabolism	8	0.1443-0.4856
Q	Secondary Structure	X	X
T	Signal Transduction	4	0.2878-0.4703
U	Intracellular trafficing and secretion	20	0.09481863-0.5305
Y	Nuclear structure	X	X
Z	Cytoskeleton	X	X
R	General Functional Prediction only	X	X
S	Function Unknown	20	0.08890642-0.6511

5. Discussion

At first a comparison of the CG content shows both genomes have a similar value with 23% for *L. polyplacis*$_{ser}$ and 23.1% for *L. polyplacis*$_{spi}$ which indicates that both genomes have undergone considerable degradation, however they are on a similar level which also can be seen in the comparison of their metabolic pathways as they are quite similar in most important pathways shown in this thesis. Also, they are quite similar in the number of degraded genes, as there are not many left, because these degraded genes disappear from the gene over time.

Neither *L. polyplacis*$_{ser}$ nor *L. polyplacis*$_{spi}$ have retained complete pathways for the amino acids pathways (Fig. 1). However, this can be attributed to the fact that they are both symbionts to blood sucking lice and all the needed amino acids can be achieved through the amino acids present in the blood (Liao et. al, 2018). This makes the degradation of these pathways logical as they obtain their amino acids over other sources. From the few pathways which have enzymes left the lysine pathway (Fig. 1) is the most functionable. Lysine residues in proteins are covalent bound to biotin (B7) which can be synthesized by both bacteria, so this could be a reason the lysine pathway is still mostly intact (Zempleni et. al, 2009). According to the previous research *L.polyplacis* should be able to synthesize glutamine and glutamic acid (Fig. 1) (Říhová et al, 2021). Yet this research shows that both pathways need help from the host organism to synthesize either of these amino acids. In the end the missing enzymes and starting points, like glucose-6-phosphate, have no big influence as the host obtains most amino acids from blood, which will be most likely provided to the endosymbiont as well. If this is the case, the pathways may vanish completely over time as already can be seen with the lysine pathways starting the degrading process.

Research shows that many host organisms with a vitamin-deficient diet, need the help from their symbionts to obtain all the needed B-vitamins (Blow et. al, 2020). Therefore, both bacteria are able to synthesize some of the vitamins like the previously mentioned biotin (B7) (Fig. 3), which is an important coenzyme in various parts like carboxylase and also as cofactor for enzymes for important reactions (Zempleni et. al, 2009; Duron & Gottlieb, 2020). Which is why the pathway is still present as it is important for the survival of the host.

For riboflavin (B2) (Fig. 2) research by Duron and Gottlieb (2020) mentioned that *L. polyplacis* should not be able to synthesize it, however Říhová et. al. (2021) showed that it is partly functionable, which aligns more with what this thesis conducted as there is only one enzyme missing in the pathway. So, for riboflavin to be synthesized help from the host is needed but the enzymes to activate riboflavin into its physiological important coenzymes are present going over FMN and FAD (Pinto & Zempleni, 2016).

Lastly for folic acid (Fig. 4) the metabolic pathway is almost fully functional with the problem that *L.polyplacis* cannot synthesize 4-Aminobenzoate which is needed for the last steps of the pathway. With having the whole pathway from GTP to folic acid still being intact the assumption can be made that the host could supply the missing 4-Aminobenzoate. With this and the host also supplying one other enzyme down the path, the pathway for folic acid could be fully functionable. In contrary to work done by Duron and Gottlieb (2020) but supporting the work of Říhová et al.(2021) none of the other B-vitamins were found in this research.

For the fatty acid pathway (Fig. 5) most enzymes are present from acetyl CoA until the acp bound fatty acid step. Only this last step to release the fatty acid from the acp is missing so it is not clear if the bacteria can release fatty acids or not, but according to Říhová et al. (2021) the pathway should be fully functionable.

Glycolysis (Fig 6.) shows a low amount of enzymes, especially in the synthetisis of pyruvate which means that *L.polyplacis* would need it to be supplied by the host otherwise many other pathways would not work. However, the pathway probably serves one purpose in producing energetic compounds for the bacterium as the reactions from glyceraldehyde 3-phosphate to 3-phosphoglycerate are present.

For the pentose phosphate pathway (Fig. 7 and Fig. 15) a difference between *L. polyplacis*$_{ser}$ and *L. polyplacis*$_{spi}$ was found in this research as the degradation of the pathway has been stronger in *L. polyplacis*$_{spi}$ with only three enzymes left which means both oxidative and non-oxidative phases are non-functional. The pathway for *L. polyplacis*$_{ser}$ however shows that there are still enzymes left for the non-oxidative pathway but as there are still some enzymes missing the pathway is probably not functionable. This also supports the work from Říhová et al.(2021).

For the citric cycle (Fig. 8) most of the enzymes are present, however as explained previously the host has to supply pyruvate as it cannot be synthesized by the bacteria. The pathway to produce acetyl-CoA is present but unfortunately is the enzyme to form citrate missing On the other hand, through the enzyme K00027 (malate dehydrogenase) the bacterium can produce malate from pyruvate. This means as the cycle can run in both directions the citric cycle is fully functional as all other enzymes for the cycle are present. So if the host supplies pyruvate, the citric cycle is fully functional in both bacteria.

For the purine synthesis (Fig 9.) both bacteria need an IMP supply from the host, with which they can form purine as well as ATP, GTP and few other products. This also applies to the synthesis of pyrimidine (Fig. 10) where the bacteria need UMP from the host to synthesise products like UTP or CTP. However, the last enzyme for either uracil, thymine or cytosine production is missing. This

is all in accord with the work of Říhová et al. (2021), as she also shows that without the help of the host these pathways would not function.

The oxidative phosphorylation pathway (Fig. 11 and Fig. 14) in both bacteria has three of the five subunit complexes working and completely functional, complex IV is missing just one enzyme in *L. polyplacis*$_{spi}$ and two enzymes in *L. polyplacis*$_{ser.}$ Only complex III is missing completely. So if the host is unable to supply ubiquinol and help out in in complex IV the ATP synthesis has to be considered non-functional.

For the Ka/Ks values this work compared the two *L. polyplacis* symbionts in *Polyplax serrata* and *Polyplax spinulosa* and also the symbiont *R. pediculicola.* This resulted in almost fully Ka/Ks values under one which indicates a purifying selection in the genomes. This also means that across the three bacteria the amount of synonymous substituents (Ks) was higher, which is contradictory to findings of Wernegreen and Moran (1999) who found Ka/Ks values mostly above one in the endosymbiont *Buchnera.* However, these results are a preparatory step which helps further functional analyses which would be too much for this thesis. The main result for this thesis is to show the variance in the value, suggesting that different genes are under different selection pressure even in these two closely related bacteria.

As *Table 2* shows most of the categories had only a few of the genes belong to them namely the following one A, C, D, E, F, G, H, I, K, L, M, O, P, T, U, S. 10 of the genes also were placed in two different categories. The most genes fell into category J (Translation) with 265 genes present. Other research on endosymbionts in insects showed similar traits with the category J being the most common and also showed that some categories, like N (cell motility) and Z (cytoskeleton) are not present as they are not needed in the controlled environment of an insect host (Kambhampati et. al, 2013; George et. al, 2020).

6. Conclusion

Both genomes *L. polyplacis*$_{ser}$ and *L. polyplacis*$_{spi}$ show reduced GC content and also both have a major reduction of their functional metabolic pathway due to degradation of unneeded pathways. These pathways are unneeded as the symbiotic between the bacteria and their host supplies the bacteria with the necessary proteins or enzymes to survive.

Both of them share a many similarities, especially in their GC content and the metabolic pathways as only four differences were found between their pathways and none of them were of major importance. This means both are on a similar level of their degradation, which is already in an advanced stage compared to other endosymbionts like *R. pediculicola.*

In comparison to other work done of the department of parasitology, namely Ms. Říhová and Ms. Zadinova, the pathways showed a big similarity to the work of Ms. Říhová as only glutamine and glutamic acid did not show a fully functional pathway in the thesis which was shown by her.

There are also open question about the functionality of the fatty acid and the oxidative phosphorylation pathways as both would need the help from the host to overcome important steps in the synthesis which cannot be shown by this thesis and would maybe be a good subject for further investigation.

Lastly the Ka/Ks values show contradictory results to published work as here the amount of synonymous substituents (Ks) was higher, which means the bacteria indicates purifying selection in their genome. This could also prove as interesting topic for the future research. Also the COG categories showed that most of the genes are in the J (Translation) category which was supported by other research.

7. Literature

Allen, J. M., Reed, D. L., Perotti, M. A., & Braig, H. R. (2007). Evolutionary relationships of "Candidatus Riesia spp.," endosymbiotic enterobacteriaceae living within hematophagous primate lice. *Applied and environmental microbiology*, *73*(5), 1659–1664. https://doi.org/10.1128/AEM.01877-06

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., … Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*, 75. https://doi.org/10.1186/1471-2164-9-75

Bennett, G. M., & Moran, N. A. (2013). Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome biology and evolution*, *5*(9), 1675–1688. https://doi.org/10.1093/gbe/evt118

Blow, F., Bueno, E., Clark, N., Zhu, D. T., Chung, S. H., Güllert, S., Schmitz, R. A., & Douglas, A. E. (2020). B-vitamin nutrition in the pea aphid-Buchnera symbiosis. *Journal of insect physiology*, *126*, 104092. https://doi.org/10.1016/j.jinsphys.2020.104092

Boyd, B. M., Allen, J. M., Nguyen, N. P., Vachaspati, P., Quicksall, Z. S., Warnow, T., Mugisha, L., Johnson, K. P., & Reed, D. L. (2017). Primates, Lice and Bacteria: Speciation and Genome Evolution in the Symbionts of Hominid Lice. *Molecular biology and evolution*, *34*(7), 1743–1757. https://doi.org/10.1093/molbev/msx117

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., 3rd, Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, *5*, 8365. https://doi.org/10.1038/srep08365

Burkhart, C. N., & Burkhart, C. G. (2006). Bacterial symbiotes, their presence in head lice, and potential treatment avenues. *Journal of cutaneous medicine and surgery*, *10*(1), 2–6. https://doi.org/10.1007/7140.2006.00003

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular biology and evolution*, msab293. Advance online publication. https://doi.org/10.1093/molbev/msab293

Duron, O., & Gottlieb, Y. (2020). Convergence of Nutritional Symbioses in Obligate Blood Feeders. *Trends in parasitology*, *36*(10), 816–825. https://doi.org/10.1016/j.pt.2020.07.007

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Fields, B. S., Benson, R. F., & Besser, R. E. (2002). Legionella and Legionnaires' disease: 25 years of investigation. *Clinical microbiology reviews*, *15*(3), 506–526. https://doi.org/10.1128/CMR.15.3.506-526.2002

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic acids research*, *49*(D1), D274–D281. https://doi.org/10.1093/nar/gkaa1018

George, E. E., Husnik, F., Tashyreva, D., Prokopchuk, G., Horák, A., Kwong, W. K., Lukeš, J., & Keeling, P. J. (2020). Highly Reduced Genomes of Protist Endosymbionts Show Evolutionary Convergence. *Current biology : CB*, *30*(5), 925–933.e3. https://doi.org/10.1016/j.cub.2019.12.070

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Hurst L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in genetics : TIG*, *18*(9), 486. https://doi.org/10.1016/s0168-9525(02)02722-1

Hypsa, V., & Krízek, J. (2007). Molecular evidence for polyphyletic origin of the primary symbionts of sucking lice (phthiraptera, anoplura). *Microbial ecology*, *54*(2), 242–251. https://doi.org/10.1007/s00248-006-9194-x

Johnson K. N. (2015). Bacteria and antiviral immunity in insects. *Current opinion in insect science*, *8*, 97–103. https://doi.org/10.1016/j.cois.2015.01.008

Kambhampati, S., Alleman, A., & Park, Y. (2013). Complete genome sequence of the endosymbiont Blattabacterium from the cockroach Nauphoeta cinerea (Blattodea: Blaberidae). *Genomics*, *102*(5-6), 479–483. https://doi.org/10.1016/j.ygeno.2013.09.003

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of molecular biology*, *428*(4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006

Kanehisa, M., & Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein science : a publication of the Protein Society*, *29*(1), 28–35. https://doi.org/10.1002/pro.3711

Kanehisa, M., Sato, Y., & Kawashima, M. (2021). KEGG mapping tools for uncovering hidden features in biological data. *Protein science : a publication of the Protein Society*, 10.1002/pro.4172. Advance online publication. https://doi.org/10.1002/pro.4172

Kikuchi Y. (2009). Endosymbiotic bacteria in insects: their diversity and culturability. *Microbes and environments*, *24*(3), 195–204. https://doi.org/10.1264/jsme2.me09140s

Kubiak, K., Sielawa, H., Chen, W., & Dzika, E. (2018). Endosymbiosis and its significance in dermatology. *Journal of the European Academy of Dermatology and Venereology : JEADV*, *32*(3), 347–354. https://doi.org/10.1111/jdv.14721

Latorre, A., & Manzano-Marín, A. (2017). Dissecting genome reduction and trait loss in insect endosymbionts. *Annals of the New York Academy of Sciences*, *1389*(1), 52–75. https://doi.org/10.1111/nyas.13222

Liao, S. F., Regmi, N., & Wu, G. (2018). Homeostatic regulation of plasma amino acid concentrations. *Frontiers in bioscience (Landmark edition)*, *23*, 640–655. https://doi.org/10.2741/4610

Li, J., Zhang, Z., Vang, S., Yu, J., Wong, G. K., & Wang, J. (2009). Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. *Journal of molecular evolution*, *68*(4), 414–423. https://doi.org/10.1007/s00239-009-9222-9

Li, L., Stoeckert, C. J., Jr, & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Manzano-Marín, A., & Latorre, A. (2016). Snapshots of a shrinking partner: Genome reduction in Serratia symbiotica. *Scientific reports*, *6*, 32590. https://doi.org/10.1038/srep32590

Marais, G. A., Calteau, A., & Tenaillon, O. (2008). Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica*, *134*(2), 205–210. https://doi.org/10.1007/s10709-007-9226-6

Nicks, T., & Rahn-Lee, L. (2017). Inside Out: Archaeal Ectosymbionts Suggest a Second Model of Reduced-Genome Evolution. *Frontiers in microbiology*, *8*, 384. https://doi.org/10.3389/fmicb.2017.00384

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, *42*(Database issue), D206–D214. https://doi.org/10.1093/nar/gkt1226

Pinto, J. T., & Zempleni, J. (2016). Riboflavin. *Advances in nutrition (Bethesda, Md.)*, *7*(5), 973–975. https://doi.org/10.3945/an.116.012716

Říhová, J., Nováková, E., Husník, F., & Hypša, V. (2017). Legionella Becoming a Mutualist: Adaptive Processes Shaping the Genome of Symbiont in the Louse Polyplax serrata. *Genome biology and evolution*, *9*(11), 2946–2957. https://doi.org/10.1093/gbe/evx217

Říhová, J., Batani, G., Rodríguez-Ruano, S. M., Martinů, J., Vácha, F., Nováková, E., & Hypša, V. (2021). A new symbiotic lineage related to Neisseria and Snodgrassella arises from the dynamic and diverse microbiomes in sucking lice. Molecular ecology, 30(9), 2178–2196. https://doi.org/10.1111/mec.15866

Rounds, M. A., Crowder, C. D., Matthews, H. E., Philipson, C. A., Scoles, G. A., Ecker, D. J., Schutzer, S. E., & Eshoo, M. W. (2012). Identification of endosymbionts in ticks by broad-range polymerase chain reaction and electrospray ionization mass spectrometry. *Journal of medical entomology*, *49*(4), 843–850. https://doi.org/10.1603/me12038

Sabater-Muñoz, B., Toft, C., Alvarez-Ponce, D., & Fares, M. A. (2017). Chance and necessity in the genome evolution of endosymbiotic bacteria of insects. *The ISME journal*, *11*(6), 1291–1304. https://doi.org/10.1038/ismej.2017.18

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, *28*(1), 33–36. https://doi.org/10.1093/nar/28.1.33

Taylor, M., Mediannikov, O., Raoult, D., & Greub, G. (2012). Endosymbiotic bacteria associated with nematodes, ticks and amoebae. *FEMS immunology and medical microbiology*, *64*(1), 21–31. https://doi.org/10.1111/j.1574-695X.2011.00916.x
Wang, C., Chuai, X., & Liang, M. (2019). Legionella feeleii: pneumonia or Pontiac fever? Bacterial virulence traits and host immune response. *Medical microbiology and immunology*, *208*(1), 25–32. https://doi.org/10.1007/s00430-018-0571-0

Wang, D., Zhang, S., He, F., Zhu, J., Hu, S., & Yu, J. (2009). How do variable substitution rates influence Ka and Ks calculations?. *Genomics, proteomics & bioinformatics*, *7*(3), 116–127. https://doi.org/10.1016/S1672-0229(08)60040-6

Wernegreen J. J. (2012). Endosymbiosis. *Current biology : CB*, *22*(14), R555–R561. https://doi.org/10.1016/j.cub.2012.06.010

Wernegreen, J. J., & Moran, N. A. (1999). Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. *Molecular biology and evolution*, *16*(1), 83–97. https://doi.org/10.1093/oxfordjournals.molbev.a026040

Zempleni, J., Wijeratne, S. S., & Hassan, Y. I. (2009). Biotin. *BioFactors (Oxford, England)*, *35*(1), 36–46. https://doi.org/10.1002/biof.8

Zytynska S. E. (2019). Cohabitation and roommate bias of symbiotic bacteria in insect hosts. *Molecular ecology*, *28*(24), 5199–5202. https://doi.org/10.1111/mec.15295