**University of South Bohemia in České Budějovice**
**Faculty of Science**

**and**

**Johannes Kepler University in Linz**
**Faculty of Engineering and Natural Sciences**

# Bachelor Thesis

**2016** **Kamila Machová**

**University of South Bohemia in České Budějovice**
**Faculty of Science**

**and**

**Johannes Kepler University in Linz**
**Faculty of Engineering and Natural Sciences**

# RNA biology of symbiotic bacteria in insects

Bachelor Thesis

**Kamila Machová**

Supervisor: RNDr. Filip Husník

České Budějovice 2016

Machová, K., 2016: RNA biology of symbiotic bacteria in insects. [Bc. Thesis, in English.]–46 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

## ANOTATION

Non-coding RNAs (ncRNAs) represent an important part of bacterial genomes. However, only few studies RNAs with limited sampling were done concerning ncRNAs of insect endosymbiotic bacteria. This study provides a broad *in silico* genome sampling of insect endosymbionts (63 lineages of 27 genera) for ncRNAs and their modifications. Most strikingly it was found out that i) genes encoding modification enzymes conserved in particular genomes differ to high extent, ii) most of tRNA and rRNA modification sites are conserved regardless whether the gene encoding the corresponding modification enzyme is conserved, iii) multiple endosymbiont lineages do not encode a full set of tRNAs. Our data imply that translation of endosymbionts is much less efficient compared to phylogenetically related free-living bacteria and that some of symbionts possibly need to cooperate with their co-symbionts or maybe even with their host to maintain translation.

I hereby declare that I have worked on my bachelor thesis independently and used only the sources listed in the bibliography.
I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full form to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages.
Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

In České Budějovice 18.4.2016

.....
Signature

# ACKNOWLEDGEMENTS

At first I would like to thank my supervisor for offering interesting thesis topics regarding endosymbiotic bacteria and supervising my thesis. Especially, I would like to appreciate his patience in my beginnings with command line, letting me work independently enough later on and correcting lots of my typing mistakes which I did not manage to find on my own.

I would also like to appreciate providing an unpublished article "tRNA editing, CCA addition, and gene loss in a set of bacterial endosymbionts" by James T. van Leuven and John P McCutcheon to myself. This article brought me multiple interesting findings to discuss.

I would also like to acknowledge all lecturers from both University of South Bohemia and Johannes Kepler University who thought me some of skills I used for work on this thesis, especially Pavel Fibich for bash, Ulrich Bodenhofer for R and knitr and Jakub Těšitel for basics of statistic. Furthermore I thank Zdeněk Paris for suggesting me to use Modomics database.

Finally, I would like to appreciate all people answering questions regarding working on computers at various stack exchange fora. They help not only the person who asked but provide detailed practical manuals for diverse tasks to many others.

# Contents

# List of Figures



# List of Tables

# 1 Introduction

## 1.1 Obligate endosymbiotic bacteria

Many eukaryotes have obligate associations with bacteria that are beneficial and heritable. Insects have numerous such alliances and their obligate endosymbiotic bacteria have originated multiple times independently from diverse bacterial groups. These bacteria are harboured inside specialized insect cells called bacteriocytes which sometimes form a distinct organ called bacteriome. Frequently, several different bacterial species (co-symbionts) coinhabit the same host and cooperate on production of nutrients provided to the host. In such cases, they are usually housed in their own distinct bacteriocytes (or even separate bacteriomes) in a single host (Baumann, 2005; Moran et al., 2008; McCutcheon and Moran, 2011; Bennett and Moran, 2013), but there are two known exeptions from plant sap-sucking insects. In whiteflies, a gammaproteobacterium *Portiera aleyrodidarum*[1] can share its bacteriocytes with several different co-symbionts (Gottlieb et al., 2008). Even more strikingly, in mealybugs, a betaproteobacterium *Tremblaya princeps* harbours its co-symbionts (called *Moranella endobia* in *Planococcus citri* mealybugs) inside its own cells (von Dohlen et al., 2001). This is the only known case of a bacterium stably residing inside another bacterium.

Only a small fraction of the original population of endosymbiotic bacteria is maternally transmitted to the next generations of their hosts. Therefore their populations undergo series of bottlenecks which cause small effective population sizes compared to free-living relative bacteria (Moran, 1996; Lambert and Moran, 1998; Woolfit, 2003). Together with their asexuality, this results in high levels of genetic drift and enables accumulation of slightly deleterious mutations. Small fitness-reducing effect of a single mutation does not prevent its fixation. Therefore this fixation is irreversible (Muller's ratchet). Cumulative effect of these mutations can be huge. In addition to protein encoding sequences, this effect was observed also in 16S rRNA gene (Lambert and Moran, 1998; Woolfit, 2003). Genetic drift and Muller's ratchet can lead to rapid sequence evolution, gene loss, lower thermal stability of proteins, codon reassignments and extreme biases in nucleotide composition (McCutcheon and Moran, 2011).

Fitness of endosymbionts depends on fitness of their hosts. Mutations benefiting endosymbionts at its host expense or deleterious ones could cause extinction of both (Wernegreen, 2002). Fixation rate of these mutations is increasing with decreasing host population sizes, but is slowed down with endodymbiont transmission causing the bottleneck effect and due to selection of endosymbionts at host level. Selection at host level works such that endosymbionts which are less useful for their host reduce their hosts fitness and are less trans-

[1] Non-cultivable bacteria should be formally described using *Candidatus* status. I use the names without this status hereafter in this thesis because vast majority of bacteria including obligate symbionts in insects is not cultivable and moreover non-cultivable species described before this taxonomic rule was suggested lack this status.

mitted to next generations (Rispe and Moran, 2000; Pettersson and Berg, 2006). Despite these mechanisms it seems that long-term endosymbiotic relationship leads to irreversible degeneration of both organisms. The host can escape it via replacing its endosymbiont by a new one (Lefevre, 2004; Conord et al., 2008; Bennett and Moran, 2015; Husnik and McCutcheon, 2016) which might also broaden its ecological opportunities (Husnik and McCutcheon, 2016).

In the initial stages of obligate endosymbiosis, bacterial genomes undergo rapid gene loss and inactivation (pseudogenization), chromosome rearrangements, and proliferation of mobile genetic elements. However, with the ongoing genome reduction, most of the accumulated pseudogenes and mobile genetic elements are purged and the genomes eventually become gene-dense. Long-term co-evolution with the host also enables loss of genes considered essential for free-living bacteria as endosymbionts are living in a stable and nutrient rich environment of the insect cell (McCutcheon and Moran, 2011; Figure 1).

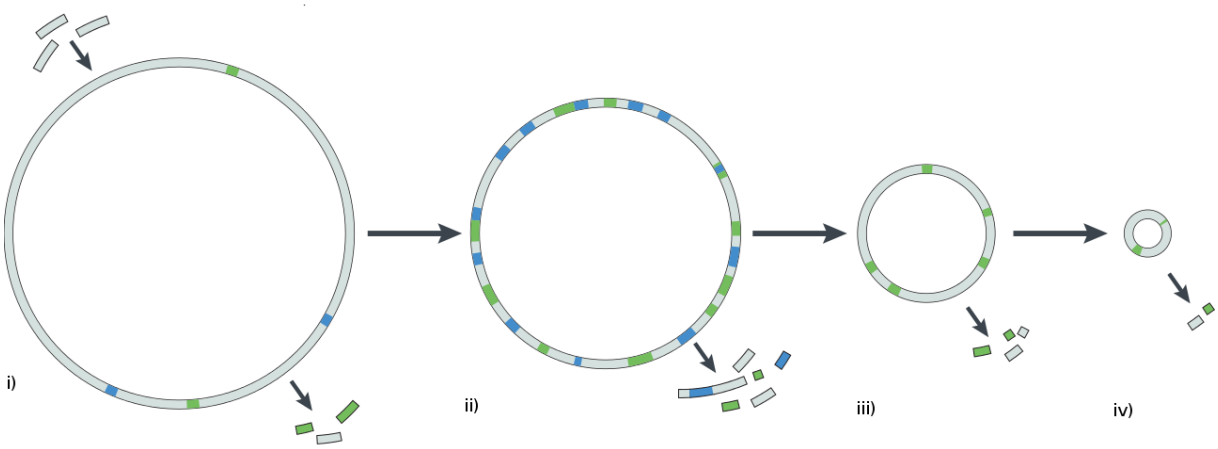


Figure 1: Stages of genome reduction in endosymbiotic bacteria: (i) a genome of a free-living bacterium, (ii) a genome of a recent endosymbiont containing numerous pseudogenes and mobile genetic elements, (iii) a highly packed genome of a long-term endosymbiont with almost no mobile elements and only a few pseudogenes, (iv) an extremely reduced genome with high coding density. Mobile genetic elements are color-coded in dark blue and pseudogenes in green. Modified from (McCutcheon and Moran, 2011)

Genome sizes of obligate endosymbionts in insects vary from 4,513 kbp for *Sodalis pierantonius* str. SOPE to 112 kbp for *Nasuia deltocephalinicola* str. PUNC (Oakeson et al., 2014; Bennett et al., 2016). The latter number is less than one quarter of genome size of *Mycobacterium genitalium*, an organism with the smallest genome which can be grown *in vitro* in absence of any other organism (McCutcheon and Moran, 2011). Coding density of these genomes can be greater than 97.3% (*Carsonella ruddii*; Nakabachi et al., 2006) and in most cases varies around 90% (McCutcheon and von Dohlen, 2011). However, there are some exceptions such as *Tremblaya princeps* PCIT with coding density 72.9% (McCutcheon and von Dohlen, 2011), *Sodalis pierantonius* SOPE with almost half of all genes pseudogenized (Oakeson et al., 2014) or two *Hodgkinia* lineages which have split from one within their host (none of them is included in this study because they were published after some of my analyses

were done and it would be complicated to include them; Campbell et al., 2015).

Mutations are universally biased towards AT in bacteria (Hershberg and Petrov, 2010). Coupled with a lack of reparation mechanisms and population genetics mechanisms described above, endosymbiont genomes tend to have extremely high AT content (the lowest known GC content among bacteria is 13.5% from *Zinderia insecticola*). Interestingly, *Tremblaya princeps* and *Hodgkinia cicadicola* seem to represent exceptions of this pattern with GC contents around 58%, although they likely still have AT mutational bias (Leuven and McCutcheon, 2011) and there are other species/strains in these two clades with much higher AT content (Husník et al., 2013; Leuven et al., 2014).

Genes of every functional category can be lost from endosymbiont genomes, but genes involved in core bacterial genetic machinery such as replication, transcription or translation (Nakabachi et al., 2006; Bennett and Moran, 2013) are usually retained. For example *Hodgkinia cicadicola* has retained only two genes (*dnaE* and *dnaQ*) coding DNA polymerase III (McCutcheon, 2010) and four endosymbionts (*Tremblaya princeps*, *Hodgkinia cicadicola*, *Carsonella ruddii* and *Zinderia insecticola*) have retained only one or even no genes involved in cell envelope synthesis implying that their membranes are likely host-derrived and lack peptidoglycan ((McCutcheon and Moran, 2011); *Nasuia deltocephalinicola*, the only remaining symbiont with a tiny genome, was not included in this study).

In addition to genes involved in core genetic processes, genes related to protein folding are the only other class of genes retained in all of the most reduced endosymbiont genomes suggesting that these proteins play a critical role in their biology (McCutcheon and Moran, 2011), perhaps by buffering low thermal stability of endosymbiont proteins (GroEL, (Fares et al., 2002)). Depending on the endosymbiont role for the insect host, remaining genes retained in its genome are mostly involved in synthesis of nutrients needed by the host. In particular, biosynthetic pathways for essential amino acids are retained in endosymbiont genomes of plant sap-sucking insects and biosynthetic pathways for B-vitamins and cofactors are retained in endosymbiont genomes of blood-sucking insects. However, the situation is not always that clear. For example, *Sodalis pierantonius* str. SOPE is needed to match a drastic host physiological need for a single non-essential aminoacid tyrosine, which is required by its weevil host as a precursor to build protective exoskeleton (Vigneron et al., 2014). There is also one psyllid endosymbiont, *Profftella armatura*, likely harboured for not nutritional, but defensive purpose (Nakabachi et al., 2013). Biosynthetic capabilities are in most cases complementarily divided among co-symbionts (with the most extreme case in mealybugs where even a single pathway can rely on gene products from both symbionts) and the very last steps of biosynthetic pathways are quite often carried out by host enzymes in the host cytoplasm (Hansen and Moran, 2013).

Unlike in organelles, none or only a few genes were transferred via horizontal gene

transfer (HGT) from obligate endosymbiotic bacteria to their hosts (Kirkness et al., 2010; Nikoh et al., 2010; Husník et al., 2013; Sloan et al., 2014; Luan et al., 2015). Two short non-functional DNA fragments were acquired by aphids from their endosymbiont *Buchnera aphidicola* (Nikoh et al., 2010) and a single gene involved in arginin biosynthesis (*argH*) was acquired by *Pachypsylla venusta* psyllid from its endosymbiont *Carsonella ruddii*. However, majority of bacterial genes transferred into insect genomes and used to support obligate endosymbionts were acquired from other bacteria, mainly facultative endosymbionts or reproductive manipulators such as *Arsenophonus*, *Cardinium*, *Rickettsia*, *Sodalis*, *Serratia*, and *Wolbachia* (Kirkness et al., 2010; Nikoh et al., 2010; Husník et al., 2013; Sloan et al., 2014; Luan et al., 2015).

## 1.2 Non-coding RNAs (ncRNA)

Ribonucleic acid (RNA) molecules are chains of nucleotides composed of a ribose sugar, a phosphate group and one of four bases: adenine (A), cytosine (C), guanine (G), or uracil (U). A non-coding RNA is an RNA molecule that is not translated into a protein. Compared to DNA, some RNA molecules called ribonucleic acid enzymes or ribozymes can achieve chemical catalysis akin to proteins (Kruger et al., 1982; Guerrier-Takada et al., 1983) and this ability was suggested to be crucial for the origin of life (Pace and Marsh, 1985; Sharp, 1985) (the RNA world hypothesis; Gilbert, 1986).

The most abundant ncRNAs in a bacterial cell are ribosomal RNAs (rRNA) and transfer RNAs (tRNA). These RNAs are involved in translation, *i.e.* synthesis of polypeptides (proteins) using messenger RNA (mRNA) as a template. First, transfer tRNAs are charged with an appropriate amino acid by enzymes called amino-acyl tRNA synthetases. Second, a complex of rRNA and proteins (ribosome) enables interactions of codons (three bases long pieces of mRNA) with anti-codons (corresponding three bases of tRNA) and addition of an amino-acid from the charged tRNA to a synthesized polypeptide chain. A single amino-acid may be coded for by more than one codon (the genetic code is redundant/degenerate) and three codons (stop codons) are specific stop signals causing end of translation instead of amino-acid addition.

### 1.2.1 Posttranscriptional ncRNA modifications

Base modifications occur in all types of ncRNAs. Presence of several modifications in all domains of life suggests their very ancient origin. Mitochondrial genomes lacks genes for RNA modification enzymes and all corresponding proteins are of nuclear origin (Motorin and Grosjean, 2005). I did not manage to find any study concerning RNA modifications in plastid genomes. The largest known plastid genomes contain approximately two and half times more genes (250 in red algae; Janouškovec et al., 2013; Smith and Keeling, 2015) than the largest

mitochondrial genomes (100 in the jakobid *Andalucia godoyi*; Burger et al., 2013; Smith and Keeling, 2015). In theory, these could contain some of genes encoding modification proteins.

There are about 100 types of modifications (Helm, 2006). According to their mechanism modifications can be divided into three separate groups: (i) simple or multiple additions, removal or substitution of chemical groups at one atom of a given base adenine (*e.g.*: methylations, thiolation, reduction, deamination, complex group additions), (ii) exchange of an encoded base by another base (*e.g.*: exchange of guanine by quenosine (Q), formation of pseudouridine ($\psi$), (iii) formation of 2'-O-derivatives of ribose (Motorin and Grosjean, 2005).

How much pariticular enzymes contribute to maintainig translation is not known. However, knockout of some genes encoding RNA modifying enzymes in the bacterium *E. coli* is lethal (*tadA*, *tilS*, *tsaC*) lethal or slows down growth and cell division. (Wolf, 2002; Soma et al., 2003; Baba et al., 2006; Yacoubi et al., 2009).

### Transfer RNA (tRNA) modifications

The largest number of modified nucleotides is found in tRNA molecules. On the other hand, tRNA molecules from organisms with reduced genomes such as *Mycoplasma* spp. and cellular organelles (mitochondria and chloroplasts) are much less modified (both in number and variety of modified residues) compared to complex organisms (Stanbridge and Reff, 1979; Motorin and Grosjean, 2005).

In all organisms, correct function of tRNA depends on its correct folding to 2D and 3D structure. Despite correct folding of most unmodified tRNA transcripts, modified nucleotides make tRNA more rigid and resistant to thermal stress and endonucleolytic degradation by reinforcing hydrogen bonds, improving base stacking, altering base flexibility, creating more binding sites for metal ions, and increasing hydrophobicity (Motorin and Grosjean, 2005).

The most frequently modified tRNA nucleosides consist of simple modifications,

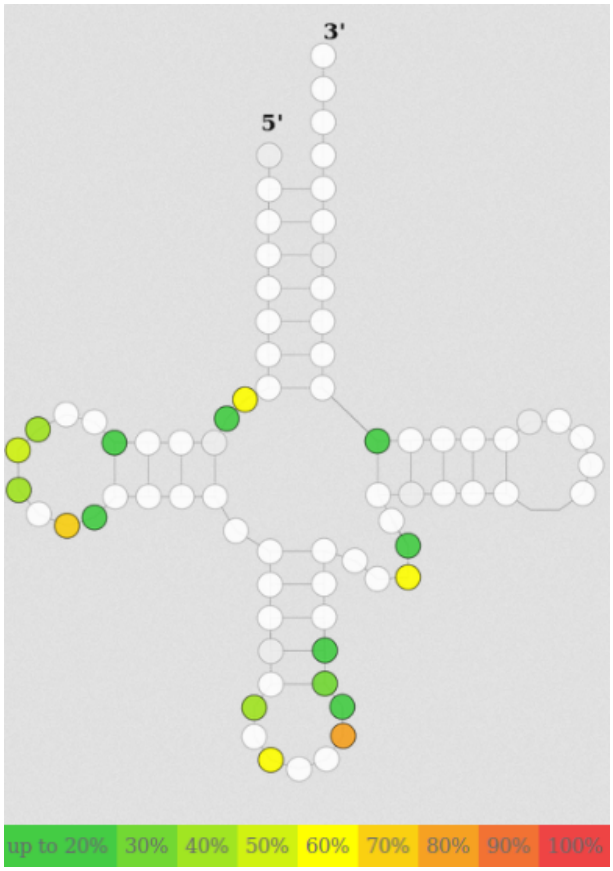


Figure 2: tRNA modification sites across tRNAs of *E. coli*. Color-coding of a site shows in how many tRNA kinds is the site modified. Grey: the site is not modified, dark green: modifications in up to 20% cases, yellow: 60%, red: 100%. Reprinted from the Modomics database [http://modomics.genesilico.pl/].

however, hyper-modified nucleotides often occur at first and last anti-codon bases. Many of them are specific for particular positions in tRNAs or even a particular tRNA (see Figure 2). These nucleotides improve efficiency and accuracy of anti-codon decoding during translation (Motorin and Grosjean, 2005).

**Ribosomal RNA (rRNA) modifications**

Modifications of rRNA stabilize its structure by the same mechanisms as in tRNA. Modifications most often occur in conserved regions of both ribosomal rRNAs (16S and 23S rRNAs) implying that they likely play a role in ribosome structure and function. rRNA regions composing parts of subunits important for specific translational events are modified the most. These are A, P and E sites on SSU (composed of 16S rRNA) and active site and tunnel on LSU (composed of 23S rRNA) providing 75% of modifications of this subunit in *E. coli*. Modifications are absent from SRL (factor-binding site that is essential for GTP-catalysed steps in translation), the lower end of polypeptide exit tunnel and areas dominated by proteins indicating most RNA proteins interactions are not affected by modifications (Decatur and Fournier, 2002; Lancaster et al., 2008).

Not a single rRNA modification has been confirmed as essential for ribosome function (but they can be essential for other reasons) implying most of modifications contribute small benefits resulting in a large one for the whole ribosome (Decatur and Fournier, 2002).

### 1.2.2 Translation, ncRNAs, and ncRNA modifications in insect endosymbionts

Studies describing translation, ncRNAs or RNA modifications in endosymbiotic bacteria are extremely rare despite the fact that ncRNAs constitute one quater of *Mycoplasma pneumoniae* (a parasitic bacterium with a tiny genome; Lluch-Senar et al., 2015) suggesting they could play an important role in bacteria with tiny genomes. According to a study by de Crécy-Lagard, Marck, and Grosjean (2012) with limited taxon sampling of endosymbionts (*Wolbachia*, *Buchnera*, *Blochmannia*, *Baumannia cicadellinicola*, *Wigglesworthia glossinidia* and *Riesia pediculicola*), these symbionts encode a set of tRNAs able to read all sense codons for the 20 canonical amino acids indicating no tRNAs from the host or a co-symbiont are needed. The total number of tRNAs per species varied from 31-32 for *Buchnera aphidicola* str. Cc to 40 for *Blochmannia pennsylvanicus*. However, endosymbionts with the tiniest genomes (*e.g. Carsonella ruddi* and *Tremblaya princeps*) were not included in this study. Several genes encoding tRNA modifying enzymes (*cmoA*, *cmoB*, *dusB*, *miaA*, *miaB*, *mnmA*, *mnmB*, *tadA*, *thilS*, *tilS*, *trmB*, *trmD*, *truA*, *truB*, and *tsaD*) were analyzed using BLAST searches and, interestingly, *R. pediculicola* was found to be missing genes for all enzymes modifying the tRNA body and retaining only a few modifications of the anticodon loop and proximal stem

(de Crécy-Lagard et al., 2012).

In other studies, transcription of all 32 tRNAs and presence of some modifications in *Buchnera aphidicola* (by MiaA (EC 2.5.1.75) and MiaB (EC 2.8.4.3); TrmD (EC 2.1.1.228); TadA (EC 3.5.4.-); MnmA(EC 2.8.1.13), MnmE, MnmG (GidA) and IscS (EC 2.8.1.7)) was confirmed by RNA-seq (Hansen and Moran, 2013). Modified tRNA nucleotides were confirmed also in *Hodgkinia*. Interestingly, its tRNAs contained cca ends not encoded by the sequences even though it lacks the CCA-adding enzyme (van Leuven and McCutcheon 2016, personal communication).

Interestingly, there are two cases of HGT involving RNA modifying genes in insects with obligate symbionts. *Planococcus citri* mealybug, the host of *Tremblaya princeps* and *Moranella endobia* (PCIT lineage), has acquired bacterial *rlmI* gene encoding 23S rRNA methyltransferase (Husník et al., 2013) and *Pachypsylla venusta* psyllid, the host of *Carsonella ruddii* PV, has acquired bacterial *rsmJ* gene encoding 16S rRNA methyltransferase (Sloan et al., 2014).

Apart from ncRNA modifications, presence of ncRNAs itselves was studied showing that families of ncRNAs which are usually conserved among free-living organisms can be lost from endosymbiotic or intracellular pathogenic bacteria. The loss is independent in particular lineages. Remarkably cis-regulatory ncRNAs localized next to the 5' end of protein-coding genes were lost, no matter whether their correponding proteins were conserved or not (Matelska et al., 2016). Matelska et al. showed that these results are not an artefact of unusual features of their genomes such as low GC content or small deletions.

Except a few pioneer studies mostly with limited sampling discussed above, genomic articles about insect symbionts generally take into account only protein-coding genes. However, ncRNAs represent the most abundant transcripts in bacterial cells and have an unquestionable role in translation. As translation is one of the few cellular processes present in all endosymbionts and is in most cases likely carried out solely by endosymbiont-coded proteins and RNAs, it provides an ideal target for a comparative approach. Here I address this omission and analyze a broad genome sampling of insect endosymbionts for presence of ncRNAs and their modifications. Results of this study can not only provide data on RNA biology of insect endosymbionts, but also give us an insight into what happens with ncRNAs and their modifications when an organism is undergoing extreme genome reduction similarly to what happened to cellular organelles, mitochondria and plastids.

# 2 Materials and methods

All analyses were done *in silico*. Free software, modified Perl scripts, my own zsh (shell) scripts and R (version 3.2.3; R Development Core Team, 2015) scripts were used. 63 genomes of endosymbiotic bacteria were used in this analysis (Table 1). For 17 of these, plasmid sequences were included (Table 5). Majority of them were downloaded from GenBank (Benson, 2004). Remaining ones, provided by my supervisor, are published but not available at GenBank yet (Nováková et al., 2015; Husnik and McCutcheon, 2016). Files of fna and faa format were created from these files using modified gbk2faa (Cai, a) and gbk2faa (Cai, b) (Cai, b) Perl scripts slightly modified by myself. *Escherichia coli* (NC_000913.3) and *Bacteroides fragilis* YCH46 (NC_006347.1) were used as free-living model organisms.

| | Lineage | Genome ID | Size (kb) | P S | Host |
|---|---|---|---|---|---|
| 1 | *Arsenophonus melophagi* | None | 1155 | P | *Melophagus ovinus* |
| 2 | *Baumannia cicadellinicola* | NC 007984 | 686 | S | *Homalodisca coagulata* |
| 3 | *Blattabacterium BGIGA* | NC 017924 | 629 | P | *Blaberus giganteus* |
| 4 | *Blattabacterium BNCIN* | NC 022550 | 623 | P | *Nauphoeta cinerea* |
| 5 | *Blattabacterium BPAA* | NC 020510 | 632 | P | *Panesthia angustipennis spadica* |
| 6 | *Blattabacterium BPLAN* | NC 013418 | 637 | P | *Periplaneta americana* |
| 7 | *Blattabacterium Cpu* | NC 016621 | 606 | P | *Cryptocercus punctulatus* |
| 8 | *Blattabacterium cuenoti Bge* | NC 013454 | 637 | P | *Blattela germanica* |
| 9 | *Blattabacterium cuenoti Tarazona* | NC 020195 | 634 | P | *Blatta orientalis* |
| 10 | *Blattabacterium MADAR* | NC 016146 | 587 | P | *Mastotermes darwiniensis* |
| 11 | *Blochmannia chromaiodes* | NC 020075 | 791 | P | *Camponotus chromaiodes* |
| 12 | *Blochmannia floridanus* | NC 005061 | 706 | P | *Camponotus floridanus* |
| 13 | *Blochmannia pennsylvanicus* | NC 007292 | 792 | P | *Camponotus pennsylvanicus* |
| 14 | *Blochmannia vafer* | NC 014909 | 723 | P | *Camponotus vafer* |
| 15 | *Buchnera aphidicola AK* | NC 017256 | 642 | P | *Acyrthosiphon kondoi* |
| 16 | *Buchnera aphidicola Bp* | NC 004545 | 616 | P | *Baizongia pistaciae* |
| 17 | *Buchnera aphidicola Cc* | NC 008513 | 416 | P | *Cinara cedri* |
| 18 | *Buchnera aphidicola Sg* | NC 004061 | 641 | P | *Schizaphis graminum* |
| 19 | *Buchnera aphidicola Ua* | NC 017259 | 615 | P | *Uroleucon ambrosiae* |
| 20 | *Carsonella ruddii CE* | NC 018414 | 163 | P | *Ctenarytaina eucalypti* |
| 21 | *Carsonella ruddii CS* | NC 018415 | 163 | P | *Ctenarytaina spatulata* |
| 22 | *Carsonella ruddii DC* | NC 021894 | 174 | P | *Diaphorina citri* |
| 23 | *Carsonella ruddii HC* | NC 018416 | 166 | P | *Heteropsylla cubana* |
| 24 | *Carsonella ruddii HT* | NC 018417 | 158 | P | *Heteropsylla texana* |
| 25 | *Carsonella ruddii PC* | NC 018418 | 160 | P | *Pachypyslla celtidis* |
| 26 | *Carsonella ruddii PV* | NC 008512 | 160 | P | *Pachypsylla venusta* |
| 44 | *Doolittlea endobia* | None | 835 | S | *Maconellicoccus hirsutus* |
| 42 | *Gullanella endobia* | None | 938 | S | *Ferrisia virgata* |
| 27 | *Hodgkinia cicadicola* | NC 012960 | 144 | S | *Diceroprocta semicincta* |
| 28 | *Ishikawaella capsulata* | AP010872 | 746 | P | *Megacopta punctatissima* |
| 45 | *Mikella endobia* | None | 353 | S | *Paracoccus marginatus* |
| 29 | *Moranella endobia PCIT* | NC 015735 | 538 | S | *Planococcus citri 1* |
| 30 | *Moranella endobia PCVAL* | NC 021057 | 538 | S | *Planococcus citri 2* |
| 31 | *Nasuia deltocephalinicola* | NC 021919 | 112 | S | *Macrosteles quadrilineatus* |
| 32 | *Portiera aleyrodidarum BT-B-HRs var1* | NC 018507 | 358 | P | *Bemisia tabaci* |
| 33 | *Portiera aleyrodidarum BT-B-HRs var2* | NC 018677 | 352 | P | *Bemisia tabaci* |

| | Lineage | Genome ID | Size (kb) | P S | Host |
|---|---|---|---|---|---|
| 34 | *Portiera aleyrodidarum BT-QVLC var1* | NC 018618 | 357 | P | *Bemisia tabaci* |
| 35 | *Portiera aleyrodidarum BT-QVLC var2* | NC 018676 | 351 | P | *Bemisia tabaci* |
| 36 | *Portiera aleyrodidarum TV* | NC 020831 | 281 | P | *Trialeurodes vaporariorum* |
| 37 | *Profftella armatura* | NC 021885 | 459 | S | *Diaphorina citri* |
| 38 | *Riesia pediculicola* | NC 014109 | 574 | P | *Pediculus humanus humanus* |
| 39 | *Serratia symbiotica* | NC 016632 | 1763 | S | *Cinara cedri* |
| 40 | *Sodalis pierantonius* | CP006568 | 4513 | P | *Sitophilus oryzae* |
| 41 | *SS Ctenarytaina eucalypti* | NC 018419 | 1441 | S | *Ctenarytaina eucalypti* |
| 43 | *SS Heteropsylla cubana* | NC 018420 | 1122 | S | *Heteropsylla cubana* |
| 46 | *Sulcia muelleri ALF* | NC 021916 | 191 | P | *Macrosteles quadrilineatus* |
| 47 | *Sulcia muelleri CARI* | NC 014499 | 277 | P | *Clastoptera arizonana* |
| 48 | *Sulcia muelleri DMIN* | NC 014004 | 244 | P | *Draeculacephala minerva* |
| 49 | *Sulcia muelleri GWSS* | NC 010118 | 246 | P | *Homalodisca coagulata* |
| 50 | *Sulcia muelleri SMDSEM* | NC 013123 | 277 | P | *Diceroprocta semicincta* |
| 51 | *Tremblaya phenacola* | NC 021555 | 171 | P | *Phenacoccus avenae* |
| 52 | *Tremblaya princeps PCIT* | NC 015736 | 139 | P | *Planococcus citri 1* |
| 53 | *Tremblaya princeps PCVAL* | NC 017293 | 139 | P | *Planococcus citri 2* |
| 54 | *Tremblaya princeps TPFVIR* | None | 142 | P | *Ferrisia virgata* |
| 55 | *Tremblaya princeps TPMHIR* | None | 138 | P | *Maconellicoccus hirsutus* |
| 56 | *Tremblaya princeps TPPLON* | None | 144 | P | *Pseudococcus longispinus* |
| 57 | *Tremblaya princeps TPPMAR* | None | 140 | P | *Paracoccus marginatus* |
| 58 | *Uzinura diaspidicola* | NC 020135 | 263 | P | *Aspidiotus nerii* |
| 59 | *Walczuchella monophlebidarum* | NZ CP006873.1 | 309 | P | *Llaveia axin axin* |
| 61 | *Westeberhardia cardiocondylae* | LN774881 | 533 | P | *Cardiocondyla obscurior* |
| 60 | *Wigglesworthia glossinidia GB* | NC 004344 | 698 | P | *Glossina brevipalpis* |
| 62 | *Wigglesworthia glossinidia GM* | NC 016893 | 720 | P | *Glossina morsitans morsitans* |
| 63 | *Zinderia insecticola* | NC 014497 | 209 | S | *Clastoptera arizonana* |

Table 1: Information about species used in this study; Genome ID = genome accession number from GenBank, Size = genome size in kilobase pairs, PS = obligate (primary - P) or co-obligate (secondary - S) symbiont

## 2.1 Searching for tRNAs and rRNAs

tRNAs were found using a script tFind (Hudson and Williams, 2014) in chromosomal genomes of endosymbionts. This script combines two different RNA finders: tRNA scan-SE (version 1.3.1; Lowe and Eddy, 1997) and Aragorn (version 1.2.36; Laslett, 2004). Algorithms of these finders differ and therefore provide different results (Ardell, 2009). Script run_tFind.sh was created to run tFind, convert data into a tabular format and prepare tRNA sequences for alignment in Infernal.

Genes for rRNAs were found according to genome annotations using genome browser Artemis (Rutherford et al., 2000). Some of genes found in this way probably can have their start and end positions, but this did not influence the result because no modification sites are close neither to the beginning nor to the end of the sequence.

## 2.2 Identifying modification sites in RNA sequences

Ribosomal DNA sequences (rDNA) of endosymbionts were aligned together with rDNAs of a model organism *E. coli* MG1655 using the Geneious alignment algorithm (Geneious version 7.1.9; Kearse et al., 2012) with automatic determination of sequence direction (other parameters were default). Modification sites were identified in Geneious according to sequence of *E. coli* which was set there as a reference sequence.

From tRNAs found in the way described above were excluded ones which were detected by one of finders only and ones which were marked as possible pseudogenes by the script. Remaining tRNAs were aligned using Infernal (Nawrocki and Eddy, 2013) according to Rfam (Nawrocki et al., 2014) database of all bacterial tRNAs. All alignments were converted from .sto format to fna format using perl script (Katz, 2012). These two steps were done using run_cmalign.sh script.

Further analysis using all tRNAs including those which were excluded for the previous step was done in R script tRNA_positions7.R: Only enzymes for which the modification site(s) and kind(s) of tRNA are known (and listed in Modomics database) were used in this part of the study. This information was used to identify modification sites in tRNAs of endosymbionts according to sequences of *E. coli*. Identification could be done only in those kinds of tRNA which were present in *E. coli* and at least one endosymbiont. Positions with a wrong base for *E. coli* sequences and some bases related to these were removed.

## 2.3 Searching for genes related to ncRNAs

83 RNA modification genes related to ncRNA were selected for this analysis using Modomics (Dunin-Horkawicz, 2006; Czerwoniec et al., 2009; Machnicka et al., 2012) and EcoCyc (Keseler et al., 2012) databases. To evaluate whether endosymbiont genomes (both chromosomal and plasmid) contain these genes, a zsh script run_orthomcl_blastp_FINAL.sh combining following methods was written: i) BlastP (an algorithm searching protein databases Altschul et al., 1990) search of *E. coli* proteins (sequences from EcoCyc or if not available from genome of *E. coli* MG1655) against a proteome database of all endosymbiotic bacteria, and ii) identification of clusters of orthologous genes from all proteomes (including *E. coli*) by OrthoMCL program (Li, 2003). OrthoMCL clusters did not include *E. coli* protein sequences in most cases on the other hand they were more precise than simple BlastP.

The rest of clusters identification was done via R script proteins_cluster_identification4.R. For each blast run I assigned corresponding cluster ID. According to this, average score and E-value for each cluster within a blast run were computed. According to these values and possible presence of *E. coli* sequence in the cluster, one or no cluster were assigned to each gene. For genes with assigned clusters which did not contain any species from Bacteroidetes

group and for ones which seemed to be lost in all endosymbionts included, it was searched for a corresponding gene in *Bacteroides fragilis* using gene annotation and other gene names found at EcoCyc. The procedure described above was repeated with these newly identified genes.

In the next step, additional genes of endosymbionts were identified using Hidden Markov Models (a type of probabilistic models; HMMER software; Sean R. Eddy and the HMMER development team, 2013). For this step a zsh script run_hmm.sh was created. For creation of HMM model, mafft alignment (Katoh, 2014) with default parameters was used. Alignments were visually checked in Geneious (version 7.1.9; Kearse et al., 2012). New genes were selected using another part of the R script proteins_cluster_identification4.R according to following criteria: genes having E-value lower than $1e^{-30}$ could be accepted according to their annotations or blast results against nr (non-redundant database; contains all sequences from ncbi, but duplicates are removed). Genes having E-value lower then $1e^{-75}$ were accepted unless their annotation did not correspond to annotations of original cluster members. In these cases blast search against nr was used to evaluate this hit. In case the E-value was between $1e^{-70}$ and $1e^{-30}$ and gene annotation did not disagree with corresponding *E. coli* gene annotation. No gene from any of those categories could be accepted if is already was identified another gene encoding a RNA modification enzyme. For this part R script using package pander (Gergely Daróczi, 2015) was used.

## 2.4 Statistics and data visualization

Statistical analyses were done in R. Statistical analysis of tRNA counts can be found in the R script Statistics_tRNA.R and of modification proteins in the Statistics_proteins.R. For statistical analyses data averaged for particular genera were used unless specified otherwise. For total counts of tRNA and modification proteins in genera and average tRNA score of particular genera basic statistical parameters (mean, minimum, first quantile, median, third quantile and maximum) were computed. Following statistical tests were performed: i) linear models (Question: is one variable linearly dependent on another one?; R function: lm), ii) F-test (Question: do variances of two groups differ significantly?, R function: var.test; t-test (Question: do means of two groups differ significantly?, R function: t-test (method="spearman")).

Additionaly, for protein presence only, differences among particular species or genera were computed such that for each possible pair of lineages within particular species (or genus) rows of table with values 1 and 0 were substracted. Absolute values of these two differences were summed together. An average was computed from these sums for particular pairs.

This document was written in knitr (Xie, 2015, 2014, 2016), an R package enabling incoporation of R scripts into latex. Pdflatex was used for compilation. Visualization of results

was done also in R. All functions used for visualization are included in code chunks of the source Rnw file of this thesis. R packages xtable (Dahl, 2016) and fields (Nychka et al., 2016) were used.

# 3 Results

| | Lineage | Lineage count | PS | Genome size | tRNAs count | tRNA score | Protein count |
|---|---|---|---|---|---|---|---|
| 1 | *Arsenophonus* | 1 | P | 1155 | 34.00 | 79.65 | 33.00 |
| 2 | *Baumannia* | 1 | S | 686 | 39.00 | 73.54 | 32.00 |
| 3 | *Blattabacterium* | 8 | P | 623 | 33.50 | 69.33 | 27.38 |
| 4 | *Blochmannia* | 4 | P | 753 | 38.50 | 68.23 | 27.75 |
| 5 | *Buchnera* | 5 | P | 586 | 31.80 | 72.78 | 35.20 |
| 6 | *Carsonella* | 7 | P | 163 | 27.86 | 55.38 | 3.00 |
| 7 | *Doolittlea* | 1 | S | 835 | 40.00 | 76.50 | 35.00 |
| 8 | *Gullanella* | 1 | S | 938 | 40.00 | 73.95 | 36.00 |
| 9 | *Hodgkinia* | 1 | S | 144 | 15.00 | 53.49 | 3.00 |
| 10 | *Ishikawaella* | 1 | P | 746 | 37.00 | 73.99 | 33.00 |
| 11 | *Mikella* | 1 | S | 353 | 41.00 | 74.17 | 32.00 |
| 12 | *Moranella* | 2 | S | 538 | 41.00 | 78.02 | 31.50 |
| 13 | *Nasuia* | 1 | S | 112 | 30.00 | 54.83 | 6.00 |
| 14 | *Portiera* | 5 | P | 340 | 33.20 | 64.94 | 11.20 |
| 15 | *Profftella* | 1 | S | 459 | 34.00 | 71.10 | 28.00 |
| 16 | *Riesia* | 1 | P | 574 | 33.00 | 66.88 | 24.00 |
| 17 | *Serratia* | 1 | S | 1763 | 36.00 | 76.21 | 41.00 |
| 18 | *Sodalis* | 1 | P | 4513 | 53.00 | 79.58 | 51.00 |
| 19 | *SS Ctenarytaina* | 1 | S | 1441 | 40.00 | 79.40 | 34.00 |
| 20 | *SS Heteropsylla* | 1 | S | 1122 | 38.00 | 74.42 | 34.00 |
| 21 | *Sulcia* | 5 | P | 247 | 30.00 | 68.08 | 13.00 |
| 22 | *Tremblaya* | 8 | P | 146 | 18.50 | 49.45 | 4.50 |
| 23 | *Uzinura* | 1 | P | 263 | 31.00 | 62.88 | 19.00 |
| 24 | *Walczuchella* | 1 | P | 309 | 34.00 | 68.53 | 19.00 |
| 25 | *Westeberhardia* | 1 | P | 533 | 35.00 | 65.40 | 27.00 |
| 26 | *Wigglesworthia* | 2 | P | 709 | 34.50 | 68.62 | 25.00 |
| 27 | *Zinderia* | 1 | S | 209 | 25.00 | 63.58 | 9.00 |

Table 2: Summary of total tRNA and genes encoding modification enzymes counts concerning data averaged for genera: i) genera, ii) obligate (primary - P) or co-obligate (secondary - S) endosymbiont, ii) genome size in kb, iii) total tRNAs count, iv) average tRNA scores (from tFind script), v) total modification protein counts (averages per genera).

## 3.1 Counts of tRNAs

Average number of tRNAs for genera of endosymbiotic bacteria is 34.22 +- 2.91 (95% confidence interval) with average score 69+-3.27 (95% confidence interval). Median is close to mean in both cases (Table 3). Bacterium with the lowest number of tRNAs is *Tremblaya princeps* PCIT (12) and the one with highest is *Sodalis pierantonius* (53). Eleven species lost all types of tRNAs for at least one amino acid (Figure 3, Table 2).

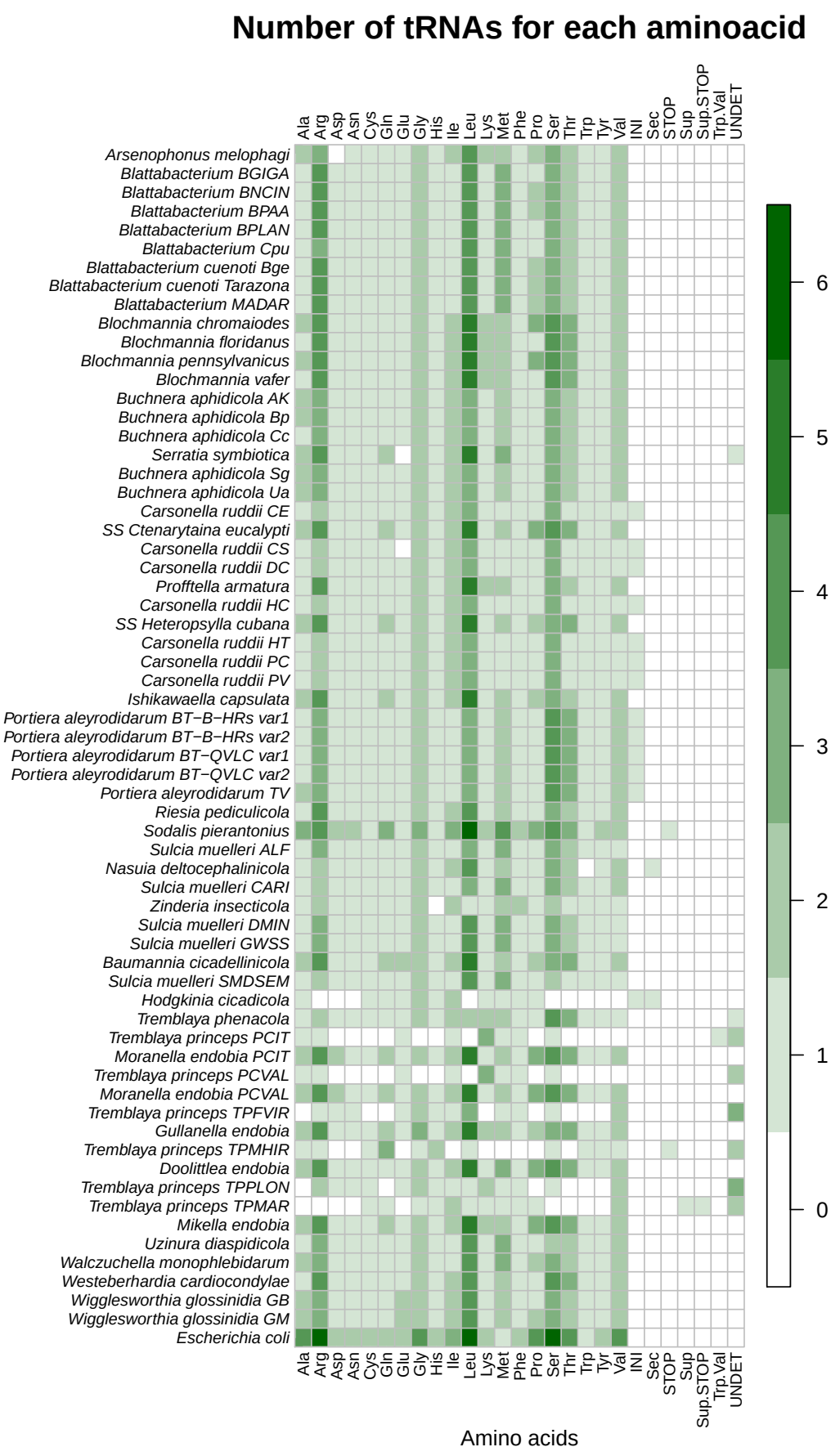


Figure 3: Counts of tRNAs encoding particular amino acids in particular bacterial species. Species are represented by rows and amino-acids by columns. Count of tRNA for each combination of row and column is represented by color, the darker, the more tRNAs there are. Lineages are sorted such that co-obligate symbionts are right below their obligate co-symbionts.

| | Mean | Min | Qu1 | Median | Qu3 | Max | SD |
|---|---|---|---|---|---|---|---|
| Genome size | 750.37 | 112.00 | 286.00 | 574.00 | 794.00 | 4513.00 | 856.02 |
| tRNAs count | 34.22 | 15.00 | 31.40 | 34.00 | 38.75 | 53.00 | 7.36 |
| tRNA score | 69.00 | 49.45 | 65.17 | 69.33 | 74.30 | 79.65 | 8.27 |
| Proteins count | 24.98 | 3.00 | 16.00 | 24.98 | 33.50 | 51.00 | 12.60 |

Table 3: Mean, minimum, first quantile, median, third quantile, maximum for genome sizes, tRNA counts, tRNA scores and protein counts. Averaged data shown in previous table were used.

Variances tRNA counts of obligate and co-obligate symbiont genera (F(14, 11) = 0.78632, p-value = 0.6614) and tRNA scores (F(14, 11) = 0.82235, p-value = 0.7185) do not differ significantly. Means for the same categories also do not differ significantly (tRNA counts t = -0.42886, df = 22.268, p-value = 0.6721; tRNA scores: t = -0.98334, df = 22.562, p-value = 0.3359). tRNA average score of tRNAs with the highest scores per genus is 90.35 and for ones with the lowest scores 42.45. (The minimal tRNA score per genus was estimated such that average from minimal scores from all species belonging to the genus was computed. The tRNAs with maximal score per genera were computed analogically.)

According to plot of linear regression neither dependency of mean counts of genera size on genome size nor dependency of mean tRNA scores are normally distributed. Fitted hits are also strongly influenced and variances of residuals are not homogeneous. After log transformations of genome sizes (the independent variable), all these three violations are much smaller. This variant was therefore used for the analysis of linear regression. There is a linear relationship between log of genome size (independent variable) and both tRNAs counts and tRNA scores (Figure 4; tRNA counts: Residual standard error: 0.5163 on 25 degrees of freedom Multiple R-squared: 0.648, Adjusted R-squared: 0.6339, F-statistic: 46.02 on 1 and 25 DF, p-value: 4.136e$^{-07}$; tRNA scores: Residual

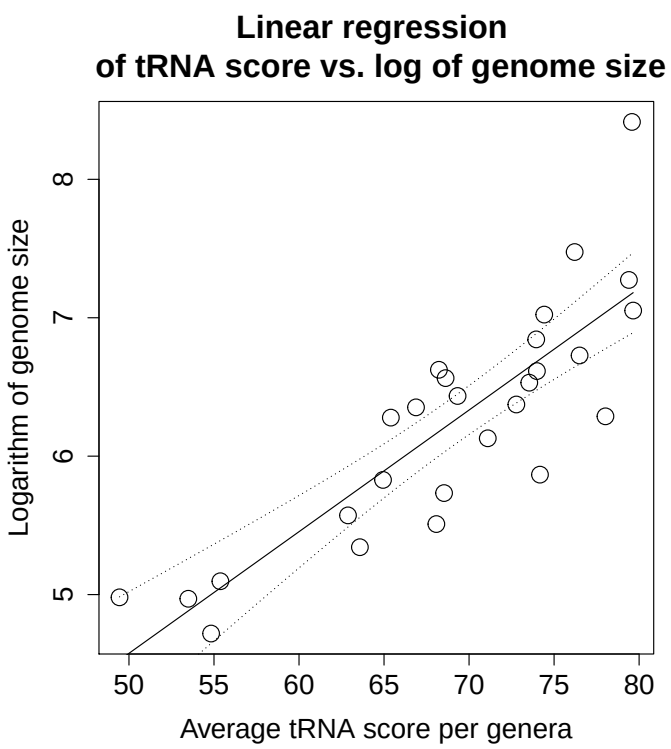


Figure 4: Linear regression of genome size, independent variable, and tRNA score, dependent variable. Average values for particular genera are used here.

standard error: 0.4553 on 25 degrees of freedom, Multiple R-squared: 0.7262, Adjusted R-squared: 0.7152, F-statistic: 66.31 on 1 and 25 DF, p-value: $1.699e^{-08}$).

## 3.2 RNA modification proteins

60 out of 84 genes encoding RNA modifications were found in at least one endosymbiont. None of these genes was found neither in all endosymbiotic species nor in all systems containing one obligate symbiont and sometimes also a co-obligate symbiont. Bacteria have conserved from 2 (*Tremblaya princeps* TPPMAR) to 51 genes (*Sodalis pierantonius*, Figure 6). OrthoMCL cluster which was identified as *tsaE* contains two sequences of *Riesia pediculicola* which is caused by misassembly of this genome (Filip Husník, personal communication). In one particular case it was not possible to distinguish whether genes in the cluster should be assigned to *rluB* or *rluF*. To this cluster of paralogous genes is refered as "RluB_RluF". None of genes for RNA modification enzymes was found in any of plasmid sequences.

In the annotation of *Bacteroides fragilis* genome, it was searched for 24 genes which were not assigned to a cluster and other 20 genes with a cluster containing no species from the Bacteroidetes phylum. After searching for these genes in the *Bacteroides fragilis* genome, only two corresponding genes were successfully identified. No cluster was assigned to these genes. Compared to the list of modification proteins, one more modification protein similar to MiaB was detected Blast search uncovered that it was MtaB protein, which catalyzes the transfer of a methylthiol group to a hyper-modified base 37 of tRNA.

In the rest of this section data averages for genera are considered. Average number of RNA modification genes is 24.95 +- 4.97 (95% confidence interval). Minimum number of genes was conserved in *Carsonella* and *Hodgkinia* genera (3) and maximum in *Sodalis* genus (51).

Five most abundant genes were, from the most abundant: *mnmG*, *mnmE*, *mnmA*, *rsmH*, *rluD*. An average gene would be conserved in 8.02 +- 1.83 (95% confidence interval) genera or in 11.23 +- 2.07 (95% confidence interval) genera if genes lost from all genomes are omitted. From the histogram we can see that genes are most frequently lost from all symbionts. The rest of histogram bars do not differ in their heights much and are neither decreasing nor increasing. From the histogram we can see that genera have conserved from 30 to 35 genes in most cases. There are only two genera with more than 40 genes conserved and counts of genera conserved between 0 and 25 genes are similar (Figure 5).

Neither variances (F(14, 11) = 1.6589, p-value = 0.4039) nor means (t = 0.36896, df = 24.989, p-value = 0.7153) of modification gene count between obligate versus co-obligate endosymbionts differ significantly. There is no linear relationship between count of conserved genes (dependent variable) and genome size (independent variable; Residual standard

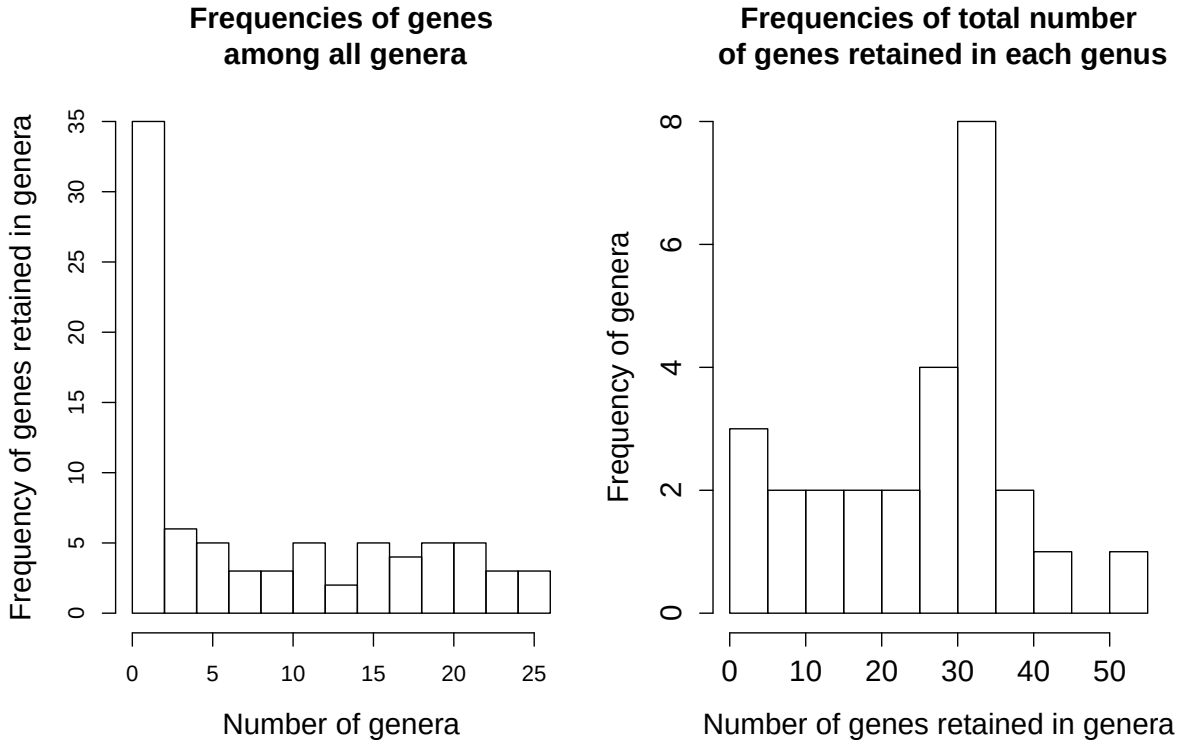


Figure 5: Left - a histogram showing frequencies of particular genes among all genera; right - a histogram showing frequencies of total numbers of genes retained in particular genera

error: 0.8704 on 25 degrees of freedom, Multiple R-squared: 0.000193, Adjusted R-squared: -0.0398, F-statistic: 0.004826 on 1 and 25 DF, p-value: 0.9452).

There is also no significant difference between variances and means of modification counts neither between tRNA modifications and rRNA modifications (variances: F = 0.84057, num df = 29, denom df = 40, p-value = 0.632; means: t = -1.1181, df = 65.487, p-value = 0.2676) nor between modifications at small ribosomal subunit and large ribosomal subunit (variances: F = 1.1228, num df = 15, denom df = 13, p-value = 0.8422; means: t = -0.44404, df = 27.82, p-value = 0.6604). The same holds when only genes present in at least one endosymbiont are included (variance rRNA vs. tRNA: F = 1.0228, num df = 19, denom df = 28, p-value = 0.9362; mean rRNA vs. tRNA: t = -1.0952, df = 40.704, p-value = 0.2799; variance large vs. small subunit: F = 1.3809, num df = 9, denom df = 9, p-value = 0.6384; mean large vs. small subunit: t = -0.20055, df = 17.551, p-value = 0.8434).

The difference in mean and variances of conservation level of genes encoding modifcation enzymes providing modifications specific for bacteria and non-specific ones are also not singnificant (Table 6; variances: F = 1.3242, num df = 20, denom df = 39, p-value = 0.4429, means: t = 1.1022, df = 36.093, p-value = 0.2777).

Two average endosymbionts genera differ in 19.52 modification proteins present (computed from averaged data of genera). Two lineages of the same genus differ in 2.68 genes.

| | Species or genera | Lineage counts | tRNAs | Proteins |
|---|---|---|---|---|
| 1 | *Blattabacterium* | 8 | 0.79 | 1.18 |
| 2 | *Blochmannia* | 4 | 2.00 | 1.50 |
| 3 | *Buchnera aphidicola* | 5 | 0.40 | 7.80 |
| 4 | *Carsonella ruddii* | 7 | 0.29 | 0.00 |
| 5 | *Moranella endobia* | 2 | 0.00 | 1.00 |
| 6 | *Portiera aleyrodidarum* | 5 | 0.40 | 2.60 |
| 7 | *Sulcia muelleri* | 5 | 2.40 | 4.40 |
| 8 | *Tremblaya princeps* | 7 | 13.27 | 1.33 |
| 9 | *Tremblaya* | 8 | 15.33 | 3.64 |
| 10 | *Wigglesworthia glossinidia* | 2 | 1.00 | 2.00 |

Table 4: In how many cases of presence/absence of proteins or tRNAs do two members of the same genus/species differ.

An average *Tremblaya princeps* differs from *T. phenacola* in 10.57 genes which is about eight times higher than do *Tremblaya princeps* lineages (1.33). Surprisingly, the average difference among *Buchnera aphidicola* lineages (7.80) is far higher than any other difference among lineages of a species or even species of a genus.

Figure 6: On the next page. Presence/absence of RNA modification considering relationships among endosymbionts. Only genes present in at least one species are included. All co-obligate symbionts and their obligate symbionts are included, from remaining symbionts two representatns of each genus are included. Color coding: white - gene is conserved neither in obligate nor in co-obligate symbiont, yellow - gene is conserved in obligate symbiont only, blue - gene is conserved in co-obligate symbiont only, green - gene is conserved in both obligate and co-obligate symbiont.


CmoA
CmoB
DusA
DusB
fmt
GluQRS
IscA
IscS
MiaA
MiaB
MtaB
MnmA
MnmE
MnmG
QueA
QueC
QueD
RlmB
RlmC
RlmE
RlmG
RlmH
RlmI
RlmKL
RlmN
RluB/RluF
RluC
RluD
RluE
RsmA
RsmB
RsmC
RsmD
RsmE
RsmG
RsmH
RsmI
RsmJ
RsuA
TadA
Tgt
ThiI
TmcA
TrmA
TrmB
TrmD
TrmH
TrmM
TruA
TruB
TsaB
TsaC
TsaD
TsaE
TtcA
TusA
TusB
TusC
TusD
TusE
Arsenophonus melophagi
Blattabacterium cuenoti Tarazona
Blattabacterium MADAR
Blochmannia chromaiodes
Blochmannia floridanus
Blochmannia pennsylvanicus
Blochmannia vafer
Buchnera aphidicola Cc + Serratia symbiotica
Buchnera aphidicola Sg
Buchnera aphidicola Ua
Carsonella ruddii CE + SS Ctenarytaina eucalypti
Carsonella ruddii DC + Profftella armatura
Carsonella ruddii HC + SS Heteropsylla cubana
Carsonella ruddii HT
Carsonella ruddii PC
Ishikawaella capsulata
Portiera aleyrodidarum BT−B−HRs1
Portiera aleyrodidarum TV
Riesia pediculicola
Sodalis pierantonius
Sulcia muelleri ALF + Nasuia deltocephalinicola
Sulcia muelleri CARI + Zinderia insecticola
Sulcia muelleri DMIN
Sulcia muelleri GWSS + Baumannia cicadellinicola
Sulcia muelleri SMDSEM + Hodgkinia cicadicola
Tremblaya phenacola
Tremblaya princeps PCIT + Moranella endobia PCIT
Tremblaya princeps PCVAL + Moranella endobia PCVAL
Tremblaya princeps TPFVIR + Gullanella endobia
Tremblaya princeps TPMHIR + Doolittlea endobia
Tremblaya princeps TPPLON
Tremblaya princeps TPPMAR + Mikella endobia
Tremblaya princeps TPTPER
Uzinura diaspidicola
Walczuchella monophlebidarum
Westeberhardia cardiocondylae
Wigglesworthia glossinidia GB
Wigglesworthia glossinidia GM

## 3.3 RNA modification sites

### 3.3.1 rRNA

25 modification sites of 19 enzymes were localized in the 23S rRNA composing large ribosomal subunit and 15 sites of 17 enzymes in 16S rRNA composing the small ribosomal subunit. In 23S rRNA, one position represeneted a modification site of two enzymes. I have detected a mistake in 23S rRNA compared to databases caused by a point insertion (for 20 sites after the 1500 position), the correct base in *E. coli* sequence was always on position which was one greater than the one listed in databases. In both subunits of rRNA, the most modification sites are conserved despite the fact that majority of other positions in rRNA sequence is not conserved. Generally, bases surrounding distinct modification sites were found to be conserved as well. All non-conserved modification sites constitute of two distinct bases. In the gene encoding 16S rRNA, at the site of RlmF (EC 2.1.1.181) thymine was substituted by adenine at rRNAs of *Hodgkinia* and *Tremblaya princeps* TPPLON and at RluC (EC 5.4.99.24) site is adenine replaced with guanine in *Nasuia* and *Hodgkinia*. In the gene encoding 16S rRNA, there are also two substitutions (guanine to adenine at RsmB site (EC 2.1.1.176) and cytidine to adenine at RsmD (EC 2.1.1.171) site modification). Interestingly, these two sites are right next to each other. Unlike the sites in 23S rRNA, they occur not only in endosymbionts with tiny genomes (all *Carsonella* lineages for RsmB *Nasuia* and all *Sulcia* lineages for RsmD), but also in others (*Baumania* for RsmB, all *Blattabacterium* lineages and *Uzinura* for RsmD and *Walczuchella* for both). In all single cases of substitutions, the corresponding modification protein was lost from all lineages with substituted sites. Moreover, one of them, RlmF, was not detected in any endosymbiont.

### 3.3.2 tRNA

Both sufficient information about their sites and the corresponding tRNA alignment were available for 39 single modifications. An incorrect base for *E. coli* was identified for seven modification sites. Five of them were modified by TruA (EC 5.4.99.12) and two were in tRNA for arginine tRNA with anticodon ACG modified by two distinct enzymes (RlmN (EC 2.1.1.192), TadA (3.5.4.-)). Therefore, all sites modified by TruA and all sites at tRNA Arg (ACG) were excluded from further analyses. For identification of TruAs (5.4.99.12) modification site, strain 294 of *E. coli* was used (Kammen et al., 1988). This strain might slightly differ from the strain MG1655 which was used in this study. Unfortunately, I did not manage to find any sequence of this strain in the NCBI database.

Only six out of 29 remaining sites were not conserved among all endosymbionts. Four of these sites were modified by the enzyme TrmB (EC 2.1.1.33, positions 46, base G) and remaining two by MiaB (EC 2.8.4.3) and RluA (EC 5.4.99.28 and EC 5.4.99.29) both at phenylalanin tRNA (anticodon GAA; positions: 37, 32; bases: A, U). No noticeable pattern was detected

among replaced base types (at one site U was replaced by C, at another C by U, at next one A by G, at other two sites vice versa and at remaining two G to both A and U). These six sites were not conserved minimally at 4.2% genera (which is one genus), maximally at 76% (19 genera) and on average at 26.5% genera. Median is 19.9%. (Data for genera were generated such that one representative for each genus was chosen.)

RluA, the enzyme with the modification site mutated in most lineages, was not detected in any of genomes. MiaB was not detected only in some of lineages which have not conserved also its modification site. Gene encoding TtcA is present in *Sodalis pierantonius* only and its site is conserved in this species. TrmB modification site was lost in four distinct tRNAs. In tRNA-Gly (GCC) and tRNA-Trp (CCA), the site was substituted in one species only (*Tremblaya princeps* TPMHIR) and the enzyme was lost from this species. Interestingly, in tRNA-Arg(CCG), where the site was lost from *Moranella endobia* and *Westeberhardia cardiocondylae*, the protein was conserved in both *Moranlela* lineages and only in nine other lineages. The last site was lost from four *Buchnera* lineages and *Nasuia*. Three of these four *Buchnera* lineages have retained the protein.

# 4 Discussion

Some of endosymbionts do not encode a full set of tRNAs implying they need to co-operate with their co-symbionts or eventually with their hosts. Genes encoding modification enzymes conserved in particular genomes differ to high extent supporting the hypothesis that each modification provides a small benefit which sums up to the total great effect. Therefore it is not that much important which single modifications are conserved within a genome. Most of tRNA and rRNA modification sites are conserved even in genomes lacking the gene corresponding to the modification enzyme which suggests that structural role of these sites is important even without being modified. My data imply that translation of endosymbionts is much less efficient[2] compared to their free-living relatives and that some of symbionts need cooperation with their co-symbionts or even with their host to maintain translation.

Average number of modification protein counts, tRNA counts and tRNA scores were computed for particular genera to fulfil the assumption of data independence. Lineages of one species are not independent, but on the other hand some of endosymbiont lineages within one genus (e.g.: *Tremblaya*, *Buchnera*; Lamelas et al., 2011; Husnik and McCutcheon, 2016) have undergone independent genome reduction in diverse host species. The same applies to tRNAs with exception of *Tremblaya*. Nevertheless, lineages within one genus are still much more similar to themselves than to other lineages (Table 4). Therefore I consider this correction necessary. It might be better to describe my data using general linear models with mixed effects (Jakub Těšitel, 2015, personal communication). Unfortunately, I have not managed to learn this advanced statistical topic yet. Another simple possibility how to deal with this issue would be to choose one sample of each lineage, but I have decided to use averages because it reduces random effect in data.

## 4.1 Most, but not all, endosymbionts retain a full set of tRNAs

Many of tRNAs have very low score implying pseudogenizations of these tRNAs. Unfortunately there is no good way to detect pseudogenes among ncRNAs because they do not have open reading frame (ORF) and we must rely on setting an arbitrary threshold of tRNA score. Moreover, some of tRNAs detected by computational methods were not present in RNAseq data from *Hodgkinia* (van Leuven and McCutcheon 2016, personal communication). It is therefore likely that some of tRNAs with low score are pseudogenized or even detected incorrectly, especially when predicted with a non-standard anticodon. Obligate symbionts which have been evolving under fast evolution accompanied with severe gene loss and genetic drift for much longer time compared to co-obligate symbionts retained less tRNAs with lower quality.

Majority of endosymbionts have retained at least one tRNA for each of standard amino

[2]Efficiency of endosymbiont translation has not been measured experimentally, but efficiency of their transcription is not lower compared to free-living bacteria (Traverse and Ochman, 2016).

acids, but eleven endosymbionts have lost all tRNAs for at least one amino acid. If these endosymbionts have a co-symbiont, the co-symbiont has always retained all these missing tRNAs. This would imply cooperation among symbionts, but all tRNAs for one amino acid are lost also from two bacteria without obligate co-symbionts (*Arsenophonus melophagi* and *Carsonella rudii* CS). I can not reject the hypothesis that these two tRNAs were not detected by mistake. On the other hand taking into account possible pseudogenization of some of tRNAs, the number od missing tRNAs could be rather higher. These missing tRNAs suggest that some endosymbionts rely on tRNAs imported from the host cytoplasm or mitochondria. An alternative explanation could be that endosymbionts can maintain translation even without these missing tRNAs.

Counts of tRNAs usually reflect codon usage (Bulmer, 1987), but this rule does not apply to endosymbionts (Hansen and Moran, 2012; van Leuven and McCutcheon 2016, personal communication). Moreover, some of endosymbionts have lost all tRNAs for one amino acid and retained multiple tRNAs for others. These two facts suggest that the tRNA loss is to some extent driven also by drift.

Some archaea species are able to compose a functional tRNA from two (Randau et al., 2005) or exceptionally even three (Fujishima et al., 2009) tRNA-split genes located separately in their genomes. It was suggested that tRNAs were split by a mobile element (Sugahara et al., 2012). Many RNA sequences shorter than tRNA genes were found among RNAseq reads of *Hodgkinia* and *Sulcia*. These reads can be either products of RNA degradation or sequencing mistakes due to modified bases or stable tRNA halves (van Leuven and McCutcheon 2016, personal communication). Proliferation of mobile elements is typical for the initial stages of endosymbiosis (McCutcheon and Moran, 2011). However all mobile elements were removed from endosymbionts in latter stages. While it is not possible to detect these split tRNA-genes with standard computational methods, they would not be detected in this study if they were present in endosymbionts genomes. It would be interesting to search for them in endosymbiont genomes with a special tool such as Splits (Sugahara et al., 2006).

Missing tRNAs for some amino acids and retaining multiple tRNAs with diverse anticodons for other ones, could be explained by reassignment of some codons from one amino-acid to another. This could happen due to loss of some genes considered essential from endosymbiont genomes. In *Escherichia coli*, knockout of one protein release factor, expression of UGA tRNA and a synonymous mutation only in seven ORFs was enough to reassign UGA codon from a stop codon to a sense codon (Mukai et al., 2010). UGA codon was reassigned to tryptophan naturally in tiny genomes such as mitochondrial (Inagaki et al., 1998), *Mycoplasma* (Inamine et al., 1990) and genomes of three endosymbionts included in my study, *Hodgkinia*, *Nasuia* and *Zinderia* (McCutcheon et al., 2009b; MartÃnez-Cano et al., 2015; *Nasuia* and *Zinderia* might constitue one lineage). Several evolutionary mechanisms for these reassignments including some related to genome reduction have been proposed (Andersson and Kurland, 1991; McCutcheon et al., 2009b). This reassigment could have been missed

in other endosymbiont genomes, because genes ending with UGA codons could have been considered sequencing errors (McCutcheon et al., 2009b).

Whether a symbiont (or two co-symbionts together) can independently incorporate a distinct amino acid to its proteins depends not only on presence of corresponding tRNA but also on presence of corresponding amino-acyl tRNA synthetase. *Tremblaya princeps* lineages retain no amino-acyl tRNA synthetase whereas their co-symbionts retain all (Husnik and McCutcheon, 2016) implying that these pairs of co-symbionts could be able to incorporate all amino-acids without using products of their hosts. *Baumannia cicadellinicola*, a co-symbiont of *Sulcia muelleri*, encodes a full set of tRNAs and all amino-acyl tRNA synthetases and *Sulcia muelleri* does not. However, four out of six tRNA-synthetases retained in *Hodgkinia*, a co-symbiont of another *Sulcia* lineage, were found in the corresponding *Sulcia* lineage and other four genes were not identified in any of these two bacteria (McCutcheon et al., 2009a). This case would support the hypothesis that translation could be host dependent.

Concerning tRNA types my results are congruent with findings by van Leuven and McCutcheon (2016, personal communication) for both *Sulcia* lineages and *Baumannia*. For *Hodgkinia* I have observed tRNA-Ile(CAT) and tRNA-Sec(TCA) instead of tRNA-Trp(UCA). I have also agreed with study by de Crécy-Lagard et al. (2012) which suggests that *Baumannia cicadellinicola*, *Blochmannia* species, *Buchnera aphidicola*, *Riesia pediculicola* and *Wigglesworthia glossinidia* have a sufficient number of tRNAs to incorporate all amino acids.

## 4.2 Random loss of RNA modifying genes from endosymbiont genomes

Genes for RNA modification proteins retained in particular species differ to high extent. Nearly four quarters of genes are present in at least one species and no gene is conserved in all species. This implies that their loss is strongly influenced by genetic drift and supports the hypothesis that none of rRNA modification genes is essential and their small benefits sum resulting in a large one for whole ribosome (Decatur and Fournier, 2002). Data from this study imply the same for genes encoding tRNA modification enzymes. However, genes retained in majority species can be slightly more important than others.

Approximately uniform distribution of particular genes counts in all endosymbiont genomes also implies differences in importance of particular genes are not large. If some of genes were be far more important than the rest, these wou;d be present in far more species that any of the others. Nothing like this is visible in the histogram (Figure 5). Alternative hypotheses to explain this pattern are i) some of modifications are provided by the host or a co-symbiont, ii) free-living ancestors of particular lineages already contained diverse modification enzymes, so the diversity of modification enzymes reflects phylogenetic origin of the particular species.

To disthinguish whether an endosymbiont gene is homologous to an *E. coli* gene can be tricky because different genes are variable to diverse extent. Related genes (e.g.: *rluA*

and *rluB*) can be very similar. In one case, I was not able to decide whether one cluster of genes corresponds to *rluB* or *rluF*. *Moranella* genomes of the same lineage differ in one gene which implies that this gene is either false positive in one lineage or false negative in the other, impling that an error in genome annotation of one of these almost identical genomes or that I did a mistake.

Six out of ten most abundant genes inferred by this study (*mnmG*, *mnmE*, *mnmA*, *gluQRS*, *truA*, *iscS* and *tsaE*) in endosymbionts encode proteins responsible for modifications of tRNA anticodon loop, mostly of the wobble position (the first position in the anticodon). Modifications of this position play an important role in codon recognition (Ikeuchi et al., 2006). This might imply that modifications increasing translation efficiency are more important than modifications involved in structure maintenance. MnmA, MnmE and MnmG enzymes modify uridine at wobble position in tRNA (the first position in the tRNA anticodon). Complex MnmEG adds carboxymethylaminomethyl or methylaminomethyl group and MnmA a sulphur group to uridine.

There are multiple ways for transfer of sulphur from cysteine and selenocysteine. Generally, some of bacteria use only one whereas others more of them (Kessler, 2006). The pathway involving MnmA is called IscS and requires also the protein of the same name and TusABCDE (Ikeuchi et al., 2006). This metabolic pathway could be used only by ten lineages of endosymbionts. However, thirty lineages could relay sulphur using just MnmA and IscS. This mechanism was detected *in vitro*, but levels of product formation were that low that authors were not sure whether this is an artefact or not (Ikeuchi et al., 2006). Remaining lineages could use some of alternative pathways (e.g.: SufS), acquire sulphurated uridine from their co-symbionts or not to have their uridine sulphurated. It was also suggested that endosymbionts lacking IscS could use SufS protein instead (McCutcheon et al., 2009a) in as predicted in the protozoan *Theileria parva* (Gardner, 2005).

Modifications provided by MnmA, MnmEG and TruA (5.4.99.12) due increase codon–anticodon affinity at the P-site and therefore prevent ribosome frameshifts(Bregeon, 2001; Urbonavicius, 2001). However, mutations in a few more genes included in this study (*tgt*, *truA*, *trmD*, *miaA*, *miaB*) also increase frameshifting (Urbonavicius, 2001), but these genes are not among the most abundant.

Some of the most abundant genes might be conserved due to their possible other functions than modification of RNA. If it is so products of these genes might not modify RNAs. For example, high level of conservation of *mnmG* could be caused also by its involving DNA replication initiation (Theisen et al., 1993) in addition to modifying uridine.

*gluQ* (*yadB*) gene encodes a protein (GluQRS) which shows high similarity to the amino terminal part of bacterial glutamyl-tRNA synthetases (called also GluQRS) but lacks the tRNA anticodon interaction domain (Campanacci et al., 2004). Its high level of conservation can be caused (i) by confusion of these two proteins or (ii) by working as amino-acyl tRNA synthetase in endosymbionts even though it is not able to fulfill this role in *E. coli* (Dubois et al., 2004;

Salazar et al., 2004).

Data imply that genes for proteins composing enzymatic complexes usually occur together in endosymbionts. My data contain three complexes: MnmEG, TusBCD and TsaBDE. Only four lineages have retained *mnmG* and not *mnmE* or *vice versa*. Both *mnmE* and *mnmG* are lost from six *Tremblaya princeps* lineages. At least in five out of these six lineages and *T. princeps* TPTPER, which retained *mnmE* gene only, have conserved both genes in their co-symbionts. For the remaining one (TPTPLON) I do not have sequence of the cresponding co-symbiont. All genes of the TusBCD complex are conserved in 15 lineages. Some but not all genes of this complex are retained in two species only. Surprisingly, one of these is *Sodalis pierantonius*, the bacterium with the most of modification genes conserved. The same pattern is less clear for the TsaBDE complex, where all three genes are conserved in 31 lineages and some but not all are lost from 12 lineages.

My data suggest that there could be cooperation among comsymbionts. There are two possible ways of cooperation. Endosymbionts with tiny genomes could use either modification proteins or already modified tRNAs from their co-symbionts. In this case the modification proteins of obligate symbionts would be either conserved due to other function than RNA modification or have no function but have not been removed yet. Both these possibilities can be used within diverse lineages of endosymbionts.

All *Carsonella* and *Tremblaya princeps* lineages retain only few modification genes whereas their diverse co-symbionts (if any) retained more than average number of genes. Co-symbionts retained all genes which are conserved in *Tremblaya phenacola* PAVE compared to *Tremblaya princeps* lineages. Only one gene retained in some *Carsonella* and *Tremblaya princeps* lineages but not in their co-symbionts. This could imply that *Carsonella* and *Tremblaya princeps* depend on their co-symbionts to high extent concerning RNA modifications. On the other hand, *Carsonella* lineages without co-symbionts retained the same number of genes as lineages with co-symbionts. This might implies either that the cooperation may be beneficial but not essential or that there is no cooperation at all.

The same applies to *Sulcia* lineages which had conserved very similar sets of RNA modification genes no matter whether they carry a co-symbiont with extremely reduced genome (*Hodgkinia*, *Nasuia* or *Zinderia*) or with more complex genome than *Sulcia* itself (*Baumannia*) or no co-symbiont. There could be reciprocal cooperation between *Sulcia* and *Baumannia*. *Hodgkinia*, *Nasuia* or *Zinderia* retained only genes which are conserved in *Sulcia* lineages and therefore can not provide any modification protein to *Sulcia*, but might take modified tRNAs or some modification proteins from *Sulcia*.

Three genes (*trmB*, *tusE* and *tsaC*) are present in all *Buchnera* lineages except *B. aphidicola* Cc harbouring an obligate co-symbiont (*Serratia symbiotica*) containing all three genes which could imply that some of genes are complementarily divided among these pairs of symbionts (Figure 7). Two more genes, *rsmJ* and *tgt*, show similar pattern, but were lost in one more *Buchnera* lineage. *miaB* is conserved in all *Buchnera* lineages and not in *Serratia*, even

though *Serratia* has conserved all genes which are conserved also in its co-symbiont except two, and eleven others. However, many more genes would be necessary to evaluate this pattern.

Some of endosymbiont RNA modifications might be provided by their hosts. Seven distinct modified tRNA sites (G9, G37, G19, C17, U34, A37 and N20) were found in tRNA of *Hodgkinia* (van Leuven and McCutcheon 2016, personal communication), but only three genes encoding tRNA modifying enzymes were identified in its genome. In *Hodgkinia* co-symbiont, *Sulcia*, only three distinct modified sites (C32, U34, G37) were found (van Leuven and McCutcheon 2016, personal communication). The same sites (according to both position and base type) as all three sites in *Sulcia* and four sites of *Hodgkinia* are modified in *E. coli* enabling to hypothesize which enzymes could be responsible for these modifications. It seems that both bacteria should be able to modify U34 by MnmEG complex. In theory, A37 in *Hodgkinia* could be modified by MiaB from *Sulcia*, however, this enzyme is known to modify a base which was already modified by MiA (2.5.1.75) lacking in both co-symbionts. Genes encoding enzymes modifying other positions (*Hodgkinia*: G37 - TrmD (2.1.1.228), N20 - DusA (1.1.1.-); *Sulcia*: C32 - TtcA (2.8.1.4), G37 - TrmD) are present neither in *Sulcia* nor in *Hodgkinia*. In both *Sulcia* and *Hodgkinia* retained other genes encoding enzymes which should modify tRNA at other sites (GluQRS (2.4.1.-) for both; TruA and MtaB for *Sulcia* only). No modifications which could be caused by these enzymes were observed. Following three hypotheses could explain this pattern: (i) Enzymes are present in *Sulcia* or *Hodgkinia* but corresponding genes were not detected in my study. I do not consider this very probable because many of genes corresponding to these enzymes were successfully detected in other related bacteria and because I did not manage to assign two sites (G9 and C17 of *Hodgkinia*) to any of enzymes from *E. coli*. (ii) The tRNA modification sites of the free-living ancestors of *Sulcia* and *Hodgkinia* were different from each other and from *E. coli*. These three lineages originate from distant phyla, so some differences in modifications are quite likely. However, considering conservation of endosymbiont modification sites based on the set of *E. coli*, I would suppose that these differences are not large. (iii) At least some of modifications are provided by host. This hypothesis would be supported by the fact that a mutation in a single nuclear gene encoding a modification enzyme can lead to lack of corresponding modification in both cytoplasmic and mitochondrial tRNAs (Hopper et al., 1982). Personally, I favour this hypothesis.

*rlmI*, one of genes included in this study, was transferred from an Enterobacteriaceae bacterium to *Planococcus citri*, the host of *Tremblaya princeps* and *Moranella endobia* (PCIT lineage; Husník et al., 2013). Interestingly, this gene was conserved only once among all endosymbiont lineages. (It was in *Sodalis pierantonius*, the bacterium with the largest number of modification proteins conserved). *Pachypsylla venusta* psyllid, the host of *Carsonella ruddii* PV, has acquired bacterial *rsmJ* (Sloan et al., 2014) which is also conserved only in few lineages. Genes encoding RNA modification proteins from *Wolbachia* were detected in *Cal-*

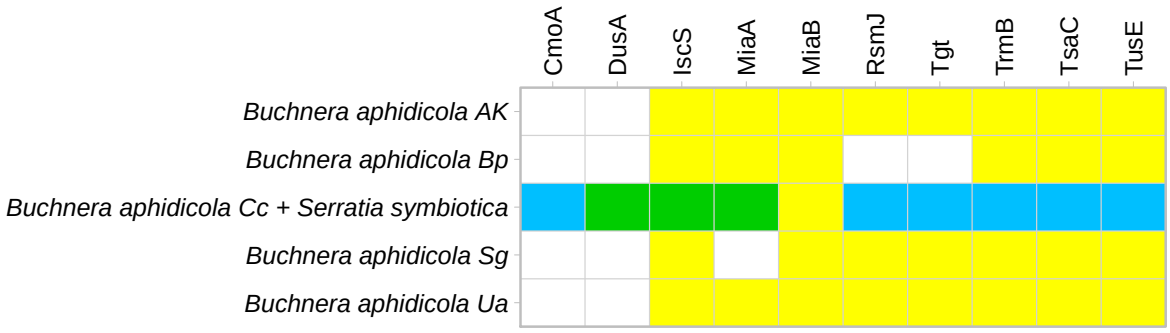


Figure 7: Example of presence/absence of modification gene in *Buchneras* strains. One of these has a co-obligate symbiont in its host. Three genes (TrmB, TsaC, TusE) were found in all *Buchnera* species except the one with the co-obligate symbiont, which could imply cooperation among symbionts. However, this pattern could easily occur by chance. Color coding: white - gene is conserved neither in obligate nor in co-obligate symbiont, yellow - gene is conserved in obligate symbiont only, blue - gene is conserved in co-obligate symbiont only, green - gene is conserved in both obligate and co-obligate symbiont.

*losobruchus chinensis* and in *Glossina morsitans*, the host of *Wigglesworthia glossinidia* GM, however, all these genes are probably pseudogenized (Nikoh et al., 2008; Attardo et al., 2014). No bacterial genes related to ncRNA metabolism were found neither in *Pediculus humanus*, the host of *Riesia pediculicola* (Kirkness et al., 2010) nor in *Acyrthosiphon pisum* cooperating with *Buchnera* (Nikoh et al., 2010) nor in *Bemisia tabaci* haboring *Portiera aleyrodidarum* BT (Luan et al., 2015), nor in four additional Hemiptera species haboring obligate endosymbionts (Machová, unpublished data). From these data it does not seem that endosymbionts commonly use a bacterial gene transferred to nuclei of the host for modification protein synthesis in the same way as organelles (Motorin and Grosjean, 2005).

A few tRNA modification genes were previously identified by BLAST analysis against the homologous genes of *E. coli* (de Crécy-Lagard et al., 2012). Compared to my study, two more genes were identified by de Crécy-Lagard et al. (*miaA* in *Buchnera aphidicola* Sg and *trmD* in *Wigglesworthia glossinidia* GB; Figure 8), nevertheless, I think that these two studies have agreed to large extent. I found a gene NP_660792.1 of *Buchnera aphidicola* Sg among blast hits of *E. coli miaA*, but it has very high E-value (0.69) and was annotated as ATP-dependent protease ATP-binding subunit ClpX. The best blast hits of this gene against nr confirmed that the annotation is correct. No significant hits were found when blasting *miaA* of other *Buchnera* lineage (AK; YP_005619727.1) against *Buchnera* Sg using the web server. Interestingly, I have not found *trmD* in *Wigglesworthia glossinidia* GB which was used in study of de Crécy-Lagard et al., but I found it in the other *Wigglesworthia* lineage, GM. Blast using web server against both lineages confirmed these results. Generally, these two lineages are very similar. *trmD* is one of two genes in which they differ.

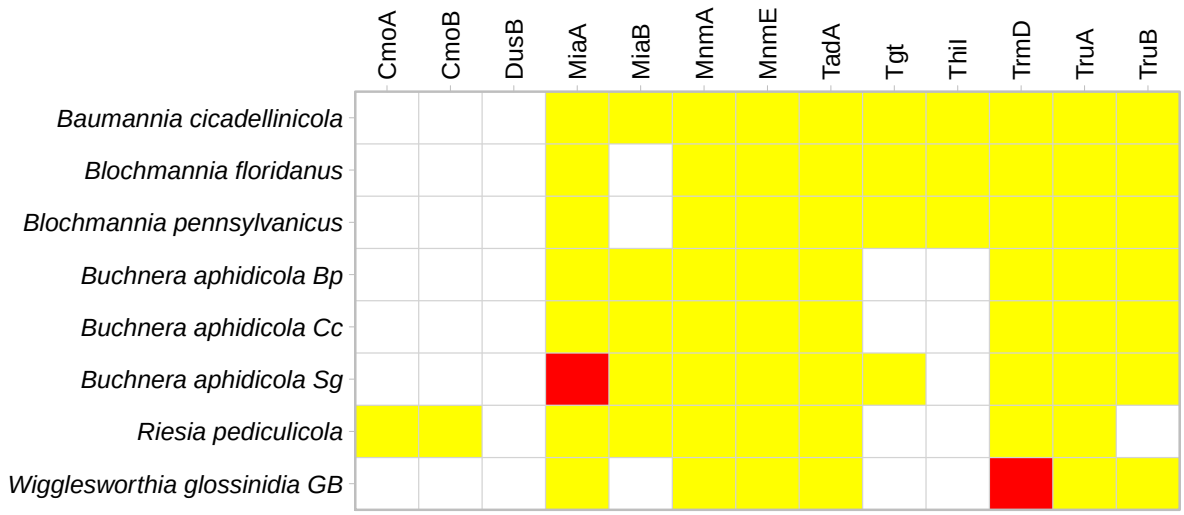


Figure 8: Comparison of presence/absence of RNA modification published by de Crécy-Lagard et al. (2012) with results of this study. Color coding: yellow: gene was observed by both study of de Crécy-Lagard et al. (2012) and this study; red gene was observed by de Crécy-Lagard et al. (2012) only.

## 4.3 RNA modification sites are generally conserved in endosymbionts

Generally, the most of rRNA modifications occur in conserved regions of rRNAs implying their importance for both ribosome function and structure (Decatur and Fournier, 2002). Surrounding conserved bases can be either also important or conserved due to strong selection pressure on modification sites (hitch-hiking). This hypothesis can be applied also on endosymbionts, because all modification sites except two of them at each ribosomal subunit were conserved in all species.

In tRNAs slightly more (seven) sites were not conserved. However, this number can be higher due to the fact that tRNAs which were not marked as "possible pseudogenes" were excluded from this part of analysis not to decrease quality of alignments that could lead into generating errors. Despite this, a few errors were detected such that a wrong base was observed also in *E. coli*. There can be some more errors, because in one case out of four a random base will be the same as the original one. There could be also a mistake in alignment of particular endosymbiont, but all alignments were checked visually and it did not seem that there were many ambiguously aligned positions.

Generally, some of sites which are considered unconserved, could be explained by a sequencing error in the genome assembly or by polymorphic population of symbionts, i.e. with multiple variants of the modification site within a single host.

Interestingly, all these non-conserved sites were modified by enzymes which should be able to modify original bases without a previous modification. This could support the hypothesis that the most important bases are hypermodified, but four cases only are not sufficient to support the hypothesis..

In rRNA, all proteins with non-conserved modification site(s) have the corresponding

gene lost. This could imply that sites which do not play very important role can be both substituted and loose their corresponding modification protein. In tRNA, this holds for five out of seven sites. The enzyme with the modification site changed at majority of lineages was not detected in any of lineages. This could imply the same consequences as for rRNA modification sites, but two modification sites of TrmB (2.1.1.33), an enzyme which is conserved in eleven lineages only, were changed mostly at species where the enzyme is conserved.

# 5 Conclusion and future prospects

This study is one of the first analyses of a broad genome sampling providing insights into RNA biology of insect endosymbionts. Several interesting patterns were found in loss of tRNAs and genes encoding RNA modification proteins and in level of conservation of their modification sites giving new ideas about translation and ncRNA modifications of endosymbiotic bacteria.

To sum up, tRNA genes and RNA modification gene loss seems to be strongly influenced by genetic drift. At least some of endosymbionts likely need to cooperate with their co-symbionts and hosts to maintain translation because they do not encode the full tRNA set. Cooperation with the host might be necessary also to modify RNA bases since they do not code tRNAs for all amino acids and retain only a handful of genes for tRNA and rRNA modifying enzymes. Unlike organelles, using modification enzymes encoded by a bacterial gene transferred to genome of their host does not seem to be a common strategy of endosymbionts. Generally, I have agreed with previously published studies to high extent.

RNA modification sites, especially those in rRNAs, seem to play an important structural role regardless whether they are modified or not. It seems that it is more important to have at least some of modifications on originally hypermodified bases than modify ones which require a single modification. Modification of tRNA wobble position might be the most important. Nevertheless, it does not seem that some group of modifications (such as those modifying tRNA or rRNA, bacteria specific or occurring in both bacteria and eukaryotes) would be significantly more important than the other. Enzymes composing enzymatic complexes tend to be conserved together.

All analyses were done *in silico* and their results could be more inaccurate than those which could be obtained experimentally. Moreover, some kinds of data can not be obtained using genome sequences only, e.g. presence of a gene does not confirm presence of its corresponding functional enzyme. Therefore, it would be good to evaluate some of results experimentally. Other interesting analyses could be done also *in silico*.

I think following analyses would be the most interesting:
(i) to find out whether endosymbionts genomes contain tRNA halves
(ii) to compare computationally found tRNAs with RNAseq data using many different species
(iii) to check which amino-acyl-tRNA synthetases are present in particular endosymbionts to find out which tRNAs can be charged by amino acids

# Bibliography


S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. doi: 10.1016/s0022-2836(05)80360-2. URL http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

G. E. Andersson and C. G. Kurland. An extreme codon preference strategy: codon reassignment. *Molecular Biology and Evolution*, 8(4):530–544, 1991. URL http://mbe.oxfordjournals.org/content/8/4/530.full.pdf+html.

D. H. Ardell. Computational analysis of tRNA identity. *FEBS Letters*, 584(2):325–333, 2009. doi: 10.1016/j.febslet.2009.11.084. URL http://dx.doi.org/10.1016/j.febslet.2009.11.084.

G. M. Attardo, P. P. Abila, J. E. Auma, A. A. Baumann, J. B. Benoit, C. L. Brelsfoard, J. M. C. Ribeiro, J. A. Cotton, D. Q. D. Pham, A. C. Darby, J. V. D. Abbeele, D. L. Denlinger, L. M. Field, S. R. G. Nyanjom, M. W. Gaunt, D. L. Geiser, L. M. Gomulski, L. R. Haines, I. A. Hansen, J. W. Jones, C. K. Kibet, J. K. Kinyua, D. M. Larkin, M. J. Lehane, R. V. M. Rio, S. J. Macdonald, R. W. Macharia, A. R. Malacrida, H. G. Marco, K. K. Marucha, D. K. Masiga, M. E. Meuti, P. O. Mireji, G. F. O. Obiero, J. J. O. Koekemoer, C. K. Okoro, I. A. Omedo, V. C. Osamor, A. S. P. Balyeidhusa, J. T. Peyton, D. P. Price, M. A. Quail, U. N. Ramphul, N. D. Rawlings, M. A. Riehle, H. M. Robertson, M. J. Sanders, M. J. Scott, Z. J. S. Dashti, A. K. Snyder, T. P. Srivastava, E. J. Stanley, M. T. Swain, D. S. T. Hughes, A. M. Tarone, T. D. Taylor, E. L. Telleria, G. H. Thomas, D. P. Walshe, R. K. Wilson, J. J. Winzerling, A. Acosta-Serrano, S. Aksoy, P. Arensburger, M. Aslett, R. Bateta, A. Benkahla, M. Berriman, K. Bourtzis, J. Caers, G. Caljon, A. Christoffels, M. Falchetto, M. Friedrich, S. Fu, G. Gade, G. Githinji, R. Gregory, N. Hall, G. Harkins, M. Hattori, C. Hertz-Fowler, W. Hide, W. Hu, T. Imanishi, N. Inoue, M. Jonas, Y. Kawahara, M. Koffi, A. Kruger, D. Lawson, S. Lehane, H. Lehvaslaiho, T. Luiz, M. Makgamathe, I. Malele, O. Manangwa, L. Manga, K. Megy, V. Michalkova, F. Mpondo, F. Mramba, A. Msangi, N. Mulder, G. Murilla, S. Mwangi, L. Okedi, S. Ommeh, C.-P. Ooi, J. Ouma, S. Panji, S. Ravel, C. Rose, R. Sakate, L. Schoofs, F. Scolari, V. Sharma, C. Sim, G. Siwo, P. Solano, D. Stephens, Y. Suzuki, S.-H. Sze, Y. Toure, A. Toyoda, G. Tsiamis, Z. Tu, M. Wamalwa, F. Wamwiri, J. Wang, W. Warren, J. Watanabe, B. Weiss, J. Willis, P. Wincker, Q. Zhang, and J.-J. Zhou. Genome sequence of the tsetse fly (*Glossina morsitans*): Vector of african trypanosomiasis. *Science*, 344(6182):380–386, 2014. doi: 10.1126/science.1249656. URL http://dx.doi.org/10.1126/science.1249656.

T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* k-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2, 2006. doi: 10.1038/msb4100050. URL http://dx.doi.org/10.1038/msb4100050.

P. Baumann. Biology of bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annual Review of Microbiology*, 59(1):155–189, 2005. doi: 10.1146/annurev.micro.59.030804.121041. URL http://dx.doi.org/10.1146/annurev.micro.59.030804.121041.

G. M. Bennett and N. A. Moran. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biology and Evolution*, 5(9):1675–1688, 2013. doi: 10.1093/gbe/evt118. URL http://dx.doi.org/10.1093/gbe/evt118.

G. M. Bennett and N. A. Moran. Heritable symbiosis: The advantages and perils of an evolutionary rabbit hole. *Proceedings of the National Academy of Sciences*, 112(33):10169–10176, 2015. doi: 10.1073/pnas.1421388112. URL http://dx.doi.org/10.1073/pnas.1421388112.

G. M. Bennett, S. Abbà, M. Kube, and C. Marzachì. Complete genome sequences of the obligate symbionts *Candidatus Sulcia muelleri* and *Ca. Nasuia deltocephalinicola* from the pestiferous leafhopper *Macrosteles quadripunctulatus* (Hemiptera: Cicadellidae). *Genome Announcements*, 4(1):e01604–15, 2016. doi: 10.1128/genomea.01604-15. URL http://dx.doi.org/10.1128/genomeA.01604-15.

D. A. Benson. GenBank. *Nucleic Acids Research*, 33(Database issue):D34–D38, 2004. doi: 10.1093/nar/gki063. URL http://dx.doi.org/10.1093/nar/gki063.

D. Bregeon. Translational misreading: a tRNA modification counteracts a 2+ ribosomal frameshift. *Genes & Development*, 15(17):2295–2306, 2001. doi: 10.1101/gad.207701. URL http://dx.doi.org/10.1101/gad.207701.

M. Bulmer. Coevolution of codon usage and transfer RNA abundance. *Nature*, 325(6106):728–730, 1987. doi: 10.1038/325728a0. URL http://dx.doi.org/10.1038/325728a0.

G. Burger, M. W. Gray, L. Forget, and B. F. Lang. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biology and Evolution*, 5(2):418–438, 2013. doi: 10.1093/gbe/evt008. URL http://dx.doi.org/10.1093/gbe/evt008.

L. Cai. *gbk2faa.pl*. Fudan University, Shanghai, China, a. URL http://cail.cn/programming.html.

L. Cai. *gbk2fna.pl*. Fudan University, Shanghai, China, b. URL http://cail.cn/programming.html.

V. Campanacci, D. Y. Dubois, H. D. Becker, D. Kern, S. Spinelli, C. Valencia, F. Pagot, A. Salomoni, S. Grisel, R. Vincentelli, C. Bignon, J. Lapointe, R. Giegé, and C. Cambillau. The *Escherichia coli* YadB gene product reveals a novel aminoacyl-tRNA synthetase like activity. *Journal of Molecular Biology*, 337(2):273–283, 2004. doi: 10.1016/j.jmb.2004.01.027. URL http://dx.doi.org/10.1016/j.jmb.2004.01.027.

M. A. Campbell, J. T. V. Leuven, R. C. Meister, K. M. Carey, C. Simon, and J. P. McCutcheon. Genome expansion via lineage splitting and genome reduction in the cicada endosymbiont hodgkinia. *Proceedings of the National Academy of Sciences*, 112(33):10192–10199, 2015. doi: 10.1073/pnas.1421386112. URL http://dx.doi.org/10.1073/pnas.1421386112.

C. Conord, L. Despres, A. Vallier, S. Balmand, C. Miquel, S. Zundel, G. Lemperiere, and A. Heddi. Long-term evolutionary stability of bacterial endosymbiosis in Curculionoidea: Additional evidence of symbiont replacement in the Dryophthoridae family. *Molecular Biology and Evolution*, 25(5): 859–868, 2008. doi: 10.1093/molbev/msn027. URL http://dx.doi.org/10.1093/molbev/msn027.

A. Czerwoniec, S. Dunin-Horkawicz, E. Purta, K. H. Kaminska, J. M. Kasprzak, J. M. Bujnicki, H. Grosjean, and K. Rother. MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Research*, 37(Database):D118–D121, 2009. doi: 10.1093/nar/gkn710. URL http://dx.doi.org/10.1093/nar/gkn710.

D. B. Dahl. *xtable: Export Tables to LaTeX or HTML*, 2016. URL https://CRAN.R-project.org/package=xtable. R package version 1.8-2.

V. de Crécy-Lagard, C. Marck, and H. Grosjean. Decoding in *Candidatus Riesia pediculicola*, close to a minimal trna modification set? *Trends Cell and Molecular Biology*, 7:11–34, 2012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3539174/.

W. A. Decatur and M. J. Fournier. rRNA modifications and ribosome function. *Trends in Biochemical Sciences*, 27(7):344–351, 2002. doi: 10.1016/s0968-0004(02)02109-6. URL http://dx.doi.org/10.1016/S0968-0004(02)02109-6.

D. Y. Dubois, M. Blaise, H. D. Becker, V. Campanacci, G. Keith, R. Giege, C. Cambillau, J. Lapointe, and D. Kern. From the cover: An aminoacyl-tRNA synthetase-like protein encoded by the *Escherichia coli* yadB gene glutamylates specifically tRNAAsp. *Proceedings of the National Academy of Sciences*, 101(20):7530–7535, 2004. doi: 10.1073/pnas.0401634101. URL http://dx.doi.org/10.1073/pnas.0401634101.

S. Dunin-Horkawicz. MODOMICS: a database of RNA modification pathways. *Nucleic Acids Research*, 34(90001):D145–D149, 2006. doi: 10.1093/nar/gkj084. URL http://dx.doi.org/10.1093/nar/gkj084.

M. A. Fares, M. X. Ruiz-González, A. Moya, S. F. Elena, and E. Barrio. Endosymbiotic bacteria GroEL buffers against deleterious mutations. *Nature*, 417(6887):398–398, 2002. doi: 10.1038/417398a. URL http://dx.doi.org/10.1038/417398a.

K. Fujishima, J. Sugahara, K. Kikuta, R. Hirano, A. Sato, M. Tomita, and A. Kanai. Tri-split tRNA is a transfer RNA made from 3 transcripts that provides insight into the evolution of fragmented tRNAs in archaea. *Proceedings of the National Academy of Sciences*, 106(8):2683–2687, 2009. doi: 10.1073/pnas.0808246106. URL http://dx.doi.org/10.1073/pnas.0808246106.

M. J. Gardner. Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, 309(5731):134–137, 2005. doi: 10.1126/science.1110439. URL http://dx.doi.org/10.1126/science.1110439.

R. T. Gergely Daróczi. *pander: An R Pandoc Writer*, 2015. URL https://CRAN.R-project.org/package=pander. R package version 0.6.0.

W. Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618–618, 1986. doi: 10.1038/319618a0. URL http://dx.doi.org/10.1038/319618a0.

Y. Gottlieb, M. Ghanim, G. Gueguen, S. Kontsedalov, F. Vavre, F. Fleury, and E. Zchori-Fein. Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. *The FASEB Journal*, 22(7):2591–2599, 2008. doi: 10.1096/fj.07-101162. URL http://dx.doi.org/10.1096/fj.07-101162.

C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983. doi: 10.1016/0092-8674(83)90117-4. URL http://dx.doi.org/10.1016/0092-8674(83)90117-4.

R. G. H Pages, P Aboyoun and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. URL https://bioconductor.org/packages/release/bioc/html/Biostrings.html. R package version 2.38.3.

A. K. Hansen and N. A. Moran. Altered tRNA characteristics and 3' maturation in bacterial symbionts with reduced genomes. *Nucleic Acids Research*, 40(16):7870–7884, 2012. doi: 10.1093/nar/gks503. URL http://dx.doi.org/10.1093/nar/gks503.

A. K. Hansen and N. A. Moran. The impact of microbial symbionts on host plant utilization by herbivorous insects. *Molecular Ecology*, 23(6):1473–1496, 2013. doi: 10.1111/mec.12421. URL http://dx.doi.org/10.1111/mec.12421.

M. Helm. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Research*, 34(2):721–733, 2006. doi: 10.1093/nar/gkj471. URL http://dx.doi.org/10.1093/nar/gkj471.

R. Hershberg and D. A. Petrov. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics*, 6(9):e1001115, 2010. doi: 10.1371/journal.pgen.1001115. URL http://dx.doi.org/10.1371/journal.pgen.1001115.

A. K. Hopper, A. H. Furukawa, H. D. Pham, and N. C. Martin. Defects in modification of cytoplasmic and mitochondrial transfer RNAs are caused by single nuclear mutations. *Cell*, 28(3):543–550, 1982. doi: 10.1016/0092-8674(82)90209-4. URL http://dx.doi.org/10.1016/0092-8674(82)90209-4.

C. M. Hudson and K. P. Williams. The tmRNA website. *Nucleic Acids Research*, 43(D1):D138–D140, 2014. doi: 10.1093/nar/gku1109. URL http://dx.doi.org/10.1093/nar/gku1109.

F. Husnik and J. P. McCutcheon. Repeated replacement of an intrabacterial symbiont in the tripartite nested mealybug symbiosis. Technical report, 2016. URL http://dx.doi.org/10.1101/042267.

F. Husník, N. Nikoh, R. Koga, L. Ross, R. P. Duncan, M. Fujie, M. Tanaka, N. Satoh, D. Bachtrog, A. C. Wilson, C. D. von Dohlen, T. Fukatsu, and J. P. McCutcheon. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, 153(7): 1567–1578, 2013. doi: 10.1016/j.cell.2013.05.040. URL http://dx.doi.org/10.1016/j.cell.2013.05.040.

Y. Ikeuchi, N. Shigi, J. ichi Kato, A. Nishimura, and T. Suzuki. Mechanistic insights into sulfur relay by multiple sulfur mediators involved in thiouridine biosynthesis at tRNA wobble positions. *Molecular Cell*, 21(1):97–108, 2006. doi: 10.1016/j.molcel.2005.11.001. URL http://dx.doi.org/10.1016/j.molcel.2005.11.001.

Y. Inagaki, M. Ehara, I. K. Watanabe, Y. Hayashi-Ishimaru, and T. Ohama. Directionally evolving genetic code: The UGA codon from stop to tryptophan in mitochondria. *Journal of Molecular Evolution*, 47(4):378–384, 1998. ISSN 1432-1432. doi: 10.1007/PL00006395. URL http://dx.doi.org/10.1007/PL00006395.

J. Inamine, K. Ho, S. Loechel, and P. Hu. Evidence that UGA is read as a tryptophan codon rather than as a stop codon by *Mycoplasma pneumoniae*, *Mycoplasma genitalium*, and *Mycoplasma gallisepticum*. *Journal of Bacteriology*, 172(1):504–506., 1990. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC208464/pdf/jbacter01043-0526.pdf.

J. Janouškovec, S.-L. Liu, P. T. Martone, W. Carré, C. Leblanc, J. Collén, and P. J. Keeling. Evolution of red algal plastid genomes: Ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PLoS ONE*, 8(3):e59001, 2013. doi: 10.1371/journal.pone.0059001. URL http://dx.doi.org/10.1371/journal.pone.0059001.

H. O. Kammen, C. C. Marvel, L. Hardy, and E. E. Penhoet. Purification, structure, and properties of *Escherichia coli* tRNA pseudouridine synthase i. *Journal of Biological Chemistry*, 263(5):2255–2263, 1988. URL http://www.jbc.org/content/263/5/2255.full.pdf+html.

K. Katoh. *MAFFT v7.215*, 2014. URL http://mafft.cbrc.jp/alignment/software/. MBE 30:772-780 (2013), NAR 30:3059-3066 (2002).

L. Katz. *Converting alignment files*. R Foundation for Statistical Computing, 2012. URL http://www.bioperl.org/wiki/Converting_alignment_files. contact: lkatz@cdc.gov.

M. Kearse, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A. Cooper, S. Markowitz, C. Duran, T. Thierer, B. Ashton, P. Meintjes, and A. Drummond. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649, 2012. doi: 10.1093/bioinformatics/bts199. URL http://dx.doi.org/10.1093/bioinformatics/bts199.

I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martinez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muniz-Rascado, Q. Ong, S. Paley, I. Schroder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss,

J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1):D605–D612, 2012. doi: 10.1093/nar/gks1027. URL http://dx.doi.org/10.1093/nar/gks1027.

D. Kessler. Enzymatic activation of sulfur for incorporation into biomolecules in prokaryotes. *FEMS Microbiology Reviews*, 30(6):825–840, 2006. doi: 10.1111/j.1574-6976.2006.00036.x. URL http://dx.doi.org/10.1111/j.1574-6976.2006.00036.x.

E. F. Kirkness, B. J. Haas, W. Sun, H. R. Braig, M. A. Perotti, J. M. Clark, S. H. Lee, H. M. Robertson, R. C. Kennedy, E. Elhaik, D. Gerlach, E. V. Kriventseva, C. G. Elsik, D. Graur, C. A. Hill, J. A. Veenstra, B. Walenz, J. M. C. Tubio, J. M. C. Ribeiro, J. Rozas, J. S. Johnston, J. T. Reese, A. Popadic, M. Tojo, D. Raoult, D. L. Reed, Y. Tomoyasu, E. Kraus, O. Mittapalli, V. M. Margam, H.-M. Li, J. M. Meyer, R. M. Johnson, J. Romero-Severson, J. P. VanZee, D. Alvarez-Ponce, F. G. Vieira, M. Aguade, S. Guirao-Rico, J. M. Anzola, K. S. Yoon, J. P. Strycharz, M. F. Unger, S. Christley, N. F. Lobo, M. J. Seufferheld, N. Wang, G. A. Dasch, C. J. Struchiner, G. Madey, L. I. Hannick, S. Bidwell, V. Joardar, E. Caler, R. Shao, S. C. Barker, S. Cameron, R. V. Bruggner, A. Regier, J. Johnson, L. Viswanathan, T. R. Utterback, G. G. Sutton, D. Lawson, R. M. Waterhouse, J. C. Venter, R. L. Strausberg, M. R. Berenbaum, F. H. Collins, E. M. Zdobnov, and B. R. Pittendrigh. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences*, 107(27):12168–12173, 2010. doi: 10.1073/pnas.1003379107. URL http://dx.doi.org/10.1073/pnas.1003379107.

K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1):147–157, 1982. doi: 10.1016/0092-8674(82)90414-7. URL http://dx.doi.org/10.1016/0092-8674(82)90414-7.

J. D. Lambert and N. A. Moran. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences*, 95(8):4458–4462, 1998. doi: 10.1073/pnas.95.8.4458. URL http://dx.doi.org/10.1073/pnas.95.8.4458.

A. Lamelas, M. J. Gosalbes, A. Moya, and A. Latorre. New clues about the evolutionary history of metabolic losses in bacterial endosymbionts, provided by the genome of *Buchnera aphidicola* from the aphid *Cinara tujafilina*. *Applied and Environmental Microbiology*, 77(13):4446–4454, 2011. doi: 10.1128/aem.00141-11. URL http://dx.doi.org/10.1128/AEM.00141-11.

L. Lancaster, N. J. Lambert, E. J. Maklan, L. H. Horan, and H. F. Noller. The sarcin-ricin loop of 23S rRNA is essential for assembly of the functional core of the 50S ribosomal subunit. *RNA*, 14(10):1999–2012, 2008. doi: 10.1261/rna.1202108. URL http://dx.doi.org/10.1261/rna.1202108.

D. Laslett. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16, 2004. doi: 10.1093/nar/gkh152. URL http://dx.doi.org/10.1093/nar/gkh152.

C. Lefevre. Endosymbiont phylogenesis in the Dryophthoridae weevils: Evidence for bacterial replacement. *Molecular Biology and Evolution*, 21(6):965–973, 2004. doi: 10.1093/molbev/msh063. URL http://dx.doi.org/10.1093/molbev/msh063.

J. T. V. Leuven and J. P. McCutcheon. An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. *Genome Biology and Evolution*, 4(1):24–27, 2011. doi: 10.1093/gbe/evr125. URL http://dx.doi.org/10.1093/gbe/evr125.

J. T. V. Leuven, R. C. Meister, C. Simon, and J. P. McCutcheon. Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. *Cell*, 158(6):1270–1280, 2014. doi: 10.1016/j.cell.2014.07.047. URL http://dx.doi.org/10.1016/j.cell.2014.07.047.

L. Li. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9): 2178–2189, 2003. doi: 10.1101/gr.1224503. URL http://dx.doi.org/10.1101/gr.1224503.

M. Lluch-Senar, J. Delgado, W.-H. Chen, V. Llorens-Rico, F. J. O'Reilly, J. A. Wodke, E. B. Unal, E. Yus, S. Martinez, R. J. Nichols, T. Ferrar, A. Vivancos, A. Schmeisky, J. Stulke, V. van Noort, A.-C. Gavin, P. Bork, and L. Serrano. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Molecular Systems Biology*, 11(1):780–780, 2015. doi: 10.15252/msb.20145558. URL http://dx.doi.org/10.15252/msb.20145558.

T. M. Lowe and S. R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997. doi: 10.1093/nar/25.5.0955. URL http://dx.doi.org/10.1093/nar/25.5.0955.

J.-B. Luan, W. Chen, D. K. Hasegawa, A. M. Simmons, W. M. Wintermantel, K.-S. Ling, Z. Fei, S.-S. Liu, and A. E. Douglas. Metabolic coevolution in the bacterial symbiosis of whiteflies and related plant sap-feeding insects. *Genome Biology and Evolution*, 7(9):2635–2647, 2015. doi: 10.1093/gbe/evv170. URL http://dx.doi.org/10.1093/gbe/evv170.

M. A. Machnicka, K. Milanowska, O. O. Oglou, E. Purta, M. Kurkowska, A. Olchowik, W. Januszewski, S. Kalinowski, S. Dunin-Horkawicz, K. M. Rother, M. Helm, J. M. Bujnicki, and H. Grosjean. MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Research*, 41(D1):D262–D267, 2012. doi: 10.1093/nar/gks1007. URL http://dx.doi.org/10.1093/nar/gks1007.

D. J. MartÃnez-Cano, M. Reyes-Prieto, E. MartÃnez-Romero, L. P. Partida-MartÃnez, A. Latorre, A. Moya, and L. Delaye. Evolution of small prokaryotic genomes. *Frontiers in Microbiology*, 5, 2015. doi: 10.3389/fmicb.2014.00742. URL http://dx.doi.org/10.3389/fmicb.2014.00742.

D. Matelska, M. Kurkowska, E. Purta, J. M. Bujnicki, and S. Dunin-Horkawicz. Loss of conserved noncoding RNAs in genomes of bacterial endosymbionts. *Genome Biology and Evolution*, 8(2): 426–438, 2016. doi: 10.1093/gbe/evw007. URL http://dx.doi.org/10.1093/gbe/evw007.

J. P. McCutcheon. The bacterial essence of tiny symbiont genomes. *Current Opinion in Microbiology*, 13(1):73–78, 2010. doi: 10.1016/j.mib.2009.12.002. URL http://dx.doi.org/10.1016/j.mib.2009.12.002.

J. P. McCutcheon and N. A. Moran. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 2011. doi: 10.1038/nrmicro2670. URL http://dx.doi.org/10.1038/nrmicro2670.

J. P. McCutcheon and C. D. von Dohlen. An interdependent metabolic patchwork in the nested symbiosis of mealybugs. *Current Biology*, 21(16):1366–1372, 2011. doi: 10.1016/j.cub.2011.06.051. URL http://dx.doi.org/10.1016/j.cub.2011.06.051.

J. P. McCutcheon, B. R. McDonald, and N. A. Moran. Convergent evolution of metabolic roles in bacterial co-symbionts of insects. *Proceedings of the National Academy of Sciences*, 106(36): 15394–15399, 2009a. doi: 10.1073/pnas.0906424106. URL http://dx.doi.org/10.1073/pnas.0906424106.

J. P. McCutcheon, B. R. McDonald, and N. A. Moran. Origin of an alternative genetic code in the extremely small and GC rich genome of a bacterial symbiont. *PLoS Genetics*, 5(7):e1000565, 2009b. doi: 10.1371/journal.pgen.1000565. URL http://dx.doi.org/10.1371/journal.pgen.1000565.

N. A. Moran. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences*, 93(7):2873–2878, 1996. doi: 10.1073/pnas.93.7.2873. URL http://dx.doi.org/10.1073/pnas.93.7.2873.

N. A. Moran, J. P. McCutcheon, and A. Nakabachi. Genomics and evolution of heritable bacterial symbionts. *Annual Review of Genetics*, 42(1):165–190, 2008. doi: 10.1146/annurev.genet.41.110306.130119. URL http://dx.doi.org/10.1146/annurev.genet.41.110306.130119.

Y. Motorin and H. Grosjean. Transfer RNA modification. In *Encyclopedia of Life Sciences - eLS*. eLS, 2005. doi: 10.1038/npg.els.0003866.

T. Mukai, A. Hayashi, F. Iraha, A. Sato, K. Ohtake, S. Yokoyama, and K. Sakamoto. Codon reassignment in the *Escherichia coli* genetic code. *Nucleic Acids Research*, 38(22):8188–8195, 2010. doi: 10.1093/nar/gkq707. URL http://dx.doi.org/10.1093/nar/gkq707.

A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, 314(5797):267–267, 2006. doi: 10.1126/science.1134196. URL http://dx.doi.org/10.1126/science.1134196.

A. Nakabachi, R. Ueoka, K. Oshima, R. Teta, A. Mangoni, M. Gurgui, N. J. Oldham, G. van Echten-Deckert, K. Okamura, K. Yamamoto, H. Inoue, M. Ohkuma, Y. Hongoh, S. ya Miyagishima, M. Hattori, J. Piel, and T. Fukatsu. Defensive bacteriome symbiont with a drastically reduced genome. *Current Biology*, 23(15):1478–1484, 2013. doi: 10.1016/j.cub.2013.06.027. URL http://dx.doi.org/10.1016/j.cub.2013.06.027.

E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935, 2013. doi: 10.1093/bioinformatics/btt509. URL http://dx.doi.org/10.1093/bioinformatics/btt509.

E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, 43(D1):D130–D137, 2014. doi: 10.1093/nar/gku1063. URL http://dx.doi.org/10.1093/nar/gku1063.

N. Nikoh, K. Tanaka, F. Shibata, N. Kondo, M. Hizume, M. Shimada, and T. Fukatsu. *Wolbachia* genome integrated in an insect chromosome: Evolution and fate of laterally transferred endosymbiont genes. *Genome Research*, 18(2):272–280, 2008. doi: 10.1101/gr.7144908. URL http://dx.doi.org/10.1101/gr.7144908.

N. Nikoh, J. P. McCutcheon, T. Kudo, S. ya Miyagishima, N. A. Moran, and A. Nakabachi. Bacterial genes in the aphid genome: Absence of functional gene transfer from *Buchnera* to its host. *PLoS Genetics*, 6(2):e1000827, 2010. doi: 10.1371/journal.pgen.1000827. URL http://dx.doi.org/10.1371/journal.pgen.1000827.

E. Nováková, F. Husník, E. Šochová, and V. Hypša. *Arsenophonus* and *Sodalis* symbionts in louse flies: an analogy to the *Wigglesworthia* and *Sodalis* system in tsetse flies. *Applied and Environmental Microbiology*, 81(18):6189–6199, 2015. doi: 10.1128/aem.01487-15. URL http://dx.doi.org/10.1128/AEM.01487-15.

D. Nychka, R. Furrer, J. Paige, and S. Sain. *fields: Tools for Spatial Data*, 2016. URL https://CRAN.R-project.org/package=fields. R package version 8.3-6.

K. F. Oakeson, R. Gil, A. L. Clayton, D. M. Dunn, A. C. von Niederhausern, C. Hamil, A. Aoyagi, B. Duval, A. Baca, F. J. Silva, A. Vallier, D. G. Jackson, A. Latorre, R. B. Weiss, A. Heddi, A. Moya, and C. Dale. Genome degeneration and adaptation in a nascent stage of symbiosis. *Genome Biology and Evolution*, 6(1):76–93, 2014. doi: 10.1093/gbe/evt210. URL http://dx.doi.org/10.1093/gbe/evt210.

N. R. Pace and T. L. Marsh. Rna catalysis and the origin of life. *Origins of Life and Evolution of the Biosphere*, 16(2):97–116, 1985. doi: 10.1007/bf01809465. URL http://dx.doi.org/10.1007/BF01809465.

M. E. Pettersson and O. G. Berg. Muller's ratchet in symbiont populations. *Genetica*, 130 (2):199–211, 2006. doi: 10.1007/s10709-006-9007-7. URL http://dx.doi.org/10.1007/s10709-006-9007-7.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL http://www.R-project.org. ISBN 3-900051-07-0.

L. Randau, R. Münch, M. J. Hohn, D. Jahn, and D. Söll. *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′- and 3′-halves. *Nature*, 433(7025):537–541, 2005. doi: 10.1038/nature03233. URL http://dx.doi.org/10.1038/nature03233.

C. Rispe and N. A. Moran. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. *The American Naturalist*, 156(4):425–441, 2000. doi: 10.1086/303396. URL http://dx.doi.org/10.1086/303396.

K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M.-A. Rajandream, and B. Barrell. Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10):944–945, 2000. doi: 10.1093/bioinformatics/16.10.944. URL http://bioinformatics.oxfordjournals.org/content/16/10/944.abstract.

J. C. Salazar, A. Ambrogelly, P. F. Crain, J. A. McCloskey, and D. Soll. From the cover: A truncated aminoacyl-tRNA synthetase modifies RNA. *Proceedings of the National Academy of Sciences*, 101(20):7536–7541, 2004. doi: 10.1073/pnas.0401982101. URL http://dx.doi.org/10.1073/pnas.0401982101.

T. J. W. Sean R. Eddy and the HMMER development team. *HMMER*. R Foundation for Statistical Computing, 3.1b1 edition, 2013. URL http://hmmer.org/.

P. A. Sharp. On the origin of RNA splicing and introns. *Cell*, 42(2):397–400, 1985. doi: 10.1016/0092-8674(85)90092-3. URL http://dx.doi.org/10.1016/0092-8674(85)90092-3.

D. B. Sloan, A. Nakabachi, S. Richards, J. Qu, S. C. Murali, R. A. Gibbs, and N. A. Moran. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular Biology and Evolution*, 31(4):857–871, 2014. doi: 10.1093/molbev/msu004. URL http://dx.doi.org/10.1093/molbev/msu004.

D. R. Smith and P. J. Keeling. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences*, 112(33):10177–10184, 2015. doi: 10.1073/pnas.1422049112. URL http://dx.doi.org/10.1073/pnas.1422049112.

A. Soma, Y. Ikeuchi, S. Kanemasa, K. Kobayashi, N. Ogasawara, T. Ote, J. ichi Kato, K. Watanabe, Y. Sekine, and T. Suzuki. An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Molecular Cell*, 12(3):689–698, 2003. doi: 10.1016/s1097-2765(03)00346-0. URL http://dx.doi.org/10.1016/S1097-2765(03)00346-0.

E. J. Stanbridge and M. E. Reff. The molecular biology of mycoplasmas. In F. B. M and S. Razin, editors, *Mycoplasmas: Cell Biology*, volume 1, chapter Transfer RNA, pages 169 – 176. Academic Pr, 1979. ISBN 0120784017. URL http://www.amazon.com/Mycoplasmas-Biology-M-F-Barile/dp/0120784017%3FSubscriptionId%3D0JYN1NVW651KCA56C102%26tag%3Dtechkie-20%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0120784017.

J. Sugahara, N. Yachie, Y. Sekine, A. Soma, M. Matsui, and M. Tomita et al. SPLITS: a new program for predicting split and intron-containing tRNA genes at the genome level. *In Silico Biology*, 6(5):411–418, 2006. URL http://content.iospress.com/articles/in-silico-biology/isb00254.

J. Sugahara, K. Fujishima, T. Nunoura, Y. Takaki, H. Takami, K. Takai, M. Tomita, and A. Kanai. Genomic heterogeneity in a natural archaeal population suggests a model of tRNA gene disruption. *PLoS ONE*, 7(3):e32504, 2012. doi: 10.1371/journal.pone.0032504. URL http://dx.doi.org/10.1371/journal.pone.0032504.

P. W. Theisen, J. E. Grimwade, A. C. Leonard, J. A. Bogan, and C. E. Helmstetter. Correlation of gene transcription with the time of initiation of chromosome replication in *Escherichia coli*. *Molecular Microbiology*, 10(3):575–584, 1993. doi: 10.1111/j.1365-2958.1993.tb00929.x. URL http://dx.doi.org/10.1111/j.1365-2958.1993.tb00929.x.

C. C. Traverse and H. Ochman. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proceedings of the National Academy of Sciences*, 113(12):3311–3316, 2016. doi: 10.1073/pnas.1525329113. URL http://dx.doi.org/10.1073/pnas.1525329113.

J. Urbonavicius. Improvement of reading frame maintenance is a common function for several tRNA modifications. *The EMBO Journal*, 20(17):4863–4873, 2001. doi: 10.1093/emboj/20.17.4863. URL http://dx.doi.org/10.1093/emboj/20.17.4863.

A. Vigneron, F. Masson, A. Vallier, S. Balmand, M. Rey, C. Vincent-Monégat, E. Aksoy, E. Aubailly-Giraud, A. Zaidman-Rémy, and A. Heddi. Insects recycle endosymbionts when the benefit is over. *Current Biology*, 24(19):2267–2273, 2014. doi: 10.1016/j.cub.2014.07.065. URL http://dx.doi.org/10.1016/j.cub.2014.07.065.

C. D. von Dohlen, S. Kohler, S. T. Alsop, and W. R. McManus. Mealybug beta-proteobacterial endosymbionts contain gamma-proteobacterial symbionts. *Nature*, 412(6845):433–436, 2001. doi: 10.1038/35086563. URL http://dx.doi.org/10.1038/35086563.

J. J. Wernegreen. Genome evolution in bacterial endosymbionts of insects. *Nature Reviews Genetics*, 3 (11):850–861, 2002. doi: 10.1038/nrg931. URL http://dx.doi.org/10.1038/nrg931.

Wickham and Hadley. Reshaping data with the reshape package. *Journal of Statistical Software*, 21 (12), 2007. URL http://www.jstatsoft.org/v21/i12/paper.

H. Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40 (1):1–29, 2011. URL http://www.jstatsoft.org/v40/i01/.

J. Wolf. tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. *The EMBO Journal*, 21(14):3841–3851, 2002. doi: 10.1093/emboj/cdf362. URL http://dx.doi.org/10.1093/emboj/cdf362.

M. Woolfit. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*, 20(9):1545–1555, 2003. doi: 10.1093/molbev/msg167. URL http://dx.doi.org/10.1093/molbev/msg167.

Y. Xie. *Implementing Reproducible Research (Chapman & Hall/CRC The R Series)*. Chapman and Hall/CRC, 2014. ISBN 1466561599.

Y. Xie. *Dynamic Documents with R and knitr, Second Edition (Chapman & Hall/CRC The R Series)*. Chapman and Hall/CRC, 2015. ISBN 1498716962.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL https://CRAN.R-project.org/package=knitr. R package version 1.12.3.

B. E. Yacoubi, B. Lyons, Y. Cruz, R. Reddy, B. Nordin, F. Agnelli, J. R. Williamson, P. Schimmel, M. A. Swairjo, and V. de Crecy-Lagard. The universal YrdC/Sua5 family is required for the formation of threonylcarbamoyladenosine in tRNA. *Nucleic Acids Research*, 37(9):2894–2909, 2009. doi: 10.1093/nar/gkp152. URL http://dx.doi.org/10.1093/nar/gkp152.

# Supplementary material

## Parsers

Note: All parsers are stored in the "Parsers" folder
Note 2: all R scripts were tested on R version 3.2.3

### gbk2faa and gbk2fna

**Language:** zsh
**Description:** Perl scripts converting files in gbk formats into faa (amino acid sequences) or fna (one nucleotide sequence) formats. They were originally written by Cai and than slightly modified by myself to get more suitable format for this study. My changes are marked with comment "NEW"
**Input file(s):** genome files in gbk format
**Output file(s):** genome files in faa (gbk2faa ) or fna (gbk2fna ) format
**Software required:** -

### run_tFind.sh

**Language:** zsh
**Description:** A zsh script which runs tFind on files in fna format, generates a summary table of all tRNAs found and prepares faa tRNA sequences for each anticodon
**Input file(s):** files with genome sequences in faa format (one file for each sequence)
**Output file(s):** tRNAs separated in distinct files by anticodons, table containing five columns: specie, amino acid, anticodon and tFind score
**Software required:** tFind (Hudson and Williams, 2014)

### run_cmalign.sh

**Language:** zsh
**Description:** aligns input tRNAs according to Rfam databases using Infernal
**Input file(s):** files with tRNAs in fna format, Rfam database of all bacterial tRNAs
**Output file(s):** RNA alignments in fna format
**Software required:** Infernal (Nawrocki and Eddy, 2013), a script for conversion of alignment format (Katz, 2012)

### tRNA_positions7.R

**Language:** R
**Description:** gets tRNA modification sites and finds out whether they are conserved or not.

**Input file(s):** alignments of tRNAs in fna format, table of modification sites
**Output file(s):** -
**Software required:** R packages Biostrings (H Pages and DebRoy) and plyr (Wickham, 2011)

**script run_orthomcl_blastp_FINAL.sh**

**Language:** zsh
**Description:** runs OrthoMCL and blastP and re-formate their outputs to tabular format
**Input file(s):** proteomes of endosymbionts in faa format, *E. coli* modification proteins in faa format
**Output file(s):** Table assigning endosymbionts proteins to their OrthoMCL cluster and two tables containing information about all blast hits
**Software required:** OrthoMCL (Li, 2003), BlastP

**proteins_cluster_identification4.R**

**Language:** R
**Description:** helps user to assign particular OrthoMCL clusters to individual query proteins using blastP results
**Input file(s):** two tables with information about blast matches, a table with all endosymbionts genes assigned to OrthoMCL clusters and a table assigning particular endosymbionts lineages
**Output file(s):** a table containing query gene names with endosymbionts genes assigned .
**Software required:** R package pander (Gergely Daróczi, 2015)

**run_hmm.sh**

**Language:** zsh
**Description:** finds additional genes of interest which were not detected by OrthoMCL. User has to check and select genes manually.
**Input file(s):** faa file containing all edosymbionts proteins, table containing information about genes already assigned to *E. coli* modification proteins
**Output file(s):** the input file extended of newly found genes
**Software required:** HMMer (Sean R. Eddy and the HMMER development team, 2013)

**Statistics_tRNA.R**

**Language:** R
**Description:** computes average data of total tRNA counts and average tRNA score for

each genus and statistically tests several hypotheses

**Input file(s):** table with information about endosymbionts tRNA and table with information about lineages

**Output file(s):** -

**Software required:** Rpackage: reshape (Wickham and Hadley, 2007) and lmconf - an unpublished function (Fibich, unpublished)

### Statistics_proteins.R

**Language:** R

**Description:** computes average data of total modification proteins counts for each genus and statistically tests several hypotheses

**Input file(s):** table with information about endosymbionts tRNA, table with information about lineages, table with information about enzymes, table informing which modifications can be done by bacteria and which both by bacteria and which enzyme corresponds to which modifications

**Output file(s):** -

**Software required:**

## Raw data

Note: These files are stored in the "Raw_data" folder

### tRNAs

**File name:** All_tRNAs.trnaraw

**Description:** A file containing information about all tRNAs identified in endosymbionts genomes including tRNAs scores, secondary structures and sequences. This file can be accessed using a text editor.

**Format:** the same as the trnaraw output of the tFind script

**Example line:** cl;307128504;man;NC_014499.1;Sulcia_Muelleri_CARI;tFind;tRNA;145121;145194;74;+;.;product=tRNA-Asp(GTC);detected_by=tRNAscan-SE_145121_145194_+_Asp,Aragorn_145121_145197_+_Asp;cove_score=77.37;amino_acid=Asp;anticodon=GTC;anticodon_center=36;t-stem_acc-stem_junction=67;structure=»»»»>..»».....««.»»>.....«««<.....»»»>.....«««««««.;sequence=GGTCTGGTAGTTCAGATGGTTAGAATACGTGCCTGTCACGCACGGGGTCACGGGTTCGAATCCCGTCCAGACCG;cca=Ctt;

### Genes encoding modification proteins

**File name:** genes_final_withHMM2.txt

**Description:** A file containing information about the final set of endosymbionts genes which were identified as ones encoding modification proteins. This file can be assesed using a text editor.

**Format:** ClusterID;Gene_ID;annotation_of_the_gene;endosymbionts_lineage

**Example line:** 755;YP_005060828.1;tRNA-dihydrouridine_synthase_A; Serratia_symbiotica;DusA_3183570

### rRNA alignments

**File name:** rRNA_alignments.geneious

**Description:** A geneious file containing alignments of genes encoding 16S and 23S rRNA of endosymbionts and *Escherichia coli* with annotated modification sites. This archive can be accesed *via* Geneious only. (The easiest way how to open it is to drag it into Geneious into some folder).

**Format:** geneious

## Tables

| | Species | Plasmid ID |
|---|---|---|
| 1 | *Blattabacterium BGIGA* | NC 017925.1 |
| 2 | *Blattabacterium BNCIN* | NC 022551.1 |
| 3 | *Blattabacterium BPLAN* | NC 013419.1 |
| 4 | *Blattabacterium Cpu* | NC 016598.1 |
| 5 | *Blattabacterium cuenoti Bge* | NC 015679.1 |
| 6 | *Blattabacterium cuenoti Tarazona* | NC 020196.1 |
| 7 | *Blattabacterium MADAR* | NC 016150.1 |
| 8 | *Buchnera aphidicola AK* | NC 017257.1 |
| 9 | *Buchnera aphidicola AK* | NC 017258.1 |
| 10 | *Buchnera aphidicola Bp* | NC 004555.1 |
| 11 | *Buchnera aphidicola Cc* | NC 011878.1 |
| 12 | *Buchnera aphidicola Ua* | NC 017261.1 |
| 13 | *Ishikawaella capsulata* | AP010873.1 |
| 14 | *Profftella armatura* | NC 021886.1 |
| 15 | *Riesia pediculicola* | NC 013962.1 |
| 16 | *Tremblaya phenacola* | NC 021553.1 |
| 17 | *Wigglesworthia glossinidia GB* | NC 003425.1 |

Table 5: Ncbi IDs of plasmids used in this study

| | Modification | Enzyme | B or E |
|---|---|---|---|
| 1 | cmnm5s2U | MnmE | B |
| 2 | cmnm5Um | MnmE | B |
| 3 | cmnm6Um | TrmL | B |
| 4 | cmo5U | CmoA | B |
| 5 | cmo5U | CmoB | B |
| 6 | m2A | RlmN | B |
| 7 | m3$\psi$ | RlmH | B |
| 8 | m3U | RsmE | B |
| 9 | m4Cm | RsmI | B |
| 10 | ms2i6A | MiaB | B |
| 11 | m6t6A | TrmO | B |
| 12 | mcmo5U | CmoA | B |
| 13 | mcmo5U | CmoB | B |
| 14 | mnm5s2U | MnmA | B |
| 15 | mnm5s2U | MnmCD | B |
| 16 | mnm5s2U | MnmE | B |
| 17 | mnm5se2U | MnmH | B |
| 18 | mnm5U | MnmE | B |
| 19 | mnm5U | MnmCD | B |
| 20 | s2C | TtcA | B |
| 21 | s4U | ThiI | B |
| 22 | ac4C | TmcA | E |
| 23 | Cm | RlmM | E |
| 24 | Cm | TrmJ | E |
| 25 | Cm | TrmL | E |
| 26 | D | DusA | E |
| 27 | D | DusB | E |
| 28 | D | DusC | E |
| 29 | Gm | RlmB | E |
| 30 | Gm | TrmH | E |
| 31 | I | TadA | E |
| 32 | k2C | TilS | E |
| 33 | m1A | NpmA | E |
| 34 | m1G | RlmA | E |
| 35 | m1G | TrmD | E |
| 36 | m5U | RlmCD | E |
| 37 | m5U | TrmA | E |
| 38 | m6A | ErmBC | E |
| 39 | m6A | RlmF | E |
| 40 | m6A | RlmJ | E |
| 41 | m6A | TrmM | E |
| 42 | m7G | ArmA | E |
| 43 | m7G | RlmKL | E |
| 44 | $\psi$ | RluA | E |
| 45 | $\psi$ | RluB | E |
| 46 | $\psi$ | RluC | E |
| 47 | $\psi$ | RluD | E |
| 48 | $\psi$ | RluE | E |
| 49 | $\psi$ | RluF | E |
| 50 | $\psi$ | RsuA | E |
| 51 | $\psi$ | TruA | E |
| 52 | $\psi$ | TruB | E |
| 53 | $\psi$ | TruC | E |
| 54 | $\psi$ | TruD | E |
| 55 | Q | QueA | E |
| 56 | Q | QueF | E |
| 57 | Q | QueG | E |
| 58 | Q | Tgt | E |
| 59 | t6A | TsaB | E |
| 60 | t6A | TsaC | E |
| 61 | t6A | TsaD | E |
| 62 | t6A | TsaE | E |
| 63 | Um | RlmE | E |
| 64 | Um | TrmJ | E |

Table 6: Bacterial specific (B) and non-specfic modifications (E; Grosjean 2009) provided by enzymes used in this study.