Trends in Cell & Molecular Biology
Vol. 7, 2012

Original Communication

Decoding in Candidatus *Riesia pediculicola,* close to a minimal tRNA modification set?

Valérie de Crécy-Lagard[1,*], Christian Marck[2], and Henri Grosjean[3]

[1]Department of Microbiology and Cell Science, University of Florida, P.O. Box 110700, Gainesville, FL 32611-0700, USA. [2]Institut de Biologie et de Technologies de Saclay (iBiTec-S) Bât 144, CEA/Saclay, F-91191 Gif-sur-Yvette Cedex, [3]Centre de Génétique Moléculaire, UPR 3404, CNRS, Associée à l'Université Paris-Sud 11, FRC 3115, 91190 Gif-sur-Yvette, France

ABSTRACT

A comparative genomic analysis of the recently sequenced human body louse unicellular endosymbiont Candidatus *Riesia pediculicola* with a reduced genome (582 Kb), revealed that it is the only known organism that might have lost all post-transcriptional base and ribose modifications of the tRNA body, retaining only modifications of the anticodon-stem-loop essential for mRNA decoding. Such a minimal tRNA modification set was not observed in other insect symbionts or in parasitic unicellular bacteria, such as *Mycoplasma genitalium* (580 Kb), that have also evolved by considerably reducing their genomes. This could be an example of a minimal tRNA modification set required for life, a question that has been at the center of the field for many years, especially for understanding the emergence and evolution of the genetic code.

KEYWORDS: tRNA, maturation, translation, modified nucleosides, comparative genomics

ABBREVIATIONS

Full names for the different acronyms used to define a given modified base can be found in [1].

INTRODUCTION

As adapters between the mRNA and the elongating peptide, tRNAs are the central decoding molecules in translation. Their overall efficiency in protein synthesis depends both on the sequence/structure of the whole set of the tRNA repertoire and on modified nucleotides that are formed during the tRNA maturation process. Depending on the organism considered, a single functional tRNA isoacceptor may contain from 2 to 17 modified nucleosides [2]. These post-transcriptional modifications are required to maintain tRNA structure, insure correct mRNA decoding, optimize translation accuracy and efficiency, and/or regulate tRNA turn-over or its cellular localization (reviewed in: [3] and [4]).

Several studies have attempted to define a minimal, possibly ancestral tRNA modification set. By comparing the modification profiles in all available sequenced tRNAs from different kingdoms (Bacteria, Eucarya and Archaea, a total of about 500 tRNA in 1998), it was predicted that eight, possibly nine, modifications were present in the putative last universal common ancestor (LUCA or Cenancestor) [5, 6]. These modifications are the Ψ residues at positions 13, 38, 39, 55, C_m at position 34 [5] or C_m at position 32 [6], Q at position 34, t^6A and m^1G adjacent to the anticodon at position 37, and m^1A at position 58 in the highly conserved sequence of the so-called TΨ-loop. Another study that combined comparative

*Corresponding author
vcrecy@ufl.edu

genomics and essentiality data predicted that LUCA harbored only three modifications [7]: Q34, Ψ13, and Ψ39. Finally Church and colleagues proposed that six modifications (k^2C34, xs^2U34 derivatives, I34, m^1G37, t^6A37, ms^2i^6A37) are required for the minimal bacterial translation set [8]. The discrepancies are due to inherent flaws in all the prediction methods used. Predictions based solely on gene essentiality can be misleading, as a dispensable tRNA modification can become essential if other modifications are missing [9, 10]. Moreover, ancient-primordial genes may have considerably diverged in different phyla of organisms, so that they are now unrecognizable by any sequence-relatedness algorithms [11]. Alternatively, distinct enzyme families can introduce the exact same modification (functional type of enzyme evolution [12, 13, 14]). These will also be missed with methods based on ortholog searches. Finally, several genes of unknown function predicted to be present in LUCA [7] have since turned out to be involved in tRNA modification [15-17] and had therefore been missed in previous searches.

The idea of defining an “absolute minimal” set when talking about tRNA modifications might be inherently flawed and probably elusive. First, parallel and convergent solutions are deployed by different organisms both for modifications involved in decoding (discussed below) and in maintaining tRNA structural integrity. For example, ribothymidine, (m^5U54) that is critical for tRNA stability in bacteria [18, 19], is replaced by $m^1Ψ$ or Um in many Archaea [2, 20, 21]. Likewise, different modified uridines are used at the wobble position of tRNA to fulfill decoding requirements in different organisms [22]. Second, one cannot separate nucleoside modifications from the sequence context of a given tRNA repertoire as there is a clear co-evolution between the two sets. For example, the *tilS* gene responsible for the k^2C (lysidine) modification at the wobble position 34 was lost in *Mycoplasma mobile*. This loss occurred with a concomitant change of the sequence of the minor $tRNA^{Ile}$ that decodes AUA codons from a CAU to a UAU anticodon [23, 24], a cellular strategy that has been experimentally verified in *B. subtilis* [25]. Third, the G+C content at the third codon position conditions the use of modified bases at the wobble position of tRNA [26]. Lastly, the requirements for modifications are going to be extremely dependent on environmental and physiological factors and will hence vary from one organism to another, for example, halophiles are predicted to require less modifications (see discussion of [24]). It is therefore not a minimal tRNA modification set but a minimal set of organism specific functional constraints that needs to be defined. An efficient and biologically relevant method to tentatively identify these minimal sets of essential tRNA modification enzymes, possibly the most reluctant to be lost during cellular evolution, is to analyze organisms with reduced genomes, such as parasitic or symbiotic intracellular and extracellular bacteria.

tRNA modification sets in Mollicutes

Mollicutes are parasitic, small unicellular bacteria normally living within eukaryotic cells. They originated from gram-positive bacteria (phylum: Firmicutes) by considerably reducing their genomes [27]. The Mollicute with the smallest genome identified so far is *Mycoplasma genitalium* (580 kb encoding 480 predicted ORFs) [28]. When cultivated in extremely rich medium, several of these Mollicutes can grow as free-living organisms, albeit very poorly and thus are considered to have minimal genomes [29]. In agreement with gene economization strategies, all Mollicutes display a minimalist, non-redundant set of tRNAs (from 28 to 35 with distinct anticodons), that is sufficient to decode all sense codons corresponding to 20 canonical amino acids [24]. In this same study, we analyzed the presence or absence of genes coding for corresponding enzymes and predicted the tRNA modifications sets in 15 Mollicutes covering the four major clades (*Spiroplasma, Pneumonia, Hominis* and *Phytoplasma)*. The genes were identified by homology with model systems such as *Escherichia coli* and *Bacillus subtilis*, and further validated from the knowledge of the modified nucleosides in the full set of 29 sequenced tRNAs of *Mycoplasma capricolum* [24]. The main conclusion was that only a few modification enzymes, all acting on nucleotides of the anticodon loop in tRNA (m^1G37, t^6A37 and $cmnm^5$U34), seemed resistant to gene loss. However, all the Mollicutes analyzed retained additional genes coding for enzymes inserting modifications in the tRNA body. For example, TruB catalyzing the Ψ55 insertion and TrmB catalyzing the methylation of G47 (m^7G47) are found in the majority of Mollicutes, and therefore resistant to loss. Inspection of 20 additional complete genome sequences of Mollicutes, made available since this study, does not fundamentally change the initial conclusion (S. Yokobori, H. Grosjean and S. Bessho, personal communication).

tRNA repertoires in insect bacterial symbionts

In the present work, we performed a similar computational analysis of 14 genomes of bacterial symbionts and endosymbionts of insects, covering *Wolbachia* (3 strains which infect arthropod species and some nematodes), *Buchnera* (6 strains, which infect aphids), Candidatus *Blochmannia* (2 strains, which infect bacteriocytes and ant ovaries), *Baumannia cicadellinicola* (infecting bacteriomes of sharpshooter leafhoppers), *Wigglesworthia glossinidia* (infecting the gut of the tsetse fly) and Candidatus *Riesia pediculicola* (infecting human body louse). All of these species are derived from gram-negative Proteobacteria, mainly gamma-proteobacteria and related to *E. coli*, with the exception of *Wolbachia* (an alpha-proteobacteria). Unlike most bacteria and Mollicutes, members of this group cannot live as free-living organisms and form an obligate relationship (intimate symbioses) with their eucaryal hosts. These symbionts are predominantly vertically transmitted along with their host, and thus extend the heritable genetic variation of the host cells [30-33].

The genome sizes of the set of organisms analyzed (Supplemental Table 1) varied from 416 kb with 371 predicted CoDing Sequences (CDSs) (*Buchnera aphidicola str. CC*) to 1,483 kb with 1586 CDSs (*Wolbachia pipientis quinquefasciatus Mel*) (numbers of CDS taken from the Rast server [34]). 557 CDSs have been predicted in C. *R. pediculicola*, but around 80 of these are very small (between 19 and 70 aa) with no homology to any known proteins. These types of small proteins are not found in the other insect symbiont genomes analyzed and might be overpredictions.

Figure 1 (right part) shows that all 14 symbionts analyzed harbor genes coding for a full set of tRNAs able to read all sense codons for the 20 canonical amino acids, indicating that no tRNAs from the host are needed. Like Mollicutes and at variance with bacteria with large genomes, these uncultivable symbionts display a quasi-non-redundant set of tRNAs, with each isoacceptor having a distinct anticodon (compare columns #1 through #14 with column #15 for *E. coli*). The total number of tRNAs varies from 31 for *Buchnera aphidicola* str Cc (#8) to 40 for C. *Blochmannia pennsylvanicus* (#10). These correspond to tRNA repertoires typically found in Bacteria and not in Eucarya and Archaea [22, 35]. For example, tRNA genes containing the wobble T34 and G34 are almost always present, while tRNA genes containing C34 are often absent (blue background in Figure 1). In both of the quartet boxes corresponding to Pro and Ala (boxed in red in Figure 1), only one tRNA gene harboring a wobble T34 is present. For the isoleucine triplet decoding box and the arginine quartet decoding box, the T34-containing tRNA genes are systematically replaced by a C34-containing $tRNA^{Ile}$ and an A34-containing $tRNA^{Arg}$, respectively (indicated with yellow and green background in Figure 1). tRNA usage is usually correlated with codon usage, which in turn controls the efficiency of decoding [36]. By comparing the relative codon usage in each of the decoding box, it appears that G34-containing tRNAs more frequently read codons ending with the wobble U3 while U34-containing tRNAs mainly read codons ending with A3 (Watson-Crick base pairing). When the C34-containing tRNA isoacceptor is absent, U34 also reads codons ending with the wobble G3 (compare information about codon usage on the left part of Figure 1 with the identity of the wobble base in the tRNA, under the column symbol ‘AC’ for anticodon). This trend reflects the low average G+C content in ORFs of insect symbionts analyzed (from 23 to 35% compared to 52% in *E. coli*; Supplemental Table 1), particularly at the third position of codons (data not shown), and reflects the type of modified nucleotide present at the wobble position of tRNA. Non-redundancy of tRNA isoacceptor may affect cellular tRNA abundance, and hence the growth rate of the symbiont [37]. Finally, in contrast to Mycoplasma [24], UGA is a genuine stop codon in these insect symbiotic organisms, correlating with the presence of Release Factor 1 and Release Factor 2 (see the “tRNA modification *E. coli*” subsystem available on the Public SEED, http://pubseed.theseed.org/SubsysEditor.cgi).

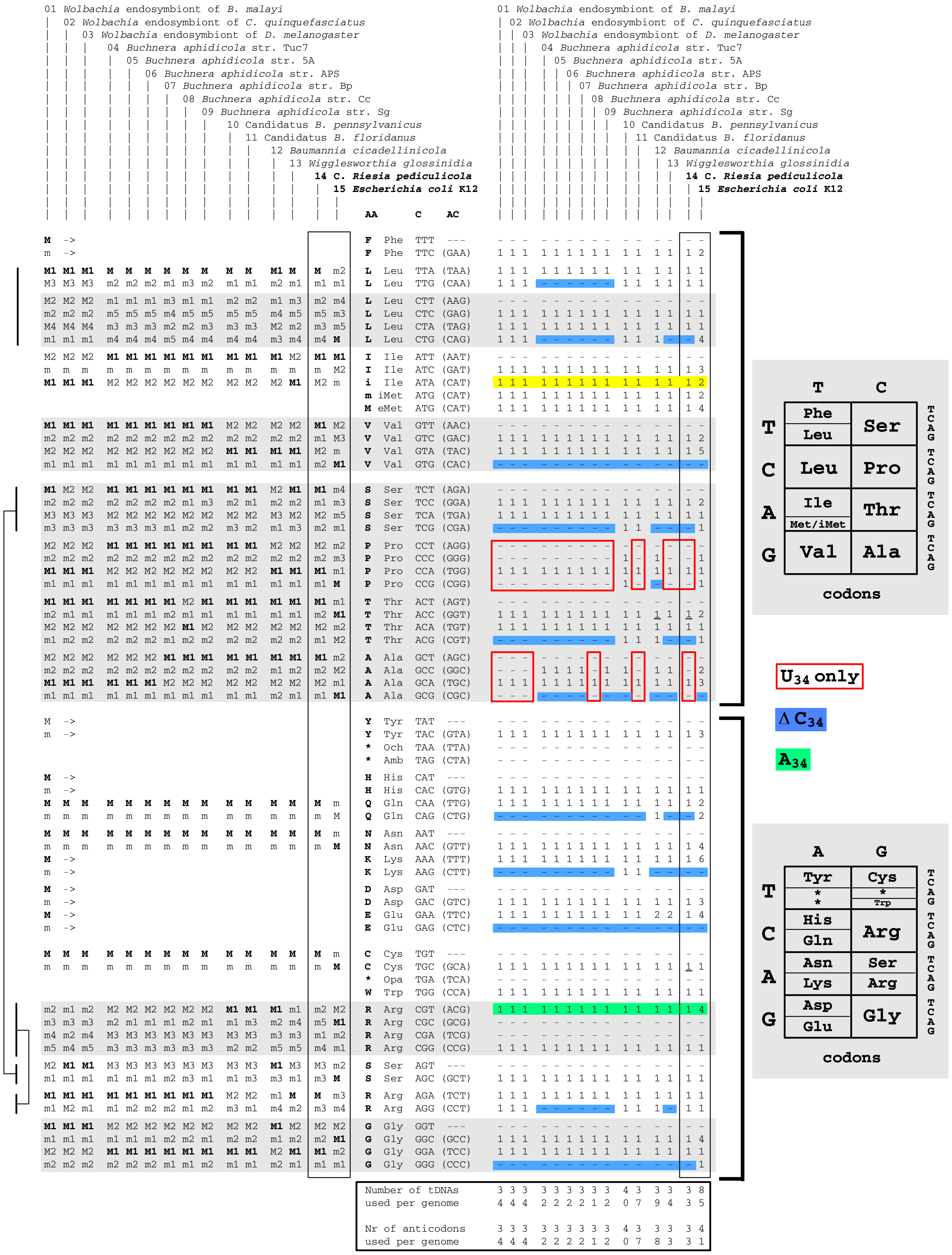
01 Wolbachia endosymbiont of B. malayi
02 Wolbachia endosymbiont of C. quinquefasciatus
03 Wolbachia endosymbiont of D. melanogaster
04 Buchnera aphidicola str. Tuc7
05 Buchnera aphidicola str. 5A
06 Buchnera aphidicola str. APS
07 Buchnera aphidicola str. Bp
08 Buchnera aphidicola str. Cc
09 Buchnera aphidicola str. Sg
10 Candidatus B. pennsylvanicus
11 Candidatus B. floridanus
12 Baumannia cicadellinicola
13 Wigglesworthia glossinidia
14 C. Riesia pediculicola
15 Escherichia coli K12
AA
C
AC
Number of tDNAs used per genome
Nr of anticodons used per genome
T
C
A
G
Phe
Ser
Leu
Pro
Ile
Thr
Met/iMet
Val
Ala
codons
U34 only
Δ C34
A34
Tyr
Cys
Trp
His
Arg
Gln
Asn
Lys
Asp
Gly
Glu

Figure 1

tRNA modifications sets in insect bacterial symbionts

Genes coding for tRNA modification enzymes in the 14 genomes analyzed were identified by BLAST analysis against the genes found in *E. coli* (see [1] and Figure 2 legend). In *E. coli*, all but four of the fully matured isoacceptor tRNAs have been sequenced, and genes coding for most of the corresponding tRNA modification enzymes have been identified. Surprisingly, the recently sequenced human louse endosymbiont C. *R. pediculicola* [38] appears to have lost all modifications of the tRNA body, retaining only a few modifications of the anticodon loop and proximal stem (Figure 2): Ψ at position 38 and 39; I, k^2C, xs^2U derivatives and xo^5U at position 34; and m^1G, t^6A, i^6A and ms^2i^6A at position 37. As it is technically impossible to extract enough tRNA from the human louse symbiont to analyze the modifications profiles, we cannot rule out that additional or unknown modifications are present in this organism. For example the acp^3U47 modification gene has not yet been identified in any organism, and could be present in C. *R. pediculicola* (Figure 2).

An identical analysis was performed on the remaining 13 symbionts (#1 to #13, Figure 3). In some genomes, additional modifications were predicted to be present: s^4U8, s^4U9, D17, 20, 20a, Q34, m^7G46 and Ψ55. However, all symbionts analyzed except C. *R. pediculicola* contain at least one modification outside the anticodon-stem loop (Figure 3).

Decoding strategy of synonymous codons in Candidatus *R. pediculicola*

Analysis of the sequences of both of the louse endosymbiont C. *R. pediculicola* and its host reveals that no eukaryotic genes, including putative tRNA modification enzymes, have been transferred to the insect bacterial genome, and that the genome reduction in C. *R. pediculicola* has not been associated with gene transfer to the host [38]. In the 14 proteobacterial symbionts analyzed, we are confident that the only genes coding for tRNAs and tRNA modification enzymes are those reported in Figures 1, 2 and 3 (except, see Figure 2 legend, for enzymes catalyzing acp^3U47 and m^2A37, for which the corresponding genes in *E. coli* are yet to be identified).

Beside the lack of some tRNA isoacceptors in the insect symbionts (discussed above and Figure 1), both nucleotide identities and post-transcriptional modifications are very similar when comparing tRNA isoacceptors from *E. coli* and C. *R. pediculicola*, attesting closely related and typical bacterial decoding strategies [22]. The differences between the two organisms are indicated in red in Figure 4. The main difference is the

Legend to Figure 1. Codon/anticodon/tDNA usage for the 20 canonical amino acids in the 14 symbiont genomes and *E. coli*. The 15 genomes investigated are listed at the top. Full names are given in Supplementary Table SS1. Codon usage within each amino acid family decoding boxes is denoted by the letters on the left: “M” corresponds to most frequently used codon and “m” to the least used ones, with “M1” > “M2” > “m1” > “m2”, etc… to indicate decreasing frequency of codon usage. Rightwards arrows indicate a similar codon usage frequency among the 15 genomes. Details about codon usage in each of the 15 bacteria analyzed can be found in Supplemental Table SS2. These were obtained from automatic determination of all non-overlapping ORFs of 100 codons or more. Vertical bars at the left indicate the six codons of Leu, Ser and Arg respectively. The four columns in the center list the amino acids (indicated as “**AA**”, one- and three-letter code), the codon (“**C**”) and anticodon (“**AC**”) at DNA level. Anticodons never used are indicated as “---”. Numbers at the right indicate the number of tRNA genes harboring the respective anticodons found in each bacteria. Dash signs indicate absence of corresponding tRNA gene. tRNA gene search was performed with tRNAscan-SE [59], and the structure of each tRNA was carefully inspected for fit to the earlier defined bacterial-type tRNA cloverleaf structure [35]. Only three cases of tRNAs (underlined) with more nucleotide than expected (+1 nt in the D-loop) were found. None of the tRNA genes were found in plasmids. The key to the color code is: light gray background denotes four-codon family boxes encoding a single amino acid; yellow background for AUA codon read by the special Ile-tRNA ($\underline{C}$AU with wobble C34 modified to $\underline{k}^2\underline{C}34$ in mature tRNA, see text); green background for the unique A34-containing $tRNA^{Arg}$ (I34 in mature tRNA, see text); red boxes correspond to ‘quartet’ decoding mode in which a single tRNA with T34 at the gene level reads the four codons; blue background denotes C-sparing strategy, the corresponding codon being read by a U_{34}–containing tRNA. The boxes to the right indicate the standard Genetic Code (split in two).

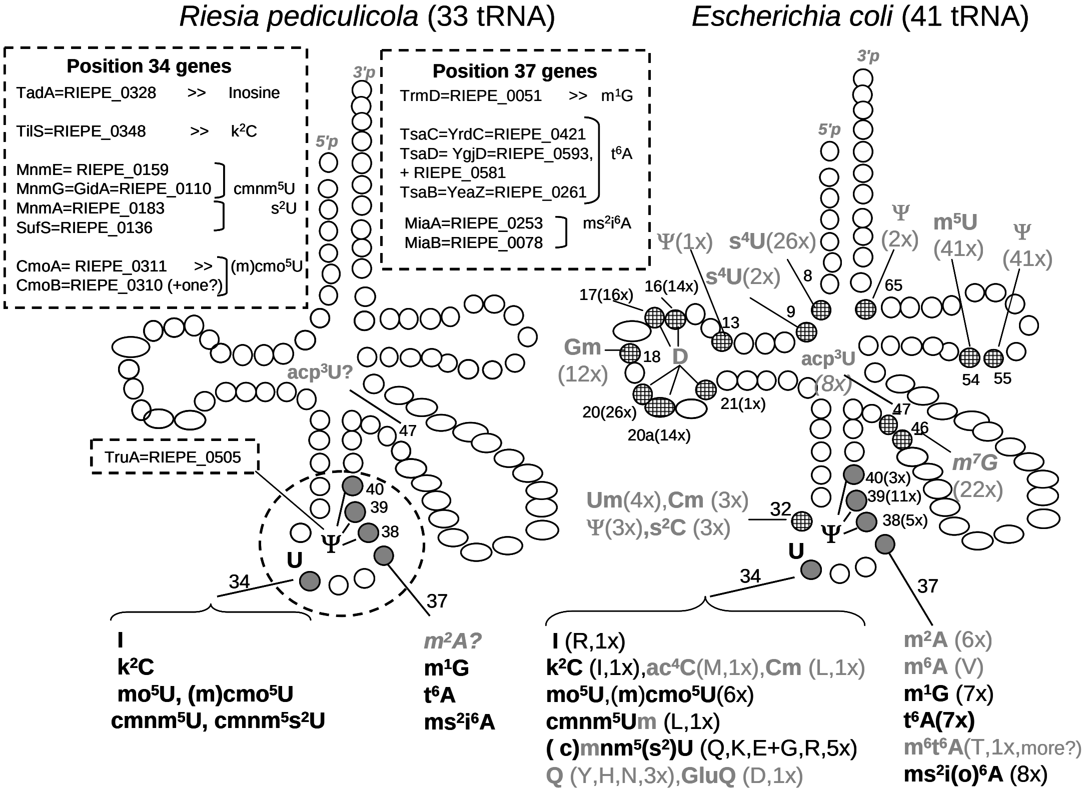

Figure 2

			Extended anticodon region modifications															
	ThiI/s^{4}U8	DusB/D17'20,20a	MnmA/s^{2}U34	MnmE/cmnm^{5}U34	Tgt/Q34	TadA/I34	TilS/k^{2}C34	CmoB/mo^{5}U34	CmoA/(m)cmo^{5}U34	TrmD /m^{1}G37	TsaD/YgjD /t^{6}A37	MiaA /i^{6}A37	MiaB/ms^{2}i^{6}A37	?/m^{2}A37	TruA/ψ39	TrmB/m^{7}G46	?/acp^{3}U47	TruB/ψ55
01 *Wolbachia* endosymbiot of *B. malayi*	□	■	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	□	□	□
02 *Wolbachia* endosymbiot of *C. quinquefasciatus*	□	■	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
03 *Wolbachia* endosymbiot of *D. melanogaster*	□	■	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	□
04 *Buchnera aphidicola* str. Tuc7	□	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
05 *Buchnera aphidicola* str. 5A	□	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
06 *Buchnera aphidicola* str. APS	□	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
07 *Buchnera aphidicola* str. BP	□	□	▨	▨	□	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
08 *Buchnera aphidicola* str. CC	□	□	▨	▨	□	▨	▨	□	□	▨	▨	▨	■	□	▨	□	□	■
09 *Buchnera aphidicola* str. Sg	□	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	■	□	■
10 Candidatus *Blochmania pensylvanicus*	■	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	□	□	▨	□	□	■
11 Candidatus *Blochmania floridanus*	■	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	□	□	▨	□	□	■
12 *Baumannia cicadellinicola*	■	□	▨	▨	■	▨	▨	□	□	▨	▨	▨	■	□	▨	□	□	■
13 *Wigglesworthia glossinidia*	□	□	▨	▨	□	▨	▨	□	□	▨	▨	▨	□	□	▨	□	□	■
14 Candidatus *Riesia pediculicola*	□	□	▨	▨	□	▨	▨	▨	▨	▨	▨	▨	■	□	▨	□	□	□
15 *Escherichia coli* K12	■	■	▨	▨	■	▨	▨	▨	▨	▨	▨	▨	■	□	▨	■	□	■

Figure 3

additional modification of several *E. coli* bases. These are 2'*O*-methylation of C32 (Cm) in tRNA^{Ser} (*UGA) and tRNA^{Trp} (CCA) or U32 (Um) in tRNAs specific for Pro, His and Gln, as well as C34 (Cm) or U34 (Um) in *E. coli* tRNA^{Leu} (*UmAA) and tRNA^{Leu} (CmAA). Many complex modifications are predicted to be missing in C. *R. pediculicola* tRNAs: the acetyl group in elongator tRNA^{Met} (ac^{4}C34), the *N6*-methylations of A37 or t^{6}A37 in tRNA^{Thr} (GGU/UGU), Q34 or GluQ34 in tRNA specific for Tyr, His, Asn and Asp, and the thio-group on C32 in several tRNA for Arg and Ser. The mnm^{5}U34 modification found in some *E. coli* Gln, Lys, Glu, Arg and Gly tRNAs, is predicted to be cmnm^{5}U34 in the corresponding C. *R. pediculicola* tRNAs, because of the presence of MnmE and MnmG and the absence of MnmC. This last enzyme normally catalyzes the stepwise decarboxylation of the '<u>cm</u>nm' group attached to *C5* of U34, followed by methylation of the resulting '<u>n</u>m' group into the final product <u>m</u>nm^{5}U34 [39]. An alternative mnm^{5}U34 biosynthetic pathway using ammonium instead of glycine as a cofactor has been demonstrated in

Legend to Figure 2. Prediction of the tRNA modifications present in C. *R. pediculicola* and comparison with *E. coli.* The analysis of the modification genes present in the genome of C. *R. pediculicola* was performed using the SEED database. We constructed a subsystem containing all known *E. coli* tRNA modifications genes (see "tRNA modification E. coli" subsystem available on the Public SEED, http://pubseed.theseed.org/SubsysEditor.cgi) and extended it to C. *R. pediculicola.* A manual search of the genome (NC_014109, NC_013962) using BlastP and tBlastN [60] with the *E. coli* proteins from the "tRNA modification *E.coli*" subsystem as input was performed. The gene list used is also found in Table 1 of [61] with the addition of the gene encoding TsaA involved in m^{6}t^{6}A formation (T. Suzuki and V. de Crécy-lagard, personal communication) and TsaD/YgjD involved in t^{6}A formation [16]. In *E. coli*, IscS and TusABCDE are required for thiol transfer [3], but no TusACDE homologs were found in C. *R. pediculicola* and SufS is the only IscS homolog in this organism. The m^{2}A37 methylase encoding gene has not been identified in any organism. The same is true for the acp^{3}U47 gene, hence the question marks. We previously predicted *yfiF* encodes the missing methylase [62], but this has not been experimentally validated. No *yfiF* homolog or no other methylase of unknown function could be identified in the C. *R. pediculicola* genome, making the presence of m^{2}A37 in this organism unlikely. Finally, to make sure no other genes had been missed, all known tRNA modification genes from *B. subtilis* and *S. cerevisiae* were queried in C. *R. pediculicola* (using the subsystems "tRNA modification Bacteria", "tRNA modification yeast cytoplasmic" and "tRNA modification yeast mitochondrial" [63]). The genes present in C. *R. pediculicola* are listed in the dashed boxes, with prediction of the resulting modification. Assuming that gene products in C. *R. pediculicola* exhibit the same specificity as the *E. coli* homologs, one can predict which modifications are found in the 33 tRNAs of the symbiont. They are all localized in the anticodon loop and proximal stem (indicated by numbered grey circles, the whole cluster of modified nucleotides being encircled by dashed line). Only acp^{3}U, normally present at position 47, cannot be excluded because the gene coding for the corresponding enzyme is unknown. For the same reason, it is not certain if m^{2}A37 is present. For comparison, the same tRNA cloverleaf is shown with all the modified nucleotides identified so far by sequencing the 37 fully mature *E. coli* tRNA, as indicated in Figure 1 (only 4 isoacceptor tRNA remain to be sequenced, see Figure 4). The modified nucleotides common to both bacteria are indicated in black, while the ones found only in *E. coli* are indicated in grey. In brackets, the number of isoacceptor tRNAs containing a given modification is indicated. When this number is low, the identity of the modified tRNA is also indicated using the one letter code for amino acid. Open circles correspond to positions in *E. coli* tRNAs where no modification has been found. This compilation was adapted from previously published data [2, 3]. Full names for the different acronyms used to define a given modified base can be found in the MODOMICS database [1].

Legend to Figure 3. Prediction of tRNA modifications present in insect symbionts. A signature gene was chosen for every modification and the distribution of the genes analyzed in all genomes listed in Figure 1 by adding them to the "tRNA_modification_*E._coli*" subsystem on the Public SEED server. Only the genes that were found in at least one of the genomes analyzed other than *E. coli* are shown, with the exception of the ones responsible for m^{2}A37 and acp^{3}U47 modifications that have yet to be identified in *E. coli*. Grey boxes denote genes present in all genomes analyzed. Black boxes denote genes present in *E. coli* and in some of the symbiotic genomes. White boxes denote that a specific gene is missing in a specific organism.

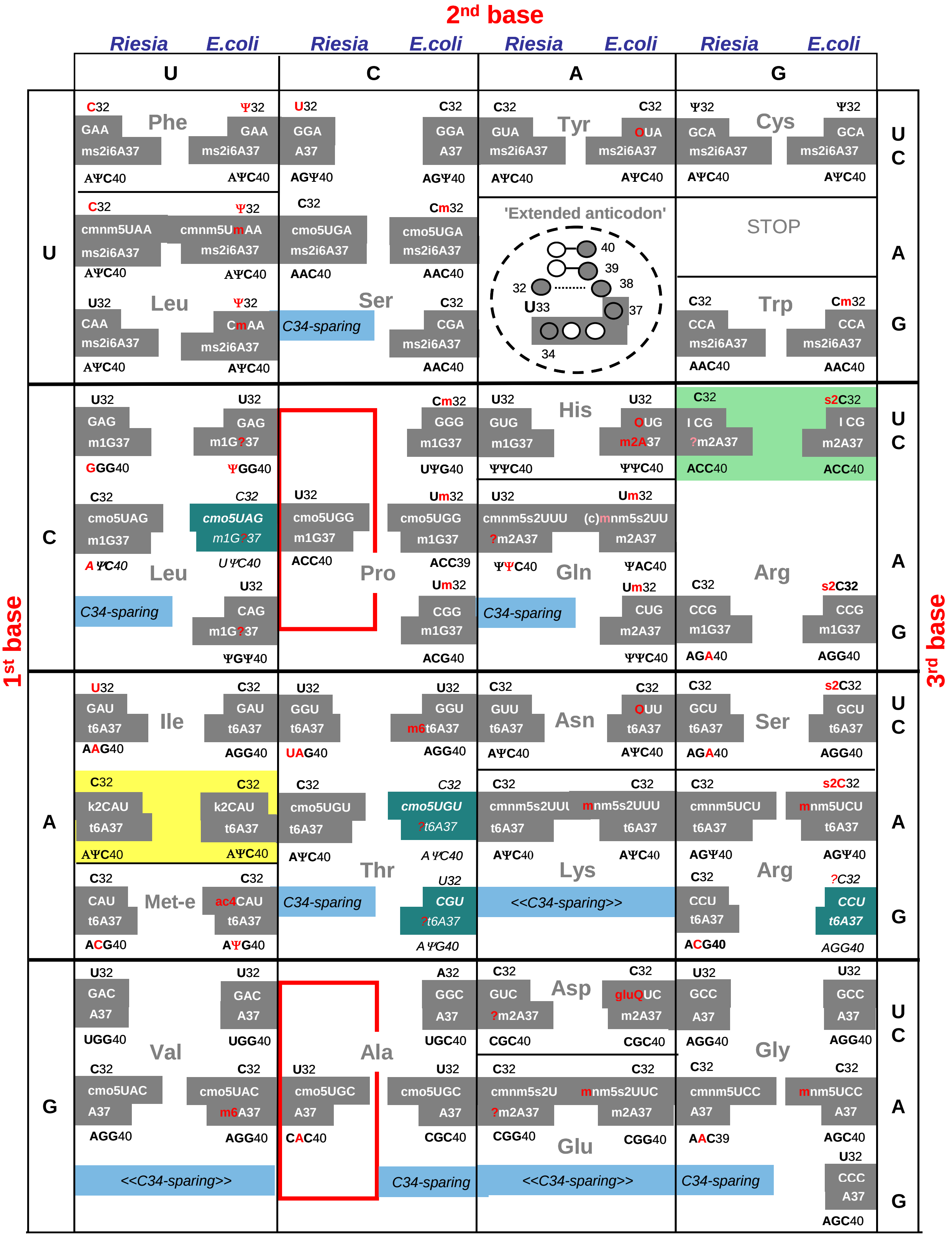
2nd base
Riesia
E.coli
U
C
A
G
1st base
3rd base
C32 Phe Ψ32
GAA GAA
ms2i6A37 ms2i6A37
AΨC40 AΨC40
C32 Ψ32
cmnm5UAA cmnm5UmAA
ms2i6A37 ms2i6A37
AΨC40 AΨC40
U32 Leu Ψ32
CAA CmAA
ms2i6A37 ms2i6A37
AΨC40 AΨC40
U32 C32
GGA GGA
A37 A37
AGΨ40 AGΨ40
C32 Cm32
cmo5UGA cmo5UGA
ms2i6A37 ms2i6A37
AAC40 AAC40
Ser
C34-sparing
C32
CGA
ms2i6A37
AAC40
C32 Tyr C32
GUA OUA
ms2i6A37 ms2i6A37
AΨC40 AΨC40
'Extended anticodon'
40
39
38
32
U33
37
34
Ψ32 Cys Ψ32
GCA GCA
ms2i6A37 ms2i6A37
AΨC40 AΨC40
STOP
C32 Trp Cm32
CCA CCA
ms2i6A37 ms2i6A37
AAC40 AAC40
U32 U32
GAG GAG
m1G37 m1G?37
GGG40 ΨGG40
C32 C32
cmo5UAG cmo5UAG
m1G37 m1G?37
AΨC40 UΨC40
Leu
U32
C34-sparing
CAG
m1G?37
ΨGΨ40
Cm32
GGG
m1G37
UΨG40
U32 Um32
cmo5UGG cmo5UGG
m1G37 m1G37
ACC40 ACC39
Pro
Um32
CGG
m1G37
ACG40
U32 His U32
GUG OUG
m1G37 m2A37
ΨΨC40 ΨΨC40
U32 Um32
cmnm5s2UUU (c)mnm5s2UU
?m2A37 m2A37
ΨΨC40 Gln ΨAC40
Um32
C34-sparing
CUG
m2A37
ΨΨC40
C32 s2C32
I CG I CG
?m2A37 m2A37
ACC40 ACC40
C32 Arg s2C32
CCG CCG
m1G37 m1G37
AGA40 AGG40
U32 C32
GAU Ile GAU
t6A37 t6A37
AAG40 AGG40
C32 C32
k2CAU k2CAU
t6A37 t6A37
AΨC40 AΨC40
C32 C32
CAU Met-e ac4CAU
t6A37 t6A37
ACG40 AΨG40
U32 U32
GGU GGU
t6A37 m6t6A37
UAG40 AGG40
C32 C32
cmo5UGU cmo5UGU
t6A37 ?t6A37
AΨC40 AΨC40
Thr
U32
C34-sparing
CGU
?t6A37
AΨG40
C32 C32
GUU Asn OUU
t6A37 t6A37
AΨC40 AΨC40
C32 C32
cmnm5s2UUU mnm5s2UUU
t6A37 t6A37
AΨC40 Lys AΨC40
<<C34-sparing>>
C32 s2C32
GCU Ser GCU
t6A37 t6A37
AGA40 AGG40
C32 s2C32
cmnm5UCU mnm5UCU
t6A37 t6A37
AGΨ40 AGΨ40
C32 Arg ?C32
CCU CCU
t6A37 t6A37
ACG40 AGG40
U32 U32
GAC GAC
A37 A37
UGG40 UGG40
Val
C32 C32
cmo5UAC cmo5UAC
A37 m6A37
AGG40 AGG40
<<C34-sparing>>
A32
GGC
A37
UGC40
Ala
U32 U32
cmo5UGC cmo5UGC
A37 A37
CAC40 CGC40
C34-sparing
C32 Asp C32
GUC gluQUC
?m2A37 m2A37
CGC40 CGC40
C32 C32
cmnm5s2U mnm5s2UUC
?m2A37 m2A37
CGG40 Glu CGG40
<<C34-sparing>>
U32 U32
GCC GCC
A37 A37
AGG40 Gly AGG40
C32 C32
cmnm5UCC mnm5UCC
A37 A37
AAC39 AGC40
U32
C34-sparing
CCC
A37
AGC40

Figure 4

E. coli MnmE/MnmG mutants *in vitro* [40]. Therefore, it is possible that such an alternative ammonium mediated biosynthetic pathway leading to the final $\underline{\text{n}}\text{m}^5$U34 derivative is used in the insect symbiont (thus by-passing the formation of $\underline{\text{cm}}\text{nm}^5$U34). The cmo^{5}U34 modification is found in *E. coli* tRNAs belonging to quartet decoding boxes for Leu (anticodon *UAG), Val (*UAC), Ser (*UGA), Pro (*UGG), Thr (*UGU) and Ala (*UGC). Its synthesis requires at least three enzymes, only two of which are known (CmoA-*yecO* and CmoB-*yecP* [41, 42]). The methylester of cmo^{5}U at position 34 ($\underline{\text{m}}\text{cmo}^5$U not mentioned in any of the cases in Figure 4) is reported to be base labile and thus only cmo^{5}U is usually detected during most analyses of modified nucleosides. In *E. coli*, tRNA$^{\text{Ser}}$ and tRNA$^{\text{Ala}}$, but not tRNA$^{\text{Val}}$ were reported to be substrates for the *E. coli* and *Salmonella* CmoA methyltransferase [3]. Remarkably, genes coding for CmoA-CmoB are found in C. *R. pediculicola* but absent in all the other 13 insect symbionts analyzed (Figure 3) as well as in Mycoplasma [24, 43]. This suggests that the cmo^{5}U34 modification is dispensable, and *cmoA/cmoB* could be the next set of modification gene lost by C. *R. pediculicola*. Alternatively, the maintenance of cmo^{5}U in several C. *R. pediculicola* tRNAs could result from subtle decoding constraints specific to that organism. Several studies exploring the function of (m)cmo^{5}U derivatives versus ho^{5}U modified U34 [3, 41, 42] concluded that the (m)cmo group added to the *C5* atom of the wobble U base enhances the ability of the tRNA to pair with all four codons, a property that was also demonstrated for a non-modified wobble U-base [44]. These observations again suggest that U34 modification of tRNA belonging to quartet decoding boxes can be dispensable, however only in certain extended anticodon contexts [45].

The situation is different in the cases of bacterial tRNAs belonging to the split/duet decoding boxes, such as tRNAs specific for Phe/Leu, His/Gln, Asn/Lys, Asp/Gly and Ser/Arg that depend strictly on the identity of modified nucleotides at wobble U base. *E. coli*, all of the insect symbionts analyzed in this work, and all of the Mollicutes analyzed earlier rely on xm^{5}U and xm^{5}s^{2}U derivatives to allow accurate and efficient discrimination of the duet codons ending with a pyrimidine U or C. The other duet codons of the same decoding box ending with a purine A or G being efficiently read by a G34 or modified G34-containing tRNA (reviewed in: [22, 46]).

Other important, modified nucleotides conserved in C. *R. pediculicola*, and possibly essential in all bacteria, are the pseudouridines at positions 38, 39 and/or 40 and the modified purine at position 37 found in tRNAs harboring an anticodon ending with A36, G36 or U36, modified into (ms^2)i^{6}A37, m^{1}G37 or m^{2}A37 and (m^6)t^{6}A37, respectively (Figure 4). Removal of these modifications has been shown to have a detrimental effect on efficiency and accuracy of decoding (reviewed in [47, 48]). However, one cannot generalize the essentiality of

Legend to Figure 4. Comparative decoding strategies of C. *Riesia pediculicola* and *E. coli.* In the standard genetic code, each decoding box contains information about identity of nucleotides present in the anticodon loop and proximal stem, as illustrated in the decoding box corresponding to codons UAA/UAG (labeled in figure as “Extended anticodon”). Shown are the nucleotides at positions 32, the three anticodon bases (34-36) and nucleotide-37 (both in grey background) and the sequence of nucleotide 38-40. On the right side of each decoding box, is listed the information for *E. coli* isoacceptor tRNAs obtained from the tRNA data banks [2, 3]. On the left side of each decoding box, is listed the information for the homologous C. *R. pediculicola* (Riesia) isoacceptor tRNAs. The identities of the nucleotides were obtained directly from the tRNA gene analysis (this work, Figure 1), while the presence of modified nucleotides was deduced by combining knowledge from the analysis done in Figure 2 with the known modifications at identical positions in the corresponding *E. coli* tRNAs. The color code is as in Figure 1. In dark green background, are the only four mature tRNAs in *E. coli* that have not yet been sequenced, only the sequence of the corresponding genes are known. Differences between the two sets of bacterial isoacceptor species are highlighted by red letters. The exact chemical nature of the hypermodified m^{1}G?37 in *E. coli* tRNA$^{\text{Leu}}$ is not known [3], so only the m^{1}G moiety was indicated for the insect symbiont tRNA. Also the presence of m^{2}A37 in C. *R. pediculicola* is questionable (see Figure 2 legend) and indicated as ?m^{2}A.

these modifications to all tRNA sets, as the naturally occurring *E. coli* tRNA^{Ser}(GGA) and/or tRNAs harboring an anticodon ending with A37 in most Mycoplasmas lack i^6A37 or ms^2i^6A37 derivatives [24, 49]. It is clear from our analysis that the genes responsible for the insertion of Ψ38-40, m^1G37, t^6A37, $cmnm^5$U34 and s^2U34 remain resistant to loss. This suggests that these genes emerged early during cellular evolution, and, once fixed in the genome, became essential for the cell.

DISCUSSION

Both the insect symbionts and Mollicutes analyzed in our work are derived from bacteria with larger genomes (gram-negative Proteobacteria and gram-positive Firmicutes, respectively). During their evolutionary adaptation to their specific eukaryotic host cell, these organisms have massively lost genes, including genes coding for many isoacceptor tRNA and tRNA modification enzymes. With their minimal genomes, and unlike more specialized organelles, they are generally considered to correspond to the simplest living, autonomous organisms. We purposely did not include in our analysis genomes of insect symbionts with extremely reduced genomes (below 300 kb), such as Candidatus *Carsonella ruddi*, the endosymbiont of the psyllid *Pachpsylla venusta* (genome size of 160 kb with 183 CDSs [50]) and the very recently sequenced Candidatus *Tremblaya princeps* str PCVAL of the citrus mealybug *Planococcus citri* (genome size of 138kb, about 110 CDSs [51]). Both C. *Carsonella ruddii* and C. *Tremblaya princeps* have lost several enzymes required for self-replication, several ribosomal RNA, and many aminoacyl-tRNA synthetases. C. *Tremblaya princeps* has even lost most of its tRNA genes. These organisms must therefore rely on host proteins and tRNAs. They resemble organelles (mitochondria and plastids) [32, 52, 53], and cannot be used in our analysis as we cannot predict the presence of modifications from the presence of the corresponding genes in the endosymbiont.

The finding that C. *R. pediculicola* has lost all modifications of the tRNA body suggests that the structural and recognition roles of modifications outside the anticodon region (reviewed in [3] and [4]) are dispensable in the context of intracellular organisms with slow growth rates and probably with limited sets of nucleases genes and whertRNA degradation might be less of an issue [9]. Indeed, one can expect that protein synthesis might not be as accurate in Mollicutes and insect symbionts as in more sophisticated free-living bacteria. However, since these organisms are not in constant competition with other bacteria, they can certainly survive with a less efficient translation system. The positions of these parasites on the bacterial phylogenetic tree suggest that these are fast evolving bacteria with elevated mutation rates ([29] and several chapters of [54]). Proteins generated by an inaccurate translation system might provide an advantage to the parasite to evolve faster than other bacteria producing a more homogeneous proteome (discussed in [55-57]) and could be an advantage for fast adaptation to the host.

The conservation of genes coding for modification enzymes acting at the wobble position as well as the proximal anticodon bases (position 37-40), at least in organisms having a relatively low G+C content (below 35%, like Mollicutes and most insect symbionts), definitively pointed out the importance of these modifications for maintaining minimalist accuracy and efficacy in reading the genetic code based on 61 sense codons for 20 amino acids. Analyzing genomes of organisms having progressively reduced the size of their genomes allows for identification of the genes more resistant to loss. Hence, from an evolutionary perspective, Mollicutes and insect symbionts constitute excellent biological specimens to identify strategies developed during evolution for reading the genetic code with a minimal set of tRNAs and modification enzymes, a situation that could correspond to what might have occurred at an early stage of life, when the genetic code was just emerging [24, 58].

Note added in proofs: It was recently found that the *E. coli rlmN* gene encodes the missing m^2A37 methyltransferase (Eugenia Armengod, personal communication). RlmN homologs are present in most insect symbiont genomes, including

C. *R. pediculicola*. A37 is therefore most certainly methylated into m^{2}A37 in a few C. *R. pediculicola* tRNAs, which fits with our general conclusion above.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (R01 GM70641) to VdC. HG hold a position of Emeritus Scientist at the Center of Molecular Genetic of the CNRS in Gif-sur-Yvette (France) in the laboratory of Dominique Fourmy. We thank David Reed and Bret Boyd for introducing us to the field of louse symbionts, Basma El Yacoubi for numerous and insightful discussions and Patrick C. Thiaville and Jennifer J. Thiaville for editing the manuscript.

REFERENCES

1. Czerwoniec, A., Dunin-Horkawicz, S., Purta, E., Kaminska, K. H., Kasprzak, J. M., Bujnicki, J. M., Grosjean, H., and Rother, K. 2009, Nucleic Acids Res., 37, D118.
2. Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Pütz, J. 2009, Nucleic Acids Res., 37, D159.
3. Björk, G. R. and Hagervall, T. G. 2005, *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology, Böck, A., Curtis, R., Kaper, J. B., Neidhardt, F. C., Nyström, T., and Squires, C. L. (Ed.), ASM. Press, Washington DC http://www.ecosal.org Module 4.6.2.
4. Phizicky, E. M. and Hopper, A. K. 2010, Genes & Dev., 24, 1832.
5. Cermakian, N. and Cedergren, R. 1998, Modification and Editing of RNA, Grosjean, H. and Benne, R. (Ed.), ASM Press, Washington DC, 535.
6. Björk, G. R. 1986, Chemica Scripta, 26B, 91.
7. Ouzounis, C. A., Kunin, V., Darzentas, N., and Goldovsky, L. 2006, Res. Microbiol., 157, 57.
8. Forster, A. C. and Church, G. M. 2006, Mol. Syst. Biol., 2.
9. Alexandrov, A., Chernyakov, I., Gu, W., Hiley, S. L., Hughes, T. R., Grayhack, E. J., and Phizicky, E. M. 2006, Mol. Cell, 21, 87.
10. Grosshans, H., Lecointe, F., Grosjean, H., Hurt, E., and Simos, G. 2001, J. Biol. Chem., 276, 46333.
11. Gerlt, J. A., Babbitt, P. C., Jacobson, M. P., and Almo, S. C. 2012, J. Biol. Chem., 287, 29.
12. Christian, T., Evilia, C., Williams, S., and Hou, Y.-M. 2004, J. Mol. Biol., 339, 707.
13. Urbonavicius, J., Skouloubris, S., Myllykallio, H., and Grosjean, H. 2005, Nucleic Acids Res., 33, 3955.
14. Galperin, M. Y. and Koonin, E. V. 2012, J. Biol. Chem., 287, 21.
15. El Yacoubi, B., Lyons, B., Cruz, Y., Reddy, R., Nordin, B., Agnelli, F., Williamson, J. R., Schimmel, P., Swairjo, M. A., and de Crécy-Lagard, V. 2009, Nucleic Acids Res., 37, 2894.
16. El Yacoubi, B., Hatin, I., Deutsch, C., Kahveci, T., Rousset, J.-P., Iwata-Reuyl, D., Murzin, A. G., and de Crécy Lagard, V. 2011, EMBO J., 30, 882.
17. Srinivasan, M., Mehta, P., Yu, Y., Prugar, E., Koonin, E. V., Karzai, A. W., and Sternglanz, R. 2011, EMBO J., 30, 873.
18. Davanloo, P., Sprinzl, M., Watanabe, K., Albani, M., and Kersten, H. 1979, Nucleic Acids Res., 6, 1571.
19. Sengupta, R., Vainauskas, S., Yarian, C., Sochacka, E., Malkiewicz, A., Guenther, R. H., Koshlap, K. M., and Agris, P. F. 2000, Nucleic Acids Res., 28, 1374.
20. Chatterjee, A. K., Blaby, I., Thiaville, P. C., Majumder, M., Grosjean, H., Yuan, Y. A., Gupta, R., and de Crécy-Lagard, V. 2012, RNA, 18, 421.
21. Wurm, J. P., Griese, M., Bahr, U., Held, M., Heckel, A., Karas, A., Soppa, J., and Wohnert, J. 2012, RNA, 18, 412.
22. Grosjean, H., de Crécy-Lagard, V., and Marck, C. 2010, FEBS Lett., 584, 252.

23. Silva, F. J., Belda, E., and Talens, S. E. 2006, Nucleic Acids Res., 34, 6015.
24. de Crécy-Lagard, V., Marck, C., Brochier-Armanet, C., and Grosjean, H. 2007, IUBMB Life, 59, 634.
25. Fabret, C., Dervyn, E., Dalmais, B., Guillot, A., Marck, C., Grosjean, H., and Noirot, P. 2011, Mol. Microbiol., 80, 1062.
26. van der Gulik, P. and Hoff, W. 2011, J. Mol. Evol., 73, 59.
27. Maniloff, J. 2000, Molecular Biology and Pathogenicity of Mycoplasmas, Razin, S. and Hermann, R. (Ed.), Kluwer Academic/ Plenum Publisher, New York, 31.
28. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., and Lucier, T. S. 1995, Science, 270, 397.
29. Sirand-Pugnet, P., Citti, C., Barré, A., and Blanchard, A. 2007, Res. Microbiol., 158, 754.
30. Moya, A., Pereto, J., Gil, R., and Latorre, A. 2008, Nat. Rev. Genet., 9, 218.
31. John, P. M. 2010, Curr. Opin. Microbiol., 13, 73.
32. McCutcheon, J. P. and Moran, N. A. 2012, Nat. Rev. Micro., 10, 13.
33. Douglas, A. E. 2011, Cell Host Microbe, 10, 359.
34. Aziz, R., Bartels, D., Best, A., DeJongh, M., Disz, T., Edwards, R., Formsma, K., Gerdes, S., Glass, E., Kubal, M., Meyer, F., Olsen, G., Olson, R., Osterman, A., Overbeek, R., McNeil, L., Paarmann, D., Paczian, T., Parrello, B., Pusch, G., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. 2008, BMC Genomics, 9, 75.
35. Marck, C. and Grosjean, H. 2002, RNA, 8, 1189.
36. Higgs, P. G. and Ran, W. 2008, Mol. Biol. Evol., 25, 2279.
37. Dong, H., Nilsson, L., and Kurland, C. G. 1996, J. Mol. Biol., 260, 649.
38. Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., Lee, S. H., Robertson, H. M., Kennedy, R. C., Elhaik, E., Gerlach, D., Kriventseva, E. V., Elsik, C. G., Graur, D., Hill, C. A., Veenstra, J. A., Walenz, B., Tubío, J. M. C., Ribeiro, J. M. C., Rozas, J., Johnston, J. S., Reese, J. T., Popadic, A., Tojo, M., Raoult, D., Reed, D. L., Tomoyasu, Y., Kraus, E., Mittapalli, O., Margam, V. M., Li, H.-M., Meyer, J. M., Johnson, R. M., Romero-Severson, J., VanZee, J. P., Alvarez-Ponce, D., Vieira, F. G., Aguadé, M., Guirao-Rico, S., Anzola, J. M., Yoon, K. S., Strycharz, J. P., Unger, M. F., Christley, S., Lobo, N. F., Seufferheld, M. J., Wang, N., Dasch, G. A., Struchiner, C. J., Madey, G., Hannick, L. I., Bidwell, S., Joardar, V., Caler, E., Shao, R., Barker, S. C., Cameron, S., Bruggner, R. V., Regier, A., Johnson, J., Viswanathan, L., Utterback, R., Sutton, G. G., Lawson, D., , Waterhouse, R. M., Venter, J. C., Strausberg, R. L., Berenbaum, M. R., Collins, F. H., Zdobnov, E. M., and Pittendrigh, B. R. 2010, Proc. Natl. Acad. Sci. USA, 107, 12168.
39. Bujnicki, J. M., Oudjama, Y., Roovers, M., Owczarek, S., Caillet, J., and Droogmans, L. 2004, RNA, 10, 1236.
40. Moukadiri, I., Prado, S., Piera, J., Velázquez-Campoy, A., Björk, G. R., and Armengod, M. E. 2009, Nucleic Acids Res., 37, 7177.
41. Näsvall, S. J., Chen, P., and Björk, G. R. 2007, RNA, 13, 2151.
42. Nasvall, S. J., Chen, P., and Björk, G. R. 2004, RNA, 10, 1662.
43. Samuelsson, T., Guindy, Y. S., Lustig, F., Boren, T., and Lagerkvist, U. 1987, Proc. Natl. Acad. Sci. USA, 84, 3166.
44. Takai, K., Okumura, S., Hosono, K., Yokoyama, S., and Takaku, H. 1999, FEBS Lett., 447, 1.

45. Ledoux, S., Olejniczak, M., and Uhlenbeck, O. C. 2009, Nat. Struct. Mol. Biol., 16, 359.

46. Takai, K. and Yokoyama, S. 2003, Nucleic Acids Res., 31, 6383.

47. Agris, P. F. 2004, Nucleic Acids Res., 32, 223.

48. Atkins, J. F. and Björk, G. R. 2009, Microbiol. Mol. Biol. Rev., 73, 178.

49. Grosjean, H., Nicoghosian, K., Haumont, E., Söll, D., and Cedergren, R. 1985, Nucleic Acids Res., 13, 5697.

50. Nakabachi, A., Yamashita, A., Toh, H., Ishikawa, H., Dunbar, H. E., Moran, N. A., and Hattori, M. 2006, Science, 314, 267.

51. López-Madrigal, S., Latorre, A., Porcar, M., Moya, A., and Gil, R. 2011, J. Bacteriol., 193, 5587.

52. Tamames, J., Gil, R., Latorre, A., Pereto, J., Silva, F., and Moya, A. 2007, BMC Evolutionary Biology, 7, 181.

53. Douglas, A. E. and Raven, J. A. 2003, Phil. Trans. R. Soc. Lond. B Biol. Sci., 358, 5.

54. Blanchard, A. and Browning G. F. 2005, Mycoplasmas : Molecular Biology, Pathogenicity and Strategies for control, Horizon Scientific Press, Norwich UK.

55. Drummond, A. D. and Wilke, C. O. 2009, Nat. Rev. Genet., 10, 715.

56. Li, L., Boniecki, M. T., Jaffe, J. D., Imai, B. S., Yau, P. M., Luthey-Schulten, Z. A., and Martinis, S. A. 2011, Proc. Natl. Acad. Sci. USA, 108, 9378.

57. Meyerovich, M., Mamou, G., and Ben-Yehuda, S. 2010, Proc. Natl. Acad. Sci. USA, 107, 11543.

58. Novozhilov, A. and Koonin, E. 2009, Biology Direct, 4, 44.

59. Lowe, T. M. and Eddy, S. R. 1997, Nucl. Acids Res., 25, 955.

60. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997, Nucleic Acids Res., 25, 3389.

61. Benítez-Páez, A., Villarroya, M., Douthwaite, S., Gabaldón, T., and Armengod, M. E. 2010, RNA, 16, 2131.

62. de Crécy-Lagard, V. 2004, Practical Bioinformatics, Bujnicki, J. (Ed.), Springer-Verlag, Berlin Heidelberg, 169.

63. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. 2005, Nucleic Acids Res. Symp. Series, 33, 5691.

Supplementary Table SS1: 14 minimal bacterial genomes studied + *Escherichia coli* K12 (as reference).

Genome #					GC %	
	Accession	**Full Name**	**Class Protobacteria**	**Length**	**global**	**ORFs**
01	NC_006833	*Wolbachia* endosymbiont strain TRS of *Brugia malayi*	α	1,080,084 nt	34.18	35.29
02	NC_010981	*Wolbachia* endosymbiont of *Culex quinquefasciatus*	α	1,482,455 nt	34.19	34.72
03	NC_002978	*Wolbachia* endosymbiont of *Drosophila melanogaster*	α	1,267,782 nt	35.23	35.45
04	NC_011834	*Buchnera aphidicola* str. Tuc7 (*Acyrthosiphon pisum)*	γ	641,895 nt	26.29	27.38
05	NC_011833	*Buchnera aphidicola* str. 5A (*Acyrthosiphon pisum*)	γ	642,122 nt	26.29	27.38
06	NC_002528	*Buchnera aphidicola* str. APS (*Acyrthosiphon pisum*)	γ	640,681 nt	26.31	27.40
07	NC_004545	*Buchnera aphidicola* str. Bp (*Baizongia pistaciae*)	γ	615,980 nt	25.34	27.02
08	NC_008513	*Buchnera aphidicola* str. Cc (*Cinara cedri*)	γ	416,380 nt	20.10	26.21
09	NC_004061	*Buchnera aphidicola* str. Sg (*Schizaphis graminum*)	γ	641,454 nt	25.33	26.27
10	NC_007292	Candidatus *Blochmannia pennsylvanicus* str. BPEN	γ	791,654 nt	29.56	32.06
11	NC_005061	Candidatus *Blochmannia floridanus*	γ	705,557 nt	27.38	32.06
12	NC_007984	*Baumannia cicadellinicola* str. Hc (*Homalodisca coagulata*)	γ	686,194 nt	33.24	34.31
13	NC_004344	*Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*	γ	697,724 nt	22.48	23.64
14	NC_014109	Candidatus *Riesia pediculicola* USDA	γ	574,390 nt	28.48	29.80
15	NC_000913	*Escherichia coli* K12	γ	4,639,675 nt	50.79	52.00

Supplementary Table SS2. Codon Usage in 14 minimal bacterial genomes + *Escherichia coli* K12.

Codon Usage deduced from the automatic determination of all non overlapping ORFs >100 codons.
First value is the number of codons, second value is the frequency with respect to 61 sense codons summing to 1.0.

01 NC_006833 *Wolbachia* endosymbiont strain TRS of *Brugia malayi*

236424 codons

TTT	phe	**F**	8820	0.03731	TCT	ser	**S**	4656	0.01969	TAT	tyr	**Y**	5880	0.02487	TGT	cys	**C**	2115	0.00895
TTC	phe	**F**	2241	0.00948	TCC	ser	**S**	1205	0.00510	TAC	tyr	**Y**	2434	0.01030	TGC	cys	**C**	1237	0.00523
TTA	leu	**L**	7851	0.03321	TCA	ser	**S**	4164	0.01761	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	4015	0.01698	TCG	ser	**S**	938	0.00397	TAG	AMB	*****	-	-	TGG	trp	**W**	1812	0.00766
CTT	leu	**L**	4515	0.01910	CCT	pro	**P**	2874	0.01216	CAT	his	**H**	3083	0.01304	CGT	arg	**R**	1669	0.00706
CTC	leu	**L**	1336	0.00565	CCC	pro	**P**	513	0.00217	CAC	his	**H**	1462	0.00618	CGC	arg	**R**	844	0.00357
CTA	leu	**L**	3432	0.01452	CCA	pro	**P**	3420	0.01447	CAA	gln	**Q**	4763	0.02015	CGA	arg	**R**	585	0.00247
CTG	leu	**L**	1883	0.00796	CCG	pro	**P**	718	0.00304	CAG	gln	**Q**	2280	0.00964	CGG	arg	**R**	263	0.00111
ATT	ile	**I**	8994	0.03804	ACT	thr	**T**	4600	0.01946	AAT	asn	**N**	9698	0.04102	AGT	ser	**S**	4529	0.01916
ATC	ile	**I**	2858	0.01209	ACC	thr	**T**	1295	0.00548	AAC	asn	**N**	3888	0.01645	AGC	ser	**S**	2617	0.01107
ATA	ile	**I**	9955	0.04211	ACA	thr	**T**	4040	0.01709	AAA	lys	**K**	13241	0.05601	AGA	arg	**R**	4030	0.01705
ATG	met	**M**	5356	0.02265	ACG	thr	**T**	1026	0.00434	AAG	lys	**K**	5822	0.02463	AGG	arg	**R**	1956	0.00827
GTT	val	**V**	6185	0.02616	GCT	ala	**A**	5345	0.02261	GAT	asp	**D**	9200	0.03891	GGT	gly	**G**	5577	0.02359
GTC	val	**V**	1267	0.00536	GCC	ala	**A**	1051	0.00445	GAC	asp	**D**	2768	0.01171	GGC	gly	**G**	2306	0.00975
GTA	val	**V**	5560	0.02352	GCA	ala	**A**	6703	0.02835	GAA	glu	**E**	10493	0.04438	GGA	gly	**G**	4586	0.01940
GTG	val	**V**	2866	0.01212	GCG	ala	**A**	1300	0.00550	GAG	glu	**E**	4538	0.01919	GGG	gly	**G**	1766	0.00747

02 NC_010981 *Wolbachia* endosymbiont of *Culex quinquefasciatus* Pel

410749 codons (34 undefined)

TTT	phe	**F**	14956	0.03641	TCT	ser	**S**	7778	0.01894	TAT	tyr	**Y**	10381	0.02527	TGT	cys	**C**	3593	0.00875
TTC	phe	**F**	3590	0.00874	TCC	ser	**S**	1687	0.00411	TAC	tyr	**Y**	3834	0.00933	TGC	cys	**C**	1978	0.00482
TTA	leu	**L**	13824	0.03366	TCA	ser	**S**	7381	0.01797	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	6551	0.01595	TCG	ser	**S**	1551	0.00378	TAG	AMB	*****	-	-	TGG	trp	**W**	3444	0.00838
CTT	leu	**L**	8035	0.01956	CCT	pro	**P**	4937	0.01202	CAT	his	**H**	5640	0.01373	CGT	arg	**R**	2809	0.00684
CTC	leu	**L**	2174	0.00529	CCC	pro	**P**	709	0.00173	CAC	his	**H**	2429	0.00591	CGC	arg	**R**	1130	0.00275

Supplementary Table SS2 continued..

CTA	leu	**L**	6039	0.01470	CCA	pro	**P**	5941	0.01446	CAA	gln	**Q**	9201	0.02240	CGA	arg	**R**	1261	0.00307
CTG	leu	**L**	3320	0.00808	CCG	pro	**P**	1238	0.00301	CAG	gln	**Q**	3923	0.00955	CGG	arg	**R**	380	0.00093
ATT	ile	**I**	15263	0.03716	ACT	thr	**T**	7888	0.01920	AAT	asn	**N**	18021	0.04387	AGT	ser	**S**	8264	0.02012
ATC	ile	**I**	4587	0.01117	ACC	thr	**T**	2226	0.00542	AAC	asn	**N**	6062	0.01476	AGC	ser	**S**	4427	0.01078
ATA	ile	**I**	15881	0.03866	ACA	thr	**T**	7439	0.01811	AAA	lys	**K**	24826	0.06044	AGA	arg	**R**	7833	0.01907
ATG	met	**M**	8555	0.02083	ACG	thr	**T**	1773	0.00432	AAG	lys	**K**	10503	0.02557	AGG	arg	**R**	3238	0.00788
GTT	val	**V**	9991	0.02432	GCT	ala	**A**	8933	0.02175	GAT	asp	**D**	17107	0.04165	GGT	gly	**G**	9110	0.02218
GTC	val	**V**	1839	0.00448	GCC	ala	**A**	1756	0.00428	GAC	asp	**D**	4102	0.00999	GGC	gly	**G**	3670	0.00893
GTA	val	**V**	9599	0.02337	GCA	ala	**A**	11667	0.02840	GAA	glu	**E**	20505	0.04992	GGA	gly	**G**	8461	0.02060
GTG	val	**V**	4270	0.01040	GCG	ala	**A**	2101	0.00512	GAG	glu	**E**	8754	0.02131	GGG	gly	**G**	2384	0.00580

03 NC_002978 *Wolbachia* endosymbiont of *Drosophila melanogaster*

334919 codons

TTT	phe	**F**	12247	0.03657	TCT	ser	**S**	6257	0.01868	TAT	tyr	**Y**	8377	0.02501	TGT	cys	**C**	2667	0.00796
TTC	phe	**F**	2921	0.00872	TCC	ser	**S**	1605	0.00479	TAC	tyr	**Y**	3181	0.00950	TGC	cys	**C**	1808	0.00540
TTA	leu	**L**	11195	0.03343	TCA	ser	**S**	5849	0.01746	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	5555	0.01659	TCG	ser	**S**	1355	0.00405	TAG	AMB	*****	-	-	TGG	trp	**W**	2889	0.00863
CTT	leu	**L**	6380	0.01905	CCT	pro	**P**	4005	0.01196	CAT	his	**H**	4342	0.01296	CGT	arg	**R**	2314	0.00691
CTC	leu	**L**	1894	0.00566	CCC	pro	**P**	675	0.00202	CAC	his	**H**	2148	0.00641	CGC	arg	**R**	1178	0.00352
CTA	leu	**L**	4723	0.01410	CCA	pro	**P**	4975	0.01485	CAA	gln	**Q**	7394	0.02208	CGA	arg	**R**	904	0.00270
CTG	leu	**L**	2728	0.00815	CCG	pro	**P**	1121	0.00335	CAG	gln	**Q**	3293	0.00983	CGG	arg	**R**	421	0.00126
ATT	ile	**I**	12032	0.03593	ACT	thr	**T**	6498	0.01940	AAT	asn	**N**	13619	0.04066	AGT	ser	**S**	6288	0.01877
ATC	ile	**I**	4048	0.01209	ACC	thr	**T**	1842	0.00550	AAC	asn	**N**	5409	0.01615	AGC	ser	**S**	3853	0.01150
ATA	ile	**I**	13203	0.03942	ACA	thr	**T**	5975	0.01784	AAA	lys	**K**	19793	0.05910	AGA	arg	**R**	5925	0.01769
ATG	met	**M**	7371	0.02201	ACG	thr	**T**	1589	0.00474	AAG	lys	**K**	8167	0.02439	AGG	arg	**R**	2719	0.00812
GTT	val	**V**	8469	0.02529	GCT	ala	**A**	7539	0.02251	GAT	asp	**D**	13463	0.04020	GGT	gly	**G**	7310	0.02183
GTC	val	**V**	1522	0.00454	GCC	ala	**A**	1528	0.00456	GAC	asp	**D**	3722	0.01111	GGC	gly	**G**	3291	0.00983
GTA	val	**V**	7435	0.02220	GCA	ala	**A**	9607	0.02868	GAA	glu	**E**	16295	0.04865	GGA	gly	**G**	6654	0.01987
GTG	val	**V**	4277	0.01277	GCG	ala	**A**	1928	0.00576	GAG	glu	**E**	6734	0.02011	GGG	gly	**G**	2413	0.00720

Supplementary Table SS2 continued..

04 NC_011834 Buchnera aphidicola str. Tuc7 (*Acyrthosiphon pisum*)

181313 codons

TTT phe **F**	8482	0.04678	TCT ser **S**	5662	0.03123	TAT tyr **Y**	5723	0.03156	TGT cys **C**	1803	0.00994
TTC phe **F**	799	0.00441	TCC ser **S**	543	0.00299	TAC tyr **Y**	878	0.00484	TGC cys **C**	392	0.00216
TTA leu **L**	11624	0.06411	TCA ser **S**	3278	0.01808	TAA OCH *	-	-	TGA OPA *	-	-
TTG leu **L**	1826	0.01007	TCG ser **S**	457	0.00252	TAG AMB *	-	-	TGG trp **W**	1664	0.00918
CTT leu **L**	2179	0.01202	CCT pro **P**	2588	0.01427	CAT his **H**	3337	0.01840	CGT arg **R**	2259	0.01246
CTC leu **L**	321	0.00177	CCC pro **P**	389	0.00215	CAC his **H**	487	0.00269	CGC arg **R**	328	0.00181
CTA leu **L**	1698	0.00937	CCA pro **P**	2094	0.01155	CAA gln **Q**	5004	0.02760	CGA arg **R**	1068	0.00589
CTG leu **L**	359	0.00198	CCG pro **P**	406	0.00224	CAG gln **Q**	760	0.00419	CGG arg **R**	95	0.00052
ATT ile **I**	11463	0.06322	ACT thr **T**	3777	0.02083	AAT asn **N**	11220	0.06188	AGT ser **S**	2718	0.01499
ATC ile **I**	1723	0.00950	ACC thr **T**	519	0.00286	AAC asn **N**	1810	0.00998	AGC ser **S**	589	0.00325
ATA ile **I**	7812	0.04309	ACA thr **T**	3540	0.01952	AAA lys **K**	16342	0.09013	AGA arg **R**	2832	0.01562
ATG met **M**	3939	0.02172	ACG thr **T**	452	0.00249	AAG lys **K**	1423	0.00785	AGG arg **R**	209	0.00115
GTT val **V**	3938	0.02172	GCT ala **A**	3539	0.01952	GAT asp **D**	6905	0.03808	GGT gly **G**	4099	0.02261
GTC val **V**	611	0.00337	GCC ala **A**	500	0.00276	GAC asp **D**	926	0.00511	GGC gly **G**	748	0.00413
GTA val **V**	3557	0.01962	GCA ala **A**	3467	0.01912	GAA glu **E**	9039	0.04985	GGA gly **G**	4395	0.02424
GTG val **V**	667	0.00368	GCG ala **A**	536	0.00296	GAG glu **E**	914	0.00504	GGG gly **G**	601	0.00331

05 NC_011833 Buchnera aphidicola str. 5A (*Acyrthosiphon pisum*)

181330 codons

TTT phe **F**	8467	0.04669	TCT ser **S**	5658	0.03120	TAT tyr **Y**	5725	0.03157	TGT cys **C**	1802	0.00994
TTC phe **F**	802	0.00442	TCC ser **S**	538	0.00297	TAC tyr **Y**	884	0.00488	TGC cys **C**	394	0.00217
TTA leu **L**	11617	0.06407	TCA ser **S**	3283	0.01811	TAA OCH *	-	-	TGA OPA *	-	-
TTG leu **L**	1836	0.01013	TCG ser **S**	455	0.00251	TAG AMB *	-	-	TGG trp **W**	1662	0.00917
CTT leu **L**	2189	0.01207	CCT pro **P**	2602	0.01435	CAT his **H**	3341	0.01842	CGT arg **R**	2257	0.01245
CTC leu **L**	313	0.00173	CCC pro **P**	381	0.00210	CAC his **H**	488	0.00269	CGC arg **R**	329	0.00181
CTA leu **L**	1687	0.00930	CCA pro **P**	2093	0.01154	CAA gln **Q**	5007	0.02761	CGA arg **R**	1061	0.00585
CTG leu **L**	362	0.00200	CCG pro **P**	407	0.00224	CAG gln **Q**	754	0.00416	CGG arg **R**	97	0.00053

Supplementary Table SS2 continued..

ATT ile **I**	11468	0.06324	ACT thr **T**	3782	0.02086	AAT asn **N**	11240	0.06199	AGT ser **S**	2716	0.01498
ATC ile **I**	1723	0.00950	ACC thr **T**	512	0.00282	AAC asn **N**	1797	0.00991	AGC ser **S**	586	0.00323
ATA ile **I**	7813	0.04309	ACA thr **T**	3549	0.01957	AAA lys **K**	16348	0.09016	AGA arg **R**	2830	0.01561
ATG met **M**	3940	0.02173	ACG thr **T**	451	0.00249	AAG lys **K**	1437	0.00792	AGG arg **R**	208	0.00115
GTT val **V**	3933	0.02169	GCT ala **A**	3539	0.01952	GAT asp **D**	6913	0.03812	GGT gly **G**	4097	0.02259
GTC val **V**	611	0.00337	GCC ala **A**	499	0.00275	GAC asp **D**	912	0.00503	GGC gly **G**	758	0.00418
GTA val **V**	3548	0.01957	GCA ala **A**	3461	0.01909	GAA glu **E**	9048	0.04990	GGA gly **G**	4389	0.02420
GTG val **V**	673	0.00371	GCG ala **A**	542	0.00299	GAG glu **E**	908	0.00501	GGG gly **G**	608	0.00335

06 NC_002528 Buchnera aphidicola str. APS (*Acyrthosiphon pisum*)

180449 codons

TTT phe **F**	8421	0.04667	TCT ser **S**	5630	0.03120	TAT tyr **Y**	5701	0.03159	TGT cys **C**	1796	0.00995
TTC phe **F**	795	0.00441	TCC ser **S**	548	0.00304	TAC tyr **Y**	881	0.00488	TGC cys **C**	390	0.00216
TTA leu **L**	11562	0.06407	TCA ser **S**	3265	0.01809	TAA OCH *****	-	-	TGA OPA *****	-	-
TTG leu **L**	1827	0.01012	TCG ser **S**	446	0.00247	TAG AMB *****	-	-	TGG trp **W**	1659	0.00919
CTT leu **L**	2166	0.01200	CCT pro **P**	2586	0.01433	CAT his **H**	3325	0.01843	CGT arg **R**	2254	0.01249
CTC leu **L**	315	0.00175	CCC pro **P**	380	0.00211	CAC his **H**	481	0.00267	CGC arg **R**	328	0.00182
CTA leu **L**	1680	0.00931	CCA pro **P**	2082	0.01154	CAA gln **Q**	4989	0.02765	CGA arg **R**	1069	0.00592
CTG leu **L**	357	0.00198	CCG pro **P**	405	0.00224	CAG gln **Q**	753	0.00417	CGG arg **R**	97	0.00054
ATT ile **I**	11421	0.06329	ACT thr **T**	3759	0.02083	AAT asn **N**	11141	0.06174	AGT ser **S**	2708	0.01501
ATC ile **I**	1709	0.00947	ACC thr **T**	516	0.00286	AAC asn **N**	1800	0.00998	AGC ser **S**	580	0.00321
ATA ile **I**	7767	0.04304	ACA thr **T**	3525	0.01953	AAA lys **K**	16241	0.09000	AGA arg **R**	2816	0.01561
ATG met **M**	3923	0.02174	ACG thr **T**	443	0.00245	AAG lys **K**	1425	0.00790	AGG arg **R**	211	0.00117
GTT val **V**	3917	0.02171	GCT ala **A**	3528	0.01955	GAT asp **D**	6881	0.03813	GGT gly **G**	4095	0.02269
GTC val **V**	609	0.00337	GCC ala **A**	498	0.00276	GAC asp **D**	917	0.00508	GGC gly **G**	740	0.00410
GTA val **V**	3530	0.01956	GCA ala **A**	3464	0.01920	GAA glu **E**	9012	0.04994	GGA gly **G**	4356	0.02414
GTG val **V**	667	0.00370	GCG ala **A**	541	0.00300	GAG glu **E**	909	0.00504	GGG gly **G**	612	0.00339

07 NC_004545 Buchnera aphidicola str. Bp (*Baizongia pistaciae*)

162548 codons

TTT phe **F**	7606	0.04679	TCT ser **S**	4684	0.02882	TAT tyr **Y**	5282	0.03250	TGT cys **C**	1936	0.01191
TTC phe **F**	672	0.00413	TCC ser **S**	399	0.00245	TAC tyr **Y**	898	0.00552	TGC cys **C**	473	0.00291

Supplementary Table SS2 continued..

TTA leu **L**	11018	0.06778	TCA ser **S**	3196	0.01966	TAA OCH *****	-	-	TGA OPA *****	-	-
TTG leu **L**	2153	0.01325	TCG ser **S**	652	0.00401	TAG AMB *****	-	-	TGG trp **W**	1482	0.00912
CTT leu **L**	1317	0.00810	CCT pro **P**	2212	0.01361	CAT his **H**	3078	0.01894	CGT arg **R**	1477	0.00909
CTC leu **L**	203	0.00125	CCC pro **P**	259	0.00159	CAC his **H**	583	0.00359	CGC arg **R**	330	0.00203
CTA leu **L**	1651	0.01016	CCA pro **P**	1978	0.01217	CAA gln **Q**	4488	0.02761	CGA arg **R**	1190	0.00732
CTG leu **L**	202	0.00124	CCG pro **P**	322	0.00198	CAG gln **Q**	613	0.00377	CGG arg **R**	143	0.00088
ATT ile **I**	10399	0.06397	ACT thr **T**	3636	0.02237	AAT asn **N**	10269	0.06318	AGT ser **S**	2559	0.01574
ATC ile **I**	1171	0.00720	ACC thr **T**	455	0.00280	AAC asn **N**	2252	0.01385	AGC ser **S**	556	0.00342
ATA ile **I**	7510	0.04620	ACA thr **T**	3229	0.01986	AAA lys **K**	13790	0.08484	AGA arg **R**	2418	0.01488
ATG met **M**	3589	0.02208	ACG thr **T**	666	0.00410	AAG lys **K**	1584	0.00974	AGG arg **R**	321	0.00197
GTT val **V**	3844	0.02365	GCT ala **A**	3162	0.01945	GAT asp **D**	5592	0.03440	GGT gly **G**	2909	0.01790
GTC val **V**	384	0.00236	GCC ala **A**	246	0.00151	GAC asp **D**	958	0.00589	GGC gly **G**	431	0.00265
GTA val **V**	3644	0.02242	GCA ala **A**	2845	0.01750	GAA glu **E**	6603	0.04062	GGA gly **G**	4524	0.02783
GTG val **V**	625	0.00385	GCG ala **A**	496	0.00305	GAG glu **E**	718	0.00442	GGG gly **G**	666	0.00410

08 NC_008513 Buchnera aphidicola str. Cc (*Cinara cedri*)

113588 codons

TTT phe **F**	6386	0.05622	TCT ser **S**	3527	0.03105	TAT tyr **Y**	4690	0.04129	TGT cys **C**	1327	0.01168
TTC phe **F**	276	0.00243	TCC ser **S**	173	0.00152	TAC tyr **Y**	270	0.00238	TGC cys **C**	112	0.00099
TTA leu **L**	8964	0.07892	TCA ser **S**	2564	0.02257	TAA OCH *****	-	-	TGA OPA *****	-	-
TTG leu **L**	549	0.00483	TCG ser **S**	132	0.00116	TAG AMB *****	-	-	TGG trp **W**	935	0.00823
CTT leu **L**	751	0.00661	CCT pro **P**	1449	0.01276	CAT his **H**	1914	0.01685	CGT arg **R**	792	0.00697
CTC leu **L**	43	0.00038	CCC pro **P**	82	0.00072	CAC his **H**	122	0.00107	CGC arg **R**	44	0.00039
CTA leu **L**	573	0.00504	CCA pro **P**	1331	0.01172	CAA gln **Q**	2852	0.02511	CGA arg **R**	534	0.00470
CTG leu **L**	56	0.00049	CCG pro **P**	176	0.00155	CAG gln **Q**	208	0.00183	CGG arg **R**	26	0.00023
ATT ile **I**	8942	0.07872	ACT thr **T**	2100	0.01849	AAT asn **N**	8600	0.07571	AGT ser **S**	1376	0.01211
ATC ile **I**	503	0.00443	ACC thr **T**	133	0.00117	AAC asn **N**	691	0.00608	AGC ser **S**	119	0.00105
ATA ile **I**	5854	0.05154	ACA thr **T**	2302	0.02027	AAA lys **K**	14878	0.13098	AGA arg **R**	2053	0.01807
GTG val **V**	160	0.00141	GCG ala **A**	164	0.00144	GAG glu **E**	205	0.00180	GGG gly **G**	188	0.00166

Supplementary Table SS2 continued..

09 NC_004061 Buchnera aphidicola str. Sg (*Schizaphis graminum*)

179622 codons

TTT	phe	**F**	8848	0.04926	TCT	ser	**S**	5584	0.03109	TAT	tyr	**Y**	5586	0.03110	TGT	cys	**C**	1810	0.01008
TTC	phe	**F**	775	0.00431	TCC	ser	**S**	362	0.00202	TAC	tyr	**Y**	838	0.00467	TGC	cys	**C**	384	0.00214
TTA	leu	**L**	12096	0.06734	TCA	ser	**S**	3396	0.01891	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	1670	0.00930	TCG	ser	**S**	405	0.00225	TAG	AMB	*****	-	-	TGG	trp	**W**	1640	0.00913
CTT	leu	**L**	2211	0.01231	CCT	pro	**P**	2611	0.01454	CAT	his	**H**	3075	0.01712	CGT	arg	**R**	2090	0.01164
CTC	leu	**L**	235	0.00131	CCC	pro	**P**	285	0.00159	CAC	his	**H**	405	0.00225	CGC	arg	**R**	277	0.00154
CTA	leu	**L**	1435	0.00799	CCA	pro	**P**	2121	0.01181	CAA	gln	**Q**	4861	0.02706	CGA	arg	**R**	986	0.00549
CTG	leu	**L**	288	0.00160	CCG	pro	**P**	304	0.00169	CAG	gln	**Q**	553	0.00308	CGG	arg	**R**	81	0.00045
ATT	ile	**I**	11715	0.06522	ACT	thr	**T**	3557	0.01980	AAT	asn	**N**	11272	0.06275	AGT	ser	**S**	2691	0.01498
ATC	ile	**I**	1413	0.00787	ACC	thr	**T**	400	0.00223	AAC	asn	**N**	1874	0.01043	AGC	ser	**S**	454	0.00253
ATA	ile	**I**	7894	0.04395	ACA	thr	**T**	3525	0.01962	AAA	lys	**K**	17502	0.09744	AGA	arg	**R**	2815	0.01567
ATG	met	**M**	3774	0.02101	ACG	thr	**T**	387	0.00215	AAG	lys	**K**	1509	0.00840	AGG	arg	**R**	230	0.00128
GTT	val	**V**	3970	0.02210	GCT	ala	**A**	3593	0.02000	GAT	asp	**D**	6715	0.03738	GGT	gly	**G**	4124	0.02296
GTC	val	**V**	510	0.00284	GCC	ala	**A**	356	0.00198	GAC	asp	**D**	855	0.00476	GGC	gly	**G**	499	0.00278
GTA	val	**V**	3425	0.01907	GCA	ala	**A**	3458	0.01925	GAA	glu	**E**	9218	0.05132	GGA	gly	**G**	4580	0.02550
GTG	val	**V**	516	0.00287	GCG	ala	**A**	369	0.00205	GAG	glu	**E**	780	0.00434	GGG	gly	**G**	430	0.00239

10 NC_007292 Candidatus *Blochmannia pennsylvanicus* str. BPEN

196557 codons

TTT	phe	**F**	7293	0.03710	TCT	ser	**S**	5085	0.02587	TAT	tyr	**Y**	6220	0.03164	TGT	cys	**C**	2314	0.01177
TTC	phe	**F**	1106	0.00563	TCC	ser	**S**	925	0.00471	TAC	tyr	**Y**	1258	0.00640	TGC	cys	**C**	748	0.00381
TTA	leu	**L**	11819	0.06013	TCA	ser	**S**	3048	0.01551	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	3253	0.01655	TCG	ser	**S**	690	0.00351	TAG	AMB	*****	-	-	TGG	trp	**W**	2146	0.01092
CTT	leu	**L**	1996	0.01015	CCT	pro	**P**	2766	0.01407	CAT	his	**H**	4686	0.02384	CGT	arg	**R**	3520	0.01791
CTC	leu	**L**	395	0.00201	CCC	pro	**P**	562	0.00286	CAC	his	**H**	857	0.00436	CGC	arg	**R**	979	0.00498
CTA	leu	**L**	1692	0.00861	CCA	pro	**P**	2710	0.01379	CAA	gln	**Q**	6048	0.03077	CGA	arg	**R**	1595	0.00811
CTG	leu	**L**	640	0.00326	CCG	pro	**P**	704	0.00358	CAG	gln	**Q**	1120	0.00570	CGG	arg	**R**	389	0.00198

Supplementary Table SS2 continued..

ATT	ile	**I**	10951	0.05571	ACT	thr	**T**	4664	0.02373	AAT	asn	**N**	10193	0.05186	AGT	ser	**S**	2628	0.01337
ATC	ile	**I**	2240	0.01140	ACC	thr	**T**	1070	0.00544	AAC	asn	**N**	2208	0.01123	AGC	ser	**S**	1059	0.00539
ATA	ile	**I**	7884	0.04011	ACA	thr	**T**	3604	0.01834	AAA	lys	**K**	11452	0.05826	AGA	arg	**R**	2168	0.01103
ATG	met	**M**	5037	0.02563	ACG	thr	**T**	876	0.00446	AAG	lys	**K**	1631	0.00830	AGG	arg	**R**	378	0.00192
GTT	val	**V**	4074	0.02073	GCT	ala	**A**	4694	0.02388	GAT	asp	**D**	7746	0.03941	GGT	gly	**G**	3815	0.01941
GTC	val	**V**	729	0.00371	GCC	ala	**A**	804	0.00409	GAC	asp	**D**	1281	0.00652	GGC	gly	**G**	1109	0.00564
GTA	val	**V**	4896	0.02491	GCA	ala	**A**	3911	0.01990	GAA	glu	**E**	7821	0.03979	GGA	gly	**G**	5454	0.02775
GTG	val	**V**	1623	0.00826	GCG	ala	**A**	1272	0.00647	GAG	glu	**E**	1397	0.00711	GGG	gly	**G**	1324	0.00674

11 NC_005061 Candidatus *Blochmannia floridanus*

191731 codons

TTT	phe	**F**	7821	0.04079	TCT	ser	**S**	5519	0.02879	TAT	tyr	**Y**	7304	0.03810	TGT	cys	**C**	2611	0.01362
TTC	phe	**F**	729	0.00380	TCC	ser	**S**	656	0.00342	TAC	tyr	**Y**	855	0.00446	TGC	cys	**C**	400	0.00209
TTA	leu	**L**	13188	0.06878	TCA	ser	**S**	3179	0.01658	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	3007	0.01568	TCG	ser	**S**	617	0.00322	TAG	AMB	*****	-	-	TGG	trp	**W**	2015	0.01051
CTT	leu	**L**	1364	0.00711	CCT	pro	**P**	2873	0.01498	CAT	his	**H**	4361	0.02275	CGT	arg	**R**	2537	0.01323
CTC	leu	**L**	251	0.00131	CCC	pro	**P**	280	0.00146	CAC	his	**H**	565	0.00295	CGC	arg	**R**	280	0.00146
CTA	leu	**L**	1196	0.00624	CCA	pro	**P**	2459	0.01283	CAA	gln	**Q**	6306	0.03289	CGA	arg	**R**	1647	0.00859
CTG	leu	**L**	299	0.00156	CCG	pro	**P**	517	0.00270	CAG	gln	**Q**	1158	0.00604	CGG	arg	**R**	316	0.00165
ATT	ile	**I**	11807	0.06158	ACT	thr	**T**	4683	0.02442	AAT	asn	**N**	11737	0.06122	AGT	ser	**S**	3010	0.01570
ATC	ile	**I**	1638	0.00854	ACC	thr	**T**	721	0.00376	AAC	asn	**N**	1520	0.00793	AGC	ser	**S**	515	0.00269
ATA	ile	**I**	8513	0.04440	ACA	thr	**T**	3304	0.01723	AAA	lys	**K**	11648	0.06075	AGA	arg	**R**	2437	0.01271
ATG	met	**M**	4854	0.02532	ACG	thr	**T**	625	0.00326	AAG	lys	**K**	1961	0.01023	AGG	arg	**R**	493	0.00257
GTT	val	**V**	4312	0.02249	GCT	ala	**A**	4309	0.02247	GAT	asp	**D**	8211	0.04283	GGT	gly	**G**	3819	0.01992
GTC	val	**V**	390	0.00203	GCC	ala	**A**	393	0.00205	GAC	asp	**D**	600	0.00313	GGC	gly	**G**	418	0.00218
GTA	val	**V**	4846	0.02527	GCA	ala	**A**	3258	0.01699	GAA	glu	**E**	7136	0.03722	GGA	gly	**G**	5640	0.02942
GTG	val	**V**	1437	0.00749	GCG	ala	**A**	754	0.00393	GAG	glu	**E**	1374	0.00717	GGG	gly	**G**	1058	0.00552

12 NC_007984 *Baumannia cicadellinicola* str. Hc(*Homalodisca coagulata*)

191706 codons (17 undefined)

TTT	phe	**F**	5945	0.03101	TCT	ser	**S**	3298	0.01720	TAT	tyr	**Y**	5531	0.02885	TGT	cys	**C**	1892	0.00987
TTC	phe	**F**	1208	0.00630	TCC	ser	**S**	502	0.00262	TAC	tyr	**Y**	1277	0.00666	TGC	cys	**C**	621	0.00324

Supplementary Table SS2 continued..

TTA	leu	**L**	10274	0.05359	TCA	ser	**S**	2465	0.01286	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	1341	0.00700	TCG	ser	**S**	543	0.00283	TAG	AMB	*****	-	-	TGG	trp	**W**	2124	0.01108
CTT	leu	**L**	2882	0.01503	CCT	pro	**P**	2686	0.01401	CAT	his	**H**	4128	0.02153	CGT	arg	**R**	4938	0.02576
CTC	leu	**L**	700	0.00365	CCC	pro	**P**	376	0.00196	CAC	his	**H**	769	0.00401	CGC	arg	**R**	1123	0.00586
CTA	leu	**L**	4934	0.02574	CCA	pro	**P**	3372	0.01759	CAA	gln	**Q**	6053	0.03157	CGA	arg	**R**	1098	0.00573
CTG	leu	**L**	750	0.00391	CCG	pro	**P**	567	0.00296	CAG	gln	**Q**	2530	0.01320	CGG	arg	**R**	423	0.00221
ATT	ile	**I**	9949	0.05190	ACT	thr	**T**	5076	0.02648	AAT	asn	**N**	8869	0.04626	AGT	ser	**S**	3519	0.01836
ATC	ile	**I**	2015	0.01051	ACC	thr	**T**	1108	0.00578	AAC	asn	**N**	2074	0.01082	AGC	ser	**S**	1336	0.00697
ATA	ile	**I**	7002	0.03652	ACA	thr	**T**	3222	0.01681	AAA	lys	**K**	9010	0.04700	AGA	arg	**R**	1479	0.00771
ATG	met	**M**	4798	0.02503	ACG	thr	**T**	756	0.00394	AAG	lys	**K**	2237	0.01167	AGG	arg	**R**	424	0.00221
GTT	val	**V**	4120	0.02149	GCT	ala	**A**	6441	0.03360	GAT	asp	**D**	7489	0.03907	GGT	gly	**G**	6669	0.03479
GTC	val	**V**	763	0.00398	GCC	ala	**A**	976	0.00509	GAC	asp	**D**	1071	0.00559	GGC	gly	**G**	1278	0.00667
GTA	val	**V**	5521	0.02880	GCA	ala	**A**	5083	0.02651	GAA	glu	**E**	7824	0.04081	GGA	gly	**G**	3055	0.01594
GTG	val	**V**	775	0.00404	GCG	ala	**A**	872	0.00455	GAG	glu	**E**	1846	0.00963	GGG	gly	**G**	699	0.00365

13 NC_004344 *Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*

196110 codons

TTT	phe	**F**	10193	0.05198	TCT	ser	**S**	7224	0.03684	TAT	tyr	**Y**	7064	0.03602	TGT	cys	**C**	1775	0.00905
TTC	phe	**F**	829	0.00423	TCC	ser	**S**	484	0.00247	TAC	tyr	**Y**	1025	0.00523	TGC	cys	**C**	713	0.00364
TTA	leu	**L**	13050	0.06654	TCA	ser	**S**	3876	0.01976	TAA	OCH	*****	-	-	TGA	OPA	*****	-	-
TTG	leu	**L**	2353	0.01200	TCG	ser	**S**	272	0.00139	TAG	AMB	*****	-	-	TGG	trp	**W**	1661	0.00847
CTT	leu	**L**	1670	0.00852	CCT	pro	**P**	2326	0.01186	CAT	his	**H**	2834	0.01445	CGT	arg	**R**	747	0.00381
CTC	leu	**L**	101	0.00052	CCC	pro	**P**	116	0.00059	CAC	his	**H**	339	0.00173	CGC	arg	**R**	109	0.00056
CTA	leu	**L**	1281	0.00653	CCA	pro	**P**	2951	0.01505	CAA	gln	**Q**	4066	0.02073	CGA	arg	**R**	139	0.00071
CTG	leu	**L**	145	0.00074	CCG	pro	**P**	200	0.00102	CAG	gln	**Q**	364	0.00186	CGG	arg	**R**	17	0.00009
ATT	ile	**I**	11816	0.06025	ACT	thr	**T**	3420	0.01744	AAT	asn	**N**	14398	0.07342	AGT	ser	**S**	2097	0.01069
ATC	ile	**I**	919	0.00469	ACC	thr	**T**	292	0.00149	AAC	asn	**N**	2195	0.01119	AGC	ser	**S**	1138	0.00580
ATA	ile	**I**	13309	0.06786	ACA	thr	**T**	3368	0.01717	AAA	lys	**K**	21359	0.10891	AGA	arg	**R**	4585	0.02338
ATG	met	**M**	3958	0.02018	ACG	thr	**T**	186	0.00095	AAG	lys	**K**	1369	0.00698	AGG	arg	**R**	426	0.00217

Supplementary Table SS2 continued..

GTT	val	**V**	3715	0.01894	GCT	ala	**A**	3304	0.01685	GAT	asp	**D**	6888	0.03512	GGT	gly	**G**	2002	0.01021
GTC	val	**V**	209	0.00107	GCC	ala	**A**	278	0.00142	GAC	asp	**D**	973	0.00496	GGC	gly	**G**	415	0.00212
GTA	val	**V**	3879	0.01978	GCA	ala	**A**	3231	0.01648	GAA	glu	**E**	9198	0.04690	GGA	gly	**G**	7097	0.03619
GTG	val	**V**	437	0.00223	GCG	ala	**A**	307	0.00157	GAG	glu	**E**	755	0.00385	GGG	gly	**G**	663	0.00338

14 NC_014109 Candidatus *Riesia pediculicola* USDA

147981 codons

TTT	phe	**F**	6625	0.04477	TCT	ser	**S**	5202	0.03515	TAT	tyr	**Y**	4157	0.02809	TGT	cys	**C**	1468	0.00992
TTC	phe	**F**	1769	0.01195	TCC	ser	**S**	1172	0.00792	TAC	tyr	**Y**	1072	0.00724	TGC	cys	**C**	403	0.00272
TTA	leu	**L**	6288	0.04249	TCA	ser	**S**	2571	0.01737	TAA	OCH	*	-	-	TGA	OPA	*	-	-
TTG	leu	**L**	2565	0.01733	TCG	ser	**S**	913	0.00617	TAG	AMB	*	-	-	TGG	trp	**W**	1239	0.00837
CTT	leu	**L**	2207	0.01491	CCT	pro	**P**	1465	0.00990	CAT	his	**H**	2511	0.01697	CGT	arg	**R**	789	0.00533
CTC	leu	**L**	565	0.00382	CCC	pro	**P**	219	0.00148	CAC	his	**H**	410	0.00277	CGC	arg	**R**	84	0.00057
CTA	leu	**L**	1600	0.01081	CCA	pro	**P**	1969	0.01331	CAA	gln	**Q**	3771	0.02548	CGA	arg	**R**	1228	0.00830
CTG	leu	**L**	582	0.00393	CCG	pro	**P**	516	0.00349	CAG	gln	**Q**	869	0.00587	CGG	arg	**R**	125	0.00084
ATT	ile	**I**	8237	0.05566	ACT	thr	**T**	2519	0.01702	AAT	asn	**N**	6774	0.04578	AGT	ser	**S**	1900	0.01284
ATC	ile	**I**	3042	0.02056	ACC	thr	**T**	564	0.00381	AAC	asn	**N**	1978	0.01337	AGC	ser	**S**	529	0.00357
ATA	ile	**I**	5671	0.03832	ACA	thr	**T**	1972	0.01333	AAA	lys	**K**	12272	0.08293	AGA	arg	**R**	4470	0.03021
ATG	met	**M**	3429	0.02317	ACG	thr	**T**	585	0.00395	AAG	lys	**K**	2976	0.02011	AGG	arg	**R**	609	0.00412
GTT	val	**V**	3638	0.02458	GCT	ala	**A**	2267	0.01532	GAT	asp	**D**	5554	0.03753	GGT	gly	**G**	1611	0.01089
GTC	val	**V**	820	0.00554	GCC	ala	**A**	364	0.00246	GAC	asp	**D**	906	0.00612	GGC	gly	**G**	189	0.00128
GTA	val	**V**	2444	0.01652	GCA	ala	**A**	1962	0.01326	GAA	glu	**E**	7253	0.04901	GGA	gly	**G**	5817	0.03931
GTG	val	**V**	784	0.00530	GCG	ala	**A**	439	0.00297	GAG	glu	**E**	1493	0.01009	GGG	gly	**G**	559	0.00378

15 NC_000913 *Escherichia coli* K12

1425506 codon

TTT	phe	**F**	32912	0.02309	TCT	ser	**S**	11774	0.00826	TAT	tyr	**Y**	21912	0.01537	TGT	cys	**C**	7667	0.00538
TTC	phe	**F**	25719	0.01804	TCC	ser	**S**	12577	0.00882	TAC	tyr	**Y**	17352	0.01217	TGC	cys	**C**	10861	0.00762
TTA	leu	**L**	19236	0.01349	TCA	ser	**S**	10640	0.00746	TAA	OCH	*	-	-	TGA	OPA	*	-	-
TTG	leu	**L**	19511	0.01369	TCG	ser	**S**	12972	0.00910	TAG	AMB	*	-	-	TGG	trp	**W**	21321	0.01496

Supplementary Table SS2 continued..

CTT leu **L**	15824	0.01110	CCT pro **P**	10016	0.00703	CAT his **H**	19563	0.01372	CGT arg **R**	28865	0.02025
CTC leu **L**	16276	0.01142	CCC pro **P**	8402	0.00589	CAC his **H**	15114	0.01060	CGC arg **R**	32523	0.02282
CTA leu **L**	5453	0.00383	CCA pro **P**	12975	0.00910	CAA gln **Q**	21401	0.01501	CGA arg **R**	5695	0.00400
CTG leu **L**	72361	0.05076	CCG pro **P**	32179	0.02257	CAG gln **Q**	42707	0.02996	CGG arg **R**	9696	0.00680
ATT ile **I**	41550	0.02915	ACT thr **T**	12561	0.00881	AAT asn **N**	25777	0.01808	AGT ser **S**	12829	0.00900
ATC ile **I**	36299	0.02546	ACC thr **T**	33728	0.02366	AAC asn **N**	31031	0.02177	AGC ser **S**	23494	0.01648
ATA ile **I**	7339	0.00515	ACA thr **T**	10300	0.00723	AAA lys **K**	45979	0.03225	AGA arg **R**	4039	0.00283
ATG met **M**	39448	0.02767	ACG thr **T**	21684	0.01521	AAG lys **K**	14603	0.01024	AGG arg **R**	2623	0.00184
GTT val **V**	26387	0.01851	GCT ala **A**	21692	0.01522	GAT asp **D**	45132	0.03166	GGT gly **G**	34941	0.02451
GTC val **V**	22331	0.01567	GCC ala **A**	37104	0.02603	GAC asp **D**	26548	0.01862	GGC gly **G**	41781	0.02931
GTA val **V**	15833	0.01111	GCA ala **A**	28093	0.01971	GAA glu **E**	53753	0.03771	GGA gly **G**	11693	0.00820
GTG val **V**	35936	0.02521	GCG ala **A**	47542	0.03335	GAG glu **E**	24548	0.01722	GGG gly **G**	15404	0.01081