# The Evolutionary Impact of Resource Availability




Emily Ann Seward

Christ Church

Department of Plant Sciences

University of Oxford

Supervised by Dr Steve Kelly

Thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

**The Evolutionary Impact of Resource Availability**

Emily Ann Seward

Christ Church

Department of Plant Sciences

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

**Abstract**

Evolution is the change in heritable characteristics in a population over successive generations that is driven by mutation, selection and random genetic drift. Understanding the factors affecting sequence evolution and the interactions between them is of fundamental importance and facilitated by the recent proliferation of genome data. One factor that is not well understood is the role that resource availability plays in determining sequence evolution. The aim of the research described in this thesis is to address this gap in our knowledge by presenting three papers on the theme of species and sequence evolution. First, I report the discovery of a new species of *Phytomonas,* a single-celled eukaryotic plant parasite. Second, using a comparative analysis of *Phytomonas* and other single-celled plant- and animal-infecting parasites, I reveal the importance of dietary nitrogen in determining biased patterns of nucleotide use in genome and transcript sequences. Third, I examine the interaction between codon biosynthetic cost and translational efficiency in a range of bacterial species. The new results reported in this thesis identify and quantify the impact of resource availability on sequence evolution. By elucidating the contribution of resource availability to sequence evolution, and investigating how selection acting on codon resource use interacts with selection acting on codon translational efficiency, the research presented in this thesis makes a significant contribution to our understanding of sequence evolution.

**Acknowledgements**

"We often take for granted the very things that most deserve our gratitude."

I would like to start by expressing my gratitude to Steve for all the support and guidance he has provided over the last few years. Thank you for giving me the flexibility to choose the projects I was interested in and the chance to acquire such a range of techniques. I would never have learnt to code without your encouragement and I almost certainly wouldn't have become so proficient at dissecting bugs. Despite the challenges, it would not have been such an enjoyable experience without your help and scientific dreams. Though I never managed to get that micropig, I'm hoping I'll come back one day and meet the lab mantis shrimp.

I would also like to thank all the members of the Kelly and Langdale labs. In particular the interloper Jessie who may not have been a Kelangdale but who always kept me sane even from across the pond. She taught me how to write like I was running out of time but also how to take a break and I am eternally grateful. I'd also like to mention my office buddies: El (who made me feel welcome and first persuaded me to try the Kelly Lab), David (a steady patient presence), Peatrie (who quite literally makes the place sparkle), Ross (who may be in a complicated time-loop but doesn't let it stop him), Michael (whose dark sense of humour kept me cheerful), Peng (who is quietly supportive), Jack (who was such a wonderful if well camouflaged companion), Rona (who may have taken the tea shrine but has more than made up for it with her cheerful presence) and Tom, Olga, Dana & Rox (who haven't had the trial of sharing an office with me but kept me going with a healthy balance of cherries and chat).

Thank you to the Gatsby Plant Science Network for the incredibly helpful training throughout my DPhil and to the BBSRC for funding me. I would also like to thank the British Society for Plant Pathology for funding my work in the Czech Republic and to all my collaborators over there who were so welcoming. Also thank you to everyone in the Plant Sciences Department, especially Gem Toes for all her help, support and general loveliness.

Last but certainly not least I need to thank Owain and my family for their unending support and patience. Without your calm encouragement this thesis would never have been written. In particular I want to thank Owain for always having my back and keeping me sane. Crunchy peanut butter is undoubtedly superior to smooth. Mom, you are an absolute star for wading through all the scientific jargon hunting for those elusive typos. Pop, thank you for the quiet reassurance and backup. You have provided much needed perspective. I would also like to thank all my other friends and colleagues whom I haven't been able to name here. You have been the light that kept me growing and I am very grateful.

**Table of Contents**

**Abstract** ..... i

**Acknowledgements** ..... ii

**Table of Contents** ..... iii

**List of Figures** ..... vi

**Abbreviations** ..... viii

**Chapter 1: General Introduction** ..... 1

**1.1 Introduction** ..... 2

**1.2 Evolution requires variation, selection and time** ..... 3

*1.2.1 Sequence redundancy facilitates variation in mRNA nucleotide use* ..... 3

*1.2.2 Selection acts to bias synonymous codon use* ..... 3

*1.2.3 Selection is not the only contributor to codon usage bias* ..... 5

*1.2.4 Codon bias correlates with an organism's generation time* ..... 6

**1.3 Selection to reduce resource requirements can bias sequence evolution** ..... 7

*1.3.1 Monomer resource requirements vary* ..... 7

*1.3.2 Restricted monomer availability can bias usage by altering mutation rates* ..... 10

*1.3.3 Selection on monomer substitution or removal can reduce resource requirements* ..... 10

**1.4 Thesis aims** ..... 13

*1.4.1 Choice of model systems used in this thesis* ..... 13

*1.4.2 Choice of Mollicutes for model bacterial parasites* ..... 15

*1.4.3 Choice of kinetoplastids for model eukaryotic parasites* ..... 16

**1.5 Thesis plan** ..... 19

**Chapter 2:** Description of *Phytomonas oxycareni* n. sp. from the salivary glands of *Oxycarenus lavaterae* ..... 21

**2.1 Chapter Introduction** ..... 22

*2.1.1 Authors* ..... 23

*2.1.2 Author Contributions* ..... 23

*2.1.3 Abstract* ..... 23

**2.2 Introduction** ..... 24

**2.3 Results** ..... 26

*2.3.1 Material collection and primary characterisation of a new trypanosomatid species from the true bug* O. lavaterae ..... 26

*2.3.2 Phylogenetic analysis places the new trypanosomatid species in the genus* Phytomonas ..... 26

*2.3.3 Parasites within the salivary gland exhibit promastigote morphology and are proliferative* ..... 32

*2.3.4* Phytomonas oxycareni *can be found inside host cells of the salivary gland* ..... 34

*2.3.5 Taxonomic summary* ..... 37

**2.4 Discussion** ..... 38

**2.5 Methods** ..... **40**
*2.5.1 Collection and dissection of bug hosts* ..... *40*
*2.5.2 Cultivation and light microscopy* ..... *40*
*2.5.3 Transmission and scanning electron microscopy* ..... *40*
*2.5.4 PCR amplification, cloning, and sequencing* ..... *41*
*2.5.5 Phylogenetic analyses* ..... *41*

**Chapter 3:** Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms ..... **45**

**3.1 Chapter Introduction** ..... **46**
*3.1.1 Authors* ..... *47*
*3.1.2 Author Contributions* ..... *47*
*3.1.3 Abstract* ..... *47*

**3.2 Introduction** ..... **48**

**3.3 Results** ..... **52**
*3.3.1 Choice of model organisms and inference of orthogroups* ..... *52*
*3.3.2 Low nitrogen availability parasites have low nitrogen content sequences and vice-versa* ..... *56*
*3.3.3 Different metabolic strategies in the same host niche cause concomitant differences in gene sequence nitrogen content* ..... *60*
*3.3.4 Differences in genome-wide patterns of synonymous codon use are explained by selection acting on codon nitrogen content.* ..... *63*
*3.3.5 Gene expression negatively correlates with selection on mRNA nitrogen content.* *70*
*3.3.6 Low nitrogen availability parasites have ribosomal RNA sequences that use the lowest amount of nitrogen* ..... *72*
*3.3.7 Nitrogen content of nucleotide sequences can predict metabolic capability* ..... *72*
*3.3.8 Selection acting on nitrogen content is independent of selection acting on translational efficiency* ..... *75*

**3.4 Discussion** ..... **78**

**3.5 Methods** ..... **82**
*3.5.1 Data sources* ..... *82*
*3.5.2 Inference of orthogroups and construction of multiple sequence alignments* ..... *82*
*3.5.3 Evaluation of nitrogen content of nucleotide sequences* ..... *83*
*3.5.4 Analysis of rRNA* ..... *83*
*3.5.5 Statistical tests* ..... *84*
*3.5.6 A model for synonymous codon use under the joint pressures of selection and mutation bias* ..... *86*
*3.5.7 Calculation of codon tRNA adaptation index (tAI) values* ..... *88*
*3.5.8 Model fitting and implementation* ..... *89*
*3.5.9 Classification of additional species using the metabolic model for synonymous codon use* ..... *90*

**Chapter 4:** Selection-driven cost-efficiency optimisation of transcripts governs evolutionary rate in bacteria ..... **93**

**4.1 Chapter Introduction** ..... **94**
*4.1.1 Authors* ..... *95*
*4.1.2 Author Contributions* ..... *95*
*4.1.3 Abstract* ..... *95*

**4.2 Introduction** ..... **96**

**4.3 Results** ..... **98**
*4.3.1 55% of bacteria exhibit a significant trade-off between codon biosynthetic cost and translational efficiency* ..... *98*
*4.3.2 Selection acting to minimise biosynthetic cost and maximise translational efficiency of transcript sequences is independent of codon cost-efficiency trade-off* ..... *100*
*4.3.3 Genes that experience the strongest selection for increased transcript translational efficiency are also under the strongest selection to minimise biosynthetic cost* ..... *103*
*4.3.4 Sequence optimisation for cost and efficiency constrains molecular evolution rate* ..... *105*

**4.4 Discussion** ..... **108**

**4.5 Methods** ..... **109**
*4.5.1 Data sources* ..... *109*
*4.5.2 Evaluation of translational efficiency (tAI)* ..... *110*
*4.5.3 Calculation of relative codon cost and efficiency* ..... *110*
*4.5.4 CodonMuSe: A fast and efficient algorithm for evaluating drivers of codon usage bias* ..... *110*
*4.5.5 Comparing selection acting on codon bias and transcript abundance levels* ..... *111*
*4.5.6 Calculating the extent to which gene sequences were jointly optimised for cost and efficiency* ..... *111*
*4.5.7 Calculation of molecular evolution rates* ..... *113*

**Chapter 5:** General Discussion ..... **115**

**5.1 Chapter Introduction** ..... **116**

**5.2 General Discussion** ..... **117**
*5.2.1 Resource availability exerts a selective pressure on genome evolution* ..... *117*
*5.2.2 Selection is stronger for genes that are more highly expressed* ..... *118*
*5.2.3 Just three factors can explain the majority of genome-wide codon bias* ..... *119*
*5.2.4 Selection acting on codon cost can be used to infer dietary nitrogen input* ..... *122*
*5.2.5 tRNA pools can alter the trade-off between codon cost and efficiency* ..... *122*
*5.2.6 Bacteria are generally under selection to reduce codon cost & increase translational efficiency* ..... *123*
*5.2.7 Selection acting on resource allocation can constrain the rate of molecular evolution* ..... *126*

**5.3 Overall Summary** ..... **127**

**References** ..... **128**

**Appendices** ..... **139**

**Appendix 1:** ..... **140**

**Appendix 2:** ..... **141**

**List of Figures**

**Figure 1.01** Simplified nucleotide biosynthetic pathways…… **8**

**Figure 1.02** Nucleotide monomers require different quantities of nitrogen atoms……....**9**

**Figure 1.03** The elemental inputs of cellular macromolecules vary ……..**9**

**Figure 2.01** New kinetoplastid parasite belongs to the subfamily Phytomonadinae…..... **28**

**Figure 2.02** Light microscopy images of *Phytomonas oxycareni* n. sp.……..…….. **31**

**Figure 2.03** Scanning electron microscopy of *Phytomonas oxycareni* n. sp.……..... **33**

**Figure 2.04** Transmission electron microscopy of an infected salivary gland containing *Phytomonas oxycareni* n. sp. ……...…… **35**

**Figure 2.05** Transmission electron micrographs of three intracellular *Phytomonas oxycareni* n. sp. within the salivary glands of the insect host *Oxycarenus lavaterae*…… **36**

**Figure S2.01** *Phytomonas oxycareni* n. sp. spliced leader (SL) RNA gene differs from other *Phytomonas* species by one nucleotide substitution.…..……... **30**

**Figure S2.02** AU test confirms that *Phytomonas oxycareni* n. sp. is a sister species to all other *Phytomonas* sp. ..……..... **43**

**Figure 3.01** Phylogenetic trees of the parasites used in this study shaded according to their host and metabolic strategy…… **53**

**Figure 3.02** Nitrogen availability influences gene sequences. ……..…..... **58**

**Figure 3.03** Analysis of gene nitrogen content of three parasites that occupy the same host niche but utilise different metabolic strategies…….…..... **61**

**Figure 3.04** A selection-mutation model for synonymous codon use in kinetoplastids explains relative synonymous codon use with ~90% accuracy and recapitulates the difference in nitrogen cost of genes…… **62**

**Figure 3.05** Model for synonymous codon use for Mollicutes explains relative synonymous codon use with ~94% accuracy and recapitulates the difference in nitrogen cost of genes**. 65**

**Figure 3.06** Selection-mutation model of codon use can predict the metabolic capacity of parasites from raw nucleotide sequences…….……..... **74**

**Figure S3.01** Phylogenetic trees and metabolic information for the parasites used in this study……..……..……..……..…..... **54**

**Figure S3.02** ATP generating metabolic pathways…….……....…..... **55**

**Figure S3.03** $L_N$ parasites use the least amount of nitrogen in their amino acid side chains compared to $M_N$ and $H_N$ parasites…….....……..... **59**

**Figure S03.4** The model for synonymous codon use under the joint pressures of selection acting on nitrogen content and mutation bias fits real codon use with > 90% accuracy..... **66**

**Figure S03.5** The model that considers both mutation bias and nitrogen content selection in combination provides a better fit than either parameter considered in isolation…….. **68**

**Figure S3.06** Boxplots showing distribution of $2N_gs$ values for individual species…… **69**

**Figure S3.07** Gene expression negatively correlates with selection acting on mRNA nitrogen content…….……..... **71**

**Figure S3.08** Gene clusters for nitrogen liberating metabolic pathways in $H_N$ Mollicute parasites…….…….... **73**

**Figure S3.09** The model parameters which provide the best fit between observed and predicted codon use also provide the best percentage of correctly predicted optimal codons and correctly ordered codon use…..……..…….…..... **77**

**Figure S3.10** Example distribution of the model when run with shuffled codon nitrogen content ……..…..... **85**

**Figure 4.01** tRNA sparing strategies alter a species' codon cost-efficiency trade-off …… **99**

**Figure 4.02** Bacterial genomes show selection to minimise nucleotide cost ($-S_c$) and maximise translational efficiency ($+S_t$) ...……...…….……..... **102**

**Figure 4.03** The genes under the strongest selection for translational efficiency ($+S_t$) are also under the strongest selection to minimise nucleotide cost ($-S_c$)..…….…... **104**

**Figure 4.04** Selection acts in proportion to mRNA abundance to decrease codon biosynthetic cost and increase codon translational efficiency in *Escherichia coli*. …… **104**

**Figure 4.05** Selection-driven optimisation of resource allocation is a critical factor that determines molecular evolution rate ……...……..…..... **107**

**Figure S4.01** Example cost-efficiency Pareto frontier for a short amino acid sequence… **112**

**Abbreviations**

| | |
|---|---|
| A/T/U/G/CTP | Adenosine/Thymidine/Uridine/Guanosine/Cytidine triphosphate |
| a.s.l | Above sea level |
| bp | base pairs |
| °C | Degrees Celsius |
| C:N | Carbon to nitrogen ratio |
| DAPI | 4',6-diamidino-2-phenylindole |
| DNA | Deoxyribonucleic acid |
| (d)NDPs/ NTPs | (deoxy)nucleotide diphosphates/ triphosphates |
| GC3 | GC-content in the third codon position |
| $K_a$ / $K_s$ | Non-synonymous/synonymous mutation rate |
| kb | kilo base pairs |
| kV | kilo volts |
| $L_N$/ $M_N$ /$H_N$ | Low/Medium/High nitrogen availability species |
| m or Mb | Mutation bias |
| μm | Micrometer |
| μL | Microlitre |
| Mbp | Mega base pairs |
| $NAD^+$ | Nicotinamide Adenine Dinucleotide |
| $N_e$ | Effective population size |
| PFR | Paraflagellar rod |
| RNA | Ribonucleic acid |
| s | seconds |
| $S_t$ | Selection on transcript translational efficiency |
| $S_c$ | Selection on transcript biosynthetic cost |
| SH | Shimodaira-Hasegawa |
| sp. | species |
| SSU | Small subunit |
| tAI | tRNA Adaptation Index |
| TPM | Transcripts per million |
| tRNA | transfer-RNA |
| tRNA-Met(CAT) | Methionine tRNA with the anticodon CAT |
| $2N_g s$ | Nitrogen dependent selection bias |

# Chapter 1: General Introduction

## 1.1 Introduction

Single cell organisms inhabit a diverse range of environments. These range from boiling sulphuric acid (*Sulfolobus sp.*) to deep-sea hydrothermal vents (*Beggiatoa sp.*) and from the mid-gut of hematophagous mosquitos (*Plasmodium sp.)* to tomato fruits (*Phytomonas sp.)* (Nelson, Wirsen, and Jannasch 1989; Mathur et al. 2007; Sinden et al. 2010; Jaskowska et al. 2015). Such varied niches are made habitable by evolutionary adaptations to those environments. For example *Sulfolobus acidocaldarius* has multiple DNA repair mechanisms that allow it to cope with elevated rates of DNA damage at high temperatures (Chen et al. 2005). While we know that exposure to environmental stresses such as high temperature and UV irradiation can alter sequence evolution (Rastogi et al. 2010; Vieira-Silva and Rocha 2010), far less is known about the contribution made by other environmental factors.

An important environmental factor, whose contribution to sequence evolution requires additional consideration, is resource availability. Due to the fundamental importance of resource availability in determining an organism's survival, multiple adaptive and acclimation responses are observed when essential resources become limited. These responses broadly fall into the categories of increasing resource acquisition or reducing resource requirements. For example nutrient-limited *Bacillus subtilis* increases resource acquisition by triggering starvation-induced fratricide and cannibalising other cells (González-Pastor 2011). By comparison, phosphorous-limited phytoplankton can reduce phosphorous requirements by substituting highly phosphorous-demanding cellular constituents such as phospholipids with sulfolipids (Van Mooy et al. 2006). In the extreme case of *Borrelia burgdorferi*, the requirement for iron has been completely circumvented by the loss of genes that encode proteins with an iron cofactor; manganese-dependent metalloproteins are used instead (Posey and Gherardini 2000). Thus, organisms experiencing resource limitation can undergo evolutionary changes to adjust the composition of their macromolecules so as to better match their environmental resource availability.

### 1.2 Evolution requires variation, selection and time

In order to understand the way in which resource limitation can alter sequence evolution, it is first necessary to consider the different components required for evolution. At the simplest level evolution requires variation in a population on which selection can act so that over time the fittest variants become predominant in that population (Darwin 1859). This applies not only to whole ecosystems or species populations but also to individual genes or cellular macromolecules such as DNA, RNA and protein. In this thesis I focus on evolution acting on DNA and RNA sequences and therefore in the subsequent sections I will focus on mRNA to exemplify the underlying requirements of evolution (variation, selection and time).

#### 1.2.1 Sequence redundancy facilitates variation in mRNA nucleotide use

The first requirement of evolution is variation in the population. In the context of mRNA, this refers to variation in the codon population used to code for a given protein sequence. Certain amino acids, such as methionine, are only coded for by one codon and therefore there is no variation for selection to act on. However, the majority of amino acids are coded for by multiple synonymous codons. For example the amino acid valine is encoded by four synonymous codons that all start 'GU' but can have A, U, C or G in the third codon position. This genetic redundancy means that the codon population can vary and selection can act on synonymous codon use to alter the underlying nucleotide sequence without altering the encoded amino acid. This non-uniform usage of synonymous codons is known as codon usage bias.

#### 1.2.2 Selection acts to bias synonymous codon use

The second requirement of evolution is a selective advantage in using one variant over another. As I shall now detail, several selective advantages of using one synonymous codon over another have been proposed. For example codon usage bias was first proposed to be influenced by variance in iso-accepting tRNA abundance altering codon translational efficiency (Ikemura 1981). Translational efficiency is the rate at which mRNA is translated

into proteins and has subsequently been found to depend both on tRNA abundance and the strength of codon-anticodon coupling (dos Reis, Wernisch, and Savva 2003; Sabi and Tuller 2014). tRNA abundance can vary as synonymous codons can be recognised by multiple tRNAs, termed iso-accepting tRNAs. These iso-accepting tRNAs facilitate 'wobble' or redundancy in the third position of a codon so that a full complement of tRNAs corresponding precisely to each codon is not required. Indeed, tRNA abundance has been found to vary between organisms and even between tissues in the same organism (Plotkin, Robins, and Levine 2004). Ikemura showed that tRNA abundance correlates with the occurrence of the corresponding codons in protein coding genes, thus linking tRNA abundance and codon bias and proposing that selection acts on synonymous codon choice to optimise translation (Ikemura 1981). Differential use of synonymous codons has also been directly linked to the speed of translation. For example the *lacZ* gene encoded by translationally efficient codons was translated three seconds faster than the same gene encoded by less translationally efficient codons (Sørensen, Kurland, and Pedersen 1989). Furthermore, selection acting on translational efficiency is expected to be stronger for more highly expressed genes (Drummond and Wilke 2009). Indeed, the use of certain codons can increase protein production by more than 1,000-fold due to altering the speed of ribosomal translation (Gustafsson, Govindarajan, and Minshull 2004). Therefore selection acting on codon translational efficiency can bias codon use and alter the rate of mRNA translation.

Selection may also act on codon bias due to alteration of translational accuracy. Translational accuracy is the fidelity with which the mRNA is translated into the correct protein sequence that folds properly within the cell (Drummond and Wilke 2008; Lee et al. 2010). A decrease in amino acid misincorporation can be achieved by using codons that increase fidelity of tRNA discrimination, facilitate more accurate proofreading or minimise the deleterious impact of a misincorporated amino acid. Selection for translational accuracy has been shown to bias synonymous codon usage in *Drosophila* and to lead to enhanced

accuracy of protein synthesis (Akashi 1994). Both translational speed and accuracy are important factors contributing to selection acting on codon bias. However, by studying codon usage in high and low expression genes and comparing conserved and variable sites within highly expressed genes, selection for translational speed was shown to be the dominant factor acting to bias codon usage in fast-growing bacteria compared to selection for accuracy (Ran and Higgs 2012).

Selection can also act to bias codon use in thermophiles (organisms living in high temperatures). Since RNA-RNA interactions are primarily entropy-driven (Lauffer 1975), unwanted RNA-RNA interactions are particularly problematic at higher temperatures where they can interfere with translation (Fire 1999). To combat unwanted RNA-RNA interactions, thermophiles have been shown to have an increased proportion of purines (A and G nucleotides) in their RNA sequences compared to non-thermophiles (Lao and Forsdyke 2000). This "purine-loading" reduces the extent to which double-stranded RNA can form, as purines are not able to pair with one another. Therefore the selective drive for thermophiles to purine-load can impact on codon choice irrespective of other constraints on protein structure and function (Lao and Forsdyke 2000; Lambros, Mortimer, and Forsdyke 2003).

Differential usage of synonymous codons can have a direct fitness effect. For example, four different synonymous mutations in the S20 ribosomal protein of *Salmonella enterica* directly affected the organism's fitness by altering growth rates (Knoppel, Nasvall, and Andersson 2016). Therefore changes in synonymous codon use can provide a competitive advantage and lead to changes in that codon's usage frequency over multiple generations.

### 1.2.3 Selection is not the only contributor to codon usage bias

Though selection undoubtedly acts to bias codon use, not all codon bias is necessarily due to selection. It has been argued that discrepancy in codon usage between taxa is due to neutral

processes and a result of mutational biases during DNA replication and repair (Eyre-Walker 1991; Francino and Ochman 1999). For example, GC-biased gene conversion, a recombination event in which the GC-rich variant is used preferentially during mismatch repair, can alter genome GC-content and bias codon use (Galtier 2003; Lassalle et al. 2015). Alternatively, it has been suggested that the loss or inefficiency of enzymes such as dUTPase can cause AT biases in the genome during replication (Williams and Pollack 1990; Pollack, Williams, and McElhaney 1997). These neutral processes, driven by inherent mechanistic biases, are sometimes argued to be the main pressures acting on codon bias (Rao et al. 2011).

### 1.2.4 Codon bias correlates with an organism's generation time

The final requirement of evolution is sufficient time, both for mutations to appear and for them to become fixed in a population. For neutral mutations in diploid sexually recombining species, the rate of mutation fixation depends purely on the effective population size ($N_e$) and is approximately equal to $4N_e$ generations (Kimura and Ohta 1969). It becomes more complex when a mutation is under selection but the mean time to fixation can be approximated as $(\frac{4}{|s|})\ln(2N_e)$ generations, where s is the selection coefficient (s > 0 is a beneficial mutation, s < 0 is deleterious) (Crow and Kimura 1970). Therefore mutations under selection will reach fixation or be lost in a population much faster than neutral mutations. Furthermore, species with larger effective population sizes or long generation times take longer to fix a mutation. This can be seen when comparing codon bias in organisms with different generation times. Codon bias in housekeeping genes is shown to have a strong inverse relationship with species generation time; species with a shorter generation time show a codon bias four orders of magnitude higher than the codon bias of organisms such as humans, which have a long generation time (Subramanian 2008). Time is therefore an important component determining sequence divergence between species.

### 1.3 Selection to reduce resource requirements can bias sequence evolution

Having introduced the requirements for evolution to alter genome sequences (variation, selection and time), I shall now highlight our current understanding of the role that resource limitation can play in determining sequence evolution of cellular macromolecules such as proteins, RNA and DNA. As noted in section 1.1, an adaptive response to resource limitation is to reduce cellular resource requirements. I will use nucleotides to exemplify how this can be achieved by biasing monomer usage.

#### 1.3.1 Monomer resource requirements vary

Nucleic acid polymers such as DNA and RNA are composed of nucleotide monomers. As noted in section 1.2.1, functional constraints dictate nucleotide use to some extent but degeneracy in the genetic code permits a degree of redundancy. This redundancy means that different nucleotides can be used to code for the same amino acid sequence. Since nucleotides are structurally similar but elementally distinct, they are synthesised via interlinked metabolic pathways that require different inputs. For example, nucleotide biosynthetic pathways differ in complexity and energetic requirements (Fig. 1.01): GTP production requires an additional $NAD^+$ compared to ATP production; and CTP is synthesised from UTP using additional ATP. Overall, this means that in DNA sequences GC base pairs require more energy to produce than AT base pairs. Similarly, purines require more inputs than pyrimidines at the RNA level (Chen et al. 2016).

Nucleotides not only differ in their energy requirements but also in their nitrogen content (Fig. 1.02 & 1.03). dTTP has the highest C:N ratio (10 carbon and 2 nitrogen atoms), (d)CTP is intermediate and (d)ATP and (d)GTP have the lowest C:N ratio (10 carbon and 5 nitrogen atoms). Thus the same amino acid sequence can be coded for by multiple mRNA sequences whose resource requirements vary. In resource-limited environments, changes in mutation rates and selection acting on nucleotide composition could therefore reduce resource requirements by preferentially using nucleotides with lower costs.

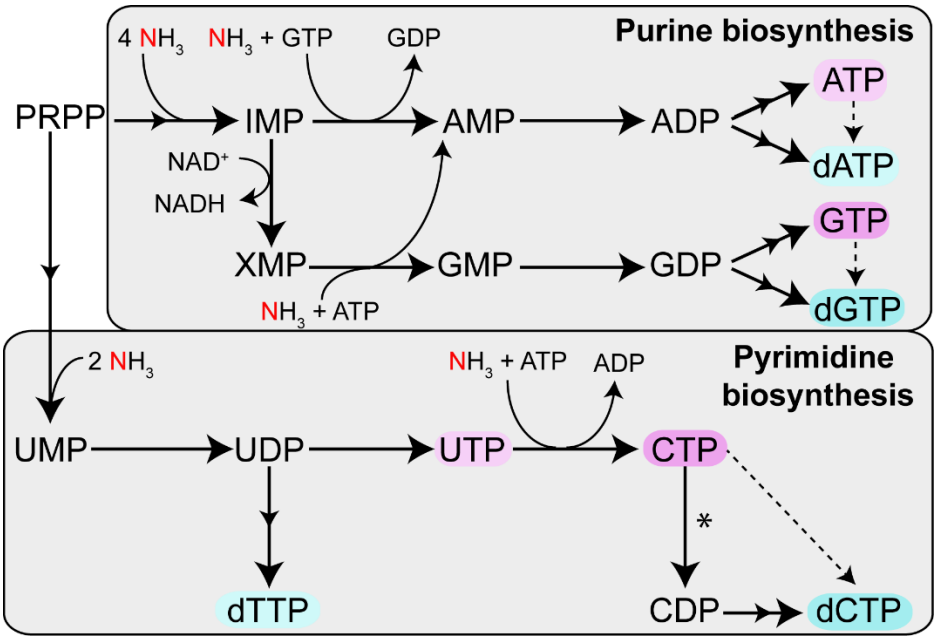



**Figure 1.01 - Simplified nucleotide biosynthetic pathways.** Synthesis of GTP and CTP demands more energy than ATP and UTP. Furthermore, the majority of organisms lack the anaerobic ribonucleoside-triphosphate reductase responsible for direct conversion of NTPs to dNTPs (dashed arrow). Instead, dNTP synthesis proceeds via NDPs; however, in the case of dCTP this can create a bottleneck (indicated by the *) as it requires CTP consumption to give CDP. This reaction proceeds in the opposite direction to enzymes such as nucleoside diphosphate kinase and thus energy is lost in futile cycling. Double arrows indicate simplified stages in the pathways where multiple enzymes are required to facilitate the conversion indicated. The nitrogen requirements of the nucleotides are also exhibited in red, emphasising the additional nitrogen required for synthesis of CTP from UTP and of the purines compared to the pyrimidines.

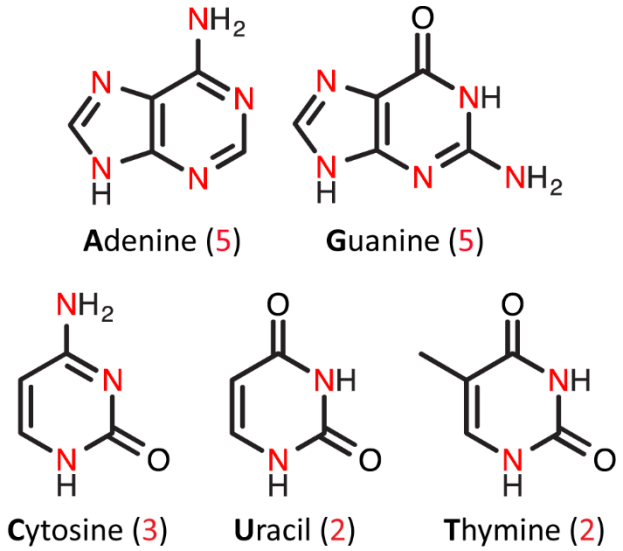



**Figure 1.02 – Nucleotide monomers require different quantities of nitrogen atoms.** Purines (Adenine and Guanine) require five nitrogen atoms per monomer compared to the pyrimidines (Cytosine, Uracil and Thymine) which require only two or three nitrogen atoms.

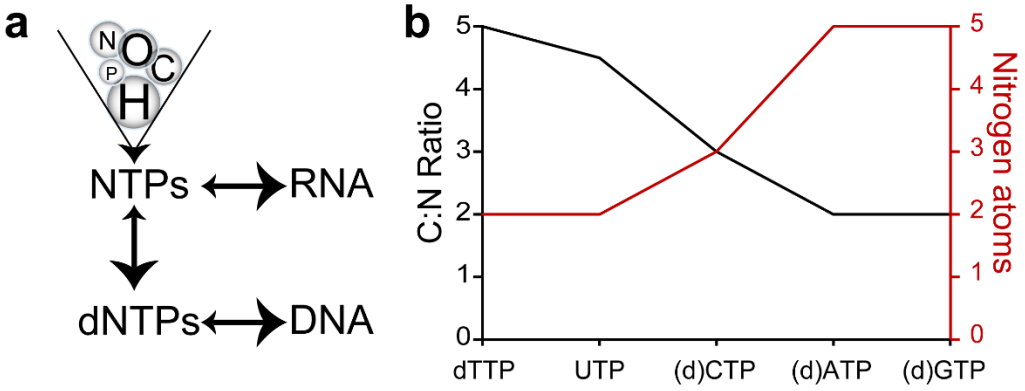



**Figure 1.03 – The elemental inputs of nucleotides** a) RNA and DNA are composed of (deoxy)nucleotides, (d)NTPs, which each require varying elemental inputs; b) The carbon to nitrogen ratio of (d)NTPs depends on the number of nitrogen atoms required per monomer.

### 1.3.2 Restricted monomer availability can bias usage by altering mutation rates

Balanced pools of cellular NTPs and dNTPs are required for accurate synthesis and replication of RNA and DNA strands (Buckland et al. 2014). However, mutations can lead to imbalanced pools of nucleotides. These imbalanced nucleotide pools increase replication errors by driving dNTP mismatch during replication (e.g. the erroneous matching of dGTP with dTTP). Furthermore imbalanced nucleotide pools can reduce proofreading capacity (e.g. due to rapid transcript extension such that the polymerase's proofreading is not able to keep up) (Buckland et al. 2014). This indicates that varying nucleotide monomer availability, as a result of a genetic mutation or due to a reduced capacity to synthesise that particular monomer, can also bias monomer usage by altering mutation rates. Therefore, differences in monomer resource requirements and availability can bias usage by altering mutation rates and can provide a platform of variation on which selection can act.

### 1.3.3 Selection on monomer substitution or removal can reduce resource requirements

Given that monomer input requirements vary and there is some redundancy as to which monomers are incorporated into a given macromolecule, cellular resource requirements can be reduced by substituting resource-intensive monomers or shortening the sequence by removing those monomers entirely. For example, it has been suggested that diploid sexual taxa with streamlined genomes have an advantage in phosphorous-limited environments compared to polyploid asexual taxa which have larger, more phosphorous-demanding genomes (Neiman, Kay, and Krist 2013). Genome streamlining has also been linked to the elemental cost of growth, with rapidly dividing bacteria reallocating phosphorous and nitrogen from DNA to RNA in resource-limited environments (Hessen et al. 2010). Therefore reduction in genome size can reduce cellular resource requirements.

In situations where it is not possible to reduce genome size, monomer substitution can reduce resource requirements. Such substitutions are manifested as usage biases in organisms that have evolved in conditions where there is a persistent elemental limitation.

For example such biases can be seen when comparing domesticated crops and wild crops (Acquisti, Elser, and Kumar 2009). This study demonstrated that domesticated crops, which have been cultivated with fertilisation for thousands of years, and nitrogen-fixing plants show increased use of nitrogen-rich nucleotides in the transcribed strand of their intergenic regions compared to wild plants, which are relatively nitrogen-starved. It has also been proposed that parasitic bacteria have more AT-rich genomes compared to free-living bacteria due to the differences in energetic cost associated with nucleotide biosynthesis (Rocha and Danchin 2002). Therefore resource restrictions can alter monomer usage in species under persistent changes in resource availability.

Alterations in monomer use due to resource restrictions are not only seen genome-wide in persistent conditions of resource limitation but also in individual sequences. For example, studies in *E. coli* and *S. cerevisiae* have shown that the bacterial and fungal enzymes required for metabolic processing of an element, such as the enzymes which constitute the carbon assimilatory pathway, have reduced quantities of that element in their sequences compared to levels found in the whole genome (Baudouin-Cornu et al. 2001). Furthermore, multiple metabolic pathways are capable of carrying out the same overall reaction, and which pathway is used depends on the organism's environmental conditions. For example *E. coli* has been shown to utilise two distinct metabolic pathways depending on nutrient availability (Carlson 2007). The first is more thermodynamically efficient but is only expressed in conditions of sufficient nutrient availability as it requires inputs that are expensive to synthesise. The second pathway has reduced thermodynamic efficiency but has a low synthesis cost; it is expressed in nutrient-limited conditions. Therefore resource limitation can act specifically on individual genes in particular pathways as well as causing genome-wide evolutionary changes to nucleotide sequences.

Evolutionary changes in response to resource limitation are not restricted to nucleotide sequences. It has been shown generally that there is a negative correlation

between the abundance of particular proteins within a cell and the atomic requirements of their constituent monomers (Li, Lv, and Niu 2009); in other words, the more abundant a protein is, the cheaper it is to synthesise. A similar example of metabolic optimisation is seen in cyanobacteria, which encode sulphur-depleted versions of their most abundant proteins; these are specifically expressed under sulphur-limiting conditions (Mazel and Marlière 1989). Another study has compared nitrogen use in plant and animal proteins. The authors found a 7.1% reduction in nitrogen use in plant-protein side chains relative to animal-protein side chains as plants are more nitrogen-limited than animals. They also found that catabolic proteins, which are generally expressed during times of nutrient limitation, were particularly nitrogen-poor in comparison to anabolic enzymes and the genome as a whole (Acquisti, Kumar, and Elser 2009). This work was followed up in a study that looked at both protein elemental sparing and codon usage bias in bacteria (Bragg et al. 2012). The authors identified significant correlations between carbon and sulfur usage and adaptive codon usage bias across 148 bacterial species, indicating a universal mechanism of selection. This lends support to the idea that nutrient scarcity influences monomer usage patterns for macromolecule synthesis. Therefore, though protein evolution is mainly determined by constraints on activity, specificity, folding and stability, other constraints such as nutritional limitations also play a role.

### 1.4 Thesis aims

Though evidence is mounting to support the importance of resource availability as an evolutionary selective pressure operating on genome sequences, current studies have been limited to comparison either between non-coding regions or between whole genes. Molecular changes in coding sequences and the alteration of evolutionary rates due to resource limitation remained less well studied. Moreover, a mathematical framework for evaluating the contribution of elemental limitation to the evolution of gene sequences was lacking. My thesis aims to address these gaps in our understanding by addressing two overarching goals: 1) To elucidate the role of resource limitation in determining sequence evolution; 2) To develop a mathematical framework to enable estimation of the contribution of elemental limitation to sequence evolution.

These overall goals can be subdivided into 4 aims:

1. Determine whether differences in nitrogen availability cause concomitant differences in DNA and protein sequence evolution.
2. Develop a mathematical framework for estimating the strength of selection acting on gene sequences.
3. Evaluate the interplay between selection acting on biosynthetic cost with selection acting on other features of gene sequences.
4. Examine how selection acting on multiple factors does so either synergistically or antagonistically to determine gene evolutionary rate.

#### 1.4.1 Choice of model systems used in this thesis

To address these four aims I decided to begin by looking at parasites. There are two main advantages to using parasites for this study. First, monophyletic lineages of parasites have evolved to obtain their energy from a limited pool of host biomolecules. This means that it is relatively easy to identify the organism's main energy source and the range of available nutrients. Second, parasites in animal hosts are subject to different levels of nitrogen

availability than plant-infecting parasites (Acquisti, Elser, and Kumar 2009). The same study conducted on free-living organisms would introduce additional uncertainty as they can be found in multiple environments with the capacity to use multiple energy sources. Parasites are therefore an ideal dataset for examining the impact of metabolic specialisation on genome evolution.

In order to undertake statistically significant comparisons of nucleotide usage bias between groups of species with different elemental availabilities, sufficient genome data was required. Furthermore, in order to determine whether the findings in one group of parasites were lineage specific or were more broadly consistent in distinct evolutionary lineages, sufficient genome data was required in both bacterial and eukaryotic groups of parasites. The bacterial class of parasites chosen were the Mollicutes due to their broad host range and large number of fully sequenced genomes. Though bacterial genomes were readily available, analogous datasets in parasitic eukaryotes were limited. For example, all of the available apicomplexan parasite genomes are intracellular. Similarly, the majority of fungal parasites that have been sequenced also survive outside of their predominant host. The kinetoplastida were selected as an ideal group to study as, from a single origin of parasitism, they have adapted to colonise plant and animal hosts and have adapted to both intracellular and extracellular lifestyles. Therefore I conducted the initial analysis on two groups of single-celled parasites; the Mollicutes (bacteria) and the kinetoplastida (eukaryotes).

I shall now introduce these two groups of parasites. In particular I want to highlight the differences in metabolic properties between the different species as this will alter a species' dietary nitrogen availability. For example, species metabolising sugars are considered nitrogen-limited in comparison to species metabolising proteins or other nitrogen-containing compounds (which release nitrogen as a by-product of energy production). These differences in metabolism and dietary nitrogen availability provide the basis for the comparison of nucleotide use in groups of species with different elemental

availabilities presented in Chapter 3. The following text (Sections 1.4.2 and 1.4.3) is adapted from the supplementary file that was originally published as part of Seward and Kelly 2016.

### 1.4.2 Choice of Mollicutes for model bacterial parasites

Mollicutes are a class of parasitic bacteria that have undergone dramatic genome reduction resulting in genomes smaller than 1.5 Mbp (Grosjean et al. 2014). For example, the plant-infective genera of Mollicute bacteria, *Phytoplasma,* have jettisoned many of the genes required for amino-acid metabolism and adapted to a phytopathogenic lifestyle by evolving specialised sugar uptake and metabolism strategies (Oshima et al. 2004; Kube et al. 2008). These include: loss of ATP-synthase genes, previously considered indispensable for life, and generation of ATP in a manner that is highly dependent on the glycolytic pathway (Oshima et al. 2004). *Candidatus P. mali* has gone one step further and jettisoned the genes for glycolysis; it has been proposed that it generates ATP through catabolism of malate and maltose (Kube et al. 2008, 2012).

The closely related genera of Mollicute bacteria, *Mycoplasma*, are animal-infective and colonise a range of tissues from the bloodstream to the synovial fluid of joints and the respiratory tract epithelium (Waites et al. 2004; Shu et al. 2011; Nascimento et al. 2012). The environments that the *Mycoplasma* evolved to occupy have shaped the parasite's metabolism. For example, *M. synoviae* is a major poultry pathogen that infects the synovial fluid and respiratory tract (Dušanic et al. 2014). Both cells on the upper respiratory tract and lubricin, a major component of the synovial fluid, have sialic acid residues. *M. synoviae* is able to cleave off, import and metabolise sialic acid in a pathway that liberates nitrogen in the form of ammonia and also feeds the glycolytic pathway for ATP generation. This pathway is not common in *Mycoplasma* and is an example of metabolic specialisation in adaptation to a specific host environment.

Host environment is not the only factor that determines the parasite's metabolism. Multiple metabolic strategies can be employed within the same ecological niche. For

example, three Mollicute bacteria (*M. hominis, M. genitalium* and *Ureaplasma parvum*) all reside in the same urogenital tract niche but have distinct energy generation strategies (Pereyre et al. 2009). *M. genitalium* and *U. parvum* metabolise glucose and urea respectively. *M. hominis,* which has the second smallest genome among self-replicating free-living organisms (only 537 coding sequences), has done away with ATP generation via glycolysis and instead generates ATP via arginine catabolism (Pereyre et al. 2009). This provides a good opportunity to study the impact of differential metabolic strategies on genome evolution, independent of environmental differences.

This type of metabolic tailoring can also be seen through comparative analysis of multiple *Mycoplasma* genomes. Differences in enzymatic pathways reflect the different genes required for carbohydrate, amino acid and nucleotide metabolism as well as host-pathogen interaction. *Mycoplasma* have different enzymatic compositions that reflect their host environment and differ between species (Arraes et al. 2007).

Interestingly, nitrogen metabolism in Mollicutes is perturbed. The canonical pathway for nitrogen assimilation, via the GS-GOGAT pathway or glutamate dehydrogenase (GDH), is not functional as the genes for glutamine synthetase (GS), glutamate synthase (GOGAT) and GDH are missing from Mollicute genomes (Amon, Titgemeyer, and Burkovski 2010). Furthermore only the *Ureaplasma* appear to have the ammonium transporters *amt1* and *amt2* (Amon, Titgemeyer, and Burkovski 2010). This indicates that these bacteria are employing unusual nitrogen utilisation strategies.

### 1.4.3 Choice of kinetoplastids for model eukaryotic parasites

The kinetoplastida are a group of widespread single-celled eukaryotic parasites that infect a broad range of hosts (Votýpka et al. 2010). These parasites cause a global health burden on both humans and crops and have a dixenous lifestyle (Jaskowska et al. 2015). This means they have to adapt to two host environments and are exposed to conditions as disparate as tomato fruits and the insect mid-gut (Jackson 2014; Jaskowska et al. 2015).

There is a plant-specific genus of kinetoplastida called *Phytomonas*. These parasites have evolved multiple strategies to live in the carbohydrate-rich environment of plants. Firstly, they utilize abundant plant sugars for their primary metabolism, feeding the glycolytic pathway via sucrose and trehalose metabolism (Porcel et al. 2014). Secondly, *Phytomonas* have also been shown, despite depending on oxidative metabolism, to be able to survive without heme (Ayala and Luke 2012). This is a unique metabolic adaptation as heme was thought to be a universally essential protein cofactor for fundamental cellular processes. Finally, unlike other kinetoplastida, *Phytomonas* do not undergo a metabolic switch from carbohydrate to amino acid metabolism when in their insect vectors (Jaskowska et al. 2015). This is possible due to the restricted feeding of their insect hosts on carbohydrate-rich plant juices. This may have facilitated the loss of a number of mitochondrial pathways such as the respiratory chain required for beta oxidation of fatty acids and the complete oxidation of amino acids (Porcel et al. 2014).

Animal-infective kinetoplastids belonging to the genera *Leishmania* and *Trypanosoma* undergo metabolic shifts between their insect and animal hosts. For example *Trypanosoma brucei*, the causative agent of sleeping sickness in humans, undergoes significant morphological and metabolic adaptations between infecting the human bloodstream and the alimentary tract and salivary glands of tsetse flies, its insect vector (Mazet et al. 2013). *T. brucei* preferentially metabolises glucose, relying exclusively on glycolysis for ATP production in the bloodstream (Spitznagel et al. 2009). However, if the glucose supply is limited or spent, such as in the tsetse fly mid-gut, *T. brucei* is able to switch metabolism to catabolism of amino acids such as proline (Mazet et al. 2013).

Host metabolism is also known to affect parasite growth. For example, proliferation of *T. cruzi,* the causative agent of human Chagas' disease has been shown to be linked to host metabolism (Caradonna et al. 2013). In particular, host nucleotide metabolism and energy production as well as fatty acid oxidation are key cellular processes that drive

intracellular *T. cruzi* growth. Targeted changes to these host metabolic pathways modulated the parasite's ability to replicate, emphasising the flexibility of this intracellular pathogen to deal with fluctuating conditions in its host (Caradonna et al. 2013).

Similar to *Trypanosoma,* there is a metabolic shift between the insect- and animal-infective stages of *Leishmania mexicana*. The change between the *L. mexicana* promastigote stage (insect-infective) and amastigote stage (infects the parasitophorous vacuole of white blood cells) produces a concomitant change in metabolism. There is a shift in the cells from using glucose and protein as their carbon source to beta-oxidation of fatty acids and increased use of amino acids (Fiebig, Kelly, and Gluenz 2015). This shift reflects the changes in substrate availability as the parasitophorous vacuole, a major site of protein degradation, is thought to be generally rich in amino acids and poor in sugars (McConville and Naderer 2011). In fact, adding exogenous arginine or ornithine to infected macrophages stimulates the growth of intracellular amastigotes. This implies that amastigote replication may be limited by the availability of these amino acids.

### 1.5 Thesis plan

Having decided to study Mollicutes and kinetoplastids, it was necessary to obtain sufficient data to conduct my analysis. Though Mollicute genomes were readily available, at the start of my project the availability of plant-infecting parasite genomes in the kinetoplastida was limited. To address this I undertook an expedition to the Czech Republic to search for new plant-infecting kinetoplastida species (*Phytomonas*) to augment the available dataset. The discovery of a new species of *Phytomonas* is reported in Chapter 2. During the initial stages of my PhD, data from an additional species (*Phytomonas françai*) became available, meaning that there was sufficient species data to conduct the comparative analysis that is presented in Chapter 3. Therefore, as soon as I had enough data to publish a paper describing the new species I terminated this work and switched focus to the main topic of the thesis. The research described in this thesis is therefore divided into five chapters:

- Chapter 1: General introduction
- Chapter 2: Description of *Phytomonas oxycareni* n. sp. from the Salivary Glands of *Oxycarenus lavaterae*
- Chapter 3: Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. This chapter addresses my first two thesis aims 1) To determine whether differences in nitrogen availability cause concomitant differences in DNA and protein sequence evolution and 2) To develop a mathematical framework for estimating the strength of selection acting on gene sequences
- Chapter 4: Selection-driven cost-efficiency optimisation of transcripts governs evolutionary rate in bacteria. This chapter addresses my third and fourth thesis aims 3) To evaluate the interplay between selection acting on biosynthetic cost with selection acting on other features of gene sequences and 4) To examine how selection acting on multiple factors can determine gene evolutionary rate
- Chapter 5: General discussion.

# **Chapter 2:** Description of *Phytomonas oxycareni* n. sp. from the salivary glands of *Oxycarenus lavaterae*

**Seward E. A., Votýpka J., Kment P., Lukeš J. & Kelly S.** 2017. Description of *Phytomonas oxycareni* n. sp. from the salivary glands of *Oxycarenus lavaterae*. *Protist*, **168:**71-79.

**Original Protist formatted paper Appendix 1**

## 2.1 Chapter Introduction

There were limited genome resources available at the start of my PhD to conduct a comparative analysis between plant and animal-infecting kinetoplastida. Consequently, I conducted fieldwork to augment my data set. This chapter of my thesis describes a new species of *Phytomonas* (*Phytomonas oxycareni*) that I discovered in collaboration with the Lukeš lab in the Czech Republic. While I was engaged in that fieldwork, the genome of another *Phytomonas* species (*Phytomonas françai*) was sequenced (Butler, Jaskowska, and Kelly 2017). This meant further characterisation of *Phytomonas oxycareni* was no longer strictly necessary to address my primary research question. Therefore, after writing up the discovery of the new species I moved on to addressing the central question of the thesis as detailed in the results chapters 3 and 4.

This chapter comprises a comparative phylogenetic and morphological analysis of the new species of parasite and reveals several key findings.

1. This new species infects *Oxycarenus lavaterae*, an invasive pest species currently migrating westwards through Europe.
2. It is found primarily in the salivary glands of the insect host.
3. It can be found both extracellularly in the gland lumen and intracellularly in the cells of the gland.
4. It is the earliest diverging species in the *Phytomonas* genus and thus, through the comparative biology described in this work, provides new insight into the conserved ancestral biology of the group.

Taken together, these findings provide significant new insight into the evolution and biology of this poorly understood group of parasites.

### 2.1.1 Authors

Emily A. Seward, Jan Votýpka, Petr Kment, Julius Lukeš, Steven Kelly

### 2.1.2 Author Contributions

EAS, JV and PK collected the bug samples in the Czech Republic. JL provided lab space in the Czech Republic and advice on the manuscript. EAS dissected the bugs and performed the light microscopy, transmission electron microscopy (TEM) and scanning electron microscopy. SK advised on the TEM. JV attempted to culture the *Phytomonas* and performed the PCR amplification and sequencing of the two gene sequences. EAS conducted the phylogenetic analysis with advice from SK. EAS wrote the manuscript with advice from SK and JV. All authors read and approved the manuscript.

### 2.1.3 Abstract

*Phytomonas* sp. (phytomonads) are a diverse and globally distributed group of unicellular eukaryotes that parasitize a wide range of plants and are transmitted by insect hosts. Here we report the discovery and characterisation of a new species of *Phytomonas*, named *Phytomonas oxycareni* n. sp., which was obtained from the salivary glands of the invasive species of true bug *Oxycarenus lavaterae* (Heteroptera). The new *Phytomonas* species exhibits a long slender promastigote morphology and can be found both within the lumen of the insect host's salivary glands as well as within the cells of the salivary gland itself. Sampling multiple individuals from the same population post-winter hibernation on two consecutive years revealed that infection was persistent over time. Finally, phylogenetic analyses of small subunit ribosomal RNA genes revealed that this species is sister to other species within the genus *Phytomonas*, providing new insight into the evolutionary history of the clade.

## 2.2 Introduction

Trypanosomatids are single-celled eukaryotic parasites that collectively cause a large burden on human health and livelihood, infecting an estimated 20 million people worldwide as well as livestock and crops (da Silva et al. 2013). One large and diverse sub-group of trypanosomatids known as *Phytomonas* (Donovan, 1909), are parasites and pathogens of plants (Camargo 1999; Jaskowska et al. 2015). *Phytomonas* are globally distributed, however little is known of their biology, host range or evolutionary history and comparatively limited sampling has been conducted outside of South America, where several species cause economically important plant pathologies (Votýpka et al. 2010; Jaskowska et al. 2015).

Species in the genus *Phytomonas* are descended from a single adaptation of monoxenous insect parasites to a plant host about 400 million years ago (Lukeš et al. 2014). Following this event different species have evolved to colonise a large diversity of plant species and can be found in multiple plant tissues including phloem, latex ducts, fruit, flowers and seeds as reviewed in (Jaskowska et al. 2015). In doing so they have evolved to inhabit both extracellular and intracellular plant environments, spanning a wide range of contrasting biochemical compositions. *Phytomonas* are transmitted between plant hosts by insect vectors of the suborder Heteroptera (order Hemiptera) and there is evidence that *Phytomonas nordicus,* a parasite of the predatory stink bug *Troilus luridus* (Fabricius, 1775) (Pentatomidae), has reverted back to a monoxenous lifestyle, completing its entire life cycle in the insect host (Frolov et al. 2016). Consistent with their colonisation of plants, *Phytomonas* have also been found to inhabit both extracellular and intracellular environments within their insect hosts (Freymuller et al. 1990; Frolov et al. 2016). Thus, given this wide range of contrasting host tissues and cellular environments, it is likely that there is a large diversity of life cycles and transmission strategies that remain unexplored in this group.

Though the genus *Phytomonas* encompasses the majority of plant-infecting trypanosomatids, the diversity of these protists is not accurately represented in the literature. To improve this the creation of the subfamily Phytomonadinae has been proposed to include the genera *Herpetomonas, Phytomonas* and *Lafontella* (Yurchenko et al. 2015). Historically, species classification of insect and plant trypanosomatids depended on morphology and host specificity (Vickerman and Preston 1976). However, these criteria are not sufficient for species descriptions as opportunistic non-*Phytomonas* trypanosomatids have been found in plants and there is extensive size and shape polymorphism within a single species depending on both host and culture conditions (Catarino et al. 2001; Wheeler, Gluenz, and Gull 2011; Jaskowska et al. 2015). Thus as there is potential uncertainty in using a morphotype-orientated approach, recent classifications rely on comparatively data-rich molecular methods for taxonomic assignment (Votýpka et al. 2015).

In this chapter we characterize a new species, *Phytomonas oxycareni* n. sp.*,* obtained from the salivary gland of the heteropteran insect *Oxycarenus lavaterae* (Fabricius, 1787). This true bug is native to the Western Mediterranean, however its range in Europe has expanded both eastwards and westwards since the 1980s (Nedvěd, Chehlarov, and Kalushkov 2014). This finding highlights the potential for the migration of herbivorous insects to be associated with the migration of associated *Phytomonas* parasites. *O. lavaterae* feeds primarily on plants in the Malvaceae family, including both herbaceous representatives of the subfamily Malvoideae (e.g., *Abelmoschus*, *Abutilon*, *Gossypium*, *Hibiscus*, *Lavatera*, *Malva*) as well as lime trees (Tilioideae: *Tilia* sp.). However, in the Mediterranean it may also feed on other plants (e.g. apricots, peaches, *Citrus* sp.) as reviewed in (Kment, Vahala, and Hradil 2006). Though the plants it feeds on include several important crops (cotton, okra, apricots, peaches) and ornamentals (hibiscus, lime trees) it is rarely reported as an agricultural pest. During the winter, the species hibernates by forming tight aggregations of several hundred individuals on the sunny side of lime tree trunks. These aggregations also

occasionally form on other structures such as buildings or fences, causing a public nuisance (Nedvěd, Chehlarov, and Kalushkov 2014). Based on phylogenetic data we classified the new parasite as *Phytomonas* and propose the species name *oxycareni* to reflect its insect host.

## 2.3 Results

### 2.3.1 Material collection and primary characterisation of a new trypanosomatid species from the true bug *O. lavaterae*

*Oxycarenus lavaterae* bugs were sampled from a single large population overwintering on the trunk of a *Tilia cordata* tree in May 2015 and March 2016 in Sedlec, Czech Republic. Individual insects within this population were dissected and examined by light microscopy to search for the presence of trypanosomatid cells. In total, trypanosomatids were found in ~80% of the salivary glands that were examined. Trypanosomatid cells failed to be detected consistently in any other tissue type that was dissected from the insect. However, trypanosomatid cells were detected in the mid-gut in a small number of dissections but not in sufficient quantities to facilitate further analysis. This very low level of detectable presence in the mid-gut could be due to the life cycle of *O. lavaterae* bugs, which at the time of collection were still aggregated on the trunk of the tree and so had not fed in many months (Nedvěd, Chehlarov, and Kalushkov 2014). Unfortunately, despite repeated attempts using different media, a culture was not established. Thus all analyses reported here were conducted using trypanosomatids isolated directly from the salivary glands of the insects.

### 2.3.2 Phylogenetic analysis places the new trypanosomatid species in the genus *Phytomonas*

Identical 18S rRNA gene sequences were obtained from trypanosomatids that originated from three infected *O. lavaterae* bugs. These bugs were collected from the same population sampled in two subsequent years. This indicates that infection by the same species of trypanosomatid was pervasive within the population and that infection was likely maintained within the population over multiple generations of the host insect. The alignable region of

the 18S rRNA gene of the new trypanosomatid (GenBank Acc. Number KX257483) shared 96% identity with both the 18S sequence of *Phytomonas serpens* (GenBank acc. Number U39577, PRJNA80957) and *P. nordicus* (GenBank Acc. No. KT2236609) (Frolov et al. 2016).

To confirm the taxonomic classification of the new species, a phylogenetic tree of 18S sequences was reconstructed using the new species as well as 40 published species sampled from across the trypanosomatids. The topology of the resultant maximum likelihood tree (Fig. 2.01) showed that the newly identified trypanosomatid species was monophyletic with previously characterised *Phytomonas* and thus part of the newly proposed subfamily, Phytomonadinae (Yurchenko et al. 2015). This grouping received 100% bootstrap support and a posterior probability of 1.

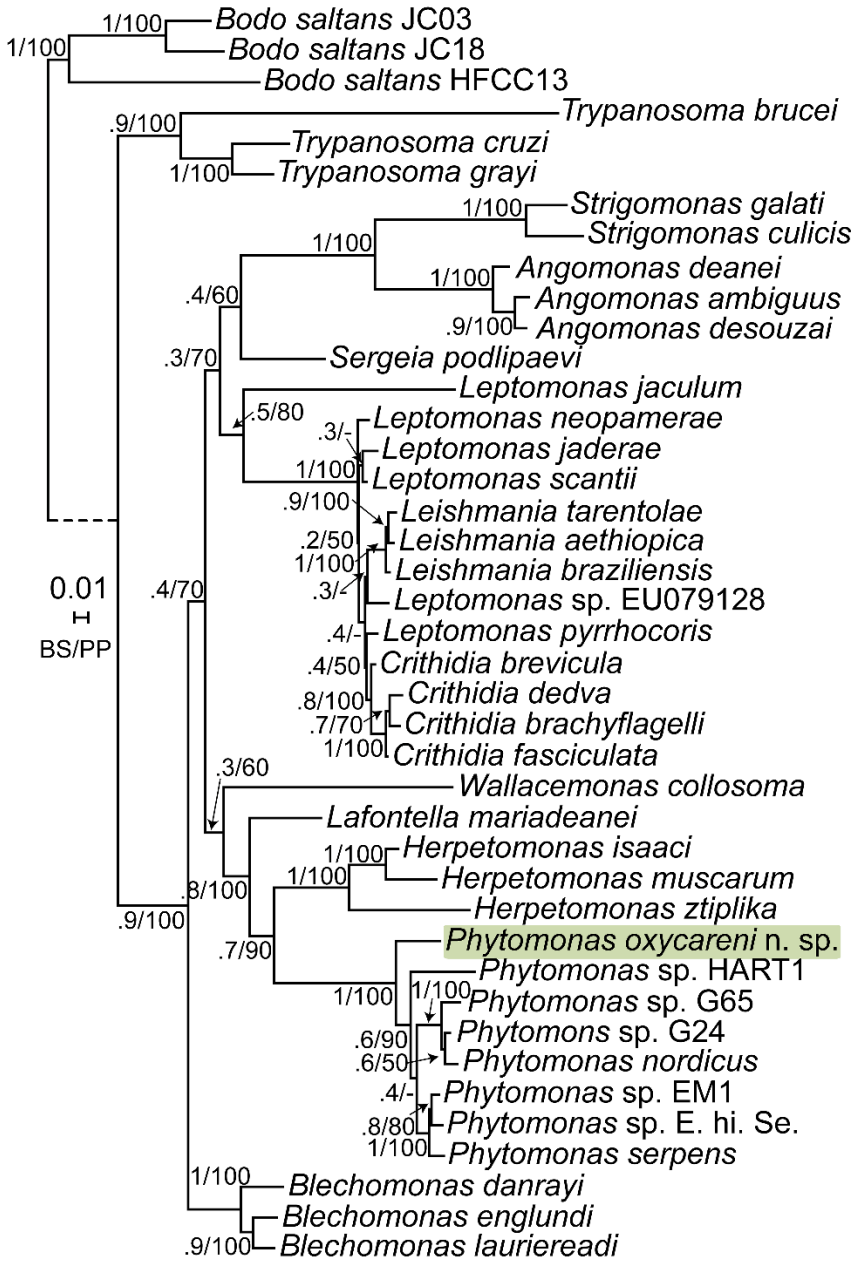



**Figure 2.01 New kinetoplastid parasite belongs to the subfamily Phytomonadinae.** Maximum likelihood phylogenetic tree of kinetoplastids obtained using RaxML on 18S ribosomal RNA gene sequences. The tree is rooted on the branch that separates *Bodo saltans* isolates and trypanosomatids. The scale bar represents the number of substitutions per site. Values on each branch represent bootstrap values/posterior probabilities for that branch. Posterior probabilities were calculated using Mr Bayes. For display purposes the dashed branch has been reduced in length by 75%. The new species described in this study (*Phytomonas oxycareni* n. sp.) is highlighted.

To independently verify the phylogenetic position of the new species, the spliced leader (SL) RNA gene from a range of representative trypanosomatid species were compared (Fig. S2.01). In support of the 18S rRNA analysis, the SL sequence showed that the newly identified trypanosomatid species had 90% bootstrap support and a posterior probability of 0.99 supporting its position as a sister species to previously characterised *Phytomonas* species. The exon of the SL sequence of the new trypanosomatid species differs from previously characterised *Phytomonas* sequences by only one nucleotide and has cytosine at position 15 of the sequence, as is the case for other *Phytomonas* species. Therefore, given the new species forms a monophyletic group with the genus *Phytomonas*, shares features such as a prolonged cell morphology and typical tissue localization (salivary glands) with other *Phytomonas* sp., and is found in a typical insect host for *Phytomonas*, we have assigned the new species to the genus *Phytomonas*. Though it is impossible to determine, this may possibly be the same flagellate that was observed previously in *Oxycarenus lavaterae* in Italy (Franchini 1922). Thus, based on the phylogenetic information, host and tissue localisation, and its cell morphology (described below), we have named the new species *Phytomonas oxycareni* n. sp.

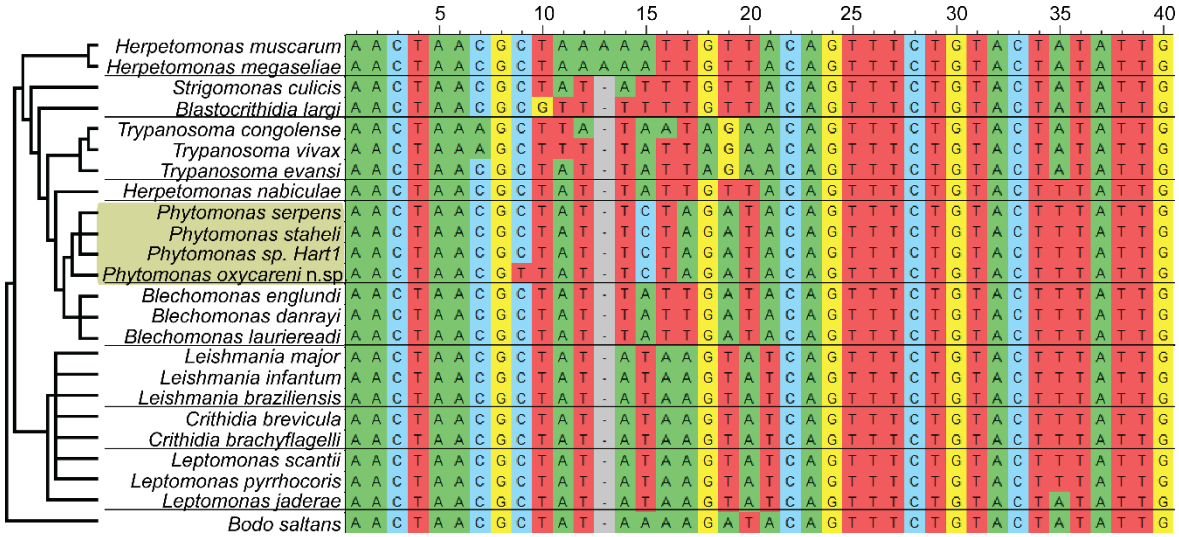



**Figure S2.01. *Phytomonas oxycareni* n. sp. spliced leader (SL) RNA gene differs from other *Phytomonas* species by one nucleotide substitution**. The spliced leader sequences from a range of trypanosomatids were downloaded for NCBI. Due to differing availability of sequences, some species from the 18S rRNA analysis are not included in this analysis. The sequences were aligned and ordered according to phylogenetic relationships. A maximum likelihood tree next to the species names shows relationships between the species based on their splice leader sequences. The new species described in this study (*Phytomonas oxycareni* n. sp.) has a sequence that differs from other *Phytomonas* species by a single nucleotide and has a cytosine at position 15 as is seen for other *Phytomonas*. In this instance, monophyly of *P. oxycareni* with other species of *Phytomonas* is supported with a boostrap value of 90%, an SH of 90 and a posterior probability of 0.98.

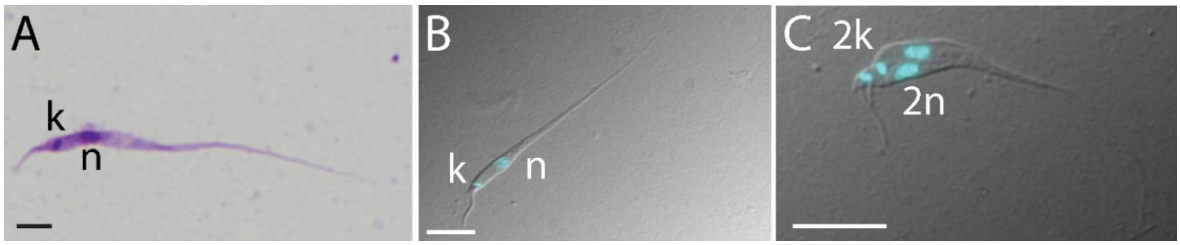



**Figure 2.02 Light microscopy images of *Phytomonas oxycareni* n. sp. A**) A Giemsa-stained cell. **B-C**) DAPI-stained cells displayed as the differential interference contrast image of the cell overlaid with the (blue) fluorescence microscopy image of the DNA stained with DAPI. **B)** Single cell with elongated slender morphology **C)** Cells in the process of dividing (2k 2n). Scale bars are 5 μm. Cells are oriented with the flagellum on the left of each image. "k" denotes the kinetoplast, "n" denotes the nucleus.

### 2.3.3 Parasites within the salivary gland exhibit promastigote morphology and are proliferative

Light microscopic examination of *P. oxycareni* cells isolated from salivary glands revealed slender promastigotes with an elongated cell body and short flagellum that is detached from the body of the cell (Fig. 2.02A-B). This morphology is similar to that of the corresponding stages observed for promastigotes of the monoxenous species *P. nordicus* (Frolov et al. 2016), and is broadly consistent with several previous morphological descriptions within this genus (Camargo 1999; Jaskowska et al. 2015; Wheeler et al. 2011). Analysis of DAPI stained slides revealed that cells within the population were undergoing cell division. That is, several cells were identified that had already completed DNA replication and segregation of the nuclei and kinetoplasts but had not yet undergone cytokinesis (Fig. 2.02C).

Higher resolution imaging of cells by scanning electron microscopy (SEM) revealed the full extent of the long slender cell body (Fig 2.03A). Although SEM cannot image nuclei and kinetoplasts, cells in late stage cytokinesis were identified as they were still joined at the posterior end (Fig. 2.03B). The observed lack of external flagella (Fig. 2.03B) is likely due to the point in cell division captured in this image rather than a diagnostic characteristic of the species (Wheeler et al. 2011).

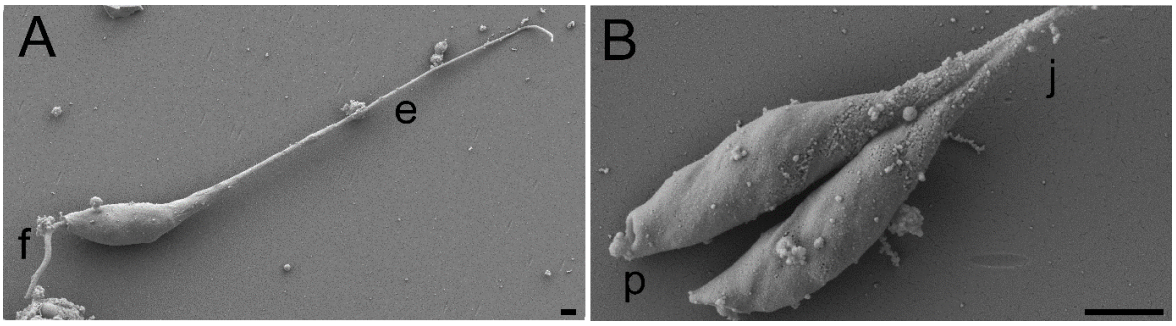



**Figure 2.03. Scanning electron microscopy of *Phytomonas oxycareni* n. sp.** Cells were obtained directly from infected salivary glands of *Oxycarenus lavaterae*, thus images contain particulate material derived from the salivary gland. **A)** Elongated slender promastigote with flagellum [f] oriented on the left of the image. [e] highlights the elongated cell body **B)** Dividing cells with no external flagella. [p] denotes the flagellar pocket. [j] shows where the cells are still joined. Scale bars are 1μm.

### *2.3.4 Phytomonas oxycareni* can be found inside host cells of the salivary gland

Transmission electron microscopy analysis of whole fixed salivary glands readily identified cells within the lumen of the gland (Fig. 2.04). The parasite cells did not appear to be attached to the surface of the lumen but were instead distributed throughout (Fig. 2.04A–C). There were also cells that appeared to be undergoing cytokinesis within these sections as indicated by the presence of two axonemal profiles within the same cell (Fig. 2.04B). Given that two different species of *Phytomonas* have previously been found to also reside within cells of the insect host (Freymuller et al. 1990; Frolov et al. 2016), the cells of the salivary gland tissue were inspected for the presence of *Phytomonas* cells. Indeed, *Phytomonas* flagellates located inside the cells of the salivary gland lumen were identified in multiple instances (Fig. 2.05). In all cases that were examined, the protists were determined to be inside two distinct membranes (Fig. 2.05A–B). This indicates that the parasite resided inside a vacuole. It was noteworthy that some intracellular *Phytomonas* cells appeared to be in the process of cytokinesis within these vacuoles (Fig. 2.05C–D). This intracellular proliferation of *Phytomonas* would be consistent with observations of dividing *P. nordicus* within the parasitophorous vacuole of a salivary gland cell (Frolov et al. 2016). However, it is unknown whether this process was initiated before or after the parasite entered the host cell, so it is unknown whether *P. oxycareni* is capable of proliferation within the insect host cells.

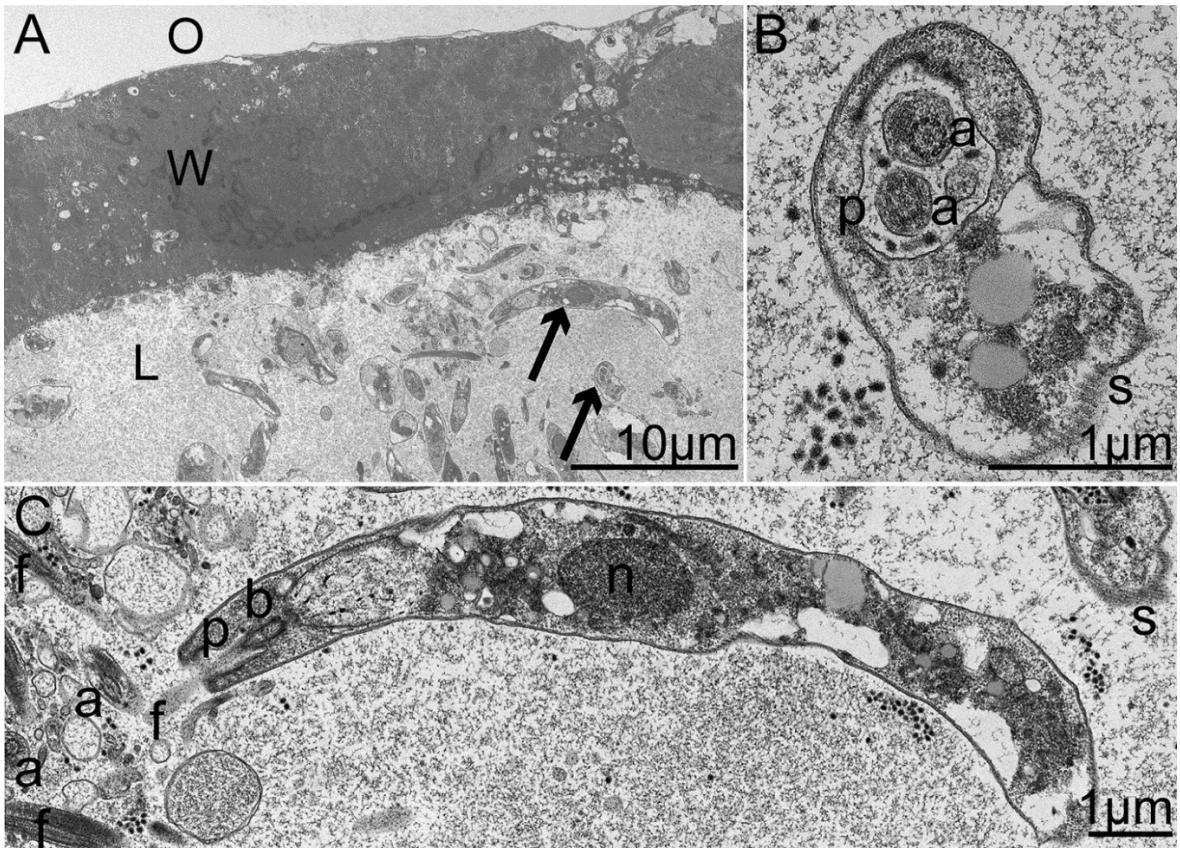



**Figure 2.04. Transmission electron microscopy of an infected salivary gland containing *Phytomonas oxycareni* n. sp. A**) A low magnification image of a section of the gland and the encompassed lumen of an infected *Oxycarenus lavaterae* salivary gland. (O) is outside the gland, (W) is the tissue of salivary gland (L) is the salivary gland lumen. Multiple parasites are visible in the salivary gland lumen and do not appear attached to the epithelium of the salivary gland. **B** & **C**) High magnification of the parasites highlighted with black arrows in (**A**) [n] nucleus. [s] subpellicular microtubules. [f] flagellum. [b] basal body. [p] flagellar pocket. [a] axoneme.

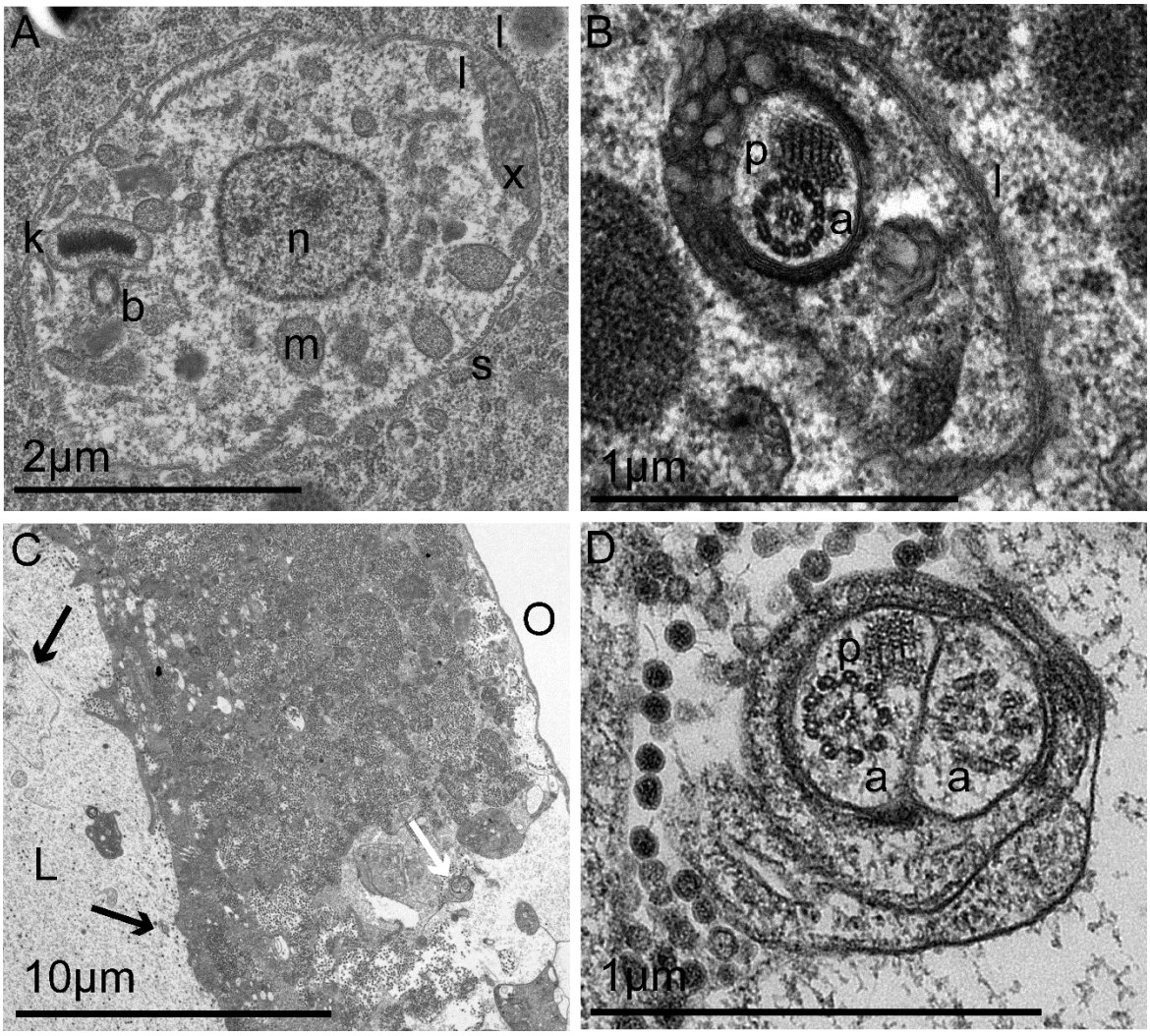



**Figure 2.05. Transmission electron micrographs of three intracellular *Phytomonas oxycareni* n. sp. within the salivary glands of the insect host *Oxycarenus lavaterae* A**) An intracellular *Phytomonas oxycareni* [k] kinetoplast [n] nucleus [s] subpellicular microtubules adjacent to two membranes. [b] basal body, [l] two lipid membranes, [x] granular material in the posterior part of the parasitophorous vacuole, [m] mitochondria. **B**) A second intracellular *P. oxycareni* [a] axoneme. [p] paraflagellar rod (PFR) with characteristic lattice [l] two membranes indicating that the parasite is inside a vacuole. **C**) A third intracellular *P. oxycareni* (white arrow) near the outside (O) of the salivary gland, multiple parasites (black arrows) are visible in the salivary gland lumen [L]. **D**) High magnification of the cell in (**C**): [a] axoneme, [p] PFR, [s] subpellicular microtubules. This is likely to be a cell in the process of cytokinesis as indicated by the presence of two axonemal profiles.

### 2.3.5 Taxonomic summary

Class: Kinetoplastea (Honigberg, 1963) Vickerman, 1976

Subclass: Metakinetoplastina Vickerman, 2004

Order: Trypanosomatida (Kent, 1880) Hollande, 1952

Family: Trypanosomatidae (Doflein, 1901) Grobben, 1905

Subfamily: Phytomonadinae Yurchenko, Kostygov, Votypka et Lukes, 2015

Genus: *Phytomonas* Donovan, 1909

***Phytomonas oxycareni* Votypka, Seward, Kment, Kelly et Lukes n. sp.**

**Diagnosis:** The species is identified by the unique sequence of 18S rRNA (GenBank accession number: KX257483) and splice leader RNA (GenBank acc. number KX611848).

**Species description:** Elongated slender promastigotes in the salivary glands were on average 30.9 ± 8.6 μm (19.7 – 52.7 μm) long and 1.76 ± 0.62 μm (0.93 – 2.99 μm) wide, with a single short external flagellum that was on average 3.7 ± 1.2 μm (2.2 – 6.4 μm). Cells had a dilated anterior portion (measuring 13-52% of the total body length) that contains the nucleus and kinetoplast. The posterior part of the cell is narrowed and pointed at the end. The cell body is twisted 0-2 times. The kinetoplast disk is compactly packed, on average 0.78 ± 0.17 μm (0.54 – 0.93 μm) in length and 0.165 μm (0.165 – 0.166 μm) in diameter and 0.68 ± 0.08 μm (0.58 – 0.83 μm) from the anterior of the cell. The nucleus was on average 2.15 ± 0.62 μm (1.34 – 2.88 μm) in length and 3.92 ± 1.06 μm (2.19 – 5.00 μm) from the anterior of the cell.

**Type host:** *Oxycarenus lavaterae* (Fabricius, 1787) (Heteroptera: Oxycarenidae). The xenotype, collected on *Tilia cordata* (Malvaceae), is deposited at the Department of Parasitology, Charles University, Prague.

**Location within host:** Found both within the midgut and lumen of the salivary gland, as well as within the cells of the salivary gland itself.

**Type locality:** Vicinity of Sedlec, Czech Republic, South Moravia (48°46'44.43"N; 16°41'55.13"E), 180 meters above sea level.

**Type material:** The name-bearing type, a hapantotype, is a Giemsa-stained slide of the dissected salivary glands, deposited in the research collection of the Department of Parasitology at Charles University in Prague. An axenic culture was not established.

**Etymology:** The specific epithet, *oxycareni*, is derived from the generic name of its host; noun in genitive case given in apposition.

### 2.4 Discussion

We report the identification of a new species of *Phytomonas*, named *Phytomonas oxycareni* that was found in the salivary glands of the heteropteran insect *Oxycarenus lavaterae*. The new species exhibits a long slender promastigote morphology and can be found both within the lumen of the salivary glands, as well as within the cells of the salivary gland itself. Furthermore, sampling different individuals from the same bug population in two consecutive years revealed that infection was likely persistent through several generations of bugs and that the protists overwinter in the insect host/vector.

Interestingly, this newly characterised species is the earliest branching *Phytomonas* species that has been identified to date. However, we were unable to confirm the presence of this species, either by cultivation or by PCR of homogenized leaves and fine branches, within the *Tilia* sp. on which the insect was found (data not shown). Therefore we could not ascertain whether it is dixenous or monoxenous, like the recently discovered *P. nordicus* (Frolov et al. 2016). While *P. nordicus* only parasitizes insect hosts and is spread between hosts via contaminated faeces and autoinfection rather than plants, *O. lavaterae* feeds primarily by drawing fluid from plant tissues such as leaves and seeds of plants in the Malvaceae family. Thus, while it is possible that *P. oxycareni* is directly transmitted between insect hosts through contact with contaminated faecal matter, it is more parsimonious to

assume that it is dixenous and is transmitted between insect hosts via a plant infective stage. In this context it is interesting to note that *P. oxycareni* does not appear to attach to the epithelia of the salivary gland like the monoxenous *P. nordicus* (Frolov et al. 2016), but can be found throughout the lumen of the gland like the promastigotes of the dixenous *P. serpens* which infects tomato plants (Jankevicius et al. 1989). However, irrespective of the confirmation of this lifestyle habit, the discovery that an early diverging species of *Phytomonas* inhabits the cells of the salivary gland suggests that this ability to survive inside host cells is ancestral and may be widespread within the genus.

Of note is the difficulty to detect parasites in the mid-gut of the insect host. As described above, this could be due to the life cycle of *O. lavaterae*. These insects form large overwintering aggregations on the trunks and branches of *Tilia* trees during which time they presumably do not feed (Nedvěd, Chehlarov, and Kalushkov 2014). The individuals from these aggregations begin to disperse in spring coincident with the flowering of lindens and they then feed on a range of plants in the Malvaceae (Nedvěd, Chehlarov, and Kalushkov 2014). Therefore it may not be the case that the putative plant host of the newly described parasite is the same species as the plant on which the insect host hibernates. Given the range of plant species that *O. lavaterae* can feed on (Kalushkov and Nedvěd 2010), and the difficulty of obtaining *Phytomonas* parasites from plant tissue (Jaskowska et al. 2015), it may be difficult to ascertain the true plant host range of this species.

In summary the work presented here identifies a new species within the *Phytomonas* clade that is a sister species to all currently sequenced phytomonad species and thus provides new insight into the ancestral biology of the genus (i.e. long slender cell morphology, host cell invasion). Furthermore identification of parasites associated with herbivorous insects that are migrating through Europe highlights the need to focus not only on the movement of insects but also on the presence of associated parasites.

## 2.5 Methods

### 2.5.1 Collection and dissection of bug hosts

Several hundred individuals of *Oxycarenus lavaterae* were collected from the trunks of several lime trees (*Tilia cordata*, and to a lesser extent *T. platyphyllos*) in the South Moravian region of the Czech Republic in May 2015 and March 2016. *O. lavaterae* remains aggregated on the trunks of trees until late May when the trees begin to flower, this facilitated ease of collection. Population infection rate varied but one bug population in particular, near Sedlec (Mikulov vicinity; 48°46′46.157″N, 16°41′54.389″E, 180 m a.s.l.) was >80% positive for parasites. This host population is the focus of this paper. The insects were euthanized in 70% ethanol, washed in 96% ethanol and a saline solution, and dissected in a saline solution to isolate the salivary glands.

### 2.5.2 Cultivation and light microscopy

Smears of infected salivary glands were fixed with methanol, hydrolysed in 5N HCl for 15 minutes at room temperature and stained with either Giemsa or 4',6-diamidino-2-pheylindole (DAPI) as has been described previously (Yurchenko et al. 2006). These stained slides were then visually inspected on a fluorescence microscope. Several attempts were made to cultivate *P. oxycareni* in different media with or without antibiotics (Amikacin) including blood agar medium, RPMI, M199, Schneider Medium, BHI Medium supplemented with FCS and Hemin, Warren's Medium, and the mix of these media (RPMI, M199, Schneider Medium and BHI Medium in a 1:1:1:1 ratio, supplemented with FCS). Although the parasites survived for several days in these media, there was no sign of growth and so no culture was obtained. The infected salivary glands were not composed of mixed infections as the same 18S rRNA and SL RNA gene amplicons were cloned from several independent insects and localities (data not shown).

### 2.5.3 Transmission and scanning electron microscopy

Dissected salivary glands were resuspended in 0.1 M phosphate-buffered saline, fixed in 2.5% (v/v) glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.4, for 1 hr at 4°C and

processed for scanning and transmission electron microscopy as described previously (Yurchenko et al. 2006). Ultrathin sections were analysed and imaged with a FEI Tecnai 12 microscope at 120kV.

### 2.5.4 PCR amplification, cloning, and sequencing

Total genomic DNA was isolated from the field samples using a DNA isolation kit for cells and tissues (Roche) according to the manufacturer's protocol. Small subunit rRNA gene (SSU rDNA, ~2100bp) was PCR amplified using the primers S762 (5'-GAC TTT TGC TTC CTC TAD TG-3')/S763 (5'-CAT ATG CTT GTT TCA AGG AC-3') (Maslov et al. 1996). The PCR thermocycler settings for DNA amplification were: denaturing at 94°C for 5 min followed by 35 cycles of 94°C for 1 min, 55°C for 90s, 72°C for 90s, and a final elongation at 72°C for 5 min. For the second round of PCR, 1μL of the previous reaction was added to 24μL of a PCR reaction and amplified using the primers TRnSSU-F2 (5'-GAR TCT GCG CAT GGC TCA TTA CAT CAG A-3') and TRnSSU-R2 (5'-CRC AGT TTG ATG AGC TGC GCC T-3'). The thermocycler settings were: denaturing at 94°C for 5 min followed by 35 cycles of 94°C for 1 min, 64°C for 90s, 72°C for 90s, and a final elongation at 72°C for 5 min. The amplified DNA sequences were subject to sequencing and the new sequence deposited in GenBank acc. no. KX257483 (18S rRNA).

The splice leader RNA gene was PCR amplified using the primers M167 (5'-GGG AAG CTT CTG ATT GGT TAC TWT A-3')/ M168 (5'-GGG AAT TCA ATA AAG TAC AGA AAC TG-3') (Westenberger et al. 2004). Amplicons were cloned into the pGEM-T Easy (Promega, Madison, USA) vector system and subject to sequencing. The new sequence was deposited in GenBank acc. no. KX611848 (Spliced leader, SL).

### 2.5.5 Phylogenetic analyses

The 18S rRNA sequences of seven different isolates of *Phytomonas* sp. and 33 insect trypanosomatid species were retrieved from GenBank (Table S2.01, online at https://github.com/easeward/Thesis) and aligned using MAFFT v7.058b (Katoh and

Standley 2013). Default parameters were used and the number of iteration steps capped at 1,000. The resulting alignment was refined manually using Cinema5 multiple sequence alignment software (Lord, Selley, and Attwood 2002) to trim the alignment file to start and end in line with the two ends of the *Phytomonas oxycareni* sequence (ie. the final dataset contained 2287 columns covering the full 1926 nucleotide sequence for *P. oxycareni*). Maximum likelihood-based phylogenetic inference was performed in RAxML version 8.2.4 using default parameters and the GTRGAMMA model with 1,000 bootstrap replicates (Stamatakis 2006). To provide additional phylogenetic support, the same sequences were analysed using Mr Bayes (Ronquist et al. 2012). The evolutionary model used was GTR with gamma-distributed rate variation across sites and a proportion of invariable sites. The covarion-like model was used and two runs, each of four chains were initiated and allowed to run for 500,000 generations sampling every 1,000 generations. Convergence was assessed through visual inspection of log-likelihood traces and through analysis of the standard deviation of split frequencies. The analysis had reached stationary phase after 5,000 generations and so this analysis was run for ample time. All other parameters used were the default.

Finally to support the position of the new species as being a sister to all other *Phytomonas* sp., an approximately unbiased (AU) test was performed. Tree 1 (Fig. S2.02) was the RAxML tree described above (Fig. 2.01). Trees 2-4 were identical to Tree 1 apart from the position of *Phytomonas oxycareni* n. sp. within the *Phytomonas* genus (Fig. S2.02). The p-value of the AU test for tree 1 was 0.92 compared to $< 0.13$ for trees 2-4, indicating that tree 1 has the greatest probability of being the true tree (Shimodaira, 2002).

Additional spliced leader sequences were downloaded from NCBI and aligned as above. Bootstrap support and posterior probabilities were calculated as above. SH values were calculated using RAxML (as above) on the maximum likelihood tree.

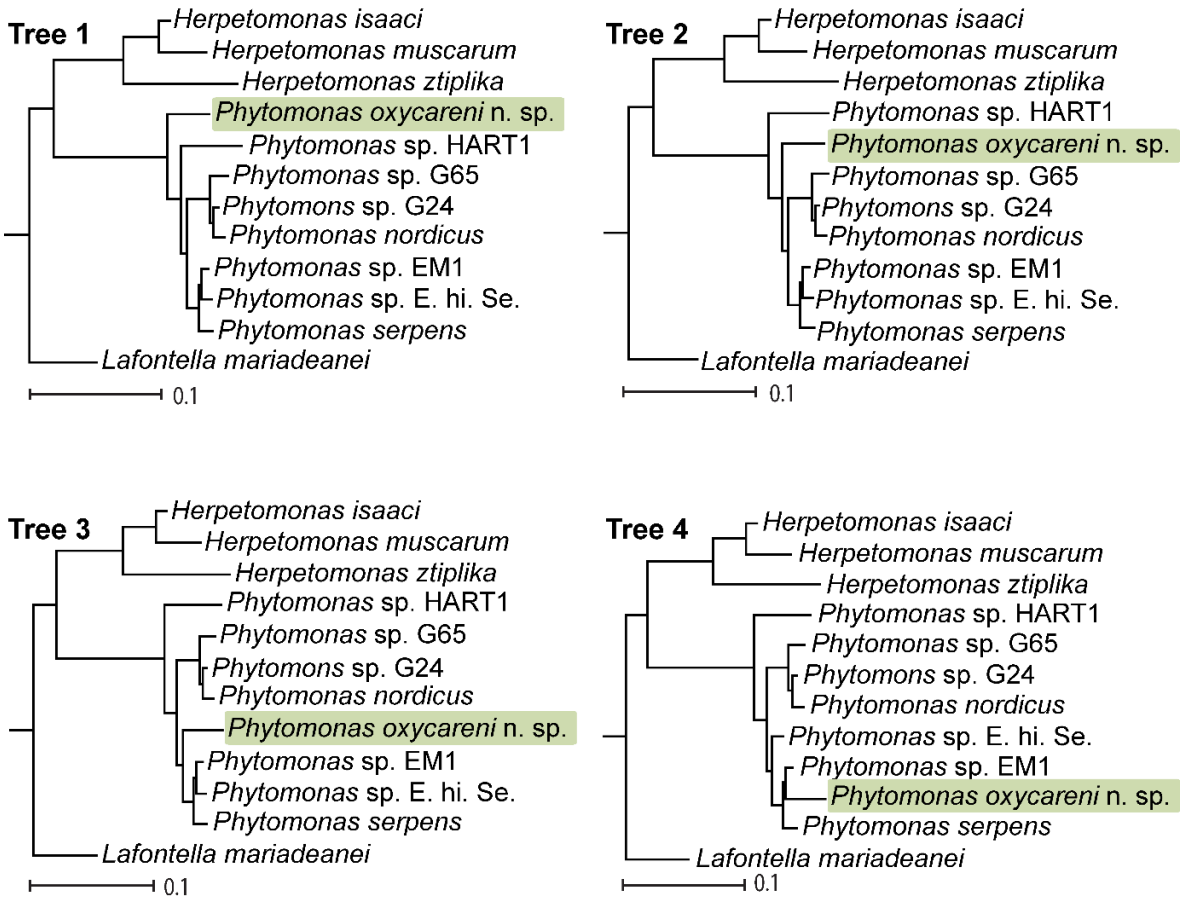



**Figure S2.02 AU test confirms that *Phytomonas oxycareni* n. sp. is a sister species to all other *Phytomonas* sp.** Phylogenetic trees used in the approximately unbiased (AU) test. Only the variable regions of the trees are displayed. The full version of tree 1 is shown in Fig. 2.01.

# **Chapter 3:** Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms

**Seward E. A. & Kelly S.** 2016. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.*, **17:**226

**Original Genome Biology formatted paper Appendix 2**

### 3.1 Chapter Introduction

In the previous chapter, I described the discovery of a new species of *Phytomonas.* In this chapter, I use genome data from *Phytomonas* and other plant- and animal-infecting parasites to compare the effect of dietary and environmental nitrogen availability on genome evolution. To do this I developed a comparative genomic approach and used a novel mathematical framework for the evaluation of how nitrogen availability affects gene evolution. The analysis presented here reveals several key findings:

1. Differences in organismal dietary nitrogen alter genome composition and are responsible for differences in synonymous codon usage between species in both bacterial and eukaryotic microorganisms.
2. The novel mathematical model developed in this chapter is able to explain relative synonymous codon use with ~92% accuracy across all species in this analysis.
3. Using the mathematical model it is possible to predict the dietary nitrogen intake of an organism from an analysis of raw nucleotide sequences.

Taken together, these findings reveal a previously hidden relationship between cellular metabolism and genome evolution. Though this study is focused on parasites, these new discoveries have clear and significant implications for the study of gene and genome evolution in all organisms.

### 3.1.1 Authors

Emily A. Seward & Steven Kelly

### 3.1.2 Author Contributions

SK conceived the study, EAS conducted the analysis, SK and EAS wrote the manuscript.

### 3.1.3 Abstract

Genomes are composed of long strings of nucleotide monomers (A, C, G, and T) that are either scavenged from the organism's environment or built from metabolic precursors. The biosynthesis of each nucleotide differs in atomic requirements with different nucleotides requiring different quantities of nitrogen atoms. However the impact of the relative availability of dietary nitrogen on genome composition and codon bias is poorly understood. Here we show that differential nitrogen availability, due to differences in environment and dietary inputs, is a major determinant of genome nucleotide composition and synonymous codon use in both bacterial and eukaryotic microorganisms. Specifically, low nitrogen availability species use nucleotides that require fewer nitrogen atoms to encode the same genes when compared to high nitrogen availability species. Furthermore we provide a novel selection-mutation framework for the evaluation of the impact of metabolism on gene sequence evolution and show that it is possible to predict the metabolic inputs of related organisms from an analysis of the raw nucleotide sequence of their genes. Taken together, these results reveal a previously hidden relationship between cellular metabolism and genome evolution and provide new insight into how genome sequence evolution can be influenced by adaptation to different diets and environment.

### 3.2 Introduction

Cells are primarily composed of a few major macromolecules (proteins, RNA, DNA, phospholipids and polysaccharides) that are constructed from monomers (amino acids, nucleotides etc.). The sequence of these monomers is important for correct molecular function, however there is often flexibility allowing for monomer usage bias. For example synonymous codons specify the same amino acid, thus different nucleotide sequences can code for the same polypeptide. Multiple competing factors have been proposed to bias the relative use of synonymous codons. These include but are not limited to: neutral drift (as a result of mutational biases during DNA replication and repair) (Eyre-Walker 1991; Francino and Ochman 1999; Rao et al. 2011), iso-accepting tRNAs (Plotkin, Robins, and Levine 2004), translational efficiency and accuracy (Sørensen, Kurland, and Pedersen 1989; Akashi 1994; Shah and Gilchrist 2011; Hu et al. 2013), altered gene splicing and protein folding (Novoa and Ribas de Pouplana 2012), mRNA purine loading as a result of temperature (Lao and Forsdyke 2000; Paz et al. 2004) and generation time (Subramanian 2008). Furthermore multiple factors such as UV radiation, nitrogen-fixation and parasitism have been proposed to explain GC variation in prokaryotes (McEwan, Gatherer, and McEwan 1998; Rocha and Feil 2010). However the impact of monomer availability (i.e. the relative availability of different nucleotides within the cell) on codon bias has been largely unexplored. We propose that differences in dietary nitrogen should cause concomitant differences in codon bias between closely related organisms whose similar lifestyles exclude alternative explanations.

Though the impact of monomer availability on synonymous codon use has yet to be elucidated, several studies have investigated the elemental composition of macromolecules (protein, DNA, RNA etc.) using genomics data and bioinformatic tools (Elser, Acquisti, and Kumar 2011). Pioneering work in this area focused on protein evolution and demonstrated that as well as the energetic costs associated with synthesising each monomer (amino acid), the monomer's elemental demands can bias usage in nutrient-limiting environments

(Baudouin-Cornu et al. 2001). Here it was shown in *E. coli* and *S. cerevisiae* that enzymes required for metabolic processing of an element have reduced quantities of that element in their sequences (Baudouin-Cornu et al. 2001). Similar studies in plants have shown that there is a 7.1% reduction in nitrogen use in amino acid side chains when plant proteins were compared to animal proteins (Acquisti, Kumar, and Elser 2009). It was proposed that this reduction was due to differences in the relative nitrogen availability of these two groups of organisms as plants are nitrogen limited in comparison to animals (Acquisti, Kumar, and Elser 2009). More generally it has also been seen that there is a negative correlation between protein abundance and the atomic requirements of its constituent monomers (Li, Lv, and Niu 2009).

Elemental limitation also has an impact on genetic sequences (DNA and RNA) which are composed of nucleotides that are either scavenged from the organism's environment or built from metabolic by-products. Like amino acids, the biosynthesis of each nucleotide differs in energetic and atomic requirements, with GC pairs consuming more ATP and requiring more nitrogen for biosynthesis than AT pairs (Rocha and Danchin 2002). The differences in energetic cost have been proposed to cause differences in the relative abundance of nucleotides within the cell, ultimately leading to nucleotide usage bias in genomic sequences (Rocha and Danchin 2002). In support of this hypothesis, it has been shown that imbalances in the relative availability of nucleotides within a cell or restrictions in nucleotide biosynthesis can lead to mutational biases that alter genome nucleotide content (Elser, Acquisti, and Kumar 2011; Buckland et al. 2014). Such differences are manifested as usage biases in organisms that have evolved in conditions where there is a persistent elemental limitation. For example domesticated crops, which have been cultivated with nitrogen fertilisation for thousands of years, and nitrogen fixing plants show increased use of nitrogen-rich nucleotides in the transcribed strand of their intergenic regions compared to wild plants, which are relatively nitrogen limited (Acquisti, Elser, and Kumar 2009).

Furthermore both protein elemental sparing and codon usage bias have been seen in 148 bacterial species with significant correlations of carbon and sulfur usage with adaptive codon usage bias (Bragg et al. 2012).

Given that changes in metabolism can lead to changes in the relative abundance of nucleotides, it follows that changes in an organism's diet (the sum of all food consumed by an organism) could have the potential to alter the nucleotide composition of the genome. Specifically, as nucleotides contain different numbers of nitrogen atoms (A/G = 5, C = 3, T/U = 2), differences in dietary nitrogen content should result in concomitant differences in the relative abundance of nucleotides within the cell and thus differences in nucleotide use between species. Moreover, these differences in nucleotide use should be detectable by comparing the nucleotide sequences for orthologous protein coding genes in organisms that share a common ancestor but have since adapted to utilise different dietary inputs. Here redundancy in the genetic code would allow differences in nucleotide use between species to manifest as changes to nucleotide sequences without necessarily altering the encoded amino acid sequence.

Microbial parasites represent an ideal model system to investigate this phenomenon and determine the effects that changes in dietary input have on the evolution and composition of genome sequences. This is because microbial parasites typically have streamlined metabolisms and often obtain energy from catabolism of a limited set of host biomolecules. Furthermore closely related parasites often utilise different metabolic strategies and obtain energy from catabolism of different host derived compounds, and even related parasites that colonise the same host niche can obtain energy from catabolism of different inputs (Pereyre et al. 2009). Thus comparative genomics between parasites that share a common ancestor but have adapted to utilise different host-derived biomolecules has the potential to reveal the effects of changes in diet on the evolution and composition of genome sequences.

Here we provide a global analysis of gene sequence evolution associated with adaptation to changes in diet. We show in two monophyletic groups of parasites (one eukaryotic and one bacterial) that adaptation to diets with differing nitrogen content produces a concomitant effect on nucleotide compositions (and hence nitrogen content) of orthologous RNA sequences. Those parasites that have adapted to low nitrogen content diets have low nitrogen content sequences while those parasites that have adapted to high nitrogen content diets have high nitrogen content sequences. We construct a novel model for synonymous codon use that is sufficient to explain the genome-wide usage of synonymous codons with >90% accuracy. We show that using this model in a predictive capacity it is able to identify the metabolic capacity of related parasites from raw nucleotide sequences. Taken together, our findings provide significant new insight into the relationship between diet, metabolism and genome evolution, and provide a novel mechanistic explanation for genome-wide patterns of synonymous codon use.

## 3.3 Results

### 3.3.1 Choice of model organisms and inference of orthogroups

To test the hypothesis that differences in diet between organisms can impact on the nucleotide composition of their genomes, a comparative genomic analysis was performed using bacterial (Mollicutes) and eukaryotic (kinetoplastida) parasites that have adapted to different host niches (Fig. 3.01, Fig. S3.01 and S3.02). These parasites were chosen for analysis because none of the species fix nitrogen and so require nitrogenous compounds obtained from their environment (Razin and Knight 1960; Creek et al. 2013). Furthermore, unlike opportunistic parasites or free-living organisms, these parasites are restricted to host niches that differ in the relative abundance of biologically available nitrogen. Specifically, parasites that colonise plant hosts are nitrogen limited in comparison to those that colonise animal hosts (Acquisti, Elser, and Kumar 2009). Additionally, the parasites' pathways for ATP generation differ in liberation of biologically available nitrogen (Fig. S3.02) (Arraes et al. 2007; Ginger, Fairlamb, and Opperdoes 2007; Pereyre et al. 2009; Kube et al. 2012; Jaskowska et al. 2015). The parasites that obtain energy through glycolysis obtain carbon skeletons and regenerate ATP, whereas the parasites that obtain energy through catabolism of arginine or amino sugars additionally obtain biologically available nitrogen (Fig. S3.02). Thus the parasites were categorised into three groups depending on host type and whether their metabolism liberates nitrogen. Low nitrogen availability ($L_N$) parasites colonise plants and obtain energy through carbohydrate catabolism, medium nitrogen availability ($M_N$) parasites colonise animals and primarily catabolise carbohydrates, and high nitrogen availability ($H_N$) parasites colonise animals and obtain energy through amino acid or amino sugar catabolism.

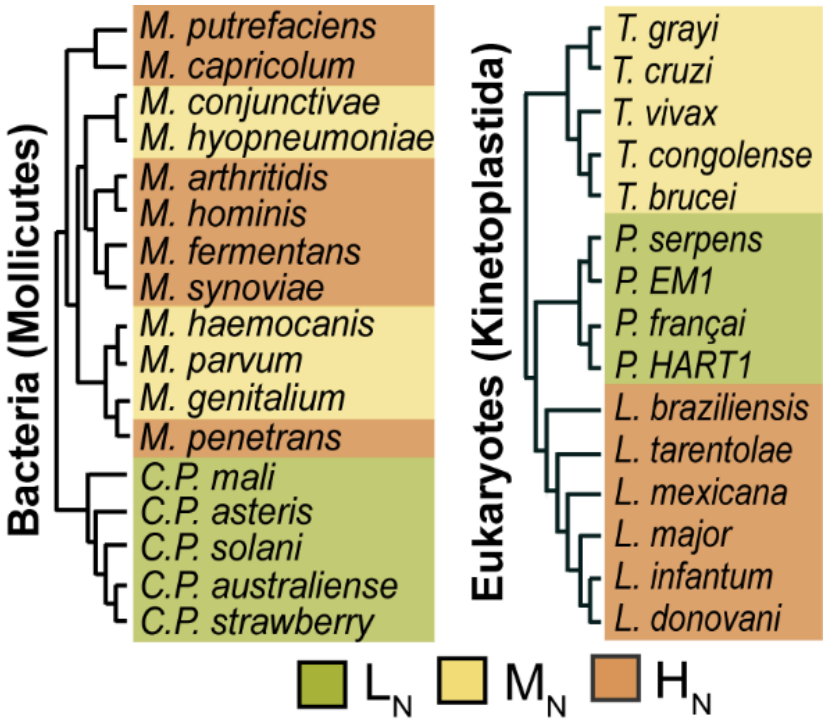



**Figure 3.01. Phylogenetic trees of the parasites used in this study shaded according to their host and metabolic strategy.** Green: plant parasites that obtain energy from carbohydrate catabolism (low nitrogen availability, $L_N$). Yellow: animal parasites that obtain energy from carbohydrate catabolism (medium nitrogen availability, $M_N$). Orange: animal parasites that obtain energy from amino acid and/or amino sugar catabolism (high nitrogen availability, $H_N$).

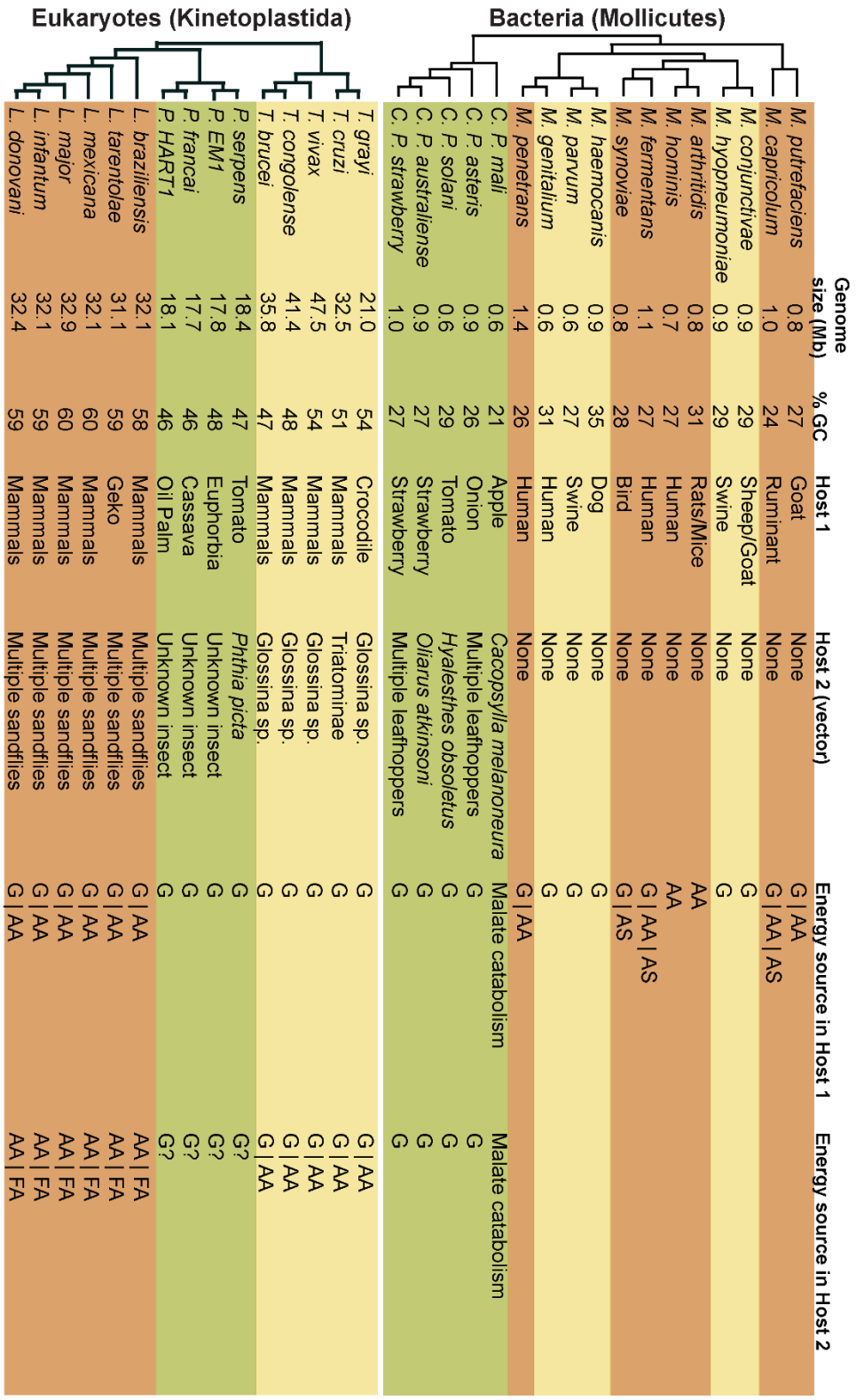



**Figure S3.01. Phylogenetic trees and metabolic information for the parasites used in this study.** Hosts and vectors are indicated where known. The major pathways used to generate ATP in each host are provided where G = glycolysis, AA = amino acid metabolism, AS = amino-sugar metabolism, FA = fatty acid metabolism. Genome size and GC-content are also provided.

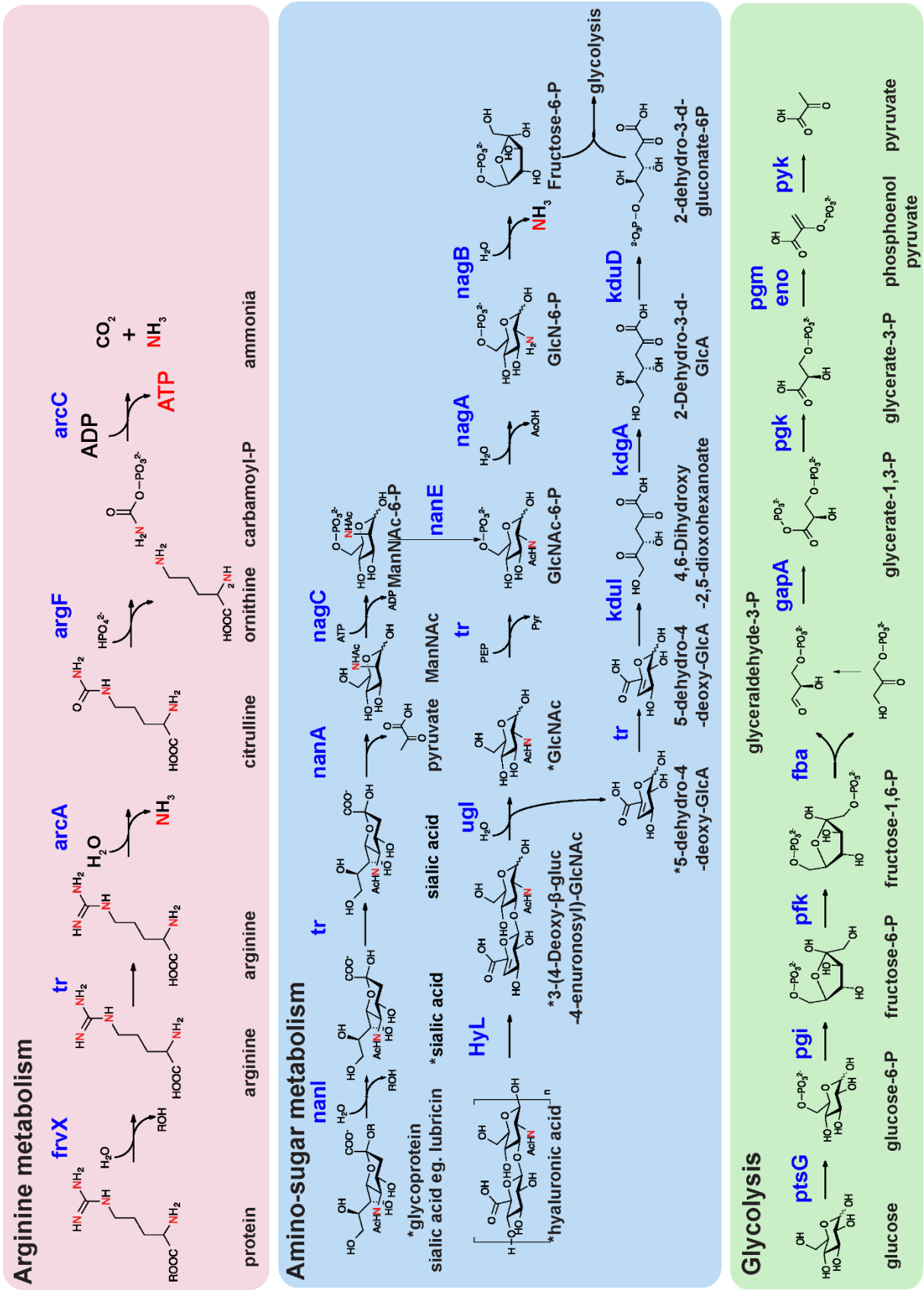



**Figure S3.02. ATP generating metabolic pathways.** Nitrogen atoms have been highlighted in red and required genes are indicated in blue. * denotes substrates that are extracellular. tr. is the abbreviation for transporter.

A set of orthologous gene groups (orthogroups) covering 15 kinetoplastid genomes (Fig. 3.01 and Table S3.01 [online at https://github.com/easeward/Thesis]) and an independent set of orthogroups covering 17 Mollicute genomes (Fig. 3.01 and Table S3.01) were inferred. Both sets of orthogroups were subject to filtering such that orthogroups were retained for further analysis only if the orthogroup comprised a single copy gene present in at least three species from each nitrogen availability group (i.e. 3 $L_N$, 3 $M_N$ and 3 $H_N$ species). In this analysis, use of orthologous protein coding genes allows direct investigation of the effect of adaptation to different metabolic strategies on nucleotide sequences that are derived from a common ancestral state. These genes may also be considered housekeeping genes as the organisms have only one tissue (unicellular) and these genes are conserved across all three groups. The same analysis cannot be done in intergenic regions where ambiguity of orthology prevents paired comparison of sites. Moreover, in the case of bacteria there are too few intergenic regions for robust statistical analyses. Of the 9,526 orthogroups identified in kinetoplastids, 3,003 satisfied the filtration criteria, encompassing ~40% of all single copy genes in these organisms. Similarly, of the 1,280 orthogroups identified in the Mollicutes, 168 satisfy the filtration criteria, encompassing 28% of all single copy genes in these organisms.

### 3.3.2 Low nitrogen availability parasites have low nitrogen content sequences and vice-versa

In the kinetoplastid parasites 878,193 orthologous codons in the 3,003 conserved single-copy orthologous genes were compared. This revealed a significant difference in the nitrogen content of mRNA between the different nitrogen availability groups (Fig. 3.02A). On average the mRNAs in $L_N$ parasites cost one fewer nitrogen atom for every fifteen codons compared to the same mRNAs in $M_N$ ($p < 0.001$) and one for every seven codons compared to $H_N$ ($p < 0.001$). This corresponds to nitrogen savings of ~0.6% and ~1.3% respectively. Given a kinetoplastid cell has ~61,000 transcripts (Kolev et al. 2010) with an average length of 630 codons, $L_N$ kinetoplastid parasites would use ~5.5 x $10^6$ fewer nitrogen atoms than

$H_N$ parasites to produce the same transcriptome. This corresponds to enough nitrogen atoms to make ~8,700 average sized proteins. The kinetoplastids also exhibit an analogous difference in the nitrogen content of dsDNA (Fig. 3.02B). Here genes in $L_N$ parasites cost one less nitrogen atom for every 4 codons compared to $H_N$ therefore saving roughly 157 nitrogen atoms per gene (~1.1%). Considering that kinetoplastids are diploid with an average of 8,000 genes, this difference in nitrogen cost means that $L_N$ parasites use ~2.5 x $10^6$ fewer nitrogen atoms to encode the exact same cohort of genes.

A similar phenomenon was observed when comparing the mRNA sequences in Mollicute parasites (Fig. 3.02C). Here comparison of 38,255 orthologous codons in 168 orthologous genes revealed that $L_N$ parasites used one fewer nitrogen atom for every 9 codons compared to the same mRNAs in $M_N$ ($p < 0.001$) and one for every 5 codons compared to $H_N$ ($p < 0.001$). This corresponds to nitrogen savings of ~1% and ~1.8% respectively. Though the Mollicutes exhibit a nitrogen dependent effect in their mRNAs, the same strong effect is not seen in their dsDNA ($p = 0.025$ when comparing $L_N$ and $H_N$, $p < 0.001$ comparing $M_N$ with either $L_N$ or $H_N$) (Fig. 3.02D). We propose that the absence of a clear nitrogen dependent effect at the DNA level is due to a strong GC to AT mutation bias thought to be caused by a lack of dUTPase coupled with a reduced ability to correct erroneous dUTP incorporation in DNA (Williams and Pollack 1990; Pollack, Williams, and McElhaney 1997). Thus though the mRNA for the same genes has a lower nitrogen cost in nitrogen-limited species, the high AT nucleotide composition of the DNA reflects the mutational bias imposed by the lack of dUTPase.

For both the kinetoplastids and Mollicutes an analogous difference is also seen in the nitrogen content of the amino acid side chains of these orthologous sites. The $L_N$ parasites use amino acids whose side chains require less nitrogen than the $M_N$ and $H_N$ parasites (Fig. S3.03). The slight discrepancy between the $M_N$ and $H_N$ parasites can be explained by the reduced use of arginine in the $H_N$ species as they primarily obtain energy from arginine

catabolism (Wanasen and Soong 2008; Pereyre et al. 2009; Fiebig, Kelly, and Gluenz 2015). This is consistent with previous studies of plant and animal proteins that observed reduced nitrogen content of amino acid side chains in the nitrogen-limited plant species (Elser et al. 2006; Acquisti, Kumar, and Elser 2009).

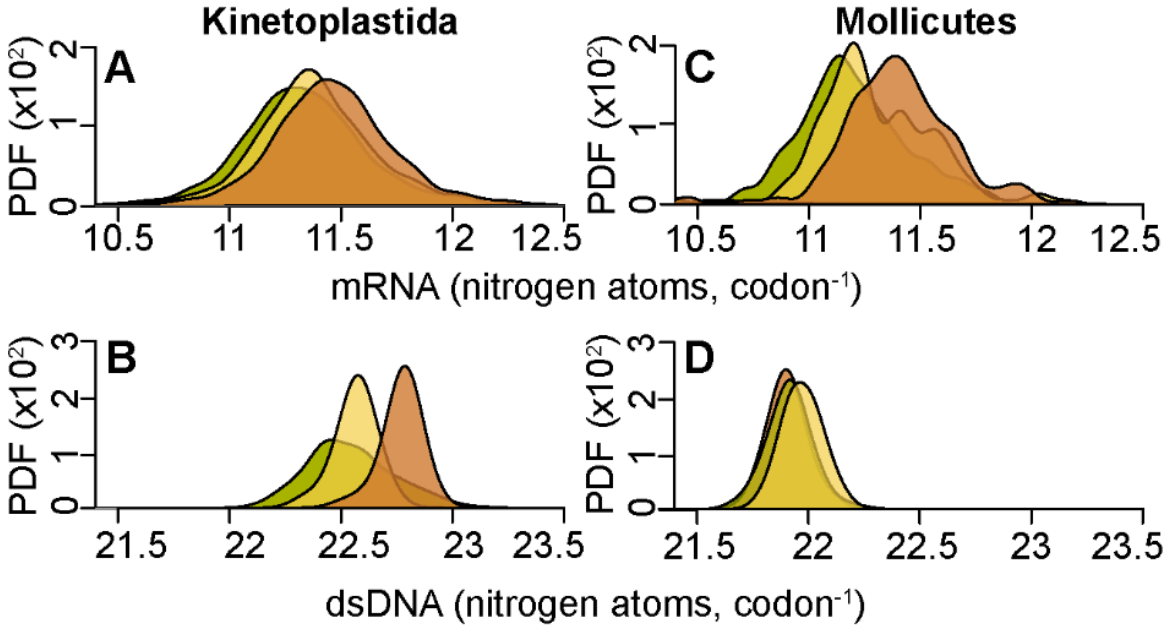



**Figure 3.02. Nitrogen availability influences gene sequences.** (*A*) The average mRNA nitrogen content per codon for 3,003 orthologous genes in the kinetoplastida. (*B*) The average nitrogen content per double stranded codon (dsDNA) for the same genes. (*C*) as in (*A*) but for 168 orthologous Mollicute genes. (*D*) As in (*B*) but for the Mollicutes. Y-axis is the probability density function (PDF) for the distributions.

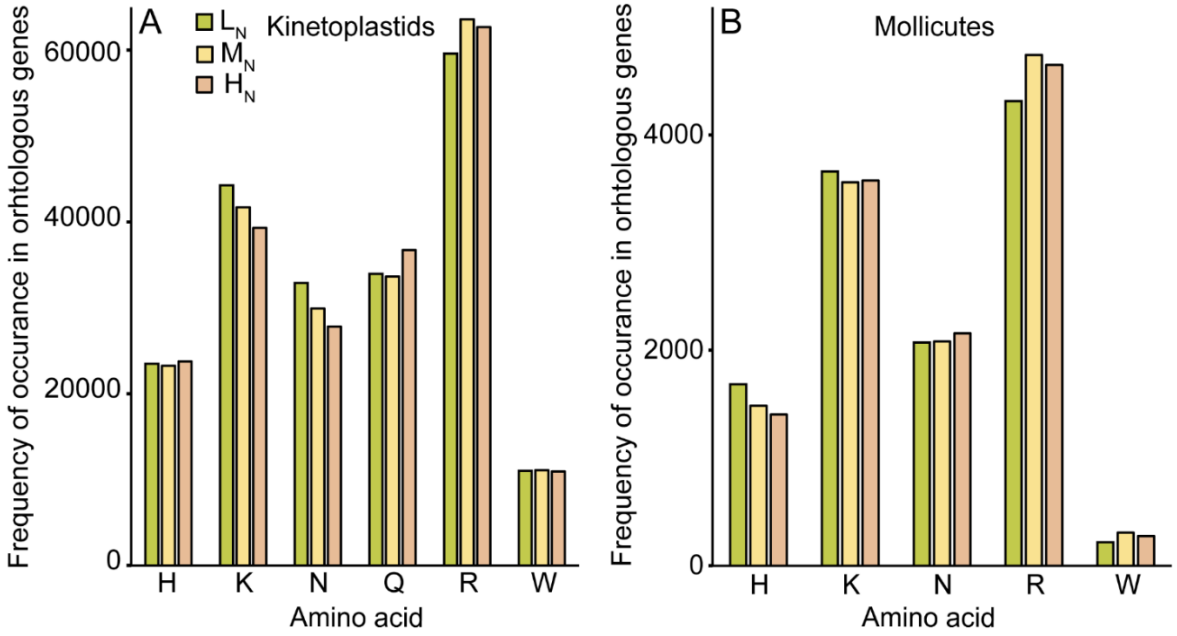



**Figure S3.03. $L_N$ parasites use the least amount of nitrogen in their amino acid side chains compared to $M_N$ and $H_N$ parasites.** (*A*) Frequency of occurrence of amino acids with nitrogen in their side chains at orthologous sites in orthologous genes in the kinetoplastid parasites. This corresponds to a total number of nitrogen atoms in amino acids side chains of $L_N$ = 347,789, $M_N$ = 353,574 and $H_N$ = 350,376. (*B*) Frequency of occurrence of amino acids with nitrogen in their side chains at orthologous sites in orthologous genes in the Mollicute parasites. This corresponds to a total number of nitrogen atoms in amino acids side chains of $L_N$ = 11,948, $M_N$ = 12,178 and $H_N$ = 12,063. Glutamine (Q) is not considered for the Mollicutes as the $M_N$ and $H_N$ groups lack glutaminyl-tRNA synthetase (GlnS). Instead a non-discriminating glutamyl-tRNA synthetase (GltX) charges both $tRNA^{Glu}$ and $tRNA^{Gln}$ with Glu. This means use of Q between *Mycoplasma* ($M_N$ and $H_N$ species) and *Phytoplasma* ($L_N$ species) is not comparable. For both the kinetoplastids and the Mollicutes, $M_N$ parasites use more nitrogen in their side chains than $H_N$ parasites. This can be explained by the reduced occurrence of arginine (R) in the $H_N$ parasites, which is expected as these parasites metabolise arginine to generate energy. In both graphs the orthologous sites are the same as for the analysis in the section "Low nitrogen availability parasites have low nitrogen content sequences and vice-versa".

### 3.3.3 Different metabolic strategies in the same host niche cause concomitant differences in gene sequence nitrogen content

To provide further insight into the relationship between metabolism and genome nucleotide composition an additional analysis was conducted on Mollicute parasites that occupy the same host niche but obtain energy through different metabolic strategies. Here, three Mollicute species, *Mycoplasma hominis, Mycoplasma genitalium* and *Ureaplasma parvum* were analysed (note *Ureaplasma parvum* is also a Mollicute but a different species to *Mycoplasma parvum* used in analyses above). Each of these three species reside in the same urogenital tract niche but obtain energy from catabolism of different biomolecules (Pereyre et al. 2009). *M. genitalium* and *U. parvum* metabolise glucose and urea respectively. However *M. hominis* has lost the ability to generate ATP via glycolysis and instead generates ATP via nitrogen-liberating arginine catabolism (Pereyre et al. 2009).

Using the same methods outlined previously, 51,998 orthologous codons in 227 conserved single-copy orthologous genes (present in each of the three species) were compared (Fig. 3.03*A*). This revealed that despite inhabiting the same niche environment, there was a significant difference ($p < 0.001$) in the nitrogen cost of genes, equating to using one fewer nitrogen atom for every 6 codons in *M. hominis* ($H_N$) compared to *M. genitalium* ($M_N$) (~1.5%). Since urea metabolism generates ammonia, one could expect *U. parvum* to be a $H_N$ parasite. However, *U. parvum* exports ammonia to drive ATP synthesis meaning energy generation is linked with export of nitrogen from the cell. Thus, analogous to *M. genitalium*, *U. parvum* is a nitrogen-limited species and uses one fewer nitrogen atom for every 5 codons as compared to *M. hominis* (~1.8%). As before, the strong mutation bias in Mollicutes means that the same nitrogen-dependent effect is not seen in their dsDNA (Fig. 3.03*B*). Taken together, this comparison reveals that in a common host niche, different metabolic strategies can result in concomitant differences in mRNA nitrogen content.

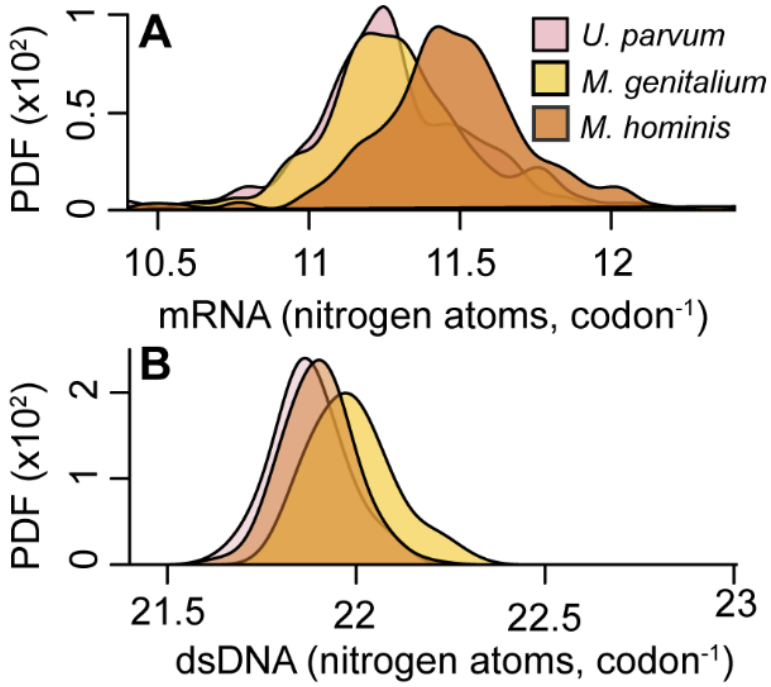



**Figure 3.03. Analysis of gene nitrogen content of three parasites that occupy the same host niche but utilise different metabolic strategies.** (*A*) The average mRNA nitrogen content per codon for 227 single copy genes present in each species. (*B*) The average nitrogen content per double stranded codon (dsDNA) for the same genes. Y-axis is the probability density function (PDF) for the distributions. Yellow and orange colour as in Fig. 3.01. Pink: animal parasite that obtains energy by exporting ammonia from the cell.

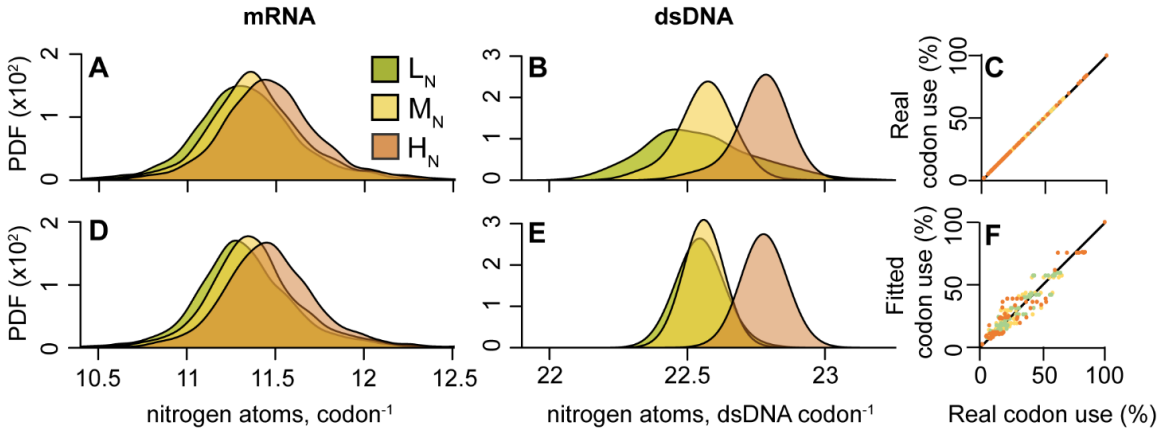



**Figure 3.04. A selection-mutation model for synonymous codon use in kinetoplastids explains relative synonymous codon use with ~90% accuracy and recapitulates the difference in nitrogen cost of genes.** (*A*) The average mRNA nitrogen content per codon for 3,003 orthologous genes in the kinetoplastida. (*B*) The average nitrogen content per double stranded codon (dsDNA) for the genes in *A*. (*C*) Empirical codon use probabilities (expressed as %) plotted against themselves. (*D*) The average mRNA nitrogen content per codon for sequences simulated using synonymous codon use probabilities derived from fitting the atomic composition model to the observed sequence data. (*E*) The average nitrogen content per double stranded codon for the genes in *D*. (*F*) The synonymous codon use probabilities inferred using the atomic composition model plotted against the empirical codon use probabilities expressed as %. Dot colour corresponds to nitrogen availability group. $L_N$ $R^2 = 0.92$, $M_N$ $R^2 = 0.88$, $H_N$ $R^2 = 0.90$.

### 3.3.4 Differences in genome-wide patterns of synonymous codon use are explained by selection acting on codon nitrogen content.

Given that there is a clear difference in the nitrogen content of genes between different nitrogen availability groups, it was assessed whether this phenomenon could be explained by differences in the nitrogen content of synonymous codons. To do this a novel model for genome-wide synonymous codon use was constructed that considers mutation bias and selection acting on the nitrogen content of codons (see methods section "A model for synonymous codon use under the joint pressures of selection and mutation bias"). Using this model the value of the nitrogen-dependent selection bias ($2N_g s$) and mutation bias ($m$) were found that best explained the real sequence data (see methods for complete model description). Here a negative value for $2N_g s$ indicates that selection is acting to decrease nitrogen content and *vice versa*.

For the kinetoplastids, application of this modelling approach was able to explain genome-wide patterns of synonymous codon use with >90% accuracy across all nitrogen availability groups (Fig. 3.04). Moreover, sequences simulated using these fitted codon use frequencies recapitulated the observed patterns of nitrogen content in mRNA (Fig. 3.04D) and dsDNA (Fig. 3.04E, Fig. S3.04 and S3.05). Consistent with nitrogen availability, the value of the selection bias for incorporation of nitrogen atoms in gene sequences was most negative in $L_N$ parasites ($2N_g s$ = -0.09), intermediate in $M_N$ parasites ($2N_g s$ = -0.06) and least negative in $H_N$ parasites ($2N_g s$ = -0.03). The distribution of $2N_g s$ parameters for individual species within each group were also significantly different between each group (ANOVA, $p < 0.01$). Thus differences in nitrogen availability between species are reflected in the relative strengths of the selection bias on codon nitrogen content. Furthermore mutation bias towards GC was lowest in $L_N$ parasites ($m$ = 0.67) and highest in $H_N$ parasites ($m$ = 0.31). Importantly, just considering selection acting on the nitrogen content of mRNA (Fig. S3.05B) or mutation bias (Fig. S3.05D) in isolation resulted in higher AIC values

(Table S3.01), indicating the dual parameter model is better. Thus the pattern of codon use and gene nitrogen content is best explained by a model that considers both selection acting on the mRNA nitrogen content of genes and mutation bias (Fig. 3.04, Fig. S3.05E). Furthermore the statistical significance of selection acting on the nitrogen content of coding sequences was assessed by a permutation test (see Methods). This showed that selection acting on the nitrogen content of the mRNA sequences was significant for $L_N$ ($p = 0.004$) and $M_N$ ($p = 0.021$) parasites but was not significant for $H_N$ kinetoplastid parasites ($p = 0.457$). This is consistent with our findings that indicate $H_N$ kinetoplastids are not under selection to minimise the nitrogen content of their coding sequences. The change in codon bias also accounts for the majority of the difference in genome-wide GC-content between species. Specifically the coding regions constitute ~ 50% of the genome in kinetoplastid parasites, and thus changes in synonymous codon use account for 61% of the observed difference in genome-wide GC-content between $H_N$ and $L_N$ species (Table S3.01).

It should be noted that simulating sequences using perfect genome-derived codon use frequencies (i.e. using 61 constrained parameters Fig. S3.05H) results in simulated sequences whose distributions are not significantly different to those obtained in our two parameter selection-mutation model. Thus the difference between the distributions of nitrogen content for the real (Fig. 3.04B) and simulated sequences (Fig. 3.04E) is a result of factors affecting codon bias in individual genes that are not encapsulated by our genome-wide model.

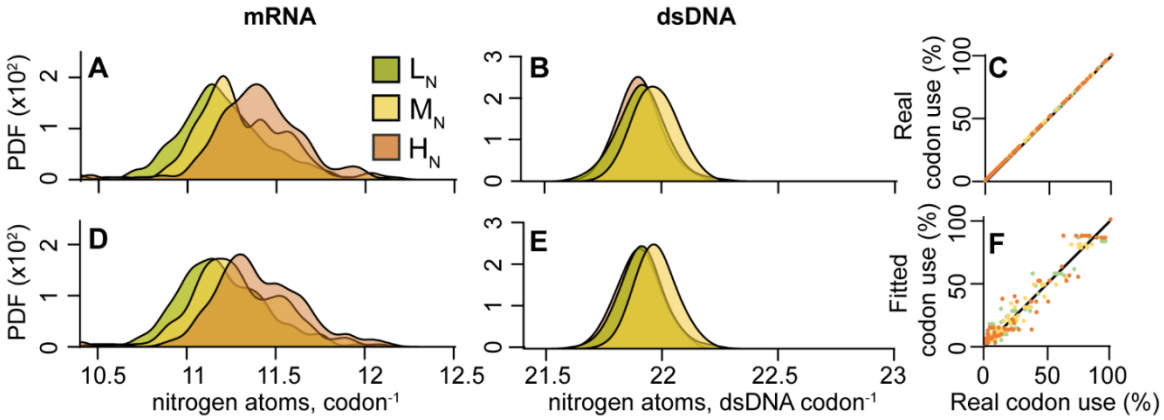
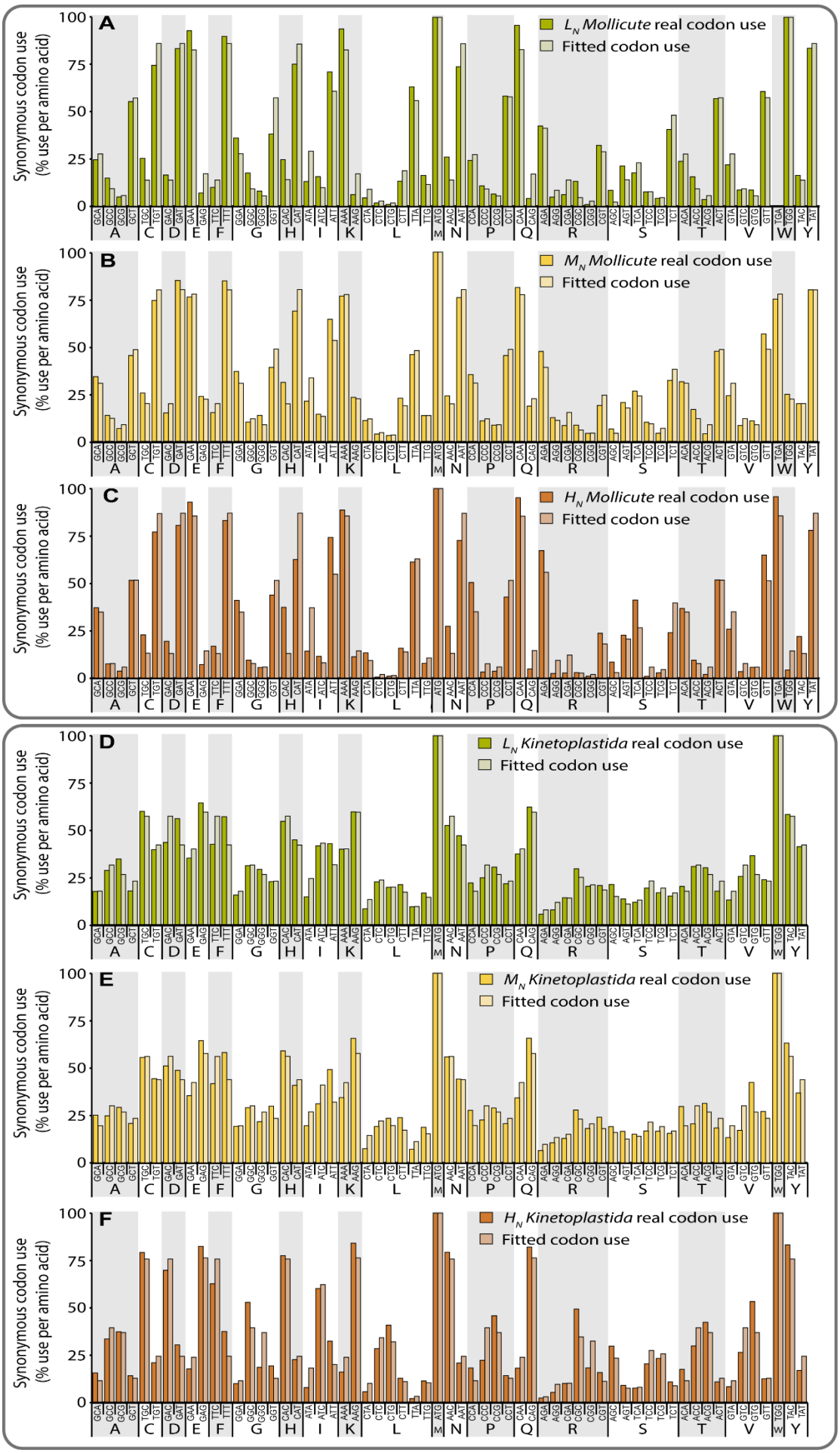



**Figure 3.05. Model for synonymous codon use for Mollicutes explains relative synonymous codon use with ~94% accuracy and recapitulates the difference in nitrogen cost of genes.** (*A*) The average mRNA nitrogen content per codon for 168 orthologous genes in the Mollicutes. (*B*) The average nitrogen content per double stranded codon (dsDNA) for the genes in *A*. (*C*) Empirical codon use probabilities (expressed as %) plotted against themselves. (*D*) The average mRNA nitrogen content per codon for sequences simulated using synonymous codon use probabilities derived from fitting the atomic composition model to the observed sequence data. (*E*) The average nitrogen content per double stranded codon (dsDNA) for the genes in *D*. (*F*) The synonymous codon use probabilities inferred using the atomic composition model plotted against the empirical codon use probabilities expressed as %. Dot colour corresponds to codon use probabilities for each nitrogen availability group $L_N$ $R^2 = 0.95$, $M_N$ $R^2 = 0.97$, $H_N$ $R^2 = 0.91$.

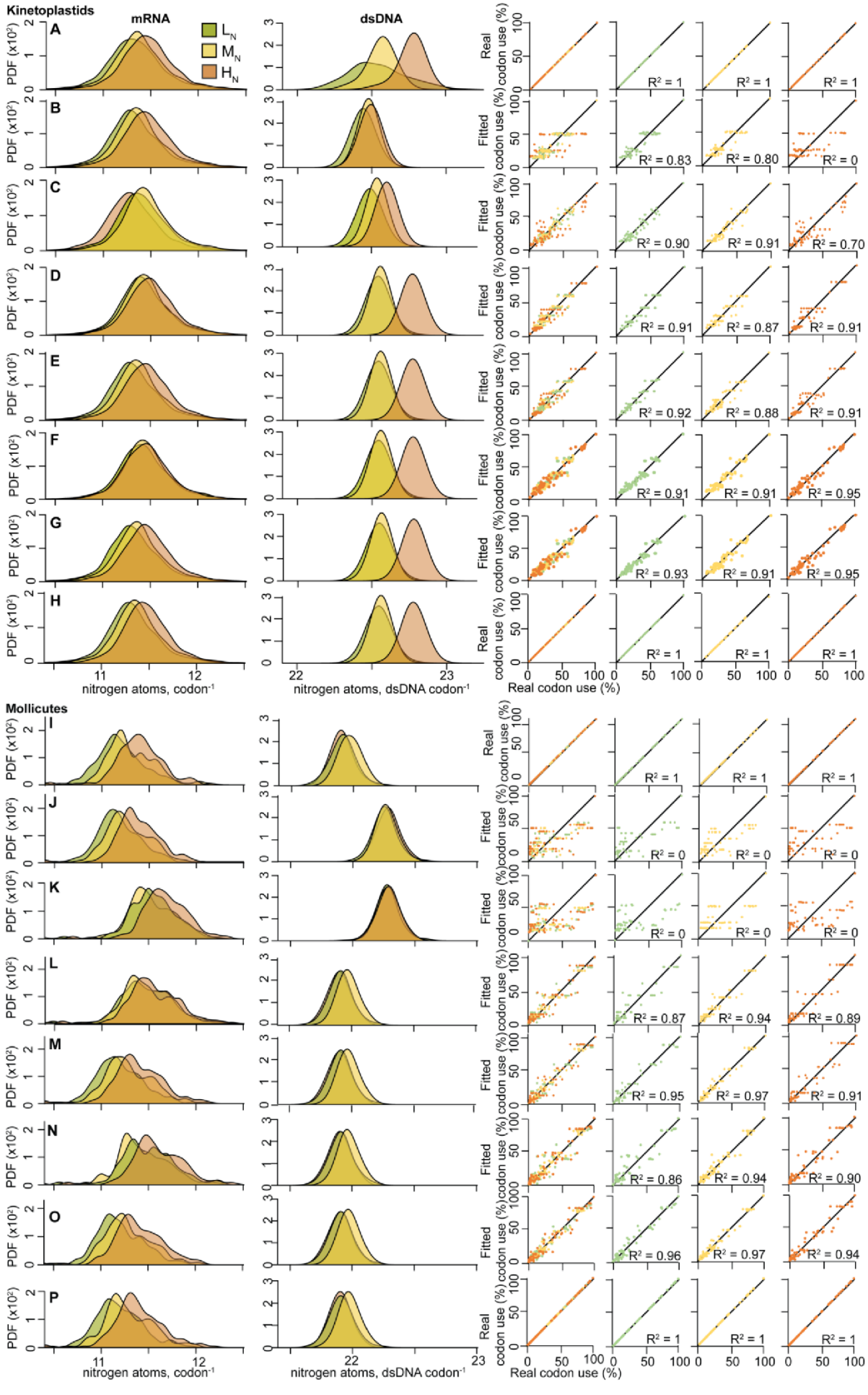

A
$L_N$ Mollicute real codon use
Fitted codon use
B
$M_N$ Mollicute real codon use
Fitted codon use
C
$H_N$ Mollicute real codon use
Fitted codon use
D
$L_N$ Kinetoplastida real codon use
Fitted codon use
E
$M_N$ Kinetoplastida real codon use
Fitted codon use
F
$H_N$ Kinetoplastida real codon use
Fitted codon use
Synonymous codon use (% use per amino acid)
0
25
50
75
100
A C D E F G H I K L M N P Q R S T V W Y

(Previous page)

**Figure S03.4. The model for synonymous codon use under the joint pressures of selection acting on nitrogen content and mutation bias fits real codon use with > 90% accuracy.** Comparison of observed (dark) and fitted (light) synonymous codon use for (A) Mollicute $L_N$ , $2N_g s$ = -0.24, $m$ = 4.8 (B) Mollicute $M_N$, $2N_g s$ = -0.15, $m$ = 3.5 (C) Mollicute $H_N$, $2N_g s$ = -0.13, $m$ = 5.9 (D) Kinetoplastid $L_N$, $2N_g s$ = -0.09, $m$ = 0.7 (E) Kinetoplastid $M_N$, $2N_g s$ = -0.06, $m$ = 0.7 and (F) Kinetoplastid $H_N$ $2N_g s$ = -0.03, $m$ = 0.3 parasites.

(Next page)

**Figure S03.5. The model that considers both mutation bias and nitrogen content selection in combination provides a better fit than either parameter considered in isolation.** (*A*) The average mRNA nitrogen content per codon for 3,003 orthologous genes in the kinetoplastida. The average nitrogen content per double stranded codon for the same set of genes. Empirical codon use probabilities (expressed as %) plotted against themselves for all groups, $L_N$, $M_N$ and $H_N$ respectively. Analogous plots are shown in panels B-H for sequences simulated using fitted values for (*B*) selection acting on codon nitrogen content (Equation 2). (C) selection acting on translational efficiency (tAI) (Equation 10). (D) mutation bias acting on the GC-content of the sequences (Equation 5). (E) both selection acting on codon nitrogen content and mutation bias (Equation 7) (F) both selection acting on translational efficiency and mutation bias (Product of equations 5 and 10). (G) all 3 parameters together i.e. Selection acting on codon nitrogen content, mutation bias and selection acting on translational efficiency (Equation 12) (H) empirical codon usage probabilities i.e. the 61 actual codon use frequencies. These simulated sequences produce symmetrical distributions with low variance that do not precisely recapitulate the data found in A. For the kinetoplastids (A-H) the best AIC values for $L_N$ and $M_N$ are 1,992,812 and 2,011,196 respectively for the 3 parameter model (G), for $H_N$ it is 1,756,308 for the two parameter model that considers both translational efficiency and mutation bias. (*I – P)* As for plots A-H but for the Mollicutes. For the Mollicutes the best AIC values for $L_N$, $M_N$ and $H_N$ are 61,440, 71,126 and 59,288 respectively for the 3 parameter model. (O) Y-axis is the probability density function (PDF) for the distributions.


Kinetoplastids
mRNA
dsDNA
$L_N$
$M_N$
$H_N$
A
B
C
D
E
F
G
H
PDF (x10²)
nitrogen atoms, codon⁻¹
nitrogen atoms, dsDNA codon⁻¹
Real codon use (%)
Fitted codon use (%)
Real codon use (%)
R² = 1
R² = 0.83
R² = 0.80
R² = 0
R² = 0.90
R² = 0.91
R² = 0.70
R² = 0.91
R² = 0.87
R² = 0.91
R² = 0.92
R² = 0.88
R² = 0.91
R² = 0.91
R² = 0.91
R² = 0.95
R² = 0.93
R² = 0.91
R² = 0.95
Mollicutes
I
J
K
L
M
N
O
P
R² = 1
R² = 0
R² = 0
R² = 0
R² = 0
R² = 0
R² = 0.87
R² = 0.94
R² = 0.89
R² = 0.95
R² = 0.97
R² = 0.91
R² = 0.86
R² = 0.94
R² = 0.90
R² = 0.96
R² = 0.97
R² = 0.94
R² = 1
R² = 1
R² = 1

A similar phenomenon is observed for the Mollicutes though the fitted mutation bias values are much larger ($m$ > 3.5), indicative of a strong GC to AT mutation bias. This high value for $m$, is consistent with the loss of dUTPase and a reduced ability to correct erroneous dUTP incorporation into the genome (Williams and Pollack 1990; Pollack, Williams, and McElhaney 1997). The selection-mutation model is capable of explaining genome-wide patterns of codon use with 94% accuracy across all nitrogen availability groups. Consistent with nitrogen availability, the value of the selection bias for incorporation of nitrogen atoms in gene sequences was most negative in $L_N$ parasites ($2N_g s$ = -0.24), intermediate in $M_N$ parasites ($2N_g s$ = -0.15), and least negative in $H_N$ parasites ($2N_g s$ = -0.13) (Fig. 3.05, Fig. S3.05M). The distribution of $2N_g s$ parameters for individual species within each group was significantly different when comparing $L_N$ species with $M_N$ or $H_N$ (ANOVA, $p < 0.01$) however the difference between $M_N$ and $H_N$ species failed to reach significance (ANOVA, $p > 0.05$) (Fig. S3.06). As for the kinetoplastids, the AIC values of the selection-mutation model were better than for the models that consider either selection or mutation bias individually (Fig. S3.05J & L, Table S3.01). Furthermore significance testing showed that selection acting on mRNA nitrogen content was significant for all Mollicute groups ($L_N$ p = 0.001, $M_N$ p = 0.001, $H_N$ p = 0.04). As coding sequences comprise the majority of these Mollicute genomes (~83%) the difference in genome wide GC-content between $M_N$ and $L_N$ species is fully attributable to differences in synonymous codon use (Table S3.01).

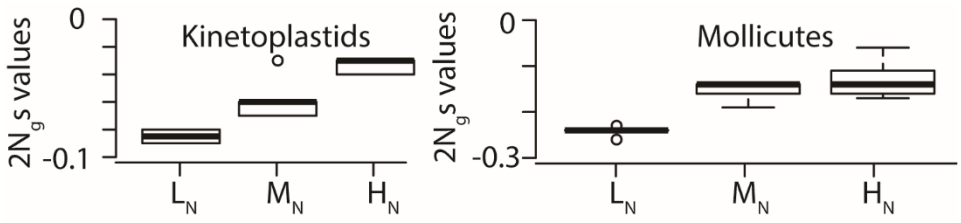



**Figure S3.06. Boxplots showing distribution of $2N_g s$ values for individual species.**

To test whether the observed bias in codon use was also seen more broadly across the genome and not just in the conserved single copy genes an additional analysis was conducted on all complete coding sequences (Table S3.01). The pattern of codon bias was recapitulated for this larger gene set. However the values obtained from the model when considering all complete coding sequences were less extreme than the values obtained when considering conserved orthologous sequences. This is expected as conserved sites in conserved genes have previously been shown to exhibit stronger codon bias (Stoletzki and Eyre-Walker 2007).

### 3.3.5 Gene expression negatively correlates with selection on mRNA nitrogen content

Selection acting on coding sequences is typically considered weak, especially given the low effective populations of the parasites in this study. However previous studies have shown that selection is detectable in highly expressed genes (Higgs and Ran 2008; Ran and Higgs 2010, 2012) and most theories of codon usage predict that the degree of bias due to selection should increase with gene expression (Drummond and Wilke 2009). Given that there is a clear signature of selection acting on nitrogen content genome wide, it was assessed whether the magnitude of this selection was a function of mRNA abundance. Here, the magnitude of selection acting on the nitrogen content of each gene was compared to the mRNA abundance of that gene. For each species there was a negative correlation between mRNA abundance and the fitted $2N_g s$ (Fig. S3.07). This shows that the strongest selection to minimise nitrogen content is observed in the most highly expressed genes. Moreover the slope of the line was greatest for the $L_N$ species, intermediate for the $M_N$ species and weakest for the $H_N$ species. This gene-level analysis is consistent with the genome-wide analysis that showed that $L_N$ species have the greatest selective pressure to minimise nitrogen use.

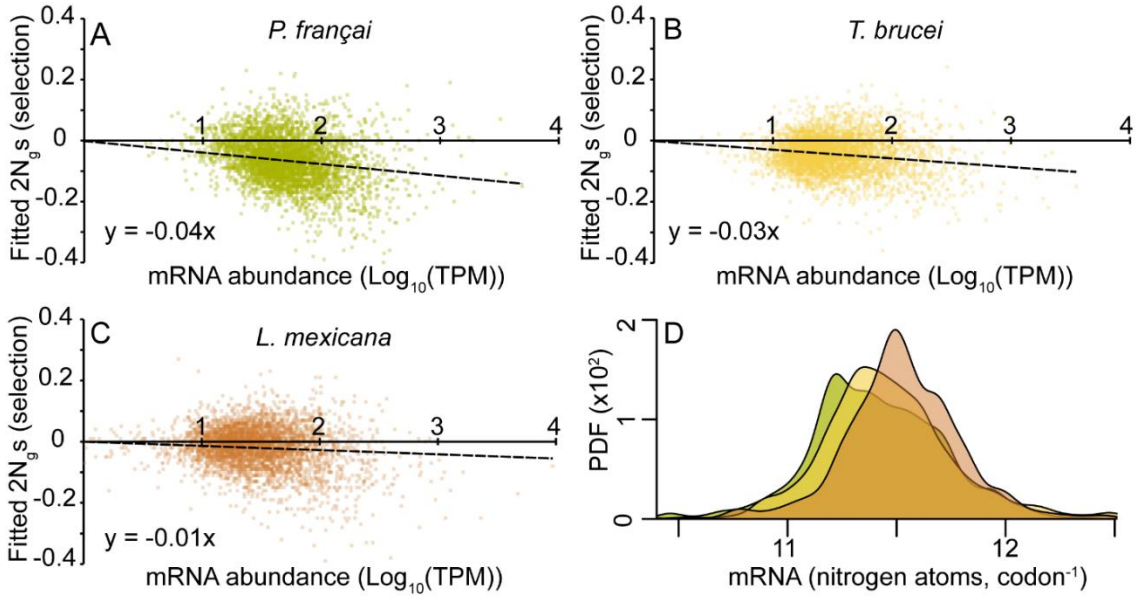



**Figure S3.07. Gene expression negatively correlates with selection acting on mRNA nitrogen content.** (*A-C*) Using the selection-mutation model, $2N_gs$ values were calculated for 4,083 orthologous genes for one species from each of the $L_N$ , $M_N$ and $H_N$ kinetoplastid groups (*P. françai, T. brucei* and *L. mexicana* respectively). Mutation bias values were fixed as constant at the overall value for each species ($L_N$ = 0.61, $M_N$ = 0.92 $H_N$ = 0.34). These values were plotted against mRNA expression data from equivalent lifecycle stages (procyclic) and data points were set to an opacity value of 50% to help judge density. Linear regressions were fitted to the data for each species. The slope of the $L_N$ line was the most negative, $M_N$ was intermediate and $H_N$ was the closest to 0. This is consistent with our other results and shows that selection to minimise nitrogen in mRNA is strongest for the species that are the most nitrogen limited. ie a gene with a TPM value of 100 would be predicted to have a $2N_gs$ value of $L_N$ = -0.08, $M_N$ = -0.06 and $H_N$ = -0.02. Two-tailed t-tests comparing the slopes showed that all of them were significantly different from one another ($p < 0.05$). $R^2$ = 0.07, 0.01 and 0.02 respectively. (*D*) Comparison of the nitrogen use of the 4,083 orthologous genes found in all of *P. françai* (green), *T. brucei* (yellow) and *L. mexicana* (orange) taking expression into account. ie. If a gene had a TPM value of 10, it is represented in the distribution 10 times. All distributions are significantly different using a Wilcoxon Signed-Ranks ($p < 0.05$).

### 3.3.6 Low nitrogen availability parasites have ribosomal RNA sequences that use the lowest amount of nitrogen

Ribosomal RNA (rRNA) typically constitutes the majority of RNA within a cell. To investigate whether selection acting on nitrogen content extends beyond coding sequences, the total nitrogen content of rRNA per ribosome was calculated. Consistent with the analysis of coding sequences, $L_N$ parasite rRNAs require the lowest amount of nitrogen. In the Mollicutes, $L_N$ parasites used 8 fewer nitrogen atoms compared to $M_N$ and 63 fewer atoms compared to $H_N$ parasites per 70S ribosome (Table S3.01). This difference is lower than expected when compared to the analysis of protein coding genes which, given the length of the rRNA sequence analysed, would have predicted differences of 77 and 140 nitrogen atoms respectively. This reduced difference is most likely due to structural constraints on rRNA and the fact that it is not composed of codons and so may lack the flexibility provided by synonymous codons.

The same analysis of rRNA sequences was carried out for the kinetoplastids. Consistent with the analysis of the Mollicutes, the RNA component of the 80S ribosome required the least amount of nitrogen in the $L_N$ kinetoplastid parasites. However, due to large insertions in *Trypanosoma cruzi* rRNAs, the $M_N$ parasites required more nitrogen than the $H_N$. These inserted regions increased the total nitrogen content in the *T. cruzi* rRNA by > 1,500 nitrogen atoms (~7% more than the other $M_N$ species, Table S3.01). Thus with one exception, the analysis of rRNA genes is consistent with the analysis of protein coding genes.

### 3.3.7 Nitrogen content of nucleotide sequences can predict metabolic capability

Given that the relative use of synonymous codons is affected by selection acting on nitrogen content, it was determined to what extent the selection-mutation model could predict the dietary nitrogen content of an organism. This was tested by analysing four additional Mollicute genomes not included in the original analysis. Each additional species was classified as $H_N$ by model selection through maximum likelihood estimation (Table S3.01). To provide support for these classifications, the parasites' genomes were searched for genes

required for amino-acid and amino-sugar catabolism. This revealed that, in contrast to $M_N$ Mollicute parasites, the genomes of the additional species each encoded complete metabolic pathways for catabolism of either arginine and/or amino-sugars (Fig. 3.06, Table S3.01). Moreover, the genes for these pathways were co-located in gene clusters, indicative of genes belonging to the same metabolic pathway (Fig. S3.08). These results demonstrate the utility of the model for providing information about the metabolic capabilities of an organism from raw nucleotide sequences.

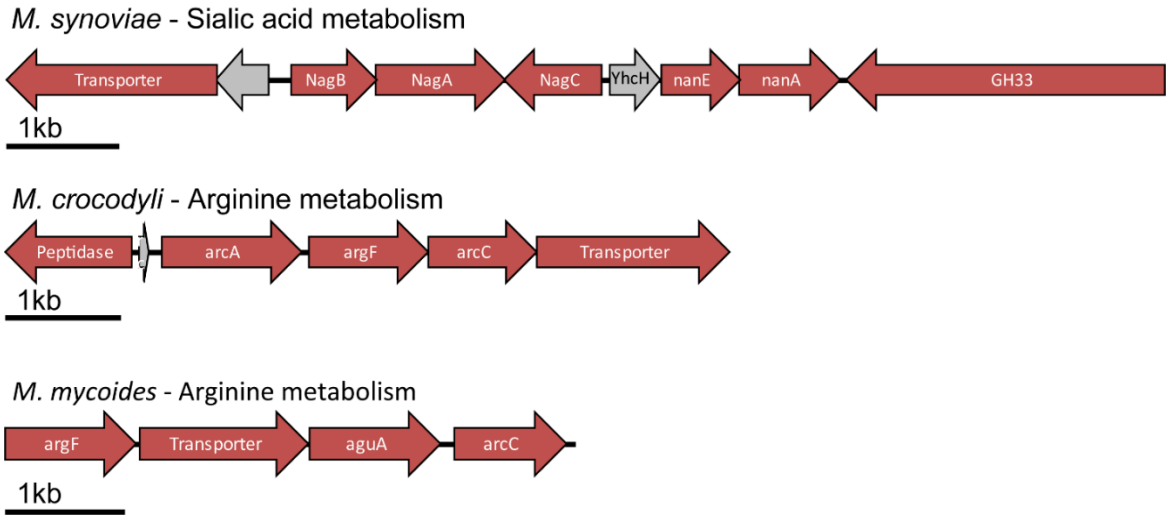



**Figure S3.08. Gene clusters for nitrogen liberating metabolic pathways in $H_N$ Mollicute parasites.**

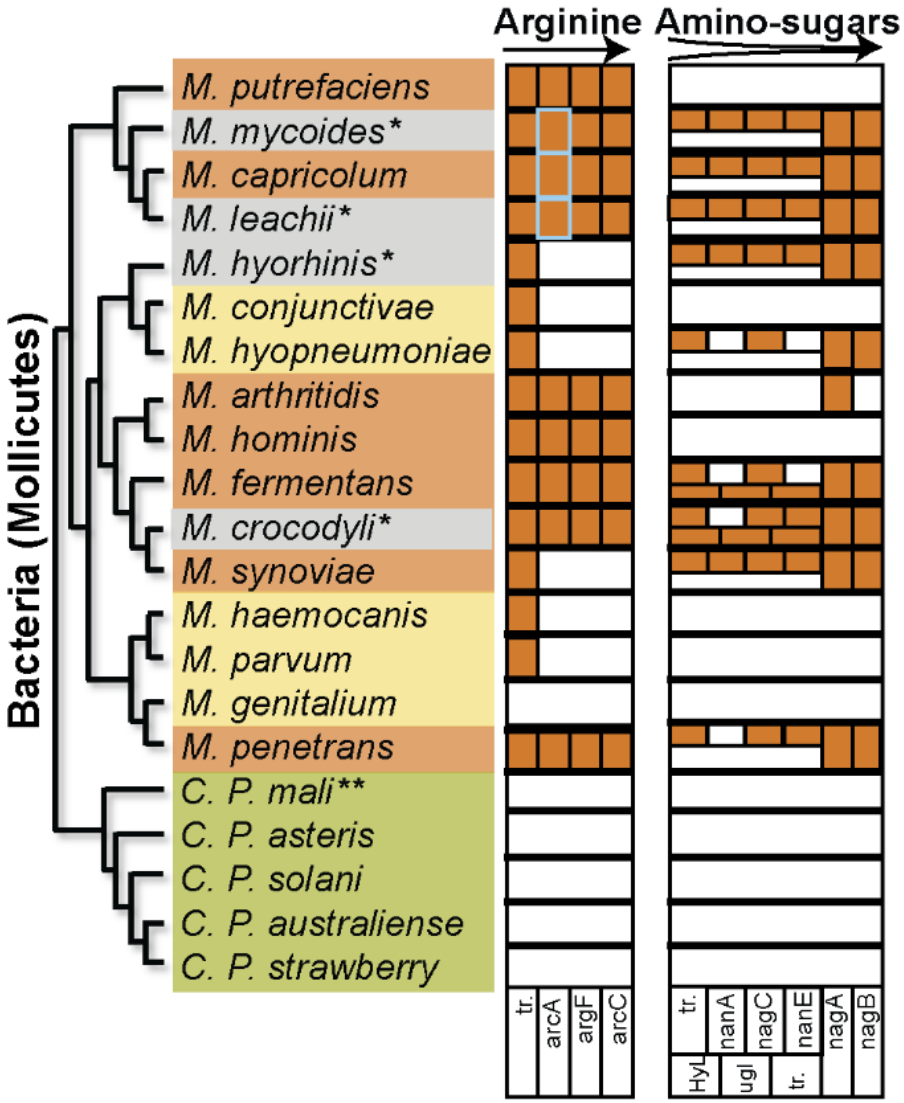



**Figure 3.06. Selection-mutation model of codon use can predict the metabolic capacity of parasites from raw nucleotide sequences.** $L_N$ (green), $M_N$ (yellow) and $H_N$ (orange) are species with low, medium and high nitrogen availability. Species denoted with * (grey) are the new species each of which is predicted to be a $H_N$ parasite. Presence of named genes involved in arginine catabolism and amino-sugar catabolism are indicated by orange boxes. Gene names are provided below the boxes and the abbreviations correspond to the following genes: tr. - transporter, *arcA* - arginine deiminase, *argF* - ornithine carbamoyltransferase, *arcC* - carbamate kinase, *nanA* - N-acetylneuraminate lyase, *nanK/nagC* - N-acetylmannosamine kinase, *nanE* - N-acetylmannosamine-6-p epimerase, *nagA* - N-acetyl-D-glucosamine-6-phosphate deacetylase, *nagB* - D-glucosamine-6-phosphate deaminase, *HyL* - hylauronate lyase, *ugl* – glucuronidase. Blue outlined boxes encode *aguA* which does the same reaction as *arcA* on an equivalent substrate.

### 3.3.8 Selection acting on nitrogen content is independent of selection acting on translational efficiency

Translational selection, which is a function of the number of iso-accepting tRNAs encoded in a genome, has long been considered a major driver of codon bias (Shah and Gilchrist 2011). To determine how selection acting on nitrogen content acts in concert with selection acting translational efficiency (tAI) the model was expanded to include tAI as an additional parameter (see Methods). For the Mollicutes, unlike the result above where selection acting on nitrogen content was significant for all three parasite groups, it was found that considering tAI values alone or in conjunction with mutation bias was not significantly better than when tAI was omitted ($p > 0.05$). However when all 3 parameters (nitrogen content, mutation bias and tAI) were considered together, the model fits the data significantly better than when considering just selection acting on nitrogen content and mutation bias for the $L_N$ and $H_N$ parasites ($p \leq 0.02$). Thus selection acting on translational efficiency is independent of selection acting on nitrogen content, and only provides a significant contribution to codon bias in $L_N$ and $H_N$ species (Fig S3.05).

In contrast, for the kinetoplastids it was found that the fit to observed patterns of codon use was significantly better with the inclusion of tAI values in conjunction with mutation bias ($L_N$ $p = 0.018$, $M_N$ and $H_N$ $p = 0$). The contribution of tAI was also significant for all three kinetoplastid parasite groups when all 3 parameters (nitrogen content, mutation bias and tAI) were considered together ($L_N$ $p = 0.006$, $M_N$ $p = 0.001$, $H_N$ $p = 0$, Fig S3.05). Thus, as for the Mollicutes, selection acting on translational efficiency is independent of selection acting on nitrogen content. Furthermore, inclusion of translational efficiency in the model improves overall fit by ~2% to give an average accuracy of 94.3%. This compares to a 3.1% improvement in overall fit when selection acting on nitrogen content is added to the model that only considers mutation bias and translational efficiency.

Selection acting on nitrogen content can explain why the most translationally optimal codons are not always the codons that are most frequently used. For example, in Mollicute parasites only 33% of the most frequently used codons for each amino acid are the most translationally efficient while 66% are those with the lowest nitrogen content (Fig. S3.09). A similar pattern occurs in the kinetoplastid parasites, however the most translationally efficient codon is the most frequently used codon more often than the most nitrogen efficient codon (74% as compared to 30% respectively). This interplay between translation and nitrogen content is also seen when the relative order of all synonymous codons is analysed in these two parasite groups (Fig. S3.09). Furthermore, these observations are consistent with the global analysis of codon use presented above which showed that selection acting on nitrogen content was more important than selection acting on translation efficiency in determining patterns of codon bias in Mollicutes, while selection acting on nitrogen content and translation efficiency were required to explain patterns of codon use in kinetoplastids.

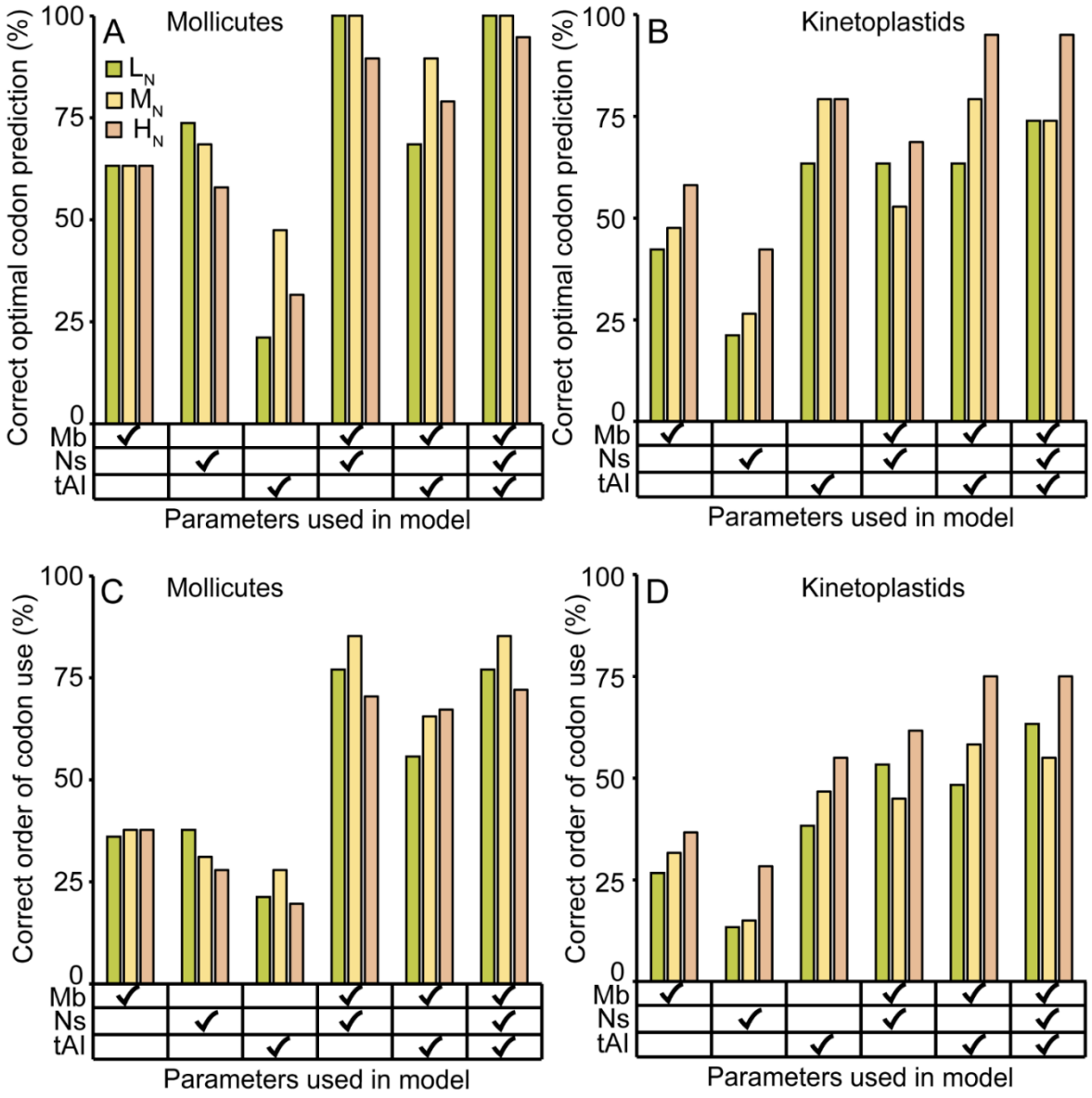



**Figure S3.09. The model parameters which provide the best fit between observed and predicted codon use also provide the best percentage of correctly predicted optimal codons and correctly ordered codon use.** (A) Percentage of correctly predicted optimal codons for Mollicute parasites using different parameter combinations. (B) As in (A) but for kinetoplastid parasites. (C) Percentage of correctly ordered relative codon use for different parameter combinations. (D) As in (C) but for kinetoplastid parasites. Mb = Mutation bias, Ns = selection acting on nitrogen content, tAI = selection acting on translational efficiency.

### 3.4 Discussion

Studies on the interactions between diet, metabolism and evolution have primarily focused on the presence or absence of individual genes in the context of specific metabolic pathways. However the impact of an organism's diet on the evolution of its genes and genome is poorly understood. Here we show that differential nitrogen availability, due to differences in host environment and metabolic inputs, alters synonymous codon usage and thus gene sequence evolution in both bacterial and eukaryotic parasites. Moreover, this impact is sufficient to enable prediction of metabolic inputs of parasites from comparative analysis of the nucleotide composition of orthologous genes.

In this work we provide a novel selection-mutation model for synonymous codon use that builds upon a strong theoretical foundation (Li 1987; Shields 1990; Bulmer 1991). In this model we have amalgamated multiple factors contributing to genome-wide GC-content into the single variable termed mutation bias ($m$, mutation bias towards AT). Such factors include the bias of an organism's DNA polymerase (Worning et al. 2006), gene conversion (Galtier 2003), differences in repair efficiency (Williams and Pollack 1990) as well as mutational biases during DNA replication (Eyre-Walker 1991; Francino and Ochman 1999; Rao et al. 2011). We also suggest that differences in nitrogen availability may contribute to differences in mutation bias through influencing the relative abundance of nucleotides (Buckland et al. 2014). Considering mutation bias alone the model was able to recapitulate the observed synonymous codon use with ~90% accuracy for both the Mollicute and kinetoplastid parasites. Furthermore the large differences in mutation bias between kinetoplastids ($m < 1$) and Mollicutes ($m > 3.5$) explain the large differences in observed patterns of codon bias between the two distantly related parasite lineages. Interestingly the kinetoplastid $m$ values are each below 1 ($L_N$ $m = 0.68$, $M_N$ $m = 0.74$, $H_N$ $m = 0.31$) and thus correspond to a bias towards GC. The differences in mutation bias between parasite groups is consistent with differences in nitrogen availability, as a high GC-content is equivalent to

high nitrogen content of the dsDNA. An analogous nitrogen-dependent difference in *m* is not seen in the Mollicutes. We propose that this is due to the strong AT mutation bias (*m* values all greater than 3.5) that constrains dsDNA nitrogen within a narrow range of values compared to the kinetoplastids.

Due to the complementary nature of DNA, a change on either DNA strand will cause a corresponding change on the other strand. Therefore mutation bias alone was unable to produce the differences in the nitrogen content of the coding strand (i.e. the mRNA) that was observed between species with different nitrogen availabilities. As shown in equation [2] (see methods), selection depends on $N_g$ (the effective number of genes at the locus in the population) which is linked to effective population size ($N_e$) of an organism (Lynch 2007). Organisms with low long-term effective population sizes have a reduced impact from selection due to the greater impact of random genetic drift. Thus $N_g$ plays an important role in determining the role of selection in biased codon usage. However as has been noted before, evaluating the long term $N_g$ value for an organism is very difficult (Sharp et al. 2005). Eukaryotes have lower $N_g$ values than prokaryotes, and parasites in general have lower $N_g$ values than their free-living counterparts due to clonal life stages and bottlenecks during transmission. Our model evaluates the selection bias acting on nitrogen content using a composite parameter ($2N_g s$). Thus the value of the selection coefficient *s* is linearly dependent on estimates of $N_g$ (ie. Increasing $N_g$ by a factor of 10 decreases *s* by a factor of 10). It is interesting to note that estimates of $N_g$ for prokaryotes and unicellular eukaryotes differ by a factor of 10 (Lynch 2007), similar to the magnitude of difference we see between $2N_g s$ for the Mollicutes and the kinetoplastids, indicating that the selection coefficient *s* may be similar for the two distantly related groups.

Previous studies investigating the role of selection in codon bias have revealed that selection acting on translational efficiency in Mollicutes is marginal (Sharp et al. 2005). Although codon biases in prokaryotic genomes are associated with gene expression levels

(Supek et al. 2010; Krisko et al. 2014), in some cases the optimal codons disagree with the tRNA composition. These observations support the results presented here which show that for Mollicutes, inclusion of tAI values does not significantly improve the fit of the model unless it is considered in conjunction with both mutation bias and selection acting on the nitrogen content of coding sequences. Our finding that selection acts on the nitrogen content of codons provides a novel mechanism that links codon usage bias to metabolism and environment. Furthermore, as the model developed here is sufficient to enable prediction of metabolic inputs from gene sequences, it may have application in interrogating metagenome data and genome data from shotgun sequencing of microbial communities where metabolic requirements are unknown.

Though the selection-mutation model provides considerable explanatory power for the species used in this analysis, it does not perfectly recapitulate the observed patterns of codon use. This is most likely due to the fact that specific sites within a gene will be under different pressures that cannot be captured by a genome-wide approach. For example factors indirectly related to protein translation such as mRNA secondary structures at the 5' region of a gene have been shown to be under selection for efficient binding of ribosomes to mRNAs and hence can have a weak effect on the frequency of codon usage at those sites (Tuller, Waldman, et al. 2010). A more complex model could include variation in codon bias between genes due to gene-specific selective pressures such as splice site conservation, mRNA stability, ribosome binding and mRNA abundance. Taken together these factors may account for the ~6% of missing variation not explained by the selection-mutation model presented here. Incorporation of these factors into the model would be an interesting avenue of future research.

Thermophilic bacteria purine-load their genomic sequences to the extent that amino acid composition is affected (Lao and Forsdyke 2000). However this effect is not seen in the mRNA of mesophilic organisms (Hurst and Merchant 2001; Lambros, Mortimer, and

Forsdyke 2003) and so would not be expected to feature in the dataset analysed here. For example the difference observed between $M_N$ and $H_N$ parasites cannot be due to temperature as both groups infect animal hosts with very similar (if not identical) temperatures. Furthermore some of the $H_N$ (*M. crocodyli*, and *L. tarentolae*) and $M_N$ (*T. grayi*) species in both the Mollicutes and the kinetoplastids infect cold-blooded reptiles and thus would have host temperatures more similar to plant infecting $L_N$ parasites than to warm blooded animals. Even though these parasites infect cold-blooded animals, their nitrogen use profiles are consistent with their metabolic group rather than their host temperature. Finally, conducting our analysis on parasites in the same ecological niche revealed that, at the same temperature, in the same micro-environment, the parasites exhibited different nucleotide nitrogen content consistent with their dietary nitrogen availability. These results indicate that while temperature may be important in extreme environments, temperature is not a factor in the comparisons presented here. This is consistent with previous analyses that showed that even at relatively freely evolving sites, mRNA GC-content did not appear to be adapted to the thermal environment (Hurst and Merchant 2001).

Elemental limitation has not previously been proposed to influence codon bias, however this analysis demonstrates via multiple complementary approaches that nitrogen availability contributes to genome composition. Taken together, the findings presented here show that differential nitrogen availability, due to differences in host environment and metabolic inputs, contributes to differences in codon bias between related species. These results reveal a previously hidden relationship between cellular metabolism and genome evolution and provide new insight into how genome sequence evolution can be influenced by adaptation to different diets.

## 3.5 Methods

### 3.5.1 Data sources

17 Mollicute genomes were obtained from NCBI genbank. These comprised 4 plant glycolytic parasite species (*C. P. asteris* (Oshima et al. 2004), *C. P. australiense* (Tran-Nguyen et al. 2008), *C. P. mali* (Kube et al. 2008), *C. P. solani* (Mitrović et al. 2014), *C. P. strawberry* (Andersen et al. 2013)), 5 animal glycolytic parasite species (*M. conjunctivae* (Calderon-Copete et al. 2009), *M. genitalium* (McGowin et al. 2012), *M. haemocanis* (Nascimento et al. 2012), *M. hyopneumoniae* (Liu et al. 2013), *M. parvum* (do Nascimento et al. 2013)) and 7 parasite species known to obtain energy from catabolism of amino acids or amino sugars (*M. arthritidis* (Dybvig et al. 2008), *M. capricolum* [PRJNA16208], *M. fermentans* (Shu et al. 2011), *M. hominis* (Pereyre et al. 2009), *M. penetrans* (Sasaki et al. 2002), *M. putrefaciens* (Calcutt and Foecking 2011), *M. synoviae* (Vasconcelos et al. 2005)). A further 4 parasite species were used for testing the predictive capacity of the model for synonymous codon use (*M. crocodyli* (Brown et al. 2011) , *M. hyorhinis* (Dabrazhynetskaya et al. 2014), *M. leachii* (Wise et al. 2012), *M. mycoides* (Wise et al. 2012)).

15 kinetoplastid genomes were obtained online from TriTrypDB (Aslett et al. 2009), NCBI genbank or the European Nucleotide Archive. These comprised four plant glycolytic parasite species (*P. EM1* [GCA_000582765] (Porcel et al. 2014)*, P. françai, P. HART1* [GCA_000982615] (Porcel et al. 2014)*, P. serpens* [PRJNA80957], 5 animal glycolytic parasite species (*T. brucei* [PRJNA15565]*, T. congolense* [PRJNA12958]*, T. cruzi* [PRJNA15540/PRJNA11755]*, T. grayi* [PRJNA258390]*, T. vivax* [PRJNA12957]) and six parasite species who obtain energy primarily from catabolism of amino acids (*L. braziliensis* [PRJNA19185]*, L. donovani* [PRJNA171503]*, L. infantum* [PRJNA19187]*, L. major* [PRJNA10724]*, L. Mexicana* [PRJNA172192]*, L. tarentolae* [PRJNA15734]).

### 3.5.2 Inference of orthogroups and construction of multiple sequence alignments

The predicted amino acid sequences for each species were subject to orthogroup inference using OrthoFinder (Emms and Kelly 2015) using the default program parameters. Single

copy genes were selected for analysis to ensure orthology and so that paired comparisons could be made. i.e. a single copy orthologous gene that is present in two different species can be treated as a paired observation. Single copy gene orthogroups were further filtered to retain those that had representation from at least 3 species per group ($L_N$, $M_N$ and $H_N$). Protein sequences for these orthogroups were aligned using MergeAlign (Collingridge and Kelly 2012). The corresponding coding sequences were re-threaded back through the aligned amino acid sequences using custom Perl scripts. These multiple sequence alignments were then filtered so that only un-gapped columns that obtained a MergeAlign column score of >0.75 were retained for further analysis. These stringent filtration criteria ensured that only high accuracy, unambiguously aligned orthologous positions were used for all analyses. The accession numbers for the full set of orthogroups are provided in Supplemental Table S3.01.

### 3.5.3 Evaluation of nitrogen content of nucleotide sequences

The filtered multiple sequence alignments above were used to calculate the number of nitrogen atoms used per codon, per gene per species. The number of nitrogen atoms per codon per gene was evaluated as the arithmetic mean of the number of nitrogen atoms in the filtered aligned codons for that gene described above. The average number of nitrogen atoms contained within the mRNA and the dsDNA were recorded for each gene. These data were plotted as probability density functions using the R density distribution plot function with the total area under each curve equal to one.

### 3.5.4 Analysis of rRNA

A database of representative rRNA sequences was generated and blasted against the genomes of all the parasites in this study to find the locations of the rRNAs. In the event of no or partial blast hits, sequences were downloaded from NCBI and the accession numbers noted in Supplemental Table S3.01. Sequences were then aligned using MAFFT (Katoh and Standley 2013) to identify the true start and end of the rRNA molecules. The nitrogen content of these sequences was calculated. Due to difficulties in sequencing and assembling

repetitive rDNA loci, some species did not have complete sequences to include in this analysis. Those were labelled NF (not found).

### 3.5.5 Statistical tests

Given that single copy orthologous genes present in different species can be treated as paired observations, Wilcoxon Signed-Ranks tests were used to compare nitrogen content between different parasite groups. In each case the null hypothesis was that the difference between the two groups was due to chance (symmetric around zero). The alternative hypothesis was that the difference in nitrogen content between each group was not due to chance. In all cases, the test used was two-tailed so that either a greater or lesser nitrogen content difference would reject the null hypothesis. Pairing of samples is justified as the paired observations (genes) are orthologous and descended from the same common ancestor under different environmental and metabolic conditions.

Goodness of fit and the statistical significance of the inclusion of additional parameters to the model were assessed by comparison of AIC values and by using a permutation test respectively. For the permutation test, the log likelihood values obtained by the model when run with real values were compared with the log likelihood values obtained by the model when it was run with shuffled values. ie. To analyse the significance of the inclusion of nitrogen selection to the model, the codon nitrogen contents were calculated and then shuffled to randomly assign the values to each codon. The model was then fitted to the data using these randomised values and the log likelihood compared to the log-likelihood obtained using the real values. This was repeated for 1,000 independently shuffled sets. The same principle was applied to significance testing of the tAI values. An example of the distributions generated with shuffled codon nitrogen content is provided (Fig. S3.10).

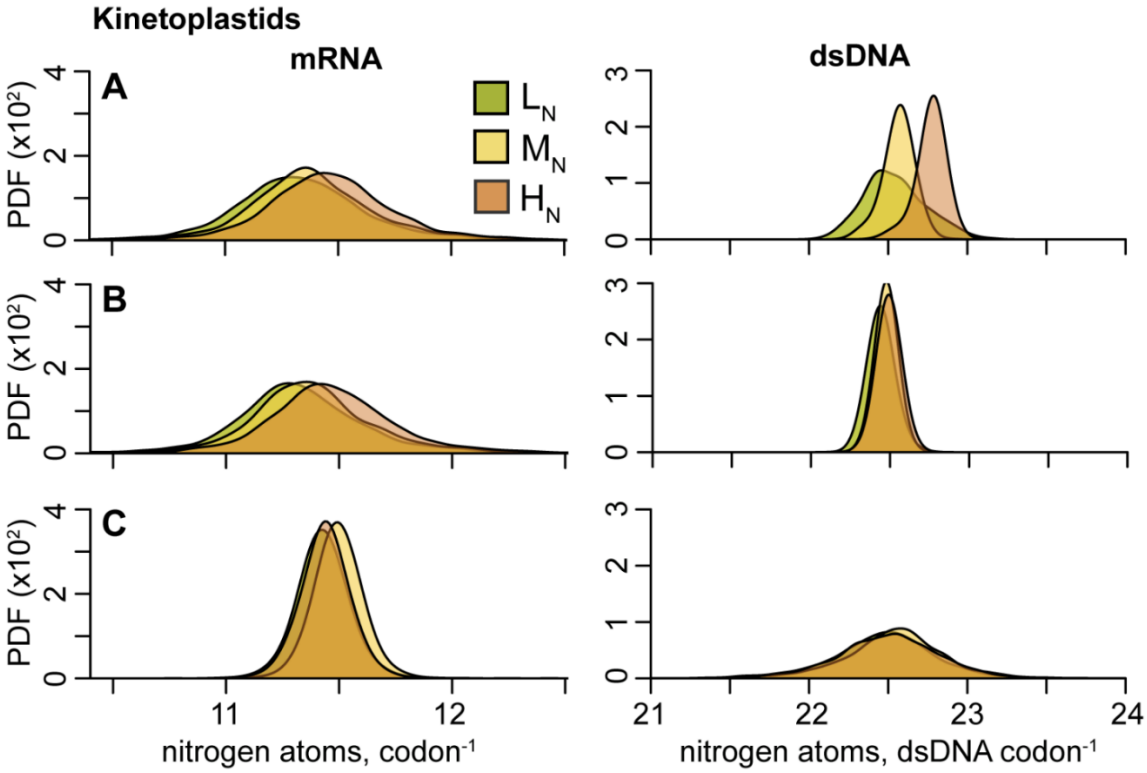



**Figure S3.10. Example distribution of the model when run with shuffled codon nitrogen content.** (*A*) Left, the average mRNA nitrogen content per codon for 3,003 orthologous genes in the kinetoplastida (observed data). Right, the average nitrogen content per double stranded codon for the same set of genes. Analogous plots are shown in panels B and C for sequences simulated using fitted values for (*B*) selection acting on codon nitrogen content (Equation 2). (C) selection acting on codon nitrogen content where the nitrogen content of the codons has been shuffled (as described in the methods). Shuffled codon nitrogen cost (C) provides distributions that do not fit observed data (A) as well as distributions that use real nitrogen content (B). $2N_gs$ values for the shuffled example were $L_N = 0.01$, $M_N = 0.02$ and $H_N = 0$ compared to -0.06, -0.04 and 0.03 for the real nitrogen content. Thus when codon nitrogen content is shuffled, the model is unable to recapitulate the observed distribution of nitrogen content in gene sequences and the influence of the model parameter is reduced towards zero.

### 3.5.6 A model for synonymous codon use under the joint pressures of selection and mutation bias

To determine whether nitrogen availability influences interspecies variation in codon use and nucleotide content, a model for synonymous codon use was constructed. This model considers the selection bias acting to modulate a codon's nitrogen content and an organism's mutation bias. The system of equations describing the model is as follows.

**Synonymous codon use considering selection acting on mRNA nitrogen content:**

Here we consider that selection acts to bias synonymous codon use in proportion to the number of nitrogen atoms contained within each codon. That is

$$[1]\ \ S(\mathcal{C}_i) = sN_{mRNA}$$

Where $S(\mathcal{C}_i)$ is a measure of the relative fitness of codon $\mathcal{C}_i$, with $N_{mRNA}$ being the number of nitrogen atoms in codon $\mathcal{C}_i$, and $s$ being the selection coefficient. Following previous published work (Li 1987; Shields 1990; Bulmer 1991), we model the selection bias towards codon $\mathcal{C}_i$ as

$$[2]\ \ \alpha(\mathcal{C}_i) = e^{2N_g S(\mathcal{C}_i)}$$

where $\alpha(\mathcal{C}_i)$ is the selection bias towards codon $\mathcal{C}_i$ and $N_g$ is the effective number of genes at a locus. Only considering this selection bias, we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as

$$[3]\ \ p(\mathcal{C}_i \mid \theta) = \frac{\alpha(\mathcal{C}_i)}{\sum_\theta \alpha(\mathcal{C})}$$

That is, the probability of observing codon $\mathcal{C}_i$ is the selection bias towards codon $\mathcal{C}_i$ divided by the sum of selection biases for all codons encoding amino acid θ. Equation [3] satisfies the law of total probability such that the sum of the probabilities of observing all the codons that encode the same amino acid sum to one.

**Synonymous codon use considering mutation bias only:**

Mutation bias is known to be influenced by a range of factors including but not limited to the bias of an organism's polymerase-α subunit (Worning et al. 2006), gene conversion (Galtier 2003), and differences in repair efficiency (Williams and Pollack 1990). We propose that nitrogen-mediated changes in nucleotide pools also contribute to this mutation bias, as changes in nucleotide pools result in changes in mutation bias (Buckland et al. 2014). For example the amount of biologically available nitrogen within a cell could alter the relative abundance of nucleotides via enzymes such as CTP synthase that catalyse nitrogen-dependent nucleotide interconversion of UTP and CTP. Here we have amalgamated these factors into the single variable $m$.

$$[4]\ \delta = \frac{m}{m+1}$$

Where $\delta$ is the probability that a particular site is A or T, given a mutation bias towards AT of $m$ as previously described (Lynch 2007). Due to base pairing, the probability of A or T is equivalent. This equation assumes that the nucleotide composition of the genome is at equilibrium and that the mutation rate per site is independent of the status of neighbouring sites (Lynch 2007). For example, if there is no mutation bias towards AT or GC $m$ will be 1 and $\delta$ will be 0.5 and thus there is an equal likelihood of any site being AT or GC. We model the mutation bias towards codon $\mathcal{C}_i$ as

$$[5]\ \ \beta(\mathcal{C}_i) = \delta^{AT}(1-\delta)^{GC}$$

Where $\beta(\mathcal{C}_i)$ is the mutation bias towards codon $\mathcal{C}_i$, $AT$ is the number of A or T nucleotides in codon $\mathcal{C}_i$ and $GC$ is the number of G or C nucleotides in codon $\mathcal{C}_i$. Considering only mutation bias, we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as

$$[6]\ \ p(\mathcal{C}_i \mid \theta) = \frac{\beta(\mathcal{C}_i)}{\sum_\theta \beta(\mathcal{C})}$$

That is, the probability of observing codon $\mathcal{C}_i$ is the mutation bias towards codon $\mathcal{C}_i$ divided by the sum of mutation biases for all codons encoding amino acid θ. Equation [6] also satisfies the law of total probability such that the sum of the probabilities of observing all the codons that encode the same amino acid sum to one. For example, if $m$ = 3 then $\delta$ = 0.75 and we consider amino acid C (encoded by codons TGC and TGT), then the mutation bias towards codon TGC = $\beta(TGC) = 0.75^1(1-0.75)^2$ = 0.047 and the mutation bias towards codon TGT = $\beta(TGT) = 0.75^2(1-0.75)^1$ = 0.141. Thus the genome-wide probability of observing codon TGC = $\frac{0.047}{0.047+0.141}$ = 0.25, and the genome-wide probability of observing codon TGT = 0.75.

**A model for synonymous codon use under the joint pressures of selection and mutation bias**

We model the bias towards codon $\mathcal{C}_i$ under the joint pressures of selection and mutation as the product of equations [2] and [5].

$$[7]\ \ \gamma(\mathcal{C}_i) = e^{2N_g S(\mathcal{C}_i)}\delta^{AT}(1-\delta)^{GC}$$

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as

$$[8]\ \ p(\mathcal{C}_i \mid \theta) = \frac{\gamma(\mathcal{C}_i)}{\sum_\theta \gamma(\mathcal{C})}$$

It should be noted selection in this model only considers the nitrogen content of a codon and does not consider other factors such as biased gene conversion (Lynch 2007). However, kinetoplastids primarily reproduce by clonal expansion and the prokaryotic genomes are haploid, thus gene conversion may have limited impact in these organisms.

### 3.5.7 Calculation of codon tRNA adaptation index (tAI) values

The tRNA adaptation index (tAI) (dos Reis, Wernisch, and Savva 2003) of a codon takes into account both the abundance of iso-accepting tRNAs and wobble-base pairing to evaluate the efficiency of translation of a given codon. Using the equation developed by dos Reis et

al (dos Reis, Savva, and Wernisch 2004) below, and the optimised $s_{ij}$ values obtained by Tuller et al (Tuller, Carmi, et al. 2010), tAI values for each codon were evaluated.

$$[9]\ \omega(C_i) = \sum_{j=1}^{n_i}\left(1 - s_{ij}\right)tGCN_{ij}$$

Where $\omega(C_i)$is the absolute adaptiveness value for each codon $C_i$ (referred to in the rest of the text as the tAI value), $n_i$ is the number of tRNA isoacceptors that recognise codon $C_i$, $tGCN_{ij}$ is the gene copy number of the $j^{\text{th}}$ tRNA that recognises codon $C_i$, and $s_{ij}$is the selective constraint on the efficiency of codon-anticodon coupling.

We model the translational selection bias towards codon $\mathcal{C}_i$ as

$$[10]\ \eta(\mathcal{C}_i) = e^{2N_g\sigma\omega(\mathcal{C}_i)}$$

Where $\omega(\mathcal{C}_i)$ is the translational selection bias towards codon $\mathcal{C}_i$, $\sigma$ is the selection coefficient and $N_g$ is the effective number of genes at a locus.

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as

$$[11]\ \ p(\mathcal{C}_i \mid \theta) = \frac{\eta(\mathcal{C}_i)}{\sum_\theta \eta(\mathcal{C})}$$

When considering all 3 parameters (mutation bias, selection acting on the nitrogen content of coding sequences and translational selection) we model the bias towards codon $\mathcal{C}_i$ as the product of equations [2], [5] and [10].

$$[12]\ \ \varepsilon(\mathcal{C}_i) = \alpha(\mathcal{C}_i)\,\beta(\mathcal{C}_i)\,\eta(\mathcal{C}_i)$$

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as

$$[13]\ \ p(\mathcal{C}_i \mid \theta) = \frac{\varepsilon(\mathcal{C}_i)}{\sum_\theta \varepsilon(\mathcal{C})}$$

### 3.5.8 Model fitting and implementation

Using the system of equations in the model, the parameters ($2N_g s$, $m$ and tAI) were estimated for each of the parasite groups using a maximum likelihood approach. The models for both

the Mollicute and kinetoplastid parasites each contain a maximum of three free parameters (selection acting on nitrogen content, mutation bias and translational efficiency) and thus a brute-force parameter search was conducted to find their optimal values. Here, the likelihood of observing the set of sequences contained within each parasite group was evaluated given the model for synonymous codon use and the values of the parameters.

It was evaluated as follows

$$\mathcal{L}(s, m|X) = \prod_{\mathcal{C}_i} p(\mathcal{C}_i|\theta)^{N_{\mathcal{C}_i}}$$

Where $X$ is the set of coding sequences for a given species, and $N_{\mathcal{C}_i}$ is the number of times that codon $\mathcal{C}_i$ occurs in the set of sequences $X$. The optimal parameter values were determined as those with the maximum likelihood. This was applied to look at both orthologous genes (the same set as those described in the section entitled 'Inference of orthogroups and construction of multiple sequence alignments') and the full set of coding sequences. Source code and data files for this analysis are available from the Zenodo research data repository https://doi.org/10.5281/zenodo.154493.

### 3.5.9 Classification of additional species using the metabolic model for synonymous codon use

Four additional Mollicute genomes not included in the initial analysis were downloaded from NCBI to test the ability of the model for synonymous codon use to predict the metabolic properties of these organisms from analysis of codon use. These species were *Mycoplasma crocodyli, Mycoplasma hyorhinis, Mycoplasma leachii* and *Mycoplasma mycoides*. Based on literature evidence and phylogeny (Fig. 3.06) it was expected that some of the additional species would be classified as $H_N$ and some as $M_N$ parasites. Using the system of equations described above and the values obtained for the dependency parameters ($2N_g s$ and $m$) for each of the $L_N$, $M_N$ and $H_N$ Mollicute parasite groups, a likelihood that each species belonged to each group was calculated (Table S3.01). The model with the highest likelihood was

determined to be $H_N$ in all instances. This classification was confirmed using a Wilcoxon Signed-Ranks test on the nitrogen cost of the mRNA. Each of the additional species were significantly different ($p < 0.001$) from the $L_N$ and $M_N$ groups and not significantly different ($p > 0.05$) from the $H_N$ group. The one exception to this was *M. crocodyli.* This species had the highest mRNA nitrogen cost of any Mollicute species in this analysis and was significantly higher than the other species in the $H_N$ group. This may indicate increased dependence on nitrogen-liberating metabolic pathways or an increased availability of nitrogen in the host environment.

**Table S3.01: [online at https://github.com/easeward/Thesis]**

Sheet 1: Accession numbers and corresponding orthogroups for all kinetoplastid species used in this analysis. Sheet 2: Accession numbers for the genes required for arginine and amino sugar metabolism in the Mollicutes. Sheet 3: The model was run on the dataset of orthologous genes and it was also run separately on all complete coding sequences. A complete coding sequence was considered as any coding sequence with start and stop codons whose length was divisible by 3 and longer than 30 nucleotides. This sheet shows the parameters obtained for each of the $L_N$, $M_N$ and $H_N$ groups for both the Mollicute and kinetoplastid parasites. Highlighted in yellow are the parameters obtained for the original 2-parameter model that only considers selection acting on nitrogen content in coding sequences and mutation bias. Highlighted in green are the values obtained for the 3-parameter model, which also considers selection acting on translational efficiency (tAI). The model parameters that produce the highest log likelihood values and the lowest AIC values are the best fit to the observed data. Sheet 4: Genome locations, nitrogen use and aligned sequences for Mollicute 5S, 16S and 23S rRNA. Sheet 5: Genome locations, nitrogen use and aligned sequences for kinetoplastid 5.8S, 18S, 23S alpha and 23S beta rRNA. Total nitrogen used in rRNA per 80S ribosome is given in the top left corner. Sheet 6: Accession numbers for the genes required for arginine and amino sugar metabolism in the Mollicutes.

# **Chapter 4:** Selection-driven cost-efficiency optimisation of transcripts governs evolutionary rate in bacteria

**Seward E. A. & Kelly S. 2017.** Selection-driven cost-efficiency optimisation of transcripts governs evolutionary rate in bacteria**.** ***bioRxiv***, **doi.org/10.1101/136861**

**This chapter is currently under review at Nature Communications.**

## 4.1 Chapter Introduction

In the previous chapter I described the role of nitrogen availability in determining sequence evolution in two groups of single-celled parasites. In this chapter I use genome data from 1,320 bacterial genomes to extend and expand that analysis. To do this I analyse the strength of selection acting on nucleotide cost and integrate it with the strength of selection acting on codon translational efficiency to provide an in-depth analysis of the interaction between these two factors. Comparing the trade-off between codon cost and efficiency places the impact that dietary nitrogen has on gene evolution in the context of other factors known to affect gene evolution. This analysis reveals three key findings:

1. Natural selection solves the combinatorial optimisation problem of reducing resource allocation to mRNA biosynthesis by simultaneously minimising the biosynthetic cost of transcript sequences and maximising the efficiency with which those transcript sequences can be translated into protein.
2. Variation in tRNA content between species creates a corresponding variation in cost-efficiency trade-off that constrains the ability of genes in some species to be both low cost and translationally efficient.
3. Selection-driven cost-efficiency optimisation can explain a significant proportion of variation in the molecular evolution rate between genes within a species.

Importantly, these findings also explain why rates of synonymous and non-synonymous mutations in genes are correlated, and thus explains a long-standing unsolved phenomenon in evolutionary biology. Taken together these findings provide significant novel insight into the economic principles of gene sequence evolution in bacteria.

### 4.1.1 Authors

Emily A. Seward & Steven Kelly

### 4.1.2 Author Contributions

SK and EAS conceived the study, EAS conducted the analysis, EAS and SK wrote the manuscript. Both authors read and approved the final manuscript.

### 4.1.3 Abstract

Due to genetic redundancy, multiple synonymous codons can code for the same amino acid. However, synonymous codons are not used equally and this biased codon use varies between different organisms. Through analysis of 1,320 bacterial genomes we show that bacteria are under genome-wide selection to reduce resource allocation to mRNA production. This is achieved by simultaneously decreasing transcript biosynthetic cost and increasing transcript translational efficiency, with highly expressed genes under the greatest selection. We show that tRNA gene copy number alters the cost-efficiency trade-off of synonymous codons such that for many species it is difficult to both minimise transcript cost and maximise transcript translational efficiency to an equal extent. Finally, we show that genes highly optimised to reduce cost and increase efficiency show reduced rates of synonymous and non-synonymous mutation. This provides a simple mechanistic explanation for variation in evolutionary rate between genes that depends on selection-driven cost-efficiency optimisation of the transcript. These findings reveal how optimisation of resource allocation to mRNA synthesis is a critical factor that determines both the evolution and composition of genes.

## 4.2 Introduction

Production of proteins is a primary consumer of cell resources (Farmer and Jones 1976). It requires allocation of cellular resources to production of RNA sequences as well as allocation of resources to production of nascent polypeptide chains. Whilst a protein's amino acid sequence is functionally constrained, redundancy in the genetic code means that multiple nucleotide sequences can code for the same protein. Since the biosynthetic cost and translational efficiency of synonymous codons varies, biased use of synonymous codons makes it possible to reduce the expenditure of cellular resources on mRNA production without altering the encoded protein sequence. Thus, it is possible to reduce resource allocation to protein synthesis without altering the encoded protein or affecting protein abundance. This is done by reducing resource allocation to transcript sequences or by increasing the efficiency with which those transcripts can be translated into protein. Consistent with this, it has been demonstrated that natural selection acts both to reduce biosynthetic cost of RNA sequences (Chen et al. 2016; Seward and Kelly 2016), and to increase the efficiency with which those RNA sequences can template the encoded polypeptide chain (Precup and Parker 1987; Sørensen, Kurland, and Pedersen 1989; Akashi 1994; Rocha 2004; Horn 2008; Shah and Gilchrist 2011; Hu et al. 2013). It should be noted here that biased patterns of codon use are also influenced by other factors not associated with cellular resource allocation. These factors include RNA structural constraints to facilitate thermal adaptation (Lao and Forsdyke 2000; Paz et al. 2004), RNA sequence constraints to preserve splice sites (Eskesen, Eskesen, and Ruvinsky 2004), and translational constraints to ensure accurate protein folding (Zhang et al. 2010; Novoa and Ribas de Pouplana 2012). These functional constraints are independent of resource allocation and are specific to individual sites or sets of sites within genes.

Cells employ different strategies to decode synonymous codons (Grosjean, de Crécy-Lagard, and Marck 2010). These strategies make use of 'wobble' base pairing between the

3rd base of the codon and the 1st base of the anticodon to facilitate translation of all 61 sense codons using a reduced set of tRNAs. As the translational efficiency of a codon is a function of the number of tRNAs that can translate that codon, and as different species encode different subsets of tRNA genes, the same codon is not necessarily equally translationally efficient in all species. In contrast, the biosynthetic cost of a codon of RNA is determined by the number and type of atoms contained within that codon and the number of high energy phosphate bonds required for their assembly. As translational efficiency varies between species but biosynthetic cost does not, it was hypothesised that this must create a corresponding variation in the codon cost-efficiency trade-off between species. For example biosynthetically cheap codons might be translationally efficient in one species but inefficient in another. It was further hypothesised that variation in the codon cost-efficiency trade-off would limit the extent to which a transcript could be optimised to be both biosynthetically inexpensive and translationally efficient.

Here, we show that natural selection acts genome-wide to reduce cellular resource allocation to mRNA synthesis by solving the multi-objective optimisation problem of minimising transcript biosynthetic cost whilst simultaneously maximising transcript translational efficiency. We show that this optimisation is achieved irrespective of the codon cost-efficiency trade-off of a species, and that the extent to which resource allocation is optimised is a function of the production demand of that gene. Finally, we reveal that selection-driven optimisation of resource allocation provides a novel mechanistic explanation for differences in evolutionary rates between genes, and for the previously unexplained correlation in synonymous and non-synonymous mutation rates of genes.

## 4.3 Results

### 4.3.1 55% of bacteria exhibit a significant trade-off between codon biosynthetic cost and translational efficiency

The biosynthetic cost of a codon is determined by the number and type of atoms contained within the codon, and the number of high energy phosphate bonds required for their assembly. Natural selection acting on biosynthetic cost, either in terms of nitrogen atoms (Seward and Kelly 2016) or energetic requirements (Chen et al. 2016), has been shown to play a role in promoting biased patterns of synonymous codon use. As the energy and nitrogen cost of a codon correlate almost perfectly (Fig. 4.01A), it is not possible to distinguish which factor is responsible for biased patterns of codon use in the absence of additional information about the biology of the organism in question. However, given the near perfect correlation, analysis of selection acting on overall codon biosynthetic cost can be approximated by analysis of either measure.

tRNA gene copy number varies between species resulting in a corresponding variation in the relative translational efficiency of their associated codons (dos Reis, Wernisch, and Savva 2003; Grosjean, de Crécy-Lagard, and Marck 2010). As the biosynthetic cost of a given codon is invariant, the relationship between codon biosynthetic cost and codon translational efficiency (referred to from here on as the codon cost-efficiency trade-off) must therefore vary between species. For example, a hypothetical species encoding a full complement of tRNAs, each present as a single copy, would have a negative correlation between cost and efficiency (Fig. 4.01B). In contrast, a hypothetical species that employed tRNA sparing strategy 1 (no ANN tRNAs) or strategy 2 (no ANN or CNN tRNAs) (Grosjean, de Crécy-Lagard, and Marck 2010), would show a positive (Fig. 4.01C) or no (Fig. 4.01D) correlation between cost and efficiency. Therefore, a broad range of codon cost-efficiency trade-offs is possible and the gradient of this trade-off is dependent on the tRNA gene copy number of a given species.

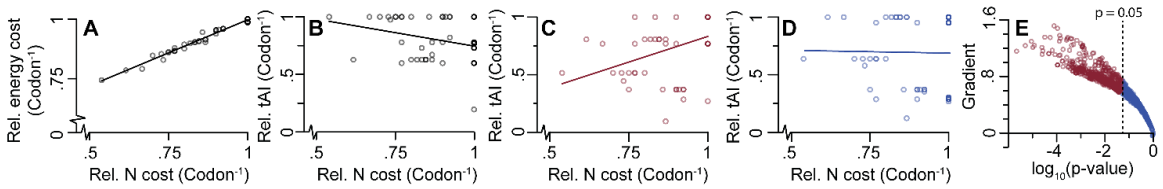



**Figure 4.01. Different tRNA sparing strategies alter a species' codon cost-efficiency trade-off.** (*A*) Codon nitrogen cost (N cost) correlates almost perfectly with codon energetic cost ($p < 0.05$, $y = 0.6x + 0.44$, $R^2 = 0.98$). (*B*) A full complement of tRNAs has a negative correlation between codon biosynthetic cost and translational efficiency (tAI) ($p< 0.05$, $y = -0.5x + 1.21$, $R^2 = 0.10$). (*C*) tRNA sparing strategy 1 (NNU codons translated by GNN anticodons) has a positive correlation between codon biosynthetic cost and translational efficiency ($p<0.05$, $y = 0.9x – 0.06$, $R^2 = 0.18$). (*D*) tRNA sparing strategy 2 (strategy 1 + NNG codons translated by UNN anticodons) has no significant correlation between codon biosynthetic cost and translational efficiency ($p > 0.05$, $y = 0.74$, $R^2 = 0$). (*E*) The 1,320 bacterial species in this analysis can be categorised into those with a significant codon cost-efficiency trade-off ($p < 0.05$, red) and those with no trade-off ($p>0.05$, blue). The y-axis is the gradient of the line of best fit between codon biosynthetic cost and translational efficiency.

Few bacterial species strictly adhere to a single sparing strategy for all synonymous codon groups (e.g. *Escherichia coli* uses strategy 2 for decoding alanine but strategy 1 for decoding glycine), and thus it is anticipated that a continuum of gradients in trade-off between cost and efficiency is possible. To assess this, the codon cost-efficiency trade-off was calculated for 1,320 bacterial species representing 730 different genera (Fig. 4.01E). Approximately 55% of species exhibited a significant codon cost-efficiency trade-off (Fig. 4.01E, $p < 0.05$). Species with a significant negative correlation between cost and efficiency were not observed; instead, the remaining species exhibited non-significant correlations between codon cost and efficiency (Fig. 4.01E, blue dots, ~45% of species). Thus for approximately half of the bacterial species in this analysis, there is a significant codon cost-efficiency trade-off whereby resource allocation to mRNA can be reduced by decreasing biosynthetic cost or increasing translational efficiency but not both simultaneously. This is because the synonymous codons that are most translationally efficient in these species are in general those that consume the most resources for biosynthesis.

### 4.3.2 Selection acting to minimise biosynthetic cost and maximise translational efficiency of transcript sequences is independent of codon cost-efficiency trade-off

Given that the codon cost-efficiency trade-off varied between species, an analysis was conducted to determine whether this trade-off was associated with concomitant differences in the strength of selection acting on the cost and efficiency of transcript sequences. For each species, genome-wide values for selection on transcript translational efficiency [$S_t$] and selection on transcript biosynthetic cost [$S_c$] were inferred. This was done using the complete set of open reading frames and tRNAs encoded in that species' genome using the SK model (Seward and Kelly 2016) implemented using CodonMuSe (see Methods). There was no significant difference in the mean or range of values for $S_c$ between the groups of species that did or did not exhibit a significant codon cost-efficiency trade-off. Instead, irrespective of codon cost-efficiency trade-off 91% of species had negative $S_c$ values (mean $S_c$ = -0.08), indicating a genome-wide selective pressure to minimise the biosynthetic cost of transcript

sequences through synonymous codon use (Fig. 4.02). This observation extends previous studies that revealed analogous effects when nitrogen or energy were limited (Chen et al. 2016; Seward and Kelly 2016). Thus irrespective of codon cost-efficiency trade-off, selection acting on codon biosynthetic cost is an important factor promoting biased patterns of codon use in bacteria.

Similarly, 78% of species had positive values for $S_t$ (mean $S_t$ = 0.1), indicating a genome wide selective pressure to increase the translational efficiency of transcript sequences (Fig. 4.02). However, species that had a significant codon cost-efficiency trade-off (Fig. 4.02A) had $S_t$ values that were significantly lower than those that had no trade-off (Fig. 4.02B, $p < 0.05$, t-test). Moreover, there was an increased variance in the observed values for both $S_c$ and $S_t$ for species that exhibited a significant codon cost-efficiency trade-off compared to those species that did not (Fig. 4.02A and 4.02B). Thus, in general the majority of species experience selection to minimise transcript biosynthetic cost while simultaneously maximising transcript translational efficiency. However, the nature of a species' codon cost-efficiency trade-off restricts a transcript's ability to be both cheap and translationally efficient, thereby limiting the extent to which resource allocation to transcript sequences can be minimised.

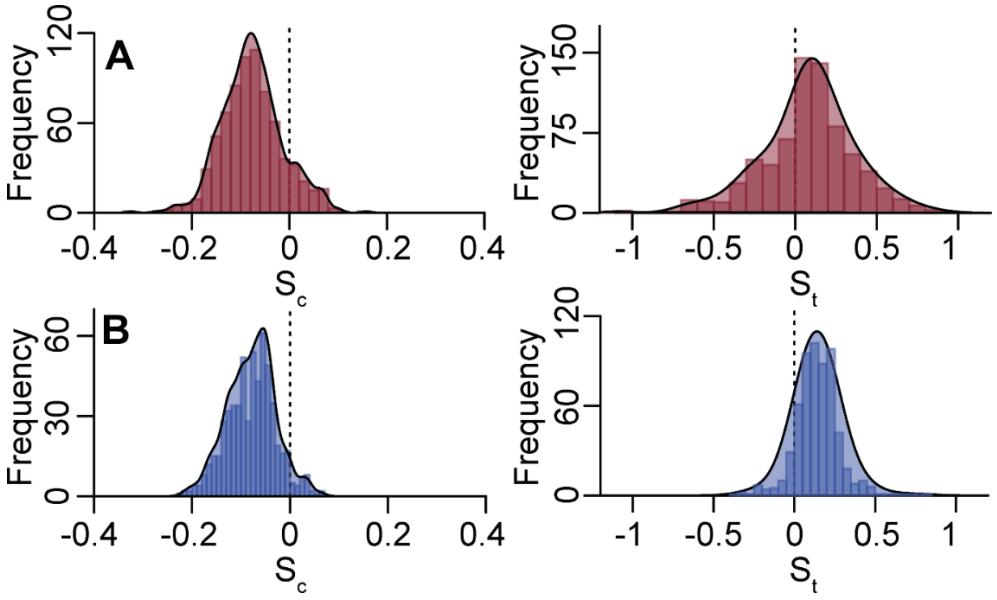



**Figure 4.02. Bacterial genomes show selection to minimise nucleotide cost ($-S_c$) and maximise translational efficiency ($+S_t$).** (*A*) Genome-wide $S_c$ and $S_t$ values for the 726 species with a codon cost-efficiency trade-off. (*B*) Genome-wide $S_c$ and $S_t$ values for the 594 species with no codon cost-efficiency trade-off.

### 4.3.3 Genes that experience the strongest selection for increased transcript translational efficiency are also under the strongest selection to minimise biosynthetic cost

Given that the majority of species exhibited selection to minimise cost and maximise translational efficiency at the genome-wide level, the extent to which this was also seen at the level of an individual gene within species was determined. Here, the strength of selection acting on transcript translational efficiency and strength of selection on transcript biosynthetic cost were inferred for each individual gene in each species. The relationship between these selection coefficients was then compared for each species. For example, in species with no codon cost-efficiency trade-off, such as *Escherichia coli,* there is a significant negative correlation between $S_c$ and $S_t$ (Fig. 4.03A). Here, the genes that experienced the greatest selection to maximise efficiency are those that experienced the greatest selection to minimise biosynthetic cost. The same phenomenon was also observed for species that exhibit significant codon cost-efficiency trade-offs, such as *Lactobacillus amylophilus* (Fig. 4.03B). Overall, significant correlations between selection coefficients for individual genes were observed for 91% of species ($p < 0.05$, Fig. 4.03C). Therefore irrespective of the extent of the codon cost-efficiency trade-off, selection is performing multi-objective optimisation of transcript sequences to reduce their biosynthetic cost while increasing their translational efficiency and thereby reducing resource allocation to mRNA production.

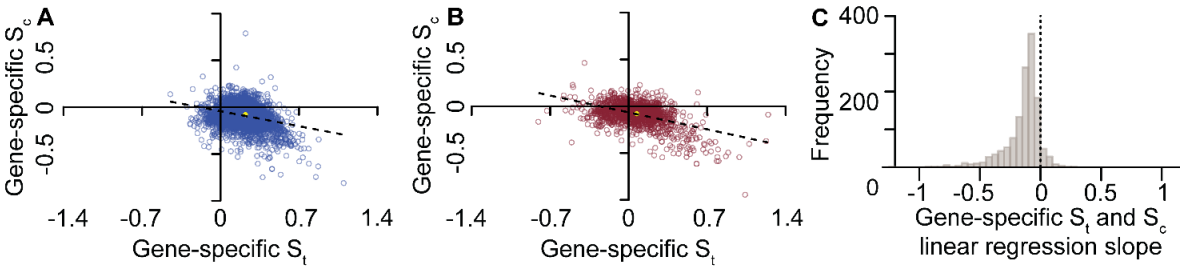



**Figure 4.03. The genes under the strongest selection for translational efficiency (+$S_t$) are also under the strongest selection to minimise nucleotide cost (-$S_c$).** Scatterplots of gene-specific $S_t$ and $S_c$ values for (*A*) *Escherichia coli (B*) *Lactobacillus amylophilus*. In both cases the line of best fit is shown and the yellow dot is the genome-wide best-fit value for each species. (*C*) Histogram of the slope between $S_c$ and $S_t$ for individual genes for each of the 1,320 bacterial species in this analysis.

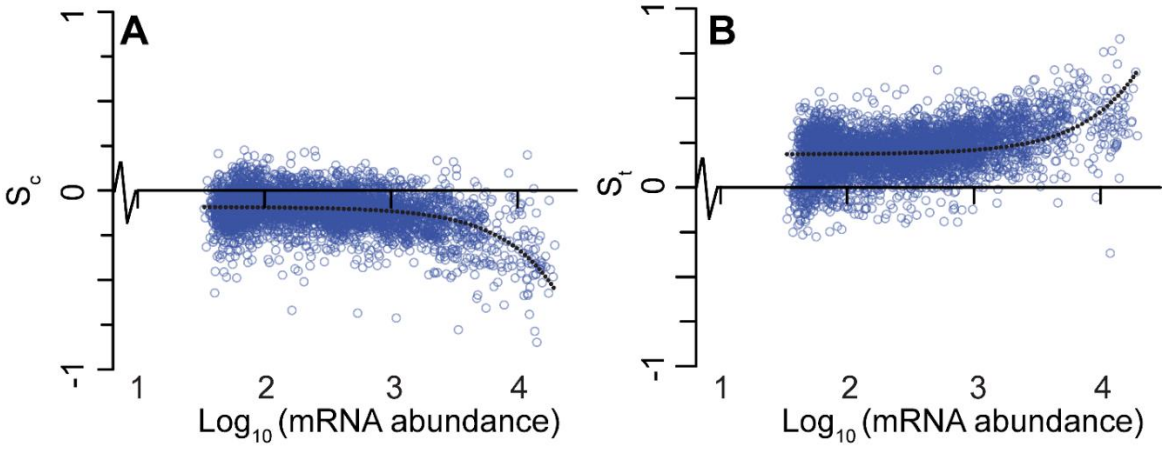



**Figure 4.04. Selection acts in proportion to mRNA abundance to decrease codon biosynthetic cost and increase codon translational efficiency in *Escherichia coli*.** (*A*) There is a negative correlation between selection acting on codon biosynthetic cost ($S_c$) and mRNA abundance. The line of best fit has an $R^2$ value of 0.18. (*B*) There is a positive correlation between selection acting to increase codon translational efficiency ($S_t$) and gene expression. The line of best fit has an $R^2$ value of 0.13.

As the most highly expressed genes in a cell comprise the largest proportion of cellular RNA, the strength of selection experienced by a gene is thought to be dependent on the mRNA abundance of that gene (Drummond and Wilke 2009; Ran and Higgs 2012). In agreement with this, evaluation of the relative mRNA abundance of genes in *E. coli* revealed that the most highly expressed genes exhibited the greatest selection to minimise transcript biosynthetic cost (Fig. 4.04A) whilst also showing the strongest selection to maximise transcript translational efficiency (Fig. 4.04B). Thus, selection acts in proportion to relative mRNA abundance to perform multi-objective optimisation of codon bias to minimise resource allocation to transcript sequences through production of low cost, high efficiency transcripts.

### 4.3.4 Sequence optimisation for cost and efficiency constrains molecular evolution rate

Given that codon choice has been shown to provide a selective advantage per codon per generation (Brandis and Hughes 2016), it was hypothesised that the extent to which a transcript is jointly optimised for codon cost and efficiency would constrain the rate at which the underlying gene sequence can evolve. Specifically, the more highly optimised a transcript is for both biosynthetic cost and translational efficiency, the higher the proportion of spontaneous mutations that would reduce the cost-efficiency optimality of the transcript sequence. Therefore, spontaneous mutations in highly optimised genes would be more likely to be disadvantageous than spontaneous mutations in less optimised genes. As deleterious mutations are lost more rapidly from the population than neutral mutations, the more highly optimised a gene sequence is, the lower its apparent evolutionary rate should be.

To test this hypothesis the complete set of gene sequences from *E.coli* was subject to stochastic *in silico* mutagenesis and the proportion of single nucleotide mutations that resulted in reduced transcript cost-efficiency optimality was evaluated. As expected, the proportion of deleterious mutations increased linearly with transcript sequence optimality. This effect was seen for both synonymous (Fig. 4.05A) and non-synonymous mutations (Fig.

4.05B). The effect in non-synonymous mutations is seen because a single base mutation from an optimal codon encoding one amino acid is unlikely to arrive at an equally optimal or better codon encoding any other amino acid. Thus as expected, the more optimal a codon is, the less likely a spontaneous mutation will result in a codon with higher optimality irrespective of whether that codon encodes the same amino acid.

The extent to which transcript sequences in *E. coli* were jointly cost-efficiency optimised was compared to the synonymous ($K_s$) and non-synonymous ($K_a$) mutation rate of that gene estimated from comparison with *Salmonella enterica*. Consistent with the hypothesis, the rate of both synonymous ($K_s$ Fig. 4.05C) and non-synonymous ($K_a$ Fig. 4.05D) changes were directly proportional to the extent to which the gene sequence had been optimised by natural selection for low biosynthetic cost and high translational efficiency (Fig. 4.05A and B). To determine if this relationship was also observed for other bacteria an additional 177 species-pairs were analysed (Fig. 4.05E). Of these species pairs, 66% were consistent with the observation for *E. coli* and *S. enterica*, such that variance in selection-driven gene sequence optimisation explained on average 8% of variance in $K_s$ rates between genes (Fig. 4.05E). Thus, the extent to which transcript sequences are jointly optimised for cost and efficiency is sufficient to explain a substantial component of variation in molecular evolutionary rate between genes within a species. Moreover, selection-driven cost-efficiency optimality is also sufficient to explain the correlation in rates of synonymous and non-synonymous mutations.

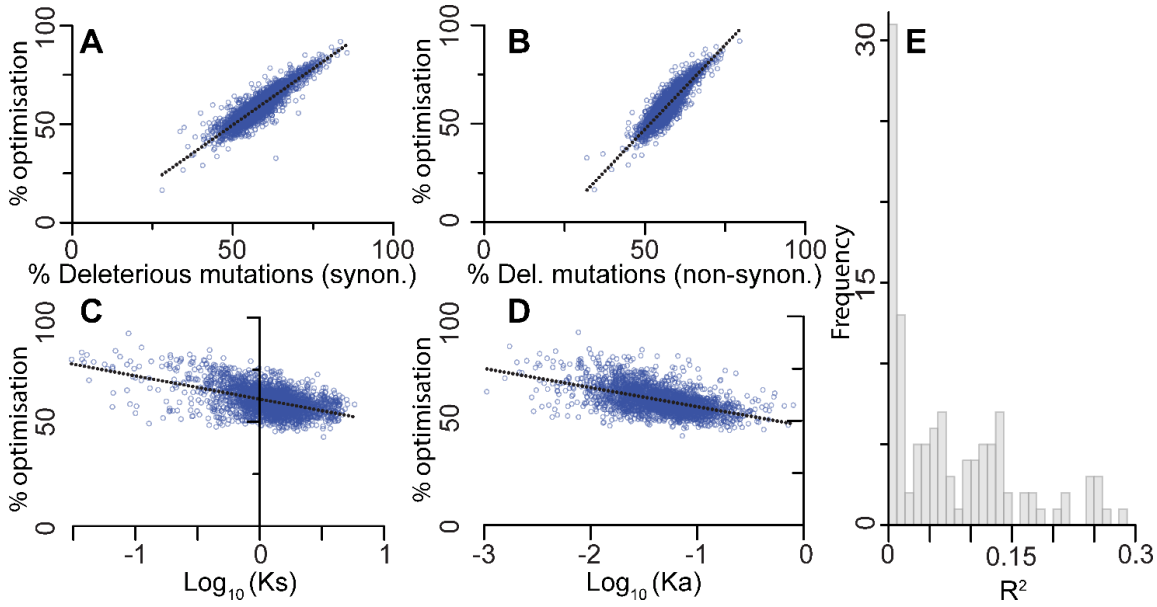



**Figure 4.05. Selection-driven optimisation of resource allocation is a critical factor that determines molecular evolution rate.** Highly cost-efficiency optimised genes have a higher proportion of deleterious (*A*) synonymous (y = 1.15x - 8, $R^2$ = 0.81) and (*B*) non-synonymous (y = 1.71x -38, $R^2$ = 0.78) mutations. Orthologous genes in *Escherichia coli* and *Salmonella enterica* show a negative correlation between sequence cost-efficiency optimisation and the rate of (*C*) synonymous mutations ($K_s$) (y = -11x + 61, $R^2$ = 0.26) and (*D*) non-synonymous mutation ($K_a$) (y = -9x + 48, $R^2$ = 0.28). (*E*) histogram of proportion of gene evolutionary rate explained by selection-driven cost-efficiency optimisation of transcript sequences.

## 4.4 Discussion

Codon use is biased across the tree-of-life, with patterns of bias varying both between species and between genes within the same species. Here we demonstrate that variation in tRNA content between species creates a corresponding variation in the codon cost-efficiency trade-off whereby codons that cost the least to biosynthesise are not equally translationally efficient in all species. We show that irrespective of the codon cost-efficiency trade-off, natural selection performs multi-objective gene sequence optimisation so that transcript sequences are optimised to be both low cost and highly translationally efficient, and that the nature of this trade-off constrains the extent of the solution. We demonstrate that this multi-objective optimisation is dependent on mRNA abundance, such that the transcripts that comprise the largest proportion of cellular mRNA are those that experience the strongest selection to be both low cost and high efficiency. Finally, we show that the extent to which a gene sequence is jointly optimised for reduced transcript cost and enhanced translational efficiency is sufficient to explain the variation in the molecular sequence rate of genes. Furthermore, it is sufficient to explain the previously unexplained phenomenon that the rate of synonymous and non-synonymous mutation for a gene is correlated (Sharp 1991).

Differences in molecular evolution rates between species are thought to be mainly due to differences in organism generation-time (Weller and Wu 2015). However, differences in evolutionary rates between genes in the same species lack a complete mechanistic explanation. Prior to the study presented here, it was known that functional constraints of the encoded protein sequence contribute to the constraint of the rate of non-synonymous changes (Zuckerkandl 1976). It had also been observed that mRNA abundance and patterns of codon bias correlated with the evolutionary rate of genes (Sharp and Li 1987; Drummond, Raval, and Wilke 2006), and that rates of synonymous and non-synonymous changes were correlated (Sharp 1991). The study presented here unifies these prior observations and provides a novel mechanistic explanation for both variation and correlation in molecular

evolution rates of genes. It proposes that stochastic mutations in gene sequences are more likely to result in deleterious alleles in proportion to the extent to which that gene sequence has been jointly optimised by natural selection for reduced transcript biosynthetic cost and enhanced translational efficiency.

The novel mechanism provided here also explains the relationship between mRNA abundance and gene evolutionary rate. Specifically, functional constraints on protein abundance stipulate the quantity of mRNA required to produce that protein. The more mRNA that is required, the larger the percentage of total cellular resources that are invested within that transcript. The mechanism simply entails that the more transcript that is present, the stronger the selective pressure will be to reduce the cellular resources committed to that transcript. Importantly, minimising these resources can be achieved both by using codons that require fewer resources for their biosynthesis, and by utilising translationally efficient codons that increase the protein to transcript ratio and therefore reduce the amount of transcript required to produce the same amount of protein. Overall, this study reveals how the economics of gene production is a critical factor determining both the evolution and composition of genes.

## 4.5 Methods

### 4.5.1 Data sources

1,320 bacterial genomes were obtained from the NCBI (www.ncbi.nlm.nih.gov). In order to avoid over-sampling of more frequently sequenced genera, the number of species from each genus was restricted to 5 with a maximum of 1 species for each genus. Therefore, the 1,320 species sampled in this study were distributed among 730 different genera. Only genes that were longer than 30 nucleotides, had no in-frame stop codons, and began and ended with start and stop codons respectively were analysed. Each species in this analysis contained a minimum of 500 genes that fit these criteria. Full details of species names, genome accession numbers, strain details and selection coefficients are provided in Table S4.01.

### 4.5.2 Evaluation of translational efficiency (tAI)

To obtain the number of tRNA genes in each genome, tRNAscan was run on each of the 1,320 bacterial genomes (Schattner, Brooks, and Lowe 2005). This (current) version of tRNAscan is unable to distinguish between tRNA-Met and tRNA-Ile with the anticodon CAT. Thus tRNA-Ile(CAT), while present, is not detected in any of the genomes. To compensate for this a single copy of tRNA-Ile with the anticodon CAT was added to the tRNA counts for each species if more than one tRNA-Met(CAT) was found. The tRNA adaptation index (tAI) (dos Reis, Wernisch, and Savva 2003), which considers both the tRNA gene copy number and wobble-base pairing when calculating the translational efficiency of a codon, was evaluated using the optimised $s_{ij}$ values for bacteria obtained by Tuller et al (Sabi and Tuller 2014) and the equation developed by dos Reis et al (dos Reis, Savva, and Wernisch 2004). $s_{uu}$ was set to 0.7 as proposed by Navon et al (Navon and Pilpel 2011) and $s_{uc}$ was set to 0.95 as $U_{34}$ has been shown to have weak codon-anticodon coupling with cytosine (Näsvall, Chen, and Björk 2004). Each species in this analysis was able to translate all codons, was not missing key tRNAs and did not require unusual tRNA-modifications.

### 4.5.3 Calculation of relative codon cost and efficiency

Codon biosynthetic cost and translational efficiency were calculated relative to other synonymous codons such that the synonymous codon with the greatest value had a relative cost or efficiency of 1. For example, the nitrogen cost of GCC is 11 atoms. The most expensive synonymous codon is GCG/GCA (13 atoms). Therefore the relative cost of GCC is 11/13 = 0.85. The same evaluation was done to calculate codon translational efficiency.

### 4.5.4 CodonMuSe: A fast and efficient algorithm for evaluating drivers of codon usage bias

The SK model (Seward and Kelly 2016) was used to infer the joint contribution of mutation bias, selection acting on codon biosynthetic cost and selection acting on codon translational efficiency to biased synonymous codon use. To facilitate the large scale comparative

application of this model a rapid, stand-alone version was implemented in python. The algorithm, instructions for use, and example files are available for download at https://github.com/easeward/CodonMuSe. For each species, the values of $M_b$, $S_c$ and $S_t$ were inferred using the complete set of protein coding genes and the tRNA copy number inferred using tRNAscan.

### 4.5.5 Comparing selection acting on codon bias and transcript abundance levels

Transcriptome data for *E. coli* str. K-12 MG1655 were downloaded from NCBI (Cho et al. 2009). The three biological replicates for the logarithmic growth phase were available, however the third replicate was inconsistent with the first two and so was excluded from this analysis. As each gene had multiple probes, the average probe value for each gene was taken. The three-parameter CodonMuSe model using the value for $M_b$ estimated from a genome-wide analysis was run for each of the 4,099 genes in *E. coli* individually, and thus values for $S_c$ and $S_t$ were obtained for each gene. The values for these selection coefficients were plotted against relative mRNA abundance data described above (Cho et al. 2009).

### 4.5.6 Calculating the extent to which gene sequences were jointly optimised for cost and efficiency

The extent to which transcript sequences were jointly optimised for both biosynthetic cost and translational efficiency was approximated by the distance of a given gene to the cost-efficiency Pareto frontier for that amino acid sequence (d1, Fig. S4.01). The cost-efficiency Pareto frontier consists of sequences that are optimised for maximal transcript translational efficiency and minimal transcript biosynthetic cost such that a gene that is 100% optimised lies on the frontier (red frontier, Fig. S4.01). Genes that are less than 100% optimised occupy the space between the cost-efficiency Pareto frontier (red frontier) and the opposite frontier (blue frontier, minimising transcript efficiency and maximising cost) for that amino acid sequence (Fig. S4.01). The percent optimality of the coding sequence is evaluated as $\left(\frac{d4}{d1+d4}\right) * 100$ (Fig. S4.01).

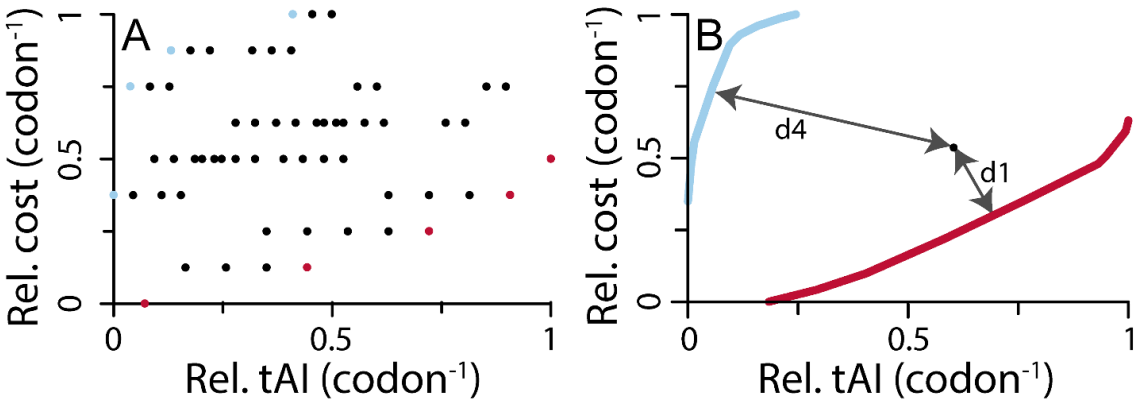



**Figure S4.01. Example cost-efficiency Pareto frontier for a short amino acid sequence.** (*A*) Scatter plot of the 64 possible coding sequences encoding the amino acid sequence MTGCD. Red dots indicate coding sequences that are positioned on the best cost-efficiency Pareto frontier (the least expensive, most translationally efficient sequences possible). Blue dots indicate coding sequences that are positioned on the worst cost-efficiency Pareto frontier (the most expensive, least translationally efficient sequences possible). (*B*) Evaluating the cost-efficiency optimality of a coding sequence. d1 is the minimum distance between a given coding sequence and the best cost-efficiency Pareto frontier (red) for that amino acid sequence. d4 is the minimum distance of the same gene to the worse cost-efficiency Pareto frontier for that amino acid sequence (blue). The percent optimality of the coding sequence is evaluated as $\left(\frac{d4}{d1+d4}\right) * 100$.

### 4.5.7 Calculation of molecular evolution rates

Molecular evolutionary rates ($K_a$ and $K_s$ values) were calculated for orthologous genes in *E. coli* and *S. enterica*. 2,468 single-copy orthologous genes were identified for *E. coli* and *S. enterica* using OrthoFinder v1.1.4 (Emms and Kelly 2015). These sequences were aligned at the amino acid level using MergeAlign (Collingridge and Kelly 2012) and this alignment was then rethreaded with the coding sequences to create codon-level nucleotide alignments. Only aligned sequences longer than 30 nucleotides with less than 10% gaps were used. Gapped regions were removed and KaKs_Calculator 2.0 (Wang et al. 2010) was run using the GMYN model to evaluate $K_a$ and $K_s$ values for each pair of aligned nucleotide sequences. As the molecular evolution rates represent the average of the mutation rates of the gene-pair since they last shared a common ancestor, these rates were compared to the average optimality of the same gene-pair in both species.

The same analysis was conducted on 1,066 additional pairs of species obtained by exhaustive pairwise comparison of all species that were within the same genus. These 1,066 pairwise comparisons were filtered to remove those with $K_s$ saturation (i.e. mean $K_s > 1$) and fewer than 1,000 genes. This filtered set contained 177 species pairs.

**Supplementary Table S4.01** Full details of species names, genome accession numbers, strain details and selection coefficients can be found online at:

https://github.com/easeward/Thesis.

# **Chapter 5:** General Discussion

## 5.1 Chapter Introduction

The overall objective of this thesis was to determine how resource limitation affects sequence evolution. This was done by looking at differences in nucleotide sequences between extant species and linking those variations with the underlying biological processes responsible for producing them. Furthermore a mathematical framework was developed that is able to evaluate the impact of multiple factors such as resource limitation on codon bias. Since each chapter already has its own discussion, I will first summarise the findings from Chapter 2 before providing a general overview of the results obtained in the remaining chapters within a broader context.

In Chapter 2, I reported the discovery of the new species *Phytomonas oxycareni,* whose position as the earliest diverging species of *Phytomonas* discovered to date provides new insight into the conserved ancestral biology of the group. For example, the discovery that *Phytomonas oxycareni* inhabits the salivary gland lumen and the cells of the salivary gland indicates that the ability to translocate through the gland wall and survive in host cells is an ancestral trait within the genus. Furthermore, the long slender promastigote morphology observed is similar to other *Phytomonas* species from similar life-cycle stages. This indicates that it is a common morphology for *Phytomonas* in the insect host salivary gland. By sampling over consecutive years, it was shown that the parasitic infection was persistent over multiple generations. Therefore, whilst the discovery of *Phytomonas oxycareni* was not directly useful for the comparative analyses conducted in the rest of my thesis, it has provided useful insight into a relatively unknown genus.

## 5.2 General Discussion

### 5.2.1 Resource availability exerts a selective pressure on genome evolution

To elucidate the role of nitrogen availability in altering codon bias and genome composition I conducted a study on eukaryotic and bacterial single-celled parasites. Considering both environmental and dietary nitrogen availability, selection was found to reduce nitrogen use in species that were nitrogen-limited. This analysis was then extended across the bacterial kingdom to show that selection generally acts to reduce transcript cost (though without additional species-specific information it was not possible to tell whether this was due to nitrogen or energy limitation). Since nitrogen and/or energy limitation have been shown to alter genome evolution, it is likely that other types of resource limitation not studied in this work might also alter genome evolution. An interesting avenue of future research would be to expand the analysis in this thesis to include additional types of resource limitation. For example it is important to consider the C:N ratio as well as total nitrogen availability when analysing their contribution to genome evolution (Francois et al. 2016). Furthermore, some elements such as phosphorous are found in equal quantities in all nucleotides and so would not be expected to exert a selective pressure on codon bias. Therefore, resource limitation has been shown to alter genome evolution, but different types of resource limitation are likely to exert different selective pressures.

Dietary nitrogen availability was shown to explain differences in codon bias between related species. An interesting extension of this work would be to determine whether the evolution of nitrogen fixation resulted in the same effect as altered dietary nitrogen availability. This could be analysed by comparing the multiple examples of species that have gained/lost the ability to fix nitrogen (Latysheva et al. 2012). However, such an analysis would be complicated by the strong correlation between codon nitrogen and energy demand. Due to the high energetic demands of nitrogen fixation, the impact of selection acting on codon nitrogen use might be masked by selection acting on codon energy use. One way to

untangle the two would be to compare the impact of nitrogen fixation on genome evolution in species that are readily able to generate energy (e.g. cyanobacteria) compared to relatively energy-limited species (non-photosynthetic bacteria). I would anticipate that photosynthetic nitrogen-fixing bacteria would be more nitrogen limited than energy limited whereas non-photosynthetic bacteria are more likely to be energy limited due to the high cost of fixing nitrogen. Therefore, though dietary nitrogen availability has been shown to alter codon bias and genome evolution, the impact of nitrogen fixation remains to be determined. Such an analysis would improve our understanding of the trade-off between energy and nitrogen limitation as well as the role of nitrogen availability in determining genome evolution.

### 5.2.2 Selection is stronger for genes that are more highly expressed

Genes that are more highly expressed were found to be under stronger selection to reduce biosynthetic cost and increase translational efficiency. In addition to how often a transcript is being transcribed, the timing of when that transcription takes place during the cell or life cycle of an organism could alter the selection that it experiences. For example, cellular resource availability and transcript expression levels are not constant over time. This might manifest as differences in codon bias between sequences that appear to have the same mRNA abundance but are expressed in different cellular conditions. For example, iso-accepting tRNA charging is sensitive to amino-acid starvation (Elf et al. 2003). This means that under varying conditions of amino-acid starvation, the translational efficiency of a given sequence can vary. This is seen under conditions of leucine starvation when the total rate of protein synthesis slows down as the major $tRNA^{Leu}$ is less efficiently charged. However, synthesis of proteins required for responding to leucine starvation is unperturbed as the mRNA is decoded by tRNAs which are insensitive to leucine starvation. Furthermore, the impact of different cellular conditions on sequence evolution could be tested by analysing transcripts that are expressed during periods of acute resource limitation and comparing them to the rest of the genome. One potential extreme example of this comes from *Plasmodium* species.

During *Plasmodium* male gametogenesis, the genome undergoes three rounds of replication concurrent with the formation of eight flagella in just 10 minutes (Sinden et al. 2010). Therefore not only is dGTP required for genome replication but also GTP is consumed during microtubule construction (Valiron, Caudron, and Job 2002). In an environment that is depleted of GTP, spontaneous replication errors are more likely to introduce AT mutations (Buckland et al. 2014; Schmidt et al. 2017). Therefore, this sudden depletion of GTP may in part explain the extremely GC-poor genomes of most *Plasmodium* species. Furthermore, we would expect genes expressed during this period of extreme GTP depletion to experience strong selection to reduce GTP requirements and experience higher rates of transcription-associated mutagenesis (Drummond and Wilke 2009; Park, Qian, and Zhang 2012). This would manifest as a significantly lower GC-content in genes expressed during male gametogenesis compared to genome-wide GC-content. Furthermore, female gametogenesis does not lead to extreme levels of GTP depletion as no flagella are produced and there is no genome replication (Delves et al. 2013). Therefore, although female gametogenesis is triggered at the same time and in the same conditions as male gametogenesis, we would not expect genes expressed during female gametogenesis to have a GC-content that is significantly lower than genome-wide GC-content (Kuehn and Pradel 2010). Therefore, local conditions at the time of transcription, combined with how often a gene is transcribed overall can contribute to differences in sequence evolution. These differences may explain some of the variation in sequence codon bias observed within a species (Chapters 3 and 4).

### 5.2.3 Just three factors can explain the majority of genome-wide codon bias

The mathematical model developed in Chapter 3 (the SK model) was able to explain genome-wide patterns of codon bias with >90% accuracy in parasitic species. The model builds on the selection-mutation-drift theory of synonymous codon use and uses a probabilistic framework to incorporate multiple factors such as translational efficiency, mutation bias and selection acting on biosynthetic cost (Bulmer 1991; dos Reis, Wernisch,

and Savva 2003; Lynch 2007). The model therefore consists of an extendable modular set of equations that can evaluate the impact of multiple factors on synonymous codon use. It was not, however, designed to explain the differences seen in codon bias along a given sequence. Since codon bias can vary along a single mRNA sequence, it would be interesting to expand the mathematical model to take into account neighbouring codons as well as a codon's position within a gene. For example it has recently been proposed that the genetic code should be considered as a triplet of triplets (Chevance and Hughes 2017). This is because the efficiency with which a particular codon can be translated is influenced by the nature of the immediately adjacent codons. Different neighbours make different contributions to the stacking energy of codon-anticodon recognition and thus alter the efficiency of translation. Furthermore, correlations between pairs of neighbouring bases in mitochondrial genomes were found to be strongly influenced by context-dependent mutation (Jia and Higgs 2008). Thus neighbouring codons can play an important role in determining codon use either through translational selection or mutational processes. Codon use also depends on where the codon is found within the gene (Agashe et al. 2013; Hockenberry et al. 2014). Previous work has shown that genes have a 'ramp' of rare codons at the 5' end and it has been proposed that this slows elongation and prevents ribosome traffic (Qin et al. 2004). This ramp of rare codons has been shown to increase protein expression ~14-fold, though this may be due to reduced RNA structure rather than codon rarity (Goodman, Church, and Kosuri 2013; Bentele et al. 2014). In particular the mRNA-structure of the first ~ 16 codons is thought to alter gene expression by changing translational efficiency but not transcriptional efficiency (Boël et al. 2016). Therefore, though a general set of codons might be favoured overall, local context is important for determining codon bias in an individual gene and it would be interesting to explore this in the context of the SK model.

As described in the introduction to this thesis, multiple factors contribute to genome evolution. This thesis focuses on just three of those factors (mutation bias, selection acting

on nucleotide cost and selection acting on translational efficiency). Ideally, it would be possible to extend the results in this thesis to include additional factors such as temperature, conservation of splice sites etc. and integrate all of the factors together to provide a more complete picture. As indicated by the complexity of comparing just two factors (Chapter 4), this is not a simple task. Instead, in this thesis I have focused on trying to control for the impact of additional factors in order to more fully understand the interaction between the three factors that were the focus of my analysis. However, there are undoubtedly additional factors that make a significant contribution to genome evolution that I have not been able to examine in this work. Due to the modular nature of the mathematical model, there is scope to expand it to include such additional factors. However, as with any model, care must be taken when handling highly correlated variables. For example, selection acting on codon nitrogen cost and selection acting on codon energetic cost cannot be incorporated at the same time due to problems of collinearity that can give rise to spurious results (Tu et al. 2005). One way to address this would be to conduct experimental work to examine not only the independent impact of additional factors on genome evolution but also the interplay between multiple factors acting together. For example, high temperatures have been shown to alter nucleotide composition and increase mRNA purine levels (Lao and Forsdyke 2000). Using the SK-model, though an increase in mRNA purine levels would logically be attributed to selection acting to increase the thermal stability of the mRNA, it would be difficult to distinguish from selection acting to increase codon biosynthetic cost (purines are more expensive than pyrimidines). This could be resolved by an *in vivo* long-term experiment that grew the same strain of bacteria in conditions of different temperatures, resource availability or both. Such combinatorial analyses would shed light on the relative importance of, and interaction between, factors contributing to genome evolution but would require a significant input of resources and time.

### 5.2.4 Selection acting on codon cost can be used to infer dietary nitrogen input

The SK model was able to predict the dietary nitrogen availability of a species from analysis of orthologous sequences in related species. This meant that not only was diet shown to affect DNA, but that DNA could be used to predict diet. Additional experimental support for the impact of diet on DNA could be provided by conducting an evolution-in-action type experiment in differing resource availability environments. For example, propagating the same bacterial strain in high- or low-nitrogen availability environments for multiple generations. In nitrogen-limited environments, based on my previous findings, I would expect to see a bias towards the accumulation of mutations that reduced sequence nitrogen requirements. For example, substitutions leading to less nitrogen-rich codons or sequence deletions. In particular, such changes might be found more often in non-coding regions or at sites where they lead to synonymous changes in codon use since those types of mutation are less likely to be deleterious. Unfortunately, this was beyond the scope of the work I could undertake due to the large time-scale required and the limited funding available for such an experiment. Though there are increasing numbers of studies that have conducted similar experiments (Gresham et al. 2010; Dettman et al. 2012; Jezequel et al. 2013; Hong and Gresham 2014), those studies were not specifically looking for differences in nucleotide use as a result of nitrogen-limitation. Therefore they focus on changes in individual genes and normally do not conduct in-depth genome-wide sequencing. Those studies that do conduct genome-wide sequencing make very few data points available and thus do not provide irrefutable experimental evidence for the work reported in this thesis. Therefore, though analysis of DNA can be used to infer diet, it would be interesting to test the impact of diet on DNA *in vivo*.

### 5.2.5 tRNA pools can alter the trade-off between codon cost and efficiency

Different tRNA sparing strategies can alter a species' codon cost-efficiency trade-off. This means that not only can selection act on codon bias but also on tRNA copy number and

diversity to alter the translational efficiency of a sequence. For example, there is a negative correlation between bacterial generation time and tRNA gene copy number (Vieira-Silva and Rocha 2010). One explanation is the coevolution of codon bias and tRNA genes (Higgs and Ran 2008). It would be interesting to further examine the link between codon cost and translational efficiency by studying sequence evolution in response to changing tRNA pools or resource availability. This could be studied by experimenting on bacteriophage, which are even simpler than single-celled organisms and can evolve rapidly. Furthermore, bacteriophage are for the most part entirely dependent on their hosts for replication, gene expression and translation. For example, the majority of phage contain no tRNA genes and have been shown to be under selective pressure to adapt their codon bias to match the translational bias of their hosts (Carbone 2008). These adaptations are most pronounced in abundant structural proteins such as capsid genes and lead to changes in patterns of GC-content in the third codon position (GC3-content) and use of host-preferred codons (Carbone 2008; Lucks et al. 2008). Based on my findings, I would hypothesise that phage infecting hosts in nutrient-limited environments would also be under selective pressure to reduce resource allocation to coding sequences. Furthermore, if the bacteriophage's host tRNA composition was altered I would expect a concomitant change in the phage's codon bias. Therefore having shown *in silico* that changes in tRNA copy numbers can alter the balance between codon cost and translational efficiency and so alter the selective pressures acting on sequence evolution, it would be interesting to confirm this with *in vivo* experimentation.

### 5.2.6 Bacteria are generally under selection to reduce codon cost & increase translational efficiency

Across the bacterial kingdom, ~80% of bacteria were shown to be under selection to reduce codon cost and increase codon translational efficiency. This indicated that cells were minimising resource allocation to transcript sequences by both reducing sequence cost and increasing sequence efficiency. This is a complex optimisation problem, as an inexpensive mRNA sequence might also be very translationally inefficient, necessitating multiple mRNA

copies. Thus, overall cellular optimisation may be increased by encoding a more energetically expensive but translationally efficient mRNA. Although studies have been conducted looking at the trade-off between translational efficiency and accuracy (Zhang 2014), no such comparison had been undertaken to integrate the effects of codon cost and translational efficiency. To address this, the study reported in Chapter 4 examined the impact on sequence evolution of the interaction between selection acting on nucleotide cost and translational efficiency. This was achieved by making use of the large number of bacterial genomes available and performing a high-throughput analysis of the joint effects of selection acting on nucleotide biosynthesis cost and selection acting on translational efficiency in determining codon biases both between species and between genes in the same species. In that chapter, it was shown that variation in tRNA gene copy number between species altered the interaction between codon cost and translational efficiency. However, irrespective of this difference in trade-off, selection acts to reduce resource allocation to mRNA biosynthesis by favouring codons that reduce transcript cost and increase efficiency through complex combinatorial optimisation.

Although the analysis in Chapter 3 was conducted on multiple different cellular constituents, such as mRNA, dsDNA, rRNA and proteins, it did not take into consideration the interactions between them. For example, there is a trade-off between the energy requirements (measured in ATP molecules required for *de novo* synthesis) of codons and their encoded amino acids (Chen et al. 2016). More energetically costly codons (increased GC-content) encode for less energetically demanding amino acids. However, though the majority of nitrogen in a cell is allocated to proteins (~80%) (Sterner and Elser 2002), functional constraints limit the interchangeability of different amino acids within a polypeptide sequence and thus limit the extent to which changes in amino acid use can reduce cellular nitrogen demands. Therefore, a future development of the work presented in

this thesis could incorporate the amino acid biosynthesis costs to better consider gene sequence optimisation in the context of the whole cell.

Though the work presented in this thesis encompasses both bacterial and eukaryotic species, it has focused purely on single-celled organisms. Given the importance of resource allocation in determining sequence evolution, it is unlikely that this is a phenomenon that is restricted to just bacteria and single-celled eukaryotes. Therefore, a natural extension of the work in this thesis would be to verify these findings by analysing multicellular organisms. Multicellularity brings additional complexity in comparison to analysis of single-celled organisms. However, one way to analyse the impact of resource availability on multicellular genome evolution would be to take an approach similar to that in Chapter 3 and compare related species that have been subject to differing amounts of dietary or environmental nitrogen for long periods of time (millions of years). For example *Ailuropoda melanoleuca* (Giant Panda) and *Ursus maritimus* (Polar bear) had a common ancestor ~ 15-20 MYA and have subsequently evolved to fill very different niches with very different dietary inputs (Kutschera et al. 2014). Similarly, bats are a relatively well-sequenced order and have both insectivorous and frugivorous (fruit-eating) species whose dietary nitrogen availability would differ. I would hypothesise that, as with the unicellular organisms, differences in dietary nitrogen availability will have had a concomitant effect on genome evolution in these species. Specifically, I would predict that Panda bears use less nitrogen in their gene sequences than Polar bears, and fruit eating bats would have less nitrogen in their gene sequences than insect eating bats. However, the comparatively long generation times of these organisms (compared to bacteria and single cell eukaryotes) and the short divergence time between these species could limit the resolution of the analysis as differences in nitrogen use on such a small time-scale would be expected to be very small.

### 5.2.7 Selection acting on resource allocation can constrain the rate of molecular evolution.

Given that genes within the same organism can experience different strengths of selection acting on resource allocation, it follows that the strength of selection acting on resource allocation should result in a concomitant difference in gene evolutionary rate. In order to better understand the link between selection acting on sequence evolution and the rate of molecular evolution, I undertook a comparative analysis that compared sequence optimisation with rates of synonymous and non-synonymous mutations. The results are reported in Chapter 4 and show that the more optimised a sequence is in terms of high translational efficiency and low biosynthetic cost, the lower the rates of both synonymous and non-synonymous mutations. This is because the more optimised a given codon is, the more likely it is that a spontaneous mutation will lead to a less optimal codon whether or not it is a synonymous or non-synonymous change. Reduced divergence of homologous sequences had previously been linked to selection acting on synonymous codons but had only taken sequence expression levels into account (Sharp 1991). Our findings build on this and provide a mechanistic explanation that links selection acting on sequence evolution and the molecular clock.

The molecular clock uses differences in molecular evolution to estimate the timing of events such as speciation by assuming that molecular evolution occurs at an approximately uniform rate over time (Zuckerkandl and Pauling 1962; Kumar 2005). Moreover, in plants, high rates of molecular evolution positively correlate with both speciation and extinction (Lancaster 2010). This is because fast rates of molecular evolution lead to higher rates of polymorphism which can lead to reproductive isolation and hence speciation. Extinction rates also increase as increased rates of molecular evolution increase the species' mutational genetic load. To put this in the context of the findings in my thesis: sequences under strong selection have reduced rates of molecular evolution. Therefore species under strong selection would be expected to have reduced rates of molecular

evolution as well as reduced rates of speciation and extinction. Since we have linked selection acting on resource allocation and rates of molecular evolution, this has implications for understanding species evolution in communities and ecosystems. For example, communities of bacteria in nutrient-limiting environments typically have very low levels of species diversity. This is attributed to the principle of competitive exclusion whereby one species becomes predominant as it out-competes all other species (Hardin 1960). Our findings can build on this and would predict that species in nutrient-limited environments will be under strong selection for efficient resource allocation. This would lead to reduced rates of molecular evolution and thus reduced species diversity.

### 5.3 Overall Summary

Genome evolution and the factors that lead to sequence divergence are fundamental areas of research that have been greatly aided by recent increases in data availability. In this thesis I contribute to this body of research in three significant ways. First, I report the discovery of a new species, *Phytomonas oxycareni,* in Chapter 2. Second, the results described in Chapter 3 show how nitrogen availability can alter sequence evolution in parasites. Finally, by examining the trade-off between codon nucleotide cost and translational efficiency in Chapter 4, I show that sequences which have been optimised for reduced resource allocation to mRNA have reduced rates of molecular evolution. This sheds light on the underlying mechanism behind variation in rates of molecular evolution between genes within a species. Overall, the results described here highlight how comparative approaches that analyse extant gene and genome sequences are able to provide new mechanistic insight into the factors driving sequence evolution.

**References**


Acquisti C., Elser J. J. & Kumar S. 2009. Ecological nitrogen limitation shapes the DNA composition of plant genomes. *Mol. Biol. Evol.*, **26**:953–956.

Acquisti C., Kumar S. & Elser J. J. 2009. Signatures of nitrogen limitation in the elemental composition of the proteins involved in the metabolic apparatus. *Proc. Biol. Sci.*, **276**:2605–2610.

Agashe D., Martinez-Gomez N. C., Drummond D. A. & Marx C. J. 2013. Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol. Evol.*, **30**:549–560.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics*, **136**:927–935.

Amon J., Titgemeyer F. & Burkovski A. 2010. Common patterns - Unique features: Nitrogen metabolism and regulation in Gram-positive bacteria. *FEMS Microbiol. Rev.*, **34**:588–605.

Andersen M. T., Liefting L. W., Havukkala I. & Beever R. E. 2013. Comparison of the complete genome sequence of two closely related isolates of *Candidatus Phytoplasma australiense* reveals genome plasticity. *BMC Genomics*, **14**:529–544.

Arraes F. B. M., de Carvalho M. J. a, Maranhão A. Q., Brígido M. M., Pedrosa F. O. & Felipe M. S. S. 2007. Differential metabolism of *Mycoplasma* species as revealed by their genomes. *Genet. Mol. Biol.*, **30**:182–189.

Aslett M., Aurrecoechea C., Berriman M., Brestelli J., Brunk B. P., Carrington M., Depledge D. P., Fischer S., Gajria B., Gao X., et al. 2009. TriTrypDB: A functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.*, **38**:457–462.

Baudouin-Cornu P., Surdin-Kerjan Y., Marliere P. & Thomas D. 2001. Molecular evolution of protein function. *Science.*, **293**:297–300.

Bentele K., Saffert P., Rauscher R., Ignatova Z. & Bluthgen N. 2014. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.*, **9**:675–675.

Boël G., Letso R., Neely H., Price W. N., Wong K., Su M., Luff J. D., Valecha M., Everett J. K., Acton T. B., et al. 2016. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*, **529**:358–363.

Bragg J. G., Quigg A., Raven J. a. & Wagner A. 2012. Protein elemental sparing and codon usage bias are correlated among bacteria. *Mol. Ecol.*, **21**:2480–2487.

Brandis G. & Hughes D. 2016. The selective advantage of synonymous codon usage bias in *Salmonella*. *PLOS Genet.*, **12**:e1005926.

Brown D. R., Farmerie W. G., May M., Benders G. a, Durkin a S., Hlavinka K., Hostetler J., Jackson J., Johnson J., Miller R. H., et al. 2011. Genome sequences of *Mycoplasma alligatoris* A21JP2T and *Mycoplasma crocodyli* MP145T. *J. Bacteriol.*, **193**:2892–2893.

Buckland R. J., Watt D. L., Chittoor B., Nilsson A. K., Kunkel T. a. & Chabes A. 2014. Increased and imbalanced dNTP pools symmetrically promote both leading and lagging strand replication infidelity. *PLoS Genet.*, **10**:e1004846.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, **129**:897–907.

Butler C. E., Jaskowska E. & Kelly S. 2017. Genome sequence of *Phytomonas françai*, a cassava (*Manihot esculenta*) latex parasite. *Genome Announc.*, **5**:e01266-16.

Calcutt M. J. & Foecking M. F. 2011. Genome sequence of *Mycoplasma putrefaciens* type strain KS1. *J. Bacteriol.*, **193**:6094–6094.

Calderon-Copete S. P., Wigger G., Wunderlin C., Schmidheini T., Frey J., Quail M. a & Falquet L. 2009. The *Mycoplasma* conjunctivae genome sequencing, annotation and analysis. *BMC Bioinformatics*, **10**:1–10.

Camargo E. P. 1999. *Phytomonas* and other trypanosomatid parasites of plants and fruit. *Adv. Parasitol.*, **42**:29–112.

Caradonna K. L., Engel J. C., Jacobi D., Lee C.-H. & Burleigh B. a. 2013. Host metabolism regulates intracellular growth of *Trypanosoma cruzi*. *Cell Host Microbe*, **13**:108–17.

Carbone A. 2008. Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J. Mol. Evol.*, **66**:210–223.

Carlson R. P. 2007. Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics*, **23**:1258–1264.

Catarino L. M., Serrano M. G., Cavazzana M., Almeida M. L., Kaneshina E. K., Campaner M., Jankevicius J. V., Teixeira M. M. G. & Itow-Jankevicius S. 2001. Classification of trypanosomatids from fruits and seeds using morphological, biochemical and molecular markers revealed several genera among fruit isolates. *FEMS Microbiol. Lett.*, **201**:65–72.

Chen L., Brügger K., Skovgaard M., She Q., Torarinsson E., Greve B., Zibat A., Klenk H.-P., Garrett R. A., Bru K., et al. 2005. The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota †. *J. Bacteriol.*, **187**:4992–4999.

Chen W.-H., Guanting L., Peer B., Songnian H. & Martin J. L. 2016. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat. comminications*, **7**:1–10.

Chevance F. F. V. & Hughes K. T. 2017. Case for the genetic code as a triplet of triplets. *Proc. Natl. Acad. Sci.*, **114**:4745–4750.

Cho B.-K., Zengler K., Qiu Y., Park Y. S., Knight E. M., Barrett C. L., Gao Y. & Palsson B. Ø. 2009. Elucidation of the transcription unit architecture of the *Escherichia coli* K-12 MG1655 genome. *Nat. Biotechnol.*, **27**:1043–1049.

Collingridge P. W. & Kelly S. 2012. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics*, **13**:1–10.

Creek D. J., Nijagal B., Kim D. H., Rojas F., Matthews K. R. & Barrett M. P. 2013. Metabolomics guides rational development of a simplified cell culture medium for drug screening against *Trypanosoma brucei*. *Antimicrob. Agents Chemother.*, **57**:2768–2779.

Crow J. F. & Kimura M. 1970. An introduction to population genetic theory. (Row H. (ed.)). New York.

Dabrazhynetskaya A., Soika V., Volokhov D., Simonyan V. & Chizhikov V. 2014. Genome sequence of *Mycoplasma hyorhinis* strain DBS 1050. **2**:e00127-14.

Darwin C. 1859. On the Origin of Species. (Murray J. (ed.)). London.

Delves M. J., Ruecker A., Straschil U., Lelièvre J., Marques S., López-Barragán M. J., Herreros E. & Sinden R. E. 2013. Male and female *Plasmodium falciparum* mature gametocytes show different responses to antimalarial drugs. *Antimicrob. Agents Chemother.*, **57**:3268–3274.

Dettman J. R., Rodrigue N., Melnyk A. H., Wong A., Bailey S. F. & Kassen R. 2012. Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Mol. Ecol.*, **21**:2058–2077.

Drummond D. A., Raval A. & Wilke C. O. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.*, **23**:327–337.

Drummond D. A. & Wilke C. O. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**:341–352.

Drummond D. A. & Wilke C. O. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet.*, **10**:715–724.

Dušanic D., Dušan B., Narat M. & Oven I. 2014. Phenotypic characterization of *Mycoplasma synoviae* induced changes in the metabolic and sensitivity profile of *in vitro* infected chicken chondrocytes. *Biomed Res. Int.*, **6**:1–10.

Dybvig K., Zuhua C., Lao P., Jordan D. S., French C. T., Tu A. H. T. & Loraine A. E. 2008. Genome of *Mycoplasma arthritidis*. *Infect. Immun.*, **76**:4000–4008.

Elf J., Nilsson D., Tenson T. & Ehrenberg M. 2003. Selective Charging of tRNA isoacceptors explains patterns of codon usage. *Science*, **300**:1718–1722.

Elser J. J., Acquisti C. & Kumar S. 2011. Stoichiogenomics: The evolutionary ecology of macromolecular elemental composition. *Trends Ecol. Evol.*, **26**:38–44.

Elser J. J., Fagan W. F., Subramanian S. & Kumar S. 2006. Signatures of ecological resource availability in the animal and plant proteomes. *Mol. Biol. Evol.*, **23**:1946–1951.

Emms D. M. & Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*, **16**:1–14.

Eskesen S. T., Eskesen F. N. & Ruvinsky A. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. *Genetics*, **167**:543–550.

Eyre-Walker a C. 1991. An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.*, **33**:442–449.

Farmer I. S. & Jones C. W. 1976. The energetics of *Escherichia coli* during aerobic growth in continuous culture. *Eur. J. Biochem.*, **67**:115–122.

Fiebig M., Kelly S. & Gluenz E. 2015. Comparative life cycle transcriptomics revises *Leishmania mexicana* genome annotation and links a chromosome duplication with parasitism of vertebrates. *PLOS Pathog.*, **11**:e1005186.

Fire A. 1999. RNA-triggered gene silencing. *Trends Genet.*, **1**:358–363.

Franchini G. 1922. Sur un flagellé de lygoeide (*Crithidia oxycareni* n. sp.). *Bull. la Socié té Pathol. Exot.*, **15**:113–116.

Francino M. P. & Ochman H. 1999. Isochores result from mutation not selection. *Nature*, **400**:30–31.

Francois C. M., Duret L., Simon L., Mermillod-Blondin F., Malard F., Konecny-Dupré L., Planel R., Penel S., Douady C. J. & Lefébure T. 2016. No evidence that nitrogen limitation influences the elemental composition of isopod transcriptomes and proteomes. *Mol. Biol. Evol.*, **33**:2605–2620.

Freymuller E., Milder R., Jankevicius J. V., Jankevicius S. & Camargo E. P. 1990. Ultrastructural studies on the genus. *J. Protozool.*, **37**:225–229.

Frolov A. O., Malysheva M. N., Yurchenko V. & Kostygov A. Y. 2016. Back to monoxeny: *Phytomonas nordicus* descended from dixenous plant parasites. *Eur. J. Protistol.*, **52**:1–10.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.*, **19**:65–68.

Ginger M., Fairlamb A. & Opperdoes F. 2007. Comparative genomics of trypanosome metabolism. *Trypanos.*, **3**:1–40.

González-Pastor J. E. 2011. Cannibalism: A social behavior in sporulating *Bacillus subtilis*. *FEMS Microbiol. Rev.*, **35**:415–424.

Goodman D. B., Church G. M. & Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science*, **342**:475–480.

Gresham D., Usaite R., Germann S. M., Lisby M., Botstein D. & Regenberg B. 2010. Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. *Proc. Natl. Acad. Sci. U. S. A.*, **107**:18551–18556.

Grosjean H., Breton M., Sirand-Pugnet P., Tardy F., Thiaucourt F., Citti C., Barré A., Yoshizawa S., Fourmy D., de Crécy-Lagard V., et al. 2014. Predicting the minimal translation apparatus: lessons from the reductive evolution of Mollicutes. *PLoS Genet.*, **10**:e1004363.

Grosjean H., de Crécy-Lagard V. & Marck C. 2010. Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. *FEBS Lett.*, **584**:252–264.

Gustafsson C., Govindarajan S. & Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.*, **22**:346–353.

Hardin G. 1960. The competitive exclusion principle. *Science*, **131**:1292–1297.

Hessen D. O., Jeyasingh P. D., Neiman M. & Weider L. J. 2010. Genome streamlining and the elemental costs of growth. *Trends Ecol. Evol.*, **25**:75–80.

Higgs P. G. & Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol. Biol. Evol.*, **25**:2279–2291.

Hockenberry A. J., Sirer M. I., Amaral L. A. N. & Jewett M. C. 2014. Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.*, **31**:1880–1893.

Hong J. & Gresham D. 2014. Molecular Specificity, Convergence and constraint shape adaptive evolution in nutrient-poor environments. *PLoS Genet.*, **10**:e1004041.

Horn D. 2008. Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC Genomics*, **9**:1–11.

Hu H., Gao J., He J., Yu B., Zheng P., Huang Z., Mao X., Yu J., Han G. & Chen D. 2013. Codon optimization significantly improves the expression level of a keratinase gene in *Pichia pastoris*. *PLoS One*, **8**:e58393.

Hurst L. D. & Merchant A. R. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. Biol. Sci.*, **268**:493–7.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**:389–409.

Jackson A. P. 2014. Genome evolution in trypanosomatid parasites. *Parasitology*, **7**:1–17.

Jankevicius J. V., Jankevicius S. I., Campaner M., Conchon I., Maeda L. A., Teixiera M. M. C., Freymuller E. & Camargo E. P. 1989. Life cycle and culturing of *Phytomonas serpens* (Gibbs), a trypanosomatid parasite of tomatoes. *J. Protozool.*, **36**:265–271.

Jaskowska E., Butler C., Preston G. & Kelly S. 2015. *Phytomonas*: Trypanosomatids adapted to plant environments. *PLOS Pathog.*, **11**:e1004484.

Jezequel N., Lagomarsino M. C., Heslot F. & Thomen P. 2013. Long-term diversity and genome adaptation of *Acinetobacter baylyi* in a minimal-medium chemostat. *Genome Biol. Evol.*, **5**:87–97.

Jia W. & Higgs P. G. 2008. Codon usage in mitochondrial genomes: Distinguishing context-dependent mutation from translational selection. *Mol. Biol. Evol.*, **25**:339–351.

Kalushkov P. & Nedvěd O. 2010. Suitability of food plants for *Oxycarenus lavaterae*

(Heteroptera: Lygaeidae), a Mediterranean bug invasive in Central and South-East Europe. *Comptes Rendus l Acad. Bulg. des Sci.*, **63**:271–276.

Katoh K. & Standley D. M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**:772–780.

Kimura M. & Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**:763–771.

Kment P., Vahala O. & Hradil K. 2006. First records of *Oxycarenus lavaterae* (Heteroptera: Oxycarenidae) from the Czech Republic, with review of its distribution and biology. *Klapalekiana*, **42**:97–127.

Knoppel A., Nasvall J. & Andersson D. I. 2016. Compensating the fitness costs of synonymous mutations. *Mol. Biol. Evol.*, **33**:1461–1477.

Kolev N. G., Franklin J. B., Carmi S., Shi H., Michaeli S. & Tschudi C. 2010. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog.*, **6**:1–15.

Koreny L., Sobotka R., Kovarova J., Gnipova A., Pavel F., Horvath A., Obornik M., Ayala F. J. & Lukeš J. 2012. Aerobic kinetoplastid flagellate *Phytomonas* does not require heme for viability. *PNAS*, **109**:3808–3813.

Krisko A., Copic T., Gabaldón T., Lehner B. & Supek F. 2014. Inferring gene function from evolutionary change in signatures of translation efficiency. *Genome Biol.*, **15**:1–17.

Kube M., Mitrovic J., Duduk B., Rabus R. & Seemüller E. 2012. Current view on *Phytoplasma* genomes and encoded metabolism. *Sci. World J.*, **20**:1–25.

Kube M., Schneider B., Kuhl H., Dandekar T., Heitmann K., Migdoll A. M., Reinhardt R. & Seemüller E. 2008. The linear chromosome of the plant-pathogenic mycoplasma "*Candidatus Phytoplasma mali*". *BMC Genomics*, **9**:306.

Kuehn A. & Pradel G. 2010. The coming-out of malaria gametocytes. *J. Biomed. Biotechnol.*, **2**:1–11.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*, **6**:654–662.

Kutschera V. E., Bidon T., Hailer F., Rodi J. L., Fain S. R. & Janke A. 2014. Bears in a forest of gene trees: Phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Mol. Biol. Evol.*, **31**:2004–2017.

Lambros R. J., Mortimer J. R. & Forsdyke D. R. 2003. Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles*, **7**:443–450.

Lancaster L. T. 2010. Molecular evolutionary rates predict both extinction and speciation in temperate angiosperm lineages. *BMC Evol. Biol.*, **10**:1–10.

Lao P. J. & Forsdyke D. R. 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.*, **10**:228–236.

Lassalle F., Périan S., Bataillon T., Nesme X., Duret L. & Daubin V. 2015. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. *PLoS Genet.*, **11**:1–20.

Latysheva N., Junker V. L., Palmer W. J., Codd G. A. & Barker D. 2012. The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics*, **28**:603–606.

Lauffer M. 1975. Entropy-driven processes in biology. Berlin, Springer.

Lee Y., Zhou T., Tartaglia G. G., Vendruscolo M. & Wilke C. O. 2010. Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, **10**:4163–71.

Li N., Lv J. & Niu D. K. 2009. Low contents of carbon and nitrogen in highly abundant proteins: Evidence of selection for the economy of atomic composition. *J. Mol. Evol.*, **68**:248–255.

Li W. H. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.*, **24**:337–345.

Liu W., Xiao S., Li M., Guo S., Li S. & Luo R. 2013. Comparative genomic analyses of *Mycoplasma hyopneumoniae* pathogenic 168 strain and its high-passaged attenuated strain. *Genomics*, **14**:1–13.

Lord P. W., Selley J. N. & Attwood T. K. 2002. CINEMA-MX: a modular multiple alignment editor. *Bioinformatics*, **18**:1402–1403.

Lucks J. B., Nelson D. R., Kudla G. R. & Plotkin J. B. 2008. Genome landscapes and bacteriophage codon usage. *PLoS Comput. Biol.*, **4**:1–17.

Lukeš J., Skalický T., Týč J., Votýpka J. & Yurchenko V. 2014. Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.*, **195**:115–122.

Lynch M. 2007. The origins of genome architecture. First Edit. Sunderland, MA, Sinauer Associates, Inc. Publishers.

Maslov D. A., Westenberger S. J., Xu X., Campbell D. A. & Sturm N. R. 2007. Discovery and barcoding by analysis of spliced leader RNA gene sequences of new isolates of trypanosomatidae from heteroptera in Costa Rica & Ecuador. *J. Euk. Microbiol.*, **54**:57–65.

Maslov D., Lukes J., Jirku M. & Simpson L. 1996. Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa. *Mol Biochem Parasitol*, **75**:197–205.

Mathur J., Bizzoco R. W., Ellis D. G., Lipson D. A., Poole A. W., Levine R. & Kelley S. T. 2007. Effects of abiotic factors on the phylogenetic diversity of bacterial communities in acidic thermal springs. *Appl. Environ. Microbiol.*, **73**:2612–2623.

Mazel D. & Marlière P. 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. *Nature*, **341**:245–248.

Mazet M., Morand P., Biran M., Bouyssou G., Courtois P., Daulouède S., Millerioux Y., Franconi J. M., Vincendeau P., Moreau P., et al. 2013. Revisiting the central metabolism of the bloodstream forms of *Trypanosoma brucei*: Production of acetate in the mitochondrion is essential for parasite viability. *PLoS Negl. Trop. Dis.*, **7**:1–14.

McConville M. J. & Naderer T. 2011. Metabolic pathways required for the intracellular survival of *Leishmania*. *Annu. Rev. Microbiol.*, **65**:543–61.

McEwan C. E., Gatherer D. & McEwan N. R. 1998. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*, **128**:173–178.

McGowin C. L., Ma L., Jensen J. S., Mancuso M. M., Hamasuna R., Adegboye D. & Martin D. H. 2012. Draft genome sequences of four axenic *Mycoplasma genitalium* strains isolated from Denmark, Japan, and Australia. *J. Bacteriol.*, **194**:6010–6011.

Mitrović J., Siewert C., Duduk B., Hecht J., Mölling K., Broecker F., Beyerlein P., Büttner C., Bertaccini A. & Kube M. 2014. Generation and analysis of draft sequences of "stolbur" *Phytoplasma* from multiple displacement amplification templates. *J. Mol. Microbiol. Biotechnol.*, **24**:1–11.

Van Mooy B. A. S., Rocap G., Fredricks H. F., Evans C. T. & Devol A. H. 2006. Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proc. Natl. Acad. Sci.*, **103**:8607–8612.

Nascimento N. C. Do, Guimaraes A. M. S., Santos A. P., SanMiguel P. J. & Messick J. B. 2012. Complete genome sequence of *Mycoplasma haemocanis* strain Illinois. *J. Bacteriol.*, **194**:1605–1606.

do Nascimento N. C., Dos Santos A. P., Chu Y., Guimaraes A. M. S., Pagliaro A. & Messick J. B. 2013. Genome sequence of *Mycoplasma parvum* (Formerly *Eperythrozoon parvum*), a diminutive hemoplasma of the pig. *Genome Announc.*, **1**:1–2.

Näsvall S. J., Chen P. & Björk G. R. 2004. The modified wobble nucleoside uridine-5-oxyacetic acid in tRNA Pro cmo 5 UGG promotes reading of all four proline codons *in vivo*. *RNA,* **10**:1662–1673.

Navon S. & Pilpel Y. 2011. The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol.*, **12**:1–10.

Nedvěd O., Chehlarov E. & Kalushkov P. 2014. Life history of the invasive bug *Oxycarenus lavaterae* ( Heteroptera : Oxycarenidae ) in Bulgaria. **66**:203–208.

Neiman M., Kay a D. & Krist a C. 2013. Can resource costs of polyploidy provide an advantage to sex? *Heredity (Edinb).*, **110**:152–9.

Nelson D. C., Wirsen C. O. & Jannasch H. W. 1989. Characterization of large, autotrophic *Beggiatoa* sp. abundant at hydrothermal vents of the Guaymas basin. *Appl. Environ. Microbiol.*, **55**:2909–2917.

Novoa E. M. & Ribas de Pouplana L. 2012. Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends Genet.*, **28**:574–581.

Oshima K., Kakizawa S., Nishigawa H., Jung H.-Y., Wei W., Suzuki S., Arashida R., Nakata D., Miyata S., Ugaki M., et al. 2004. Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat. Genet.*, **36**:27–29.

Park C., Qian W. & Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep.*, **13**:1123–1129.

Paz A., Mester D., Baca I., Nevo E. & Korol A. 2004. Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, **101**:2951–2956.

Pereyre S., Sirand-Pugnet P., Beven L., Charron A., Renaudin H., Barré A., Avenaud P., Jacob D., Couloux A., Barbe V., et al. 2009. Life on arginine for *Mycoplasma hominis*: Clues from its minimal genome and comparison with other human urogenital mycoplasmas. *PLoS Genet.*, **5**:1–4.

Plotkin J. B., Robins H. & Levine A. J. 2004. Tissue-specific codon usage and the expression of human genes. *Proc. Natl. Acad. Sci. U. S. A.*, **101**:12588–12591.

Pollack J. D., Williams M. V & McElhaney R. N. 1997. The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.*, **23**:269–354.

Porcel B. M., Denoeud F., Opperdoes F., Noel B., Madoui M.-A., Hammarton T. C., Field M. C., Da Silva C., Couloux A., Poulain J., et al. 2014. The streamlined genome of *Phytomonas* sp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants.McDowell J. M. (ed.). *PLoS Genet.*, **10**:e1004007.

Posey J. E. & Gherardini F. C. 2000. Lack of a role for iron in the Lyme disease pathogen. *Science*, **288**:1651–1653.

Precup J. & Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J. Biol. Chem.*, **262**:11351–11355.

Qin H., Wu W. B., Comeron J. M., Kreitman M. & Li W. H. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, **168**:2245–2260.

Ran W. & Higgs P. G. 2010. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol. Biol. Evol.*, **27**:2129–2140.

Ran W. & Higgs P. G. 2012. Contributions of speed and accuracy to translational selection in bacteria. *PLoS One*, **7**:1–7.

Rao Y., Wu G., Wang Z., Chai X., Nie Q. & Zhang X. 2011. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Res.*, **18**:499–512.

Rastogi R. P., Richa, Kumar A., Tyagi M. B. & Sinha R. P. 2010. Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair. *J. Nucleic Acids*, **1**:1–32.

Razin S. & Knight B. C. 1960. A partially defined medium for the growth of *Mycoplasma*. *J. Gen. Microbiol.*, **22**:492–503.

dos Reis M., Savva R. & Wernisch L. 2004. Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.*, **32**:5036–5044.

dos Reis M., Wernisch L. & Savva R. 2003. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.*, **31**:6976–6985.

Rocha E. P. C. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, & efficient decoding for translation optimization. *Genome Res.*, **14**:2279–86.

Rocha E. P. C. & Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**:291–294.

Rocha E. P. C. & Feil E. J. 2010. Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet.*, **6**:1–4.

Ronquist F., Teslenko M., Van Der Mark P., Ayres D. L., Darling A., Höhna S., Larget B., Liu L., Suchard M. A. & Huelsenbeck J. P. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**:539–42.

Sabi R. & Tuller T. 2014. Modelling the efficiency of codon – tRNA interactions based on codon usage bias. *DNA Res.*, **21**:511–525.

Sasaki Y., Ishikawa J., Yamashita A., Oshima K., Kenri T., Furuya K., Yoshino C., Horino A., Shiba T., Sasaki T., et al. 2002. The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. *Nucleic Acids Res.*, **30**:5293–5300.

Schattner P., Brooks A. N. & Lowe T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**:686–689.

Schmidt T. T., Reyes G., Gries K., Ceylan C. Ü., Sharma S., Meurer M., Knop M., Chabes A. & Hombauer H. 2017. Alterations in cellular metabolism triggered by URA7 or GLN3 inactivation cause imbalanced dNTP pools and increased mutagenesis. *Proc. Natl. Acad. Sci.*, **1**:1–12.

Seward E. A. & Kelly S. 2016. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.*, **17**:1–15.

Shah P. & Gilchrist M. a. 2011. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proc. Natl. Acad. Sci. U. S. A.*, **108**:10231–10236.

Sharp P. M. 1991. Determinants of DNA sequence divergence between *E. coli* & *Salmonella typhimurium*: Codon usage, map position, & concerted evolution. *J. Mol. Evol.*, **33**:23–33.

Sharp P. M., Bailes E., Grocock R. J., Peden J. F. & Sockett R. E. 2005. Variation in the

strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, **33**:1141–1153.

Sharp P. M. & Li W. H. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.*, **4**:222–230.

Shields D. C. 1990. Switches in species-specific codon preferences: The influence of mutation biases. *J. Mol. Evol.*, **31**:71–80.

Shu H. W., Liu T. T., Chan H. I., Liu Y. M., Wu K. M., Shu H. Y., Tsai S. F., Hsiao K. J., Hu W. S. & Ng W. V. 2011. Genome sequence of the repetitive-sequence-rich *Mycoplasma fermentans* strain M64. *J. Bacteriol.*, **193**:4302–4303.

da Silva R. V, Malvezi A. D., Augusto L. D. S., Kian D., Tatakihara V. L. H., Yamauchi L. M., Yamada-Ogatta S. F., Rizzo L. V, Schenkman S. & Pinge-Filho P. 2013. Oral exposure to *Phytomonas serpens* attenuates thrombocytopenia and leukopenia during acute infection with *Trypanosoma cruzi*. *PLoS One*, **8**:e68299.

Sinden R. E., Talman A., Marques S. R., Wass M. N. & Sternberg M. J. 2010. The flagellum in malarial parasites. *Curr Opin Microbiol*, **13**:491–500.

Sørensen M. a, Kurland C. G. & Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.*, **207**:365–377.

Spitznagel D., Ebikeme C., Biran M., Nicabháird N., Bringaud F., Henehan G. T. M. & Nolan D. P. 2009. Alanine aminotransferase of *Trypanosoma brucei*- a key role in proline metabolism in procyclic life forms. *FEBS J.*, **276**:7187–7199.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**:2688–2690.

Sterner R. W. & Elser J. J. 2002. Ecological stoichiometry: The biology of elements from molecules to the biosphere. Princeton (NJ), Princeton University Press.

Stoletzki N. & Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. *Mol. Biol. Evol.*, **24**:374–381.

Subramanian S. 2008. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics*, **178**:2429–2432.

Supek F., Škunca N., Repar J., Vlahoviček K. & Šmuc T. 2010. Translational selection is ubiquitous in prokaryotes. *PLoS Genet.*, **6**:1–13.

Tran-Nguyen L. T. T., Kube M., Schneider B., Reinhardt R. & Gibb K. S. 2008. Comparative genome analysis of "*Candidatus Phytoplasma australiense*" and "*Ca. phytoplasma asteris*" strains OY-M and AY-WB. *J. Bacteriol.*, **190**:3979–3991.

Tu Y.-K., Kellett M., Clerehugh V. & Gilthorpe M. S. 2005. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *Br. Dent. J.*, **199**:457–461.

Tuller T., Carmi A., Vestsigian K., Navon S., Dorfan Y., Zaborske J., Pan T., Dahan O., Furman I. & Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**:344–354.

Tuller T., Waldman Y. Y., Kupiec M. & Ruppin E. 2010. Translation efficiency is determined by both codon bias & folding energy. *Proc. Natl. Ac. Sci. U. S. A.*, **107**:3645–50.

Valiron O., Caudron N. & Job D. 2002. Microtubule dynamics. *Curr. Opin. Cell Biol.*, **4**:2069–2084.

Vasconcelos A. T. R., Vasconcelos A. T. R., Ferreira H. B., Ferreira H. B., Bizarro C. V, Bizarro C. V, Bonatto S. L., Bonatto S. L., Carvalho M. O., Carvalho M. O., et al. 2005. Swine and poultry pathogens: the complete genome sequences of two strains. *Microbiology*,

**187**:5568–5577.

Vickerman K. & Preston T. M. 1976. Comparative cell biology of the kinetoplastid flagellates. *Biol. Kinetoplastida*, **1**:35–130.

Vieira-Silva S. & Rocha E. P. C. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.*, **6**:e1000808.

Votýpka J., d'Avila-Levy C. M., Grellier P., Maslov D. A., Lukeš J. & Yurchenko V. 2015. New approaches to systematics of Trypanosomatidae: Criteria for taxonomic (re)description. *Trends Parasitol.*, **31**:460–469.

Votýpka J., Maslov D. a., Yurchenko V., Jirků M., Kment P., Lun Z. R. & Lukeš J. 2010. Probing into the diversity of trypanosomatid flagellates parasitizing insect hosts in South-West China reveals both endemism and global dispersal. *Mol. Phylogen. Evol.*, **54**:243–53.

Waites K. B., Waites K. B., Talkington D. F. & Talkington D. F. 2004. *Mycoplasma pneumoniae*. *Society*, **17**:697–728.

Wanasen N. & Soong L. 2008. L-arginine metabolism and its impact on host immunity against *Leishmania* infection. *Immunol. Res.*, **41**:15–25.

Wang D., Zhang Y., Zhang Z., Zhu J. & Yu J. 2010. KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics Bioinforma.*, **8**:77–80.

Weller C. & Wu M. 2015. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution (N. Y).*, **69**:643–652.

Westenberger S. J., Sturm N. R., Yanega D., Podlipaev S. A., Zeledon R., Campbell D. A. & Maslov D. A. 2004. Trypanosomatid biodiversity in Costa Rica: genotyping of parasites from Heteroptera using the spliced leader RNA gene. *Parasitology*, **129**:537–547.

Wheeler R. J., Gluenz E. & Gull K. 2011. The cell cycle of *Leishmania*: Morphogenetic events and their implications for parasite biology. *Mol. Microbiol.*, **79**:647–662.

Williams M. V. & Pollack J. D. 1990. A mollicute (Mycoplasma) DNA repair enzyme: Purification and characterization of uracil-DNA glycosylase. *J. Bacteriol.*, **172**:2979–2985.

Wise K. S., Calcutt M. J., Foecking M. F., Madupu R., DeBoy R. T., Röske K., Hvinden M. L., Martin T. R., Durkin a. S., Glass J. I., et al. 2012. Complete genome sequences of *Mycoplasma leachii* strain PG50T and the pathogenic *Mycoplasma mycoides* subsp. mycoides small colony biotype strain Gladysdale. *J. Bacteriol.*, **194**:4448–4449.

Worning P., Jensen L. J., Hallin P. F., Stærfeldt H. & Ussery D. W. 2006. Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.*, **8**:353–361.

Yurchenko V., Kostygov A., Havlová J., Grybchuk-Ieremenko A., Ševčíková T., Lukeš J., Ševčík J. & Votýpka J. 2015. Diversity of Trypanosomatids in cockroaches and the description of *Herpetomonas tarakana* sp. n. *J. Eukaryot. Microbiol.*, **63**:198–209.

Yurchenko V. Y., Lukes J., Jirku M., Zeledon R. & Maslov D. a. 2006. *Leptomonas costaricensis* sp. n., a member of the novel phylogenetic group of insect trypanosomatids closely related to the genus *Leishmania*. *Parasitology*, **133**:537–546.

Zhang F., Saha S., Shabalina S. A. & Kashina A. 2010. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science*, **329**:1534–1537.

Zuckerkandl E. 1976. Evolutionary processes and evolutionary noise at the molecular level. *J. Mol. Evol.*, **7**:167–183.

Zuckerkandl E. & Pauling L. B. 1962. Molecular disease, evolution, and genic heterogeneity. (Kasha M. & Pullman B. (eds.)). New York, Horizons in Biochemistry.

# Appendices

**Appendix 1: Seward E. A., Votýpka J., Kment P., Lukeš J. & Kelly S.** 2017. Description of *Phytomonas oxycareni* n. sp. from the Salivary Glands of *Oxycarenus lavaterae*. *Protist*, **168:**71-79.

Protist, Vol. 168, 71–79, xx 2017
http://www.elsevier.de/protis
Published online date 30 November 2016

Protist

ORIGINAL PAPER

# Description of *Phytomonas oxycareni* n. sp. from the Salivary Glands of *Oxycarenus lavaterae*

**Emily A. Seward[a], Jan Votýpka[b,c], Petr Kment[d], Julius Lukeš[b,e,f], and Steven Kelly[a,1]**

[a]Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK
[b]Institute of Parasitology, Biology Centre, Czech of the Academy of Sciences, 370 05 České Budějovice, Czechia
[c]Department of Parasitology, Faculty of Science, Charles University, 128 44 Prague, Czechia
[d]Department of Entomology, National Museum, Cirkusova 1790, 193 00 Prague, Czechia
[e]Faculty of Sciences, University of South Bohemia, 370 05 České Budějovice, Czechia
[f]Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada

Submitted May 20, 2016; Accepted November 12, 2016
Monitoring Editor: Dmitri Maslov

***Phytomonas*** **spp. (phytomonads) are a diverse and globally distributed group of unicellular eukaryotes that parasitize a wide range of plants and are transmitted by insect hosts. Here we report the discovery and characterisation of a new species of *Phytomonas*, named *Phytomonas oxycareni* n. sp., which was obtained from the salivary glands of the invasive species of true bug *Oxycarenus lavaterae* (Heteroptera). The new *Phytomonas* species exhibits a long slender promastigote morphology and can be found both within the lumen of the insect host's salivary glands as well as within the cells of the salivary gland itself. Sampling multiple individuals from the same population post-winter hibernation on two consecutive years revealed that infection was persistent over time. Finally, phylogenetic analyses of small subunit ribosomal RNA genes revealed that this species is sister to other species within the genus *Phytomonas*, providing new insight into the evolutionary history of the clade.**
**© 2016 Elsevier GmbH. All rights reserved.**

**Key words:** Kinetoplastid; *Phytomonas*; phylogeny; intracellular; vector; trypanosomatid.

## Introduction

Trypanosomatids are single-celled eukaryotic parasites that collectively cause a large burden on human health and livelihood, infecting an estimated 20 million people worldwide as well as livestock and crops (da Silva et al. 2013). One large and diverse sub-group of trypanosomatids known as *Phytomonas* (Donovan, 1909), are parasites and pathogens of plants (Camargo 1999; Jaskowska et al. 2015). *Phytomonas* are globally distributed, however little is known of their biology, host range or evolutionary history and comparatively limited sampling has been conducted outside of South America, where several species cause economically important plant pathologies (Jaskowska et al. 2015; Votýpka et al. 2010).

Species in the genus *Phytomonas* are descended from a single adaptation of monoxe-

[1]Corresponding author; fax +44 1865-285-691
e-mail steven.kelly@plants.ox.ac.uk (S. Kelly).

http://dx.doi.org/10.1016/j.protis.2016.11.002
1434-4610/© 2016 Elsevier GmbH. All rights reserved.

nous insect parasites to a plant host about 400 million years ago (Lukeš et al. 2014). Following this event different species have evolved to colonise a large diversity of plant species and can be found in multiple plant tissues including phloem, latex ducts, fruit, flowers and seeds as reviewed in (Jaskowska et al. 2015). In doing so they have evolved to inhabit both extracellular and intracellular plant environments, spanning a wide range of contrasting biochemical compositions. *Phytomonas* are transmitted between plant hosts by insect vectors of the suborder Heteroptera (order Hemiptera) and there is evidence that *Phytomonas nordicus,* a parasite of the predatory stink bug *Troilus luridus* Fabricius, 1775 (Pentatomidae), has reverted back to a monoxenous lifestyle, completing its entire life cycle in the insect host (Frolov et al. 2016). Consistent with their colonisation of plants, *Phytomonas* have also been found to inhabit both extracellular and intracellular environments within their insect hosts (Freymuller et al. 1990; Frolov et al. 2016). Thus, given this wide range of contrasting host tissues and cellular environments, it is likely that there is a large diversity of life cycles and transmission strategies that remain unexplored in this group.

Though the genus *Phytomonas* encompasses the majority of plant-infecting trypanosomatids, the diversity of these protists is not accurately represented in the literature. To improve this the creation of the subfamily Phytomonadinae has been proposed to include the genera *Herpetomonas, Phytomonas* and *Lafontella* (Yurchenko et al. 2015). Historically, species classification of insect and plant trypanosomatids depended on morphology and host specificity (Vickerman and Preston 1976). However, these criteria are not sufficient for species descriptions as opportunistic non-*Phytomonas* trypanosomatids have been found in plants and there is extensive size and shape polymorphism within a single species depending on both host and culture conditions (Catarino et al. 2001; Jaskowska et al. 2015; Wheeler et al. 2011). Thus as there is potential uncertainty in using a morphotype-orientated approach, recent classifications rely on comparatively data-rich molecular methods for taxonomic assignment (Votýpka et al. 2015).

In this paper we characterize a new species, *Phytomonas oxycareni* n. sp., obtained from the salivary gland of the heteropteran insect *Oxycarenus lavaterae* Fabricius, 1787. This true bug is native to the Western Mediterranean, however its range in Europe has expanded both eastwards and westwards since the 1980s (Nedvěd et al. 2014). This finding highlights the potential for the migration of herbivorous insects to be associated with the migration of associated *Phytomonas* parasites. *O. lavaterae* feeds primarily on plants in the Malvaceae family, including both herbaceous representatives of the subfamily Malvoideae (e.g., *Abelmoschus*, *Abutilon*, *Gossypium*, *Hibiscus*, *Lavatera*, *Malva*) as well as lime trees (Tilioideae: *Tilia* spp.). However, in the Mediterranean it may also feed on other plants (e.g. apricots, peaches, *Citrus* spp.) as reviewed in (Kment et al. 2006). Though the plants it feeds on include several important crops (cotton, okra, apricots, peaches) and ornamentals (hibiscus, lime trees) it is rarely reported as an agricultural pest. During the winter, the species hibernates by forming tight aggregations of several hundred individuals on the sunny side of lime tree trunks. These aggregations also occasionally form on other structures such as buildings or fences, causing a public nuisance (Nedvěd et al. 2014).

Based on phylogenetic data we classified the new parasite as *Phytomonas* and propose the species name *oxycareni* to reflect its insect host.

## Results

### Material Collection and Primary Characterisation of a New Trypanosomatid Species from the True Bug *O. lavaterae*

*Oxycarenus lavaterae* bugs were sampled from a single large population overwintering on the trunk of a *Tilia cordata* tree in May 2015 and March 2016 in Sedlec, Czech Republic. Individual insects within this population were dissected and examined by light microscopy to search for the presence of trypanosomatid cells. In total, trypanosomatids were found in ~80% of the salivary glands that were examined. Trypanosomatid cells failed to be detected consistently in any other tissue type that was dissected from the insect. However, trypanosomatid cells were detected in the mid-gut in a small number of dissections but not in sufficient quantities to facilitate further analysis. This very low level of detectable presence in the mid-gut could be due to the life cycle of *O. lavaterae* bugs, which at the time of collection were still aggregated on the trunk of the tree and so had not fed in many months (Nedvěd et al. 2014). Unfortunately, despite repeated attempts using different media, a culture was not established. Thus all analyses reported here were conducted using trypanosomatids isolated directly from the salivary glands of the insects.

## Phylogenetic Analysis Places the New Trypanosomatid Species in the Genus *Phytomonas*

Identical 18S rRNA gene sequences were obtained from trypanosomatids that originated from three infected *O. lavaterae* bugs. These bugs were collected from the same population sampled in two subsequent years. This indicates that infection by the same species of trypanosomatid was pervasive within the population and that infection was likely maintained within the population over multiple generations of the host insect. The alignable region of the 18S rRNA gene of the new trypanosomatid (GenBank Acc. Number KX257483) shared 96% identity with both the 18S sequence of *Phytomonas serpens* (GenBank acc. Number U39577, PRJNA80957) and *P. nordicus* (GenBank Acc. No. KT2236609) (Frolov et al. 2016).

To confirm the taxonomic classification of the new species, a phylogenetic tree of 18S sequences was reconstructed using the new species as well as 40 published species sampled from across the trypanosomatids. The topology of the resultant maximum likelihood tree (Fig. 1) showed that the newly identified trypanosomatid species was monophyletic with previously characterised *Phytomonas* and thus part of the newly proposed subfamily, Phytomonadinae (Yurchenko et al. 2015). This grouping received 100% bootstrap support and a posterior probability of 1.

To independently verify the phylogenetic position of the new species, the spliced leader (SL) RNA gene from a range of representative trypanosomatid species were compared (Supplementary Material Fig. S1). In support of the 18S rRNA analysis, the SL sequence showed that the newly identified trypanosomatid species had 90% bootstrap support and a posterior probability of 0.99 supporting its position as a sister species to previously characterised *Phytomonas* species. The exon of the SL sequence of the new trypanosomatid species differs from previously characterised *Phytomonas* sequences by only one nucleotide and has cytosine at position 15 of the sequence, as is the case for other *Phytomonas* species. Therefore, given the new species forms a monophyletic group with the genus *Phytomonas*, shares features such as a prolonged cell morphology and typical tissue localization (salivary glands) with other *Phytomonas* spp., and is found in a typical insect host for *Phytomonas*, we have assigned the new species to the genus *Phytomonas*. Though it is impossible to determine, this may possibly be the same flagellate that was observed previously in *Oxycarenus*

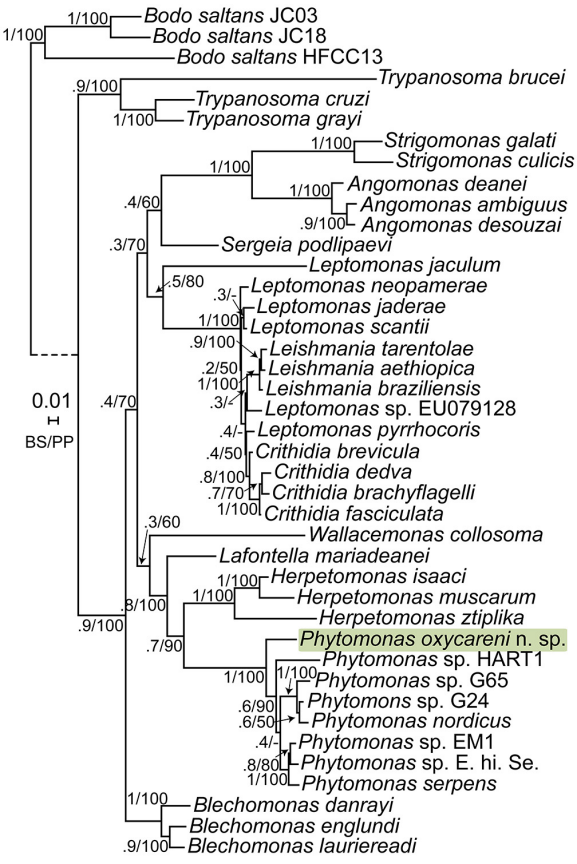



**Figure 1.** The new kinetoplastid parasite belongs to the subfamily Phytomonadinae. Maximum likelihood phylogenetic tree of kinetoplastids obtained using RaxML on 18S ribosomal RNA gene sequences. The tree is rooted on the branch that separates *Bodo saltans* isolates and trypanosomatids. The scale bar represents the number of substitutions per site. Values on each branch represent bootstrap values/posterior probabilities for that branch. Posterior probabilities were calculated using Mr Bayes. For display purposes the dashed branch has been reduced in length by 75%. The new species described in this study (*Phytomonas oxycareni* n. sp.) is highlighted.

*lavaterae* in Italy (Franchini 1922). Thus, based on the phylogenetic information, host and tissue localisation, and its cell morphology (described below), we have named the new species *Phytomonas oxycareni* n. sp.

## Parasites Within the Salivary Gland Exhibit Promastigote Morphology and are Proliferative

Light microscopic examination of *P. oxycareni* cells isolated from salivary glands revealed slender

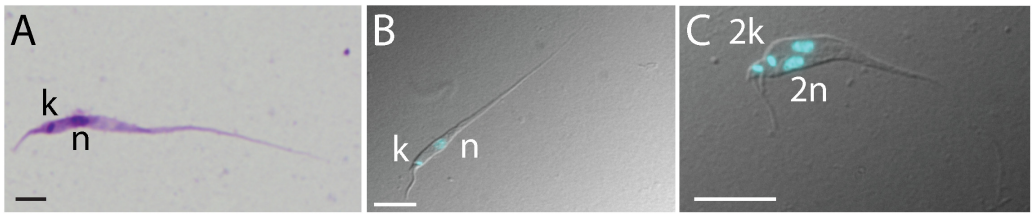



**Figure 2.** Light microscopy images of *Phytomonas oxycareni* n. sp. **A)** A Giemsa-stained cell. **B-C)** DAPI-stained cells displayed as the differential interference contrast image of the cell overlaid with the (blue) fluorescence microscopy image of the DNA stained with DAPI. **B)** Single cell with elongated slender morphology **C)** Cells in the process of dividing (2k 2n). Scale bars are 5 μm. Cells are oriented with the flagellum on the left of each image. "k" denotes the kinetoplast, "n" denotes the nucleus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

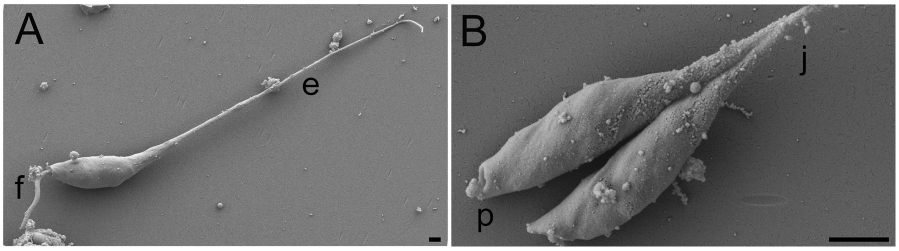



**Figure 3.** Scanning electron microscopy of *Phytomonas oxycareni* n. sp. Cells were obtained directly from infected salivary glands of *Oxycarenus lavaterae*, thus images contain particulate material derived from the salivary gland. **A)** Elongated slender promastigote with flagellum [f] oriented on the left of the image. [e] highlights the elongated cell body **B)** Dividing cells with no external flagella. [p] denotes the flagellar pocket. [j] shows where the cells are still joined. Scale bars are 1 μm.

promastigotes with an elongated cell body and short flagellum that is detached from the body of the cell (Fig. 2A-B). This morphology is similar to that of the corresponding stages observed for promastigotes of the monoxenous species *P. nordicus* (Frolov et al. 2016), and is broadly consistent with several previous morphological descriptions within this genus (Camargo 1999; Jaskowska et al. 2015; Wheeler et al. 2011). Analysis of DAPI stained slides revealed that cells within the population were undergoing cell division. That is, several cells were identified that had already completed DNA replication and segregation of the nuclei and kinetoplasts but had not yet undergone cytokinesis (Fig. 2C).

Higher resolution imaging of cells by scanning electron microscopy (SEM) revealed the full extent of the long slender cell body (Fig. 3A). Although SEM cannot image nuclei and kinetoplasts, cells in late stage cytokinesis were identified as they were still joined at the posterior end (Fig. 3B). The observed lack of external flagella (Fig. 3B) is likely due to the point in cell division captured in this image rather than a diagnostic characteristic of the species (Wheeler et al. 2011).

## *Phytomonas oxycareni* Can be Found Inside Host Cells of the Salivary Gland

Transmission electron microscopy analysis of whole fixed salivary glands readily identified cells within the lumen of the gland (Fig. 4). The parasite cells did not appear to be attached to the surface of the lumen but were instead distributed throughout (Fig. 4A–C). There were also cells that appeared to be undergoing cytokinesis within these sections as indicated by the presence of two axonemal profiles within the same cell (Fig. 4B). Given that two different species of *Phytomonas* have previously been found to also reside within cells of the insect host (Freymuller et al. 1990; Frolov et al. 2016), the cells of the salivary gland tissue were inspected for the presence of *Phytomonas* cells. Indeed, *Phytomonas* flagellates located inside the cells of the salivary gland lumen were identified in multiple instances (Fig. 5). In all cases that were examined, the protists were determined to be inside two distinct membranes (Fig. 5A–B). This indicates that the parasite resided inside a vacuole. It was noteworthy that some intracellular *Phytomonas* cells

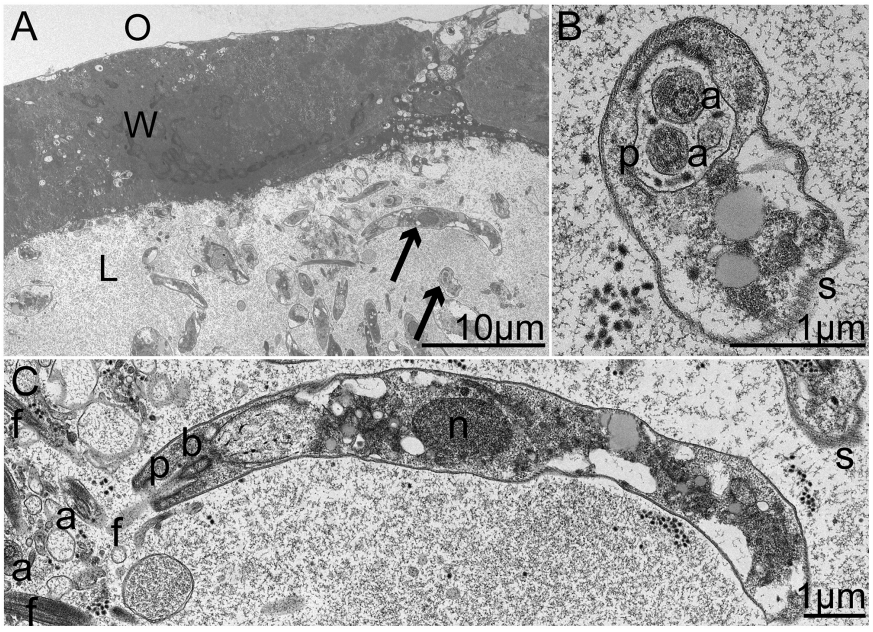



**Figure 4.** Transmission electron microscopy of an infected salivary gland containing *Phytomonas oxycareni* n. sp. **A)** A low magnification image of a section of the gland and the encompassed lumen of an infected *Oxycarenus lavaterae* salivary gland. (O) is outside the gland, (W) is the tissue of salivary gland (L) is the salivary gland lumen. Multiple parasites are visible in the salivary gland lumen and do not appear attached to the epithelium of the salivary gland. **B, C)** High magnification of the parasites highlighted with black arrows in (A) [n] nucleus. [s] subpellicular microtubules. [f] flagellum. [b] basal body. [p] flagellar pocket.

appeared to be in the process of cytokinesis within these vacuoles (Fig. 5C–D). This intracellular proliferation of *Phytomonas* would be consistent with observations of dividing *P. nordicus* within the parasitophorous vacuole of a salivary gland cell (Frolov et al. 2016). However, it is unknown whether this process was initiated before or after the parasite entered the host cell, so it is unknown whether *P. oxycareni* is capable of proliferation within the insect host cells.

## Taxonomic Summary

Class: Kinetoplastea (Honigberg, 1963) Vickerman, 1976

Subclass: Metakinetoplastina Vickerman, 2004

Order: Trypanosomatida (Kent, 1880) Hollande, 1952

Family: Trypanosomatidae (Doflein, 1901) Grobben, 1905

Subfamily: Phytomonadinae Yurchenko, Kostygov, Votypka et Lukes, 2015

Genus: *Phytomonas* Donovan, 1909

## *Phytomonas oxycareni* Votypka, Seward, Kment, Kelly et Lukes n. sp.

### Diagnosis

The species is identified by the unique sequence of 18S rRNA (GenBank accession number: KX257483) and splice leader RNA (GenBank accession number KX611848).

### Species description

Elongated slender promastigotes in the salivary glands were on average 30.9 ± 8.6 µm (19.7–52.7 µm) long and 1.76 ± 0.62 µm (0.93–2.99 µm) wide, with a single short external flagellum that was on average 3.7 ± 1.2 µm (2.2–6.4 µm). Cells had a dilated anterior portion (measuring 13-52% of the total body length) that contains the nucleus and kinetoplast. The posterior part of the cell is narrowed and pointed at the end. The cell body is twisted 0-2 times. The kinetoplast disk is compactly packed, on average 0.78 ± 0.17 µm (0.54–0.93 µm) in length and 0.165 µm (0.165–0.166 µm) in diameter and 0.68 ± 0.08 µm (0.58–0.83 µm) from the anterior of the cell. The nucleus was on average 2.15 ± 0.62 µm (1.34–2.88 µm) in length and

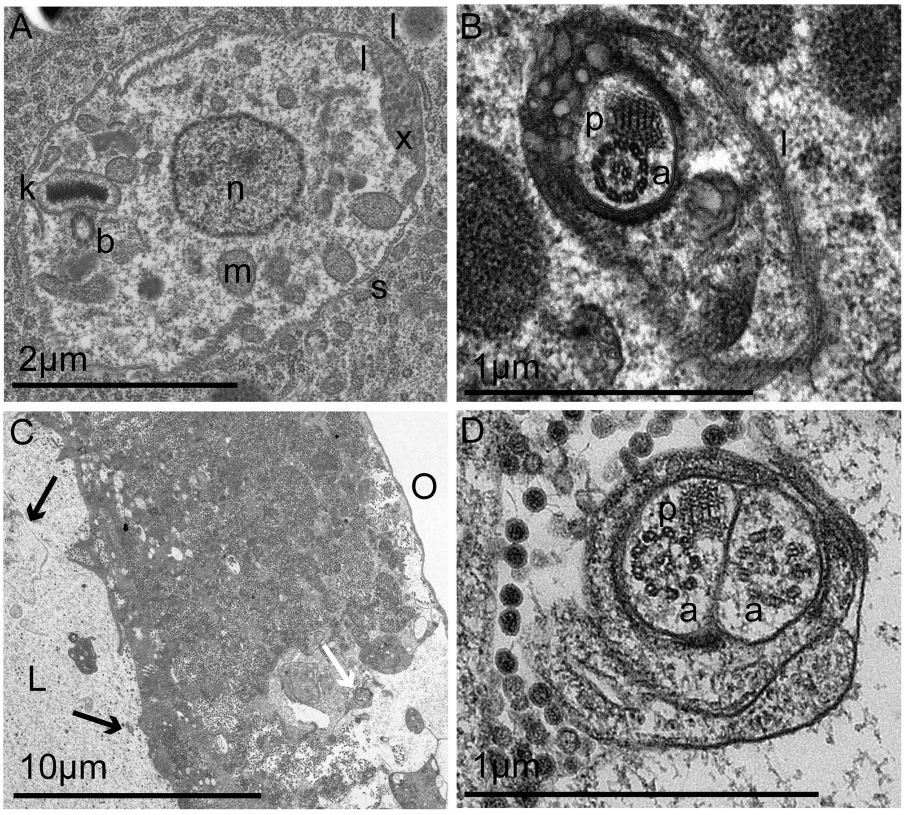



**Figure 5.** Transmission electron micrographs of three intracellular *Phytomonas oxycareni* n. sp. within the salivary glands of the insect host *Oxycarenus lavaterae*. **A)** An intracellular *Phytomonas oxycareni* [k] kinetoplast [n] nucleus [s] subpellicular microtubules adjacent to two membranes. [b] basal body, [l] two lipid membranes, [x] granular material in the posterior part of the parasitophorous vacuole, [m] mitochondria. **B)** A second intracellular *P. oxycareni* [a] axoneme. [p] paraflagellar rod (PFR) with characteristic lattice [l] two membranes indicating that the parasite is inside a vacuole. **C)** A third intracellular *P. oxycareni* (white arrow) near the outside (O) of the salivary gland, multiple parasites (black arrows) are visible in the salivary gland lumen [L]. **D)** High magnification of the cell in (C): [a] axoneme, [p] PFR, [s] subpellicular microtubules. This is likely to be a cell in the process of cytokinesis as indicated by the presence of two axonemal profiles.

$3.92 \pm 1.06\,\mu m$ (2.19–5.00 $\mu$m) from the anterior of the cell.

## Type host

*Oxycarenus lavaterae* (Fabricius, 1787) (Heteroptera: Oxycarenidae). The xenotype, collected on *Tilia cordata* (Malvaceae), is deposited at the Department of Parasitology, Charles University, Prague.

## Location within host

Found both within the midgut and lumen of the salivary gland, as well as within the cells of the salivary gland itself.

## Type locality

Vicinity of Sedlec, Czech Republic, South Moravia (48°46′44.43″N; 16°41′55.13″E), 180 meters above sea level.

## Type material

The name-bearing type, a hapantotype, is a Giemsa-stained slide of the dissected salivary glands, deposited in the research collection of the Department of Parasitology at Charles University in Prague. An axenic culture was not established.

## Etymology

The specific epithet, *oxycareni*, is derived from the generic name of its host; noun in genitive case given in apposition.

## Discussion

We report the identification of a new species of *Phytomonas*, named *Phytomonas oxycareni* that was found in the salivary glands of the heteropteran insect *Oxycarenus lavaterae*. The new species exhibits a long slender promastigote morphology and can be found both within the lumen of the salivary glands, as well as within the cells of the salivary gland itself. Furthermore, sampling different individuals from the same bug population in two consecutive years revealed that infection was likely persistent through several generations of bugs and that the protists overwinter in the insect host/vector.

Interestingly, this newly characterised species is the earliest branching *Phytomonas* species that has been identified to date. However, we were unable to confirm the presence of this species, either by cultivation or by PCR of homogenized leaves and fine branches, within the *Tilia* spp. on which the insect was found (data not shown). Therefore we could not ascertain whether it is dixenous or monoxenous, like the recently discovered *P. nordicus* (Frolov et al. 2016). While *P. nordicus* only parasitizes insect hosts and is spread between hosts via contaminated faeces and autoinfection rather than plants, *O. lavaterae* feeds primarily by drawing fluid from plant tissues such as leaves and seeds of plants in the Malvaceae family. Thus, while it is possible that *P. oxycareni* is directly transmitted between insect hosts through contact with contaminated faecal matter, it is more parsimonious to assume that it is dixenous and is transmitted between insect hosts via a plant infective stage. In this context it is interesting to note that *P. oxycareni* does not appear to attach to the epithelia of the salivary gland like the monoxenous *P. nordicus* (Frolov et al. 2016), but can be found throughout the lumen of the gland like the promastigotes of the dixenous *P. serpens* which infects tomato plants (Jankevicius et al. 1989). However, irrespective of the confirmation of this lifestyle habit, the discovery that an early diverging species of *Phytomonas* inhabits the cells of the salivary gland suggests that this ability to survive inside host cells is ancestral and may be widespread within the genus.

Of note is the difficulty to detect parasites in the mid-gut of the insect host. As described above, this could be due to the life cycle of *O. lavaterae*. These insects form large overwintering aggregations on the trunks and branches of *Tilia* trees during which time they presumably do not feed (Nedvěd et al. 2014). The individuals from these aggregations begin to disperse in spring, coincident with the flowering of lindens, and they then feed on a range of plants in the Malvaceae (Nedvěd et al. 2014). Therefore it may not be the case that the putative plant host of the newly described parasite is the same species as the plant on which the insect host hibernates. Given the range of plant species that *O. lavaterae* can feed on (Kalushkov and Nedvěd 2010), and the difficulty of obtaining *Phytomonas* parasites from plant tissue (Jaskowska et al. 2015), it may be difficult to ascertain the true plant host range of this species.

In summary the work presented here identifies a new species within the *Phytomonas* clade that is a sister species to all currently sequenced phytomonad species and thus provides new insight into the ancestral biology of the genus (i.e. long slender cell morphology, host cell invasion). Furthermore identification of parasites associated with herbivorous insects that are migrating through Europe highlights the need to focus not only on the movement of insects but also on the presence of associated parasites.

## Methods

**Collection and dissection of bug hosts:** Several hundred individuals of *Oxycarenus lavaterae* were collected from the trunks of several lime trees (*Tilia cordata*, and to a lesser extent *T. platyphyllos*) in the South Moravian region of the Czech Republic in May 2015 and March 2016. *O. lavaterae* remains aggregated on the trunks of trees until late May when the trees begin to flower, this facilitated ease of collection. Population infection rate varied but one bug population in particular, near Sedlec (Mikulov vicinity; 48°46′46.157″N, 16°41′54.389″E, 180 m a.s.l.) was >80% positive for parasites. This host population is the focus of this paper. The insects were euthanized in 70% ethanol, washed in 96% ethanol and a saline solution, and dissected in a saline solution to isolate the salivary glands.

**Cultivation and light microscopy:** Smears of infected salivary glands were fixed with methanol, hydrolysed in 5N HCl for 15 minutes at room temperature and stained with either Giemsa or 4′,6-diamidino-2-pheylindole (DAPI) as has been described previously (Yurchenko et al. 2006). These stained slides were then visually inspected on a fluorescence microscope. Several attempts were made to cultivate *P. oxycareni* in different media with or without antibiotics (Amikacin) including blood agar medium, RPMI, M199, Schneider Medium, BHI Medium supplemented with FCS and Hemin, Warren's Medium, and the mix of these media (RPMI, M199, Schneider Medium and BHI Medium in a 1:1:1:1 ratio, supplemented with FCS). Although the parasites survived for several days in these media, there was no sign of growth and so no culture was obtained. The infected salivary glands were not composed of mixed infections as the same 18S rRNA and SL RNA gene amplicons were cloned from several independent insects and localities (data not shown).

**Transmission and scanning electron microscopy:** Dissected salivary glands were resuspended in 0.1 M phosphate-buffered saline, fixed in 2.5% (v/v) glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.4, for 1 hr at 4 °C and processed for scanning and transmission electron microscopy as

described previously (Yurchenko et al. 2006). Ultrathin sections were analysed and imaged with a FEI Tecnai 12 microscope at 120 kV.

**PCR amplification, cloning, and sequencing:** Total genomic DNA was isolated from the field samples using a DNA isolation kit for cells and tissues (Roche) according to the manufacturer's protocol. Small subunit rRNA gene (SSU rDNA, ~2100 bp) was PCR amplified using the primers S762 (5′-GAC TTT TGC TTC CTC TAD TG-3′)/S763 (5′-CAT ATG CTT GTT TCA AGG AC-3′) (Maslov et al. 1996). The PCR thermocycler settings for DNA amplification were: denaturing at 94 °C for 5 min followed by 35 cycles of 94 °C for 1 min, 55 °C for 90 s, 72 °C for 90 s, and a final elongation at 72 °C for 5 min. For the second round of PCR, 1 μL of the previous reaction was added to 24 μL of a PCR reaction and amplified using the primers TRnSSU-F2 (5′-GAR TCT GCG CAT GGC TCA TTA CAT CAG A-3′) and TRnSSU-R2 (5′-CRC AGT TTG ATG AGC TGC GCC T-3′). The thermocycler settings were: denaturing at 94 °C for 5 min followed by 35 cycles of 94 °C for 1 min, 64 °C for 90 s, 72 °C for 90 s, and a final elongation at 72 °C for 5 min. The amplified DNA sequences were subject to sequencing and the new sequence deposited in GenBank acc. no. KX257483 (18S rRNA).

The splice leader RNA gene was PCR amplified using the primers M167 (5′-GGG AAG CTT CTG ATT GGT TAC TWT A-3′)/M168 (5′-GGG AAT TCA ATA AAG TAC AGA AAC TG-3′) (Westenberger et al. 2004). Amplicons were cloned into the pGEM-T Easy (Promega, Madison, USA) vector system and subject to sequencing. The new sequence was deposited in GenBank acc. no. KX611848 (Spliced leader, SL).

**Phylogenetic analyses:** The 18S rRNA sequences of seven different isolates of *Phytomonas* spp. and 33 insect trypanosomatid species were retrieved from GenBank (Supplementary Material Table S1) and aligned using MAFFT v7.058b (Katoh and Standley 2013). Default parameters were used and the number of iteration steps capped at 1000. The resulting alignment was refined manually using Cinema5 multiple sequence alignment software (Lord et al. 2002) to trim the alignment file to start and end in line with the two ends of the *P. oxycareni* sequence (ie. the final dataset contained 2287 columns covering the full 1926 nucleotide sequence for *P. oxycareni*). Maximum likelihood-based phylogenetic inference was performed in RAxML version 8.2.4 using default parameters and the GTRGAMMA model with 1000 bootstrap replicates (Stamatakis 2006).

To provide additional phylogenetic support, the same sequences were analysed using Mr Bayes (Ronquist et al. 2012). The evolutionary model used was GTR with gamma-distributed rate variation across sites and a proportion of invariable sites. The covarion-like model was used and two runs, each of four chains were initiated and allowed to run for 500,000 generations sampling every 1,000 generations. Convergence was assessed through visual inspection of log-likelihood traces and through analysis of the standard deviation of split frequencies. The analysis had reached stationary phase after 5,000 generations and so this analysis was run for ample time. All other parameters used were the default.

Finally to support the position of the new species as being a sister to all other *Phytomonas* spp., an approximately unbiased (AU) test was performed. Tree 1 (Supplementary Material Fig. S2) was the RAxML tree described above (Fig. 1). Trees 2-4 were identical to Tree 1 apart from the position of *Phytomonas oxycareni* n. sp. within the *Phytomonas* genus (Supplementary Material Fig. S2). The p-value of the AU test for tree 1 was 0.92 compared to <0.13 for trees 2-4, indicating that tree 1 has the greatest probability of being the true tree (Shimodaira 2002).

Additional spliced leader sequences were downloaded from NCBI and aligned as above. Bootstrap support and posterior probabilities were calculated as above. SH values were calculated using RAxML (as above) on the maximum likelihood tree.

## Acknowledgements

EAS is supported by a BBSRC studentship through BB/J014427/1 and a Junior Research Fellowship from the BSPP. SK is a Royal Society University Research Fellow. This work was partially funded by Czech Grant Agency (14-23986S) to JL and by the Ministry of Culture of the Czech Republic (DKRVO 2016/14, National Museum, 00023272) to PK, and by a project from the Czech Ministry of Education (Czech-BioImaging LM2015062 and COST-CZ LD14076). The authors would like to thank Mr and Mrs Kment for hosting EAS, JV and PK during the field expeditions. Finally we would like to thank the anonymous reviewers for their comments and corrections which have improved the manuscript.

## Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.protis.2016.11.002.

## References

**Camargo EP** (1999) *Phytomonas* and other trypanosomatid parasites of plants and fruit. Adv Parasitol **42**:29–112

**Catarino LM, Serrano MG, Cavazzana M, Almeida ML, Kaneshina EK, Campaner M, Jankevicius JV, Teixeira MMG, Itow-Jankevicius S** (2001) Classification of trypanosomatids from fruits and seeds using morphological, biochemical and molecular markers revealed several genera among fruit isolates. FEMS Microbiol Lett **201**:65–72

**Donovan C** (1909) Kala-azar in Madras, especially with regard to its connexion with the dog and the bug (*Conorrhinus*). The Lancet **174**:1495–1496

**Franchini G** (1922) Sur un flagellé de lygoeide (*Crithidia oxycareni* n. sp.). Bulletin de la Socié té de Pathologie Exotique **15**:113–116

**Freymuller E, Milder R, Jankevicius JV, Jankevicius SI, Camargo EP** (1990) Ultrastructural studies on the trypanosomatid *Phytomonas serpens* in the salivary glands of a phytophagous hemipteran. J Protozool **37**:225–229

**Frolov AO, Malysheva MN, Yurchenko V, Kostygov AY** (2016) Back to monoxeny: *Phytomonas nordicus* descended from dixenous plant parasites. Europ J Protistol **52**:1–10

**Jankevicius JV, Jankevicius SI, Campaner M, Conchon I, Maeda LA, Teixiera MMC, Freymuller E, Camargo EP** (1989)

Life cycle and culturing of *Phytomonas serpens* (Gibbs), a trypanosomatid parasite of tomatoes. J Protozool **36**:265–271

**Jaskowska E, Butler C, Preston G, Kelly S** (2015) *Phytomonas*: Trypanosomatids adapted to plant environments. PLOS Pathog **11**:e1004484

**Kalushkov P, Nedvěd O** (2010) Suitability of food plants for *Oxycarenus lavaterae* (Heteroptera: Lygaeidae), a Mediterranean bug invasive in Central and South-East Europe. Comptes Rendus l Acad Bulg des Sci **63**:271–276

**Katoh K, Standley DM** (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol **30**:772–780

**Kment P, Vahala O, Hradil K** (2006) First records of *Oxycarenus lavaterae* (Heteroptera: Oxycarenidae) from the Czech Republic, with review of its distribution and biology. Klapalekiana **42**:97–127

**Lord PW, Selley JN, Attwood TK** (2002) CINEMA-MX: a modular multiple alignment editor. Bioinformatics **18**:1402–1403

**Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V** (2014) Evolution of parasitism in kinetoplastid flagellates. Mol Biochem Parasitol **195**:115–122

**Maslov D, Lukes J, Jirku M, Simpson L** (1996) Phylogeny of trypanosomes as inferred from the small and large subunit rRNAs: implications for the evolution of parasitism in the trypanosomatid protozoa. Mol Biochem Parasitol **75**:197–205

**Nedvěd O, Chehlarov E, Kalushkov P** (2014) Life history of the invasive bug *Oxycarenus lavaterae* (Heteroptera: Oxycarenidae) in Bulgaria. Acta Zool Bulg **66**:203–208

**Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP** (2012) Mrbayes 3. 2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol **61**:539–542

**Shimodaira H** (2002) An approximately unbiased test of phylogenetic tree selection. Syst Biol **51**:492–508

**da Silva RV, Malvezi AD, Augusto LDS, Kian D, Tatakihara VLH, Yamauchi LM, Yamada-Ogatta SF, Rizzo LV, Schenkman S, Pinge-Filho P** (2013) Oral exposure to *Phytomonas serpens* attenuates thrombocytopenia and leukopenia during acute infection with *Trypanosoma cruzi*. PLoS ONE **8**:e68299

**Stamatakis A** (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**:2688–2690

**Vickerman K, Preston TM** (1976) Comparative cell biology of the kinetoplastid flagellates. Biol Kinetoplastida **1**:35–130

**Votýpka J, d'Avila-Levy CM, Grellier P, Maslov DA, Lukeš J, Yurchenko V** (2015) New approaches to systematics of Trypanosomatidae: Criteria for taxonomic (re)description. Trends Parasitol **31**:460–469

**Votýpka J, Maslov DA, Yurchenko V, Jirků M, Kment P, Lun ZR, Lukeš J** (2010) Probing into the diversity of trypanosomatid flagellates parasitizing insect hosts in South-West China reveals both endemism and global dispersal. Mol Phylogenet Evol **54**:243–253

**Westenberger SJ, Sturm NR, Yanega D, Podlipaev SA, Zeledon R, Campbell DA, Maslov DA** (2004) Trypanosomatid biodiversity in Costa Rica: genotyping of parasites from Heteroptera using the spliced leader RNA gene. Parasitology **129**:537–547

**Wheeler RJ, Gluenz E, Gull K** (2011) The cell cycle of *Leishmania*: Morphogenetic events and their implications for parasite biology. Mol Microbiol **79**:647–662

**Yurchenko V, Kostygov A, Havlová J, Grybchuk-Ieremenko A, Ševčíková T, Lukeš J, Ševčík J, Votýpka J** (2015) Diversity of trypanosomatids in cockroaches and the description of *Herpetomonas tarakana* sp. n. J Eukaryot Microbiol **63**:198–209

**Yurchenko VY, Lukes J, Jirku M, Zeledon R, Maslov DA** (2006) *Leptomonas costaricensis* sp. n. (Kinetoplastea: Trypanosomatidae), a member of the novel phylogenetic group of insect trypanosomatids closely related to the genus *Leishmania*. Parasitology **133**:537–546

Available online at www.sciencedirect.com

ScienceDirect

**Appendix 2: Seward E. A. & Kelly S.** 2016. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol.*, **17:**226

Seward and Kelly *Genome Biology* (2016) 17:226
DOI 10.1186/s13059-016-1087-9

Genome Biology

RESEARCH Open Access

# Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms

Emily A. Seward and Steven Kelly*

## Abstract

**Background:** Genomes are composed of long strings of nucleotide monomers (A, C, G and T) that are either scavenged from the organism's environment or built from metabolic precursors. The biosynthesis of each nucleotide differs in atomic requirements with different nucleotides requiring different quantities of nitrogen atoms. However, the impact of the relative availability of dietary nitrogen on genome composition and codon bias is poorly understood.

**Results:** Here we show that differential nitrogen availability, due to differences in environment and dietary inputs, is a major determinant of genome nucleotide composition and synonymous codon use in both bacterial and eukaryotic microorganisms. Specifically, low nitrogen availability species use nucleotides that require fewer nitrogen atoms to encode the same genes compared to high nitrogen availability species. Furthermore, we provide a novel selection-mutation framework for the evaluation of the impact of metabolism on gene sequence evolution and show that it is possible to predict the metabolic inputs of related organisms from an analysis of the raw nucleotide sequence of their genes.

**Conclusions:** Taken together, these results reveal a previously hidden relationship between cellular metabolism and genome evolution and provide new insight into how genome sequence evolution can be influenced by adaptation to different diets and environments.

**Keywords:** Genome evolution, Mutation bias, Elemental selection, Nitrogen metabolism, Synonymous codon use, Comparative genomics, Codon bias, Kinetoplastids, Mollicutes, Stoichiogenomics

## Background

Cells are primarily composed of a few major macromolecules (proteins, RNA, DNA, phospholipids and polysaccharides) that are constructed from monomers (amino acids, nucleotides, etc.). The sequence of these monomers is important for correct molecular function, although there is often flexibility allowing for monomer usage bias. For example, synonymous codons specify the same amino acid and different nucleotide sequences can thus code for the same polypeptide. Multiple competing factors have been proposed to bias the relative use of synonymous codons. These include, but are not limited to, neutral drift (as a result of mutational biases during DNA replication and repair) [1–3], iso-accepting tRNAs [4], translational efficiency and accuracy [5–8], altered gene splicing and protein folding [9], mRNA purine loading as a result of temperature [10, 11] and generation time [12]. Furthermore, multiple factors such as UV radiation, nitrogen fixation and parasitism have been proposed to explain GC variation in prokaryotes [13, 14]. However, the impact of monomer availability (i.e. the relative availability of different nucleotides within the cell) on codon bias has been largely unexplored. We propose that differences in dietary nitrogen should cause concomitant differences in codon bias between closely related organisms whose similar lifestyles exclude alternative explanations.

Though the impact of monomer availability on synonymous codon use has yet to be elucidated, several studies have investigated the elemental composition of macromolecules (protein, DNA, RNA, etc.) using genomics data and bioinformatics tools [15]. Pioneering work in this area focused on protein evolution and demonstrated that as well as the energetic costs associated with synthesising each monomer (amino

* Correspondence: steven.kelly@plants.ox.ac.uk
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

BioMed Central

© The Author(s). 2016 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

acid), the monomer's elemental demands can bias usage in nutrient-limiting environments [16]. Here it was shown in *Escherichia coli* and *Saccharomyces cerevisiae* that enzymes required for metabolic processing of an element have reduced quantities of that element in their sequences [16]. Similar studies in plants have shown that there was a 7.1 % reduction in nitrogen use in amino acid side chains when plant proteins were compared to animal proteins [17]. It was proposed that this reduction was due to differences in the relative nitrogen availability of these two groups of organisms as plants are nitrogen limited in comparison to animals [17]. More generally it has also been seen that there is a negative correlation between protein abundance and the atomic requirements of its constituent monomers [18].

Elemental limitation also has an impact on genetic sequences (DNA and RNA), which are composed of nucleotides that are either scavenged from the organism's environment or built from metabolic by-products. Like amino acids, the biosynthesis of each nucleotide differs in energetic and atomic requirements, with GC pairs consuming more ATP and requiring more nitrogen for biosynthesis than AT pairs [19]. The differences in energetic cost have been proposed to cause differences in the relative abundance of nucleotides within the cell, ultimately leading to nucleotide usage bias in genomic sequences [19]. In support of this hypothesis, it has been shown that imbalances in the relative availability of nucleotides within a cell or restrictions in nucleotide biosynthesis can lead to mutational biases that alter genome nucleotide content [15, 20]. Such differences are manifested as usage biases in organisms that have evolved in conditions where there is a persistent elemental limitation. For example, domesticated crops, which have been cultivated with nitrogen fertilisation for thousands of years, and nitrogen-fixing plants show increased use of nitrogen-rich nucleotides in the transcribed strand of their intergenic regions compared to wild plants, which are relatively nitrogen limited [21]. Furthermore, both protein elemental sparing and codon usage bias have been seen in 148 bacterial species, with significant correlations between carbon and sulfur usage and adaptive codon usage bias [22].

Given that changes in metabolism can lead to changes in the relative abundance of nucleotides, it follows that changes in an organism's diet (the sum of all food consumed by an organism) could have the potential to alter the nucleotide composition of the genome. Specifically, as nucleotides contain different numbers of nitrogen atoms (A/G = 5, C = 3, T/U = 2), differences in dietary nitrogen content should result in concomitant differences in the relative abundance of nucleotides within the cell and thus differences in nucleotide use between species. Moreover, these differences in nucleotide use should be detectable by comparing the nucleotide sequences for orthologous protein-coding genes in organisms that share a common ancestor but have since adapted to utilise different dietary inputs. Here redundancy in the genetic code would allow differences in nucleotide use between species to manifest as changes to nucleotide sequences without necessarily altering the encoded amino acid sequence.

Microbial parasites represent an ideal model system to investigate this phenomenon and determine the effects that changes in dietary input have on the evolution and composition of genome sequences. This is because microbial parasites typically have streamlined metabolisms and often obtain energy from catabolism of a limited set of host biomolecules. Furthermore, closely related parasites often utilise different metabolic strategies and obtain energy from catabolism of different host-derived compounds and even related parasites that colonise the same host niche can obtain energy from catabolism of different inputs [23]. Thus, comparative genomics between parasites that share a common ancestor but have adapted to utilise different host-derived biomolecules has the potential to reveal the effects of changes in diet on the evolution and composition of genome sequences.

Here we provide a global analysis of gene sequence evolution associated with adaptation to changes in diet. We show in two monophyletic groups of parasites (one eukaryotic and one bacterial) that adaptation to diets with differing nitrogen content produces a concomitant effect on nucleotide compositions (and hence nitrogen content) of orthologous RNA sequences. Those parasites that have adapted to low nitrogen content diets have low nitrogen content sequences while those parasites that have adapted to high nitrogen content diets have high nitrogen content sequences. We construct a novel model for synonymous codon use that is sufficient to explain the genome-wide usage of synonymous codons with >90 % accuracy. We show that this model used in a predictive capacity is able to identify the metabolic capacity of related parasites from raw nucleotide sequences. Taken together our findings provide significant new insight into the relationship between diet, metabolism and genome evolution and provide a novel mechanistic explanation for genome-wide patterns of synonymous codon use.

## Results

### Choice of model organisms and inference of orthogroups

To test the hypothesis that differences in diet between organisms can impact on the nucleotide composition of their genomes, a comparative genomic analysis was performed using bacterial (Mollicutes) and eukaryotic (Kinetoplastida) parasites that have adapted to different host niches (Fig. 1; Additional file 1: Figures S1 and S2).

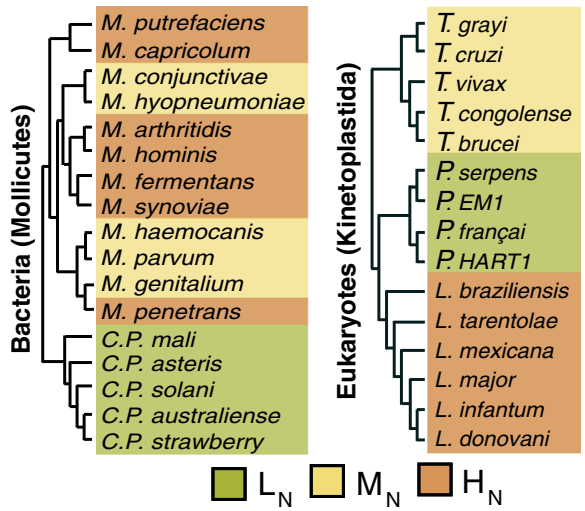



**Fig. 1** Phylogenetic trees of the parasites used in this study shaded according to their host and metabolic strategy. *Green* indicates plant parasites that obtain energy from carbohydrate catabolism (low nitrogen availability, $L_N$). *Yellow* indicates animal parasites that obtain energy from carbohydrate catabolism (medium nitrogen availability, $M_N$). *Orange* indicates animal parasites that obtain energy from amino acid and/or amino sugar catabolism (high nitrogen availability, $H_N$)

These parasites were chosen for analysis because none of the species fix nitrogen and so require nitrogenous compounds obtained from their environment [24, 25]. Furthermore, unlike opportunistic parasites or free-living organisms, these parasites are restricted to host niches that differ in the relative abundance of biologically available nitrogen. Specifically, parasites that colonise plant hosts are nitrogen limited in comparison to those that colonise animal hosts [21]. Additionally, the parasites' pathways for ATP generation differ in liberation of biologically available nitrogen (Additional file 1: Figure S2) [23, 26–29]. The parasites that obtain energy through glycolysis obtain carbon skeletons and re-generate ATP, whereas the parasites that obtain energy through catabolism of arginine or amino sugars additionally obtain biologically available nitrogen (Additional file 1: Figure S2). Thus, the parasites were categorised into three groups depending on host type and whether their metabolism liberates nitrogen. Low nitrogen availability ($L_N$) parasites colonise plants and obtain energy through carbohydrate catabolism, medium nitrogen availability ($M_N$) parasites colonise animals and primarily catabolise carbohydrates, and high nitrogen availability ($H_N$) parasites colonise animals and obtain energy through amino acid or amino sugar catabolism. For further details on the metabolic properties of these parasites see Additional file 2.

A set of orthologous gene groups (orthogroups) covering 15 kinetoplastid genomes (Fig. 1; Additional file 3: Table S1) and an independent set of orthogroups covering 17 Mollicute genomes (Fig. 1; Additional file 3: Table S1) were inferred. Both sets of orthogroups were subject to filtering such that orthogroups were retained for further analysis only if the orthogroup comprised a single-copy gene present in at least three species from each nitrogen availability group (i.e. three $L_N$, three $M_N$ and three $H_N$ species). In this analysis, use of orthologous protein-coding genes allows direct investigation of the effect of adaptation to different metabolic strategies on nucleotide sequences that are derived from a common ancestral state. These genes may also be considered house-keeping genes as the organisms have only one tissue (unicellular) and these genes are conserved across all three groups. The same analysis cannot be done in intergenic regions where ambiguity of orthology prevents paired comparison of sites. Moreover, in the case of bacteria there are too few intergenic regions for robust statistical analyses. Of the 9526 orthogroups identified in kinetoplastids, 3003 satisfied the filtration criteria, encompassing ~40 % of all single-copy genes in these organisms. Similarly, of the 1280 orthogroups identified in the Mollicutes, 168 satisfy the filtration criteria, encompassing 28 % of all single-copy genes in these organisms.

### Low nitrogen availability parasites have low nitrogen content sequences and vice versa

In the kinetoplastid parasites 878,193 orthologous codons in the 3003 conserved single-copy orthologous genes were compared. This revealed a significant difference in the nitrogen content of mRNA between the different nitrogen availability groups (Fig. 2a). On average the mRNAs in $L_N$ parasites cost one fewer nitrogen atom for every 15 codons compared to the same mRNAs in $M_N$ ($p < 0.001$) and one for every seven codons compared to $H_N$ ($p < 0.001$). This corresponds to nitrogen savings of ~0.6 % and ~1.3 %, respectively. Given a kinetoplastid cell has ~61,000 transcripts [30] with an average length of 630 codons, $L_N$ kinetoplastid parasites would use $\sim 5.5 \times 10^6$ fewer nitrogen atoms than $H_N$ parasites to produce the same transcriptome. This is enough nitrogen atoms to make ~8700 average sized proteins. The kinetoplastids also exhibit an analogous difference in the nitrogen content of double-stranded DNA (dsDNA; Fig. 2b). Here genes in $L_N$ parasites cost one less nitrogen atom for every four codons compared to $H_N$, saving roughly 157 nitrogen atoms per gene (~1.1 %). Considering that kinetoplastids are diploid with an average of 8000 genes, this difference in nitrogen cost means that $L_N$ parasites use $\sim 2.5 \times 10^6$ fewer nitrogen atoms to encode the exact same cohort of genes.

A similar phenomenon was observed when comparing the mRNA sequences in Mollicutes parasites (Fig. 2c). Here comparison of 38,255 orthologous codons in 168 orthologous genes revealed that $L_N$ parasites used one fewer nitrogen atom for every nine codons compared to

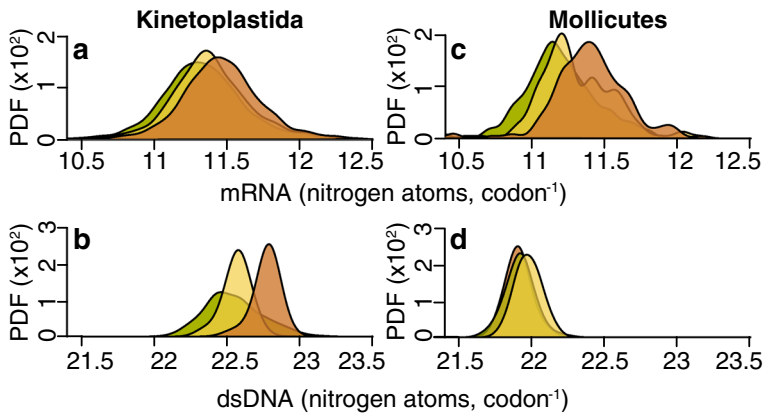



**Fig. 2** Nitrogen availability influences gene sequences. **a** The average mRNA nitrogen content per codon for 3003 orthologous genes in the Kinetoplastida. **b** The average nitrogen content per double stranded codon (*dsDNA*) for the same genes. **c** As in **a** but for 168 orthologous Mollicute genes. **d** As in **b** but for the Mollicutes. The *y-axis* is the probability density function (*PDF*) for the distributions

the same mRNAs in $M_N$ ($p < 0.001$) and one for every five codons compared to $H_N$ ($p < 0.001$). This corresponds to nitrogen savings of ~1 % and ~1.8 %, respectively. Though the Mollicutes exhibit a nitrogen-dependent effect in their mRNAs, the same strong effect is not seen in their dsDNA ($p = 0.025$ when comparing $L_N$ and $H_N$, $p < 0.001$ comparing $M_N$ with either $L_N$ or $H_N$; Fig. 2d). We propose that the absence of a clear nitrogen-dependent effect at the DNA level is due to a strong GC to AT mutation bias thought to be caused by a lack of dUTPase coupled with a reduced ability to correct erroneous dUTP incorporation in DNA [31, 32]. Thus, though the mRNA for the same genes has a lower nitrogen cost in nitrogen-limited species, the high AT nucleotide composition of the DNA reflects the mutational bias imposed by the lack of dUTPase.

For both the kinetoplastids and Mollicutes an analogous difference is also seen in the nitrogen content of the amino acid side chains of these orthologous sites. The $L_N$ parasites use amino acids whose side chains require less nitrogen than the $M_N$ and $H_N$ parasites (Additional file 1: Figure S3). The slight discrepancy between the $M_N$ and $H_N$ parasites can be explained by the reduced use of arginine in the $H_N$ species as they primarily obtain energy from arginine catabolism [23, 33, 34]. This is consistent with previous studies of plant and animal proteins that observed reduced nitrogen content of amino acid side chains in the nitrogen-limited plant species [17, 35].

### Different metabolic strategies in the same host niche cause concomitant differences in gene sequence nitrogen content

To provide further insight into the relationship between metabolism and genome nucleotide composition, an additional analysis was conducted on Mollicutes parasites that occupy the same host niche but obtain energy through different metabolic strategies. Here, three Mollicutes species, *Mycoplasma hominis*, *Mycoplasma genitalium* and *Ureaplasma parvum* were analysed (note *Ureaplasma parvum* is also a Mollicute but a different species to *Mycoplasma parvum* used in the analyses above). Each of these three species reside in the same urogenital tract niche but obtain energy from catabolism of different biomolecules [23]. *M. genitalium* and *U. parvum* metabolise glucose and urea, respectively. However, *M. hominis* has lost the ability to generate ATP via glycolysis and instead generates ATP via nitrogen-liberating arginine catabolism [23].

Using the same methods outlined previously, 51,998 orthologous codons in 227 conserved single-copy orthologous genes (present in each of the three species) were compared (Fig. 3a). This revealed that despite inhabiting the same niche environment, there was a significant difference ($p < 0.001$) in the nitrogen cost of genes, equating to using one fewer nitrogen atom for every six codons in *M. hominis* ($H_N$) compared to *M. genitalium* ($M_N$) (~1.5 %). Since urea metabolism generates ammonia, one could expect *U. parvum* to be a $H_N$ parasite. However, *U. parvum* exports ammonia to drive ATP synthesis, meaning energy generation is linked with export of nitrogen from the cell. Thus, analogous to *M. genitalium*, *U. parvum* is a nitrogen-limited species and uses one fewer nitrogen atom for every five codons compared to *M. hominis* (~1.8 %). As before, the strong mutation bias in Mollicutes means that the same nitrogen-dependent effect is not seen in their dsDNA (Fig. 3b). Taken together, this comparison reveals that, in a common host niche, different metabolic strategies can result in concomitant differences in mRNA nitrogen content.

### Differences in genome-wide patterns of synonymous codon use are explained by selection acting on codon nitrogen content

Given that there is a clear difference in the nitrogen content of genes between different nitrogen availability groups, it was assessed whether this phenomenon could

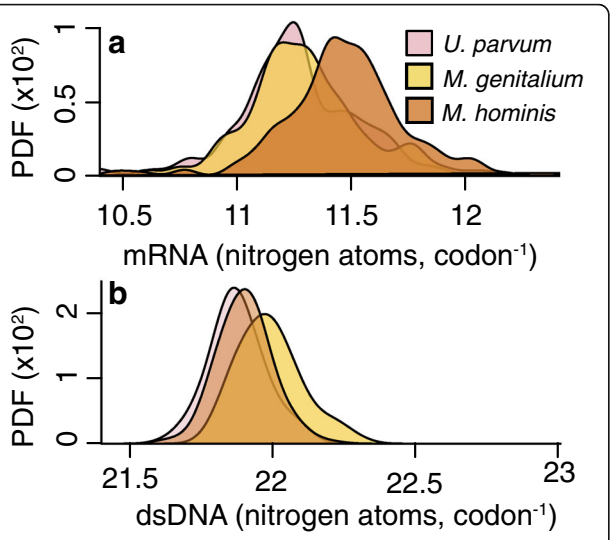



**Fig. 3** Analysis of gene nitrogen content of three parasites that occupy the same host niche but utilise different metabolic strategies. **a** The average mRNA nitrogen content per codon for 227 single copy genes present in each species. **b** The average nitrogen content per double-stranded codon (*dsDNA*) for the same genes. The *y-axis* is the probability density function (*PDF*) for the distributions. *Yellow* indicates an animal parasite that obtains energy from carbohydrate catabolism (medium nitrogen availability, $M_N$). *Orange* indicates an animal parasite that obtains energy from amino acid and/or amino sugar catabolism (high nitrogen availability, $H_N$). *Pink* indicates an animal parasite that obtains energy by exporting ammonia from the cell

be explained by differences in the nitrogen content of synonymous codons. To do this a novel model for genome-wide synonymous codon use was constructed that considers mutation bias and selection acting on the nitrogen content of codons (see the "A model for synonymous codon use under the joint pressures of selection and mutation bias" section in the "Methods"). Using this model, the value of the nitrogen-dependent selection bias ($2N_gs$) and mutation bias ($m$) were found that best explained the real sequence data (see "Methods" for complete model description). Here a negative value for $2N_gs$ indicates that selection is acting to decrease nitrogen content and vice versa.

For the kinetoplastids, application of this modelling approach was able to explain genome-wide patterns of synonymous codon use with >90 % accuracy across all nitrogen availability groups (Fig. 4). Moreover, sequences simulated using these fitted codon use frequencies recapitulated the observed patterns of nitrogen content in mRNA (Fig. 4d) and dsDNA (Fig. 4e (Additional file 1: Figures S4 and S5). Consistent with nitrogen availability, the value of the selection bias for incorporation of nitrogen atoms in gene sequences was most negative in $L_N$ parasites ($2N_gs = -0.09$), intermediate in $M_N$ parasites ($2N_gs = -0.06$) and least negative in $H_N$ parasites ($2N_gs = -0.03$). The distribution of $2N_gs$ parameters for individual species within each group were also significantly different between each group (ANOVA, $p < 0.01$). Thus, differences in nitrogen availability between species are reflected in the relative strengths of the selection bias on codon nitrogen content. Furthermore, mutation bias towards GC was lowest in $L_N$ parasites ($m = 0.67$) and highest in $H_N$ parasites ($m = 0.31$). Importantly, just considering selection acting on the nitrogen content of mRNA (Additional file 1: Figure S5b) or mutation bias (Additional file 1: Figure S5d) in isolation resulted in higher AIC values (Additional file 3: Table S1), indicating the dual parameter model is better. Thus, the pattern of codon use and gene nitrogen content is best explained by a model that considers both selection acting on the mRNA nitrogen content of genes and mutation bias (Fig. 4; Additional file 1: Figure S5e). Furthermore, the statistical significance of selection acting on the nitrogen content of coding sequences was assessed by a permutation test (see "Methods"). This showed that selection acting on the nitrogen content of the mRNA sequences was significant for $L_N$ ($p = 0.004$) and $M_N$ ($p = 0.021$) parasites but was not significant for the $H_N$ kinetoplastid parasites ($p = 0.457$). This is consistent with our findings that indicate $H_N$ kinetoplastids are not under selection to minimise the nitrogen content of their coding sequences. The change in codon bias also accounts for the majority of the difference in genome-wide GC content between species. Specifically, the coding regions constitute ~50 % of the genome in kinetoplastid parasites and thus changes in synonymous codon use account for 61 % of the observed difference in genome-wide GC content between $H_N$ and $L_N$ species (Additional file 3: Table S1).

It should be noted that simulating sequences using perfect genome-derived codon use frequencies (i.e. using 61 constrained parameters; Additional file 1: Figure S5h) results in simulated sequences whose distributions are not significantly different to those obtained in our two-parameter selection-mutation model. Thus, the difference between the distributions of nitrogen content for the real (Fig. 4b) and simulated sequences (Fig. 4e) is a result of factors affecting codon bias in individual genes that are not encapsulated by our genome-wide model.

A similar phenomenon is observed for the Mollicutes, though the fitted mutation bias values are much larger ($m > 3.5$), indicative of a strong GC to AT mutation bias. This high value for $m$ is consistent with the loss of dUTPase and a reduced ability to correct erroneous dUTP incorporation into the genome [31, 32]. The selection-mutation model is capable of explaining genome-wide patterns of codon use with 94 % accuracy across all nitrogen availability groups. Consistent with nitrogen availability, the value of the selection bias for incorporation of nitrogen atoms in gene sequences was

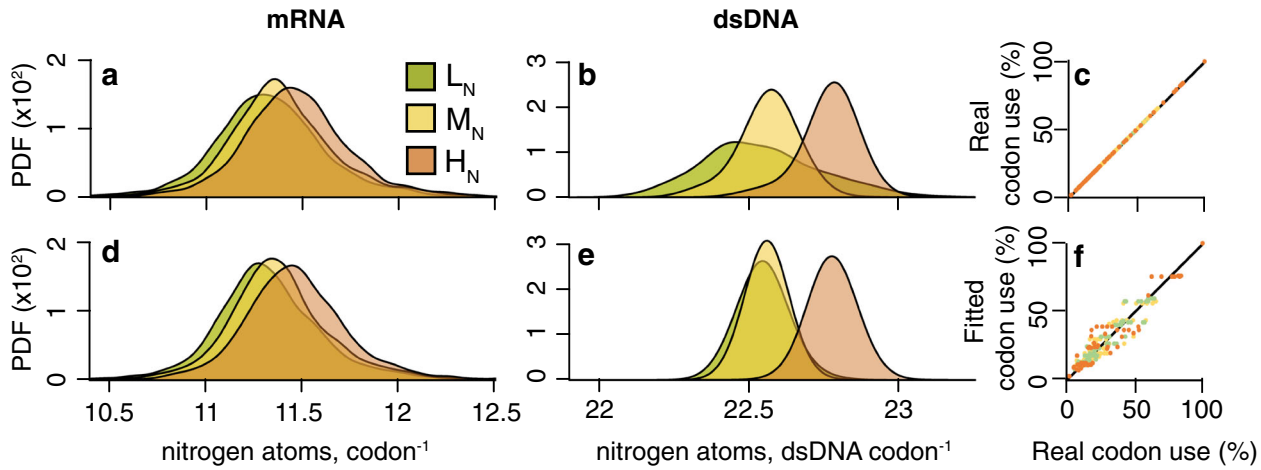



**Fig. 4** A selection-mutation model for synonymous codon use in kinetoplastids explains relative synonymous codon use with ~90 % accuracy and recapitulates the difference in the nitrogen cost of genes. **a** The average mRNA nitrogen content per codon for 3003 orthologous genes in the Kinetoplastida. **b** The average nitrogen content per double-stranded codon (*dsDNA*) for the genes in **a. c** Empirical codon use probabilities (expressed as percentages) plotted against themselves. **d** The average mRNA nitrogen content per codon for sequences simulated using synonymous codon use probabilities derived from fitting the selection-mutation model to the observed sequence data. **e** The average nitrogen content per double-stranded codon for the genes in **d. f** The synonymous codon use probabilities inferred using the selection-mutation model plotted against the empirical codon use probabilities expressed as percentages. *Dot colour* corresponds to nitrogen availability group. $L_N$ $R^2 = 0.92$, $M_N$ $R^2 = 0.88$, $H_N$ $R^2 = 0.90$

most negative in $L_N$ parasites ($2N_es = -0.24$), intermediate in $M_N$ parasites ($2N_es = -0.15$) and least negative in $H_N$ parasites ($2N_es = -0.13$) (Fig. 5; Additional file 1: Figure S5m). The distribution of $2N_es$ parameters for individual species within each group was significantly different when comparing $L_N$ species with $M_N$ or $H_N$ (ANOVA, $p < 0.01$); however, the difference between $M_N$ and $H_N$ species failed to reach significance (ANOVA, $p > 0.05$) (Additional file 1: Figure S6). As for the kinetoplastids, the AIC values of the selection-mutation model were better than for the models that consider either selection or mutation bias individually (Additional file 1: Figure S5J, L; Additional file 3: Table S1). Furthermore, significance testing showed that selection acting on mRNA nitrogen content was significant for all Mollicutes groups ($L_N$ $p = 0.001$, $M_N$ $p = 0.001$, $H_N$ $p = 0.04$). As coding sequences comprise the majority of these Mollicutes genomes (~83 %) the difference in genome-

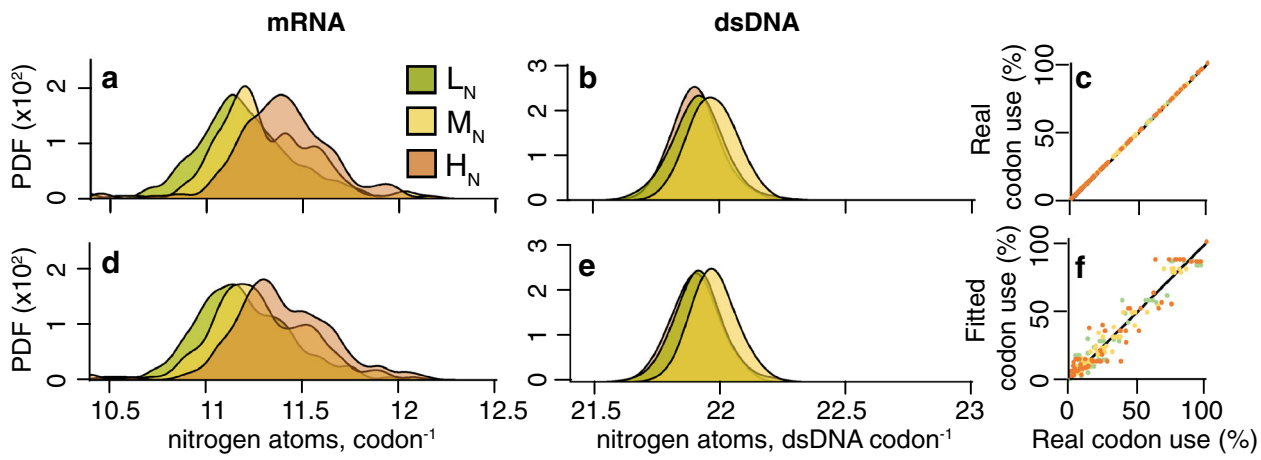



**Fig. 5** Model for synonymous codon use for Mollicutes explains relative synonymous codon use with ~94 % accuracy and recapitulates the difference in the nitrogen cost of genes. **a** The average mRNA nitrogen content per codon for 168 orthologous genes in the Mollicutes. **b** The average nitrogen content per double-stranded codon (dsDNA) for the genes in **a. c** Empirical codon use probabilities (expressed as percentages) plotted against themselves. **d** The average mRNA nitrogen content per codon for sequences simulated using synonymous codon use probabilities derived from fitting the selection-mutation model to the observed sequence data. **e** The average nitrogen content per double-stranded codon (dsDNA) for the genes in **d. f** The synonymous codon use probabilities inferred using the selection-mutation model plotted against the empirical codon use probabilities expressed as percentages. *Dot colour* corresponds to codon use probabilities for each nitrogen availability group. $L_N$ $R^2 = 0.95$, $M_N$ $R^2 = 0.97$, $H_N$ $R^2 = 0.91$

wide GC content between $M_N$ and $L_N$ species is fully attributable to differences in synonymous codon use (Additional file 3: Table S1).

To test whether the observed bias in codon use was also seen more broadly across the genome and not just in the conserved single copy genes, an additional analysis was conducted on all complete coding sequences (Additional file 3: Table S1). The pattern of codon bias was recapitulated for this larger gene set. However, the values obtained from the model when considering all complete coding sequences were less extreme than the values obtained when considering conserved orthologous sequences. This is expected as conserved sites in conserved genes have previously been shown to exhibit stronger codon bias [36].

### Gene expression negatively correlates with selection acting on mRNA nitrogen content

Selection acting on coding sequences is typically considered weak, especially given the low effective populations of the parasites in this study. However, previous studies have shown that selection is detectable in highly expressed genes [37–39] and most theories of codon usage predict that the degree of bias due to selection should increase with gene expression [40]. Given that there is a clear signature of selection acting on nitrogen content genome-wide, it was assessed whether the magnitude of this selection was a function of mRNA abundance. Here, the magnitude of selection acting on the nitrogen content of each gene was compared to the mRNA abundance of that gene. For each species there was a negative correlation between mRNA abundance and the fitted $2N_gs$ (Additional file 1: Figure S7). This shows that the strongest selection to minimise nitrogen content is observed in the most highly expressed genes. Moreover, the slope of the line was greatest for the $L_N$ species, intermediate for the $M_N$ species and weakest for the $H_N$ species. This gene-level analysis is consistent with the genome-wide analysis that showed that $L_N$ species have the greatest selective pressure to minimise nitrogen use.

### Low nitrogen availability ($L_N$) parasites have ribosomal RNA sequences that use the lowest amount of nitrogen

Ribosomal RNA (rRNA) typically constitutes the majority of RNA within a cell. To investigate whether selection acting on nitrogen content extends beyond coding sequences, the total nitrogen content of rRNA per ribosome was calculated. Consistent with the analysis of coding sequences, $L_N$ parasite rRNAs require the lowest amount of nitrogen. In the Mollicutes, $L_N$ parasites used eight fewer nitrogen atoms compared to $M_N$ and 63 fewer atoms compared to $H_N$ parasites per 70S ribosome (Additional file 3: Table S1). This difference is lower than expected when compared to the analysis of protein-coding genes. Given the length of the rRNA sequence analysed, a difference of 77 and 140 nitrogen atoms would have been predicted. This reduced difference is most likely due to structural constraints on rRNA and the fact that it is not composed of codons and so may lack the flexibility provided by synonymous codons.

The same analysis of rRNA sequences was carried out for the kinetoplastids. Consistent with the analysis of the Mollicutes, the RNA component of the 80S ribosome required the least amount of nitrogen in the $L_N$ kinetoplastid parasites. However, due to large insertions in *Trypanosoma cruzi* rRNAs, the $M_N$ parasites required more nitrogen than the $H_N$. These inserted regions increased the total nitrogen content in the *T. cruzi* rRNA by >1500 nitrogen atoms (~7 % more than the other $M_N$ species; Additional file 3: Table S1). Thus, with one exception, the analysis of rRNA genes is consistent with the analysis of protein-coding genes.

### Nitrogen content of nucleotide sequences can predict metabolic capability

Given that the relative use of synonymous codons is affected by selection acting on nitrogen content, it was determined to what extent the selection-mutation model could predict the dietary nitrogen content of an organism. This was tested by analysing four additional Mollicute genomes not included in the original analysis. Each additional species was classified as $H_N$ by model selection through maximum likelihood estimation (Additional file 3: Table S1). To provide support for these classifications, the parasites' genomes were searched for genes required for amino acid and amino sugar catabolism. This revealed that, in contrast to $M_N$ Mollicutes parasites, the genomes of the additional species each encoded complete metabolic pathways for catabolism of either arginine and/or amino sugars (Fig. 6; Additional file 3: Table S1). Moreover, the genes for these pathways were co-located in gene clusters, indicative of genes belonging to the same metabolic pathway (Additional file 1: Figure S8). These results demonstrate the utility of the model for providing information about the metabolic capabilities of an organism from raw nucleotide sequences.

### Selection acting on nitrogen content is independent of selection acting on translational efficiency

Translational selection, which is a function of the number of iso-accepting tRNAs encoded in a genome, has long been considered a major driver of codon bias [8]. To determine how selection acting on nitrogen content acts in concert with selection acting on translational efficiency (tAI), the model was expanded to include tAI as an additional parameter (see "Methods"). For the Mollicutes, unlike the result above where selection acting on nitrogen

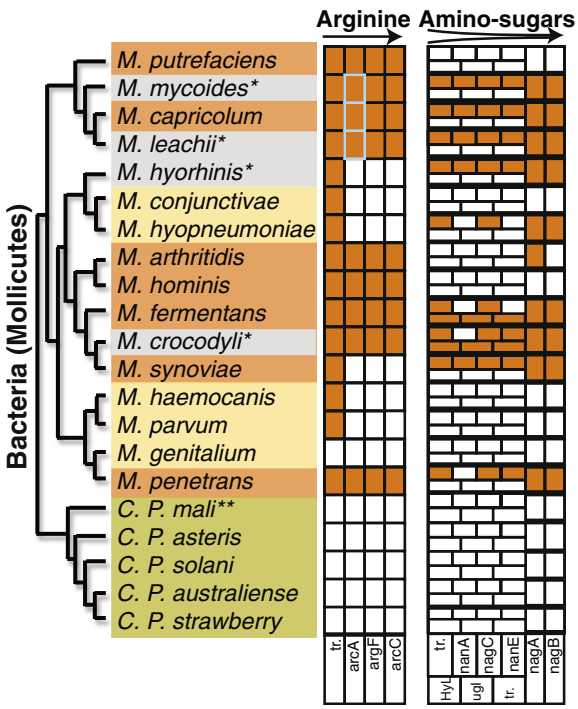



**Fig. 6** Selection-mutation model of codon use can predict the metabolic capacity of parasites from raw nucleotide sequences. $L_N$ (*green*), $M_N$ (*yellow*) and $H_N$ (*orange*) are species with low, medium and high nitrogen availability, respectively. Species denoted with an *asterisk* (*grey*) are the new species, each of which is predicted to be a $H_N$ parasite. The presence of named genes involved in arginine catabolism and amino sugar catabolism is indicated by *orange boxes*. Gene names are provided below the boxes and the abbreviations correspond to the following genes: *tr.* transporter, *arcA* arginine deiminase, *argF* ornithine carbamoyltransferase, *arcC* carbamate kinase, *nanA* N-acetylneuraminate lyase, *nanK/nagC* N-acetylmannosamine kinase, *nanE* N-acetylmannosamine-6-p epimerase, *nagA* N-acetyl-D-glucosamine-6-phosphate amidohydrolase, *nagB* D-glucosamine-6-phosphate aminohydrolase, *HyL* hylauronate lyase, *ugl* glucuronidase. *Grey outlined boxes* encode *aguA*, whose protein product catalyses the same reaction as the product of *arcA* on an equivalent substrate

content was significant for all three parasite groups, it was found that considering tAI values alone or in conjunction with mutation bias was not significantly better than when tAI was omitted ($p > 0.05$). However, when all three parameters (nitrogen content, mutation bias and tAI) were considered together, the model fits the data significantly better than when considering just selection acting on nitrogen content and mutation bias for the $L_N$ and $H_N$ parasites ($p \leq 0.02$). Thus, selection acting on translational efficiency is independent of selection acting on nitrogen content and only provides a significant contribution to codon bias in $L_N$ and $H_N$ species (Additional file 1: Figure S5).

In contrast, for the kinetoplastids it was found that the fit to observed patterns of codon use was significantly better with the inclusion of tAI values in conjunction with mutation bias ($L_N$ $p = 0.018$, $M_N$ and $H_N$ $p = 0$). The contribution of tAI was also significant for all three kinetoplastid parasite groups when all three parameters (nitrogen content, mutation bias and tAI) were considered together ($L_N$ $p = 0.006$, $M_N$ $p = 0.001$, $H_N$ $p = 0$; Additional file 1: Figure S5). Thus, as for the Mollicutes, selection acting on translational efficiency is independent of selection acting on nitrogen content. Furthermore, inclusion of translational efficiency in the model improves overall fit by ~2 % to give an average accuracy of 94.3 %. This compares to a 3.1 % improvement in overall fit when selection acting on nitrogen content is added to the model that only considers mutation bias and translational efficiency.

Selection acting on nitrogen content can explain why the most translationally optimal codons are not always the codons that are most frequently used. For example, in Mollicutes parasites only 33 % of the most frequently used codons for each amino acid are the most translationally efficient while 66 % are those with the lowest nitrogen content (Additional file 1: Figure S9). A similar pattern occurs in the kinetoplastid parasites, although the most translationally efficient codon is the most frequently used codon more often than the most nitrogen-efficient codon (74 % compared to 30 %, respectively). This interplay between translation and nitrogen content is also seen when the relative order of all synonymous codons is analysed in these two parasite groups (Additional file 1: Figure S9). Furthermore, these observations are consistent with the global analysis of codon use presented above which showed that selection acting on nitrogen content was more important than selection acting on translation efficiency in determining patterns of codon bias in Mollicutes, while selection acting on nitrogen content and translation efficiency was required to explain patterns of codon use in kinetoplastids.

## Discussion

Studies on the interactions between diet, metabolism and evolution have primarily focused on the presence or absence of individual genes in the context of specific metabolic pathways. However, the impact of an organism's diet on the evolution of its genes and genome is poorly understood. Here we show that differential nitrogen availability, due to differences in host environment and metabolic inputs, alters synonymous codon usage and thus gene sequence evolution in both bacterial and eukaryotic parasites. Moreover, this impact is sufficient to enable prediction of metabolic inputs of parasites from comparative analysis of the nucleotide composition of orthologous genes.

In this work we provide a novel selection-mutation model for synonymous codon use that builds upon a strong theoretical foundation [41–43]. In this model we have amalgamated multiple factors contributing to

genome-wide GC content into the single variable termed mutation bias (*m*, mutation bias towards AT). Such factors include the bias of an organism's DNA polymerase [44], gene conversion [45], differences in repair efficiency [32] as well as mutational biases during DNA replication [1–3]. We also suggest that differences in nitrogen availability may also contribute to differences in mutation bias through influencing the relative abundance of nucleotides [20]. Considering mutation bias alone was able to recapitulate the observed synonymous codon use with ~90 % accuracy for both the Mollicutes and kinetoplastid parasites. Furthermore, the large differences in mutation bias between kinetoplastids ($m < 1$) and Mollicutes ($m > 3.5$) is able to explain the large differences in observed patterns of codon bias between the two distantly related parasite lineages. Interestingly, the kinetoplastid *m* values are each below 1 ($L_N$ $m = 0.68$, $M_N$ $m = 0.74$, $H_N$ $m = 0.31$) and thus correspond to a bias towards GC. The differences in mutation bias between parasite groups is consistent with differences in nitrogen availability, as a high GC content is equivalent to high nitrogen content of the dsDNA. An analogous nitrogen-dependent difference in *m* is not seen in the Mollicutes. We propose that this is due to the strong AT mutation bias (*m* values all greater than 3.5) that constrains dsDNA nitrogen within a narrow range of values compared to the kinetoplastids.

Due to the complementary nature of DNA, a change on either DNA strand will cause a corresponding change on the other strand. Therefore, mutation bias alone was unable to produce the differences in the nitrogen content of the coding strand (i.e. the mRNA) that was observed between species with different nitrogen availabilities. As shown in Eq. 2, selection depends on $N_g$, the effective number of genes at the locus in the population, which is linked to the effective population size ($N_e$) of an organism [46]. Organisms with low long-term effective population sizes have a reduced impact from selection due to the greater impact of random genetic drift. Thus, $N_g$ plays an important role in determining the role of selection in biased codon usage. As has been noted before, however, evaluating the long-term $N_g$ value for an organism is very difficult [47]. Eukaryotes have lower $N_g$ values than prokaryotes and parasites in general have lower $N_g$ values than their free-living counterparts due to clonal life stages and bottlenecks during transmission. Our model evaluates the selection bias acting on nitrogen content using a composite parameter ($2N_gs$). Thus, the value of the selection coefficient *s* is linearly dependent on estimates of $N_g$ (i.e. increasing $N_g$ by a factor of 10 decreases *s* by a factor of 10). It is interesting to note that estimates of $N_g$ for prokaryotes and unicellular eukaryotes differ by a factor of 10 [46], similar to the magnitude of difference we see between $2N_gs$ for the Mollicutes and the kinetoplastids, indicating that the selection coefficient *s* may be similar for the two distantly related groups.

Previous studies investigating the role of selection in codon bias have revealed that selection acting on translational efficiency in Mollicutes is marginal [47]. Although codon biases in prokaryotic genomes are associated with gene expression levels [48, 49], in some cases the optimal codons disagree with the tRNA composition. These observations support the results presented here which show that, for Mollicutes, inclusion of tAI values does not significantly improve the fit of the model unless it is considered in conjunction with both mutation bias and selection acting on the nitrogen content of coding sequences. Our finding that selection acts on the nitrogen content of codons provides a novel mechanism that links codon usage bias to metabolism and environment. Furthermore, as the model developed here is sufficient to enable prediction of metabolic inputs from gene sequences, it may have application in interrogating metagenome data and genome data from shotgun sequencing of microbial communities where metabolic requirements are unknown.

Though the selection-mutation model provides considerable explanatory power for the species used in this analysis, it does not perfectly re-capitulate the observed patterns of codon use. This is most likely due to the fact that specific sites within a gene will be under different pressures that cannot be captured by a genome-wide approach. For example, factors indirectly related to protein translation, such as mRNA secondary structures at the 5′ region of a gene, have been shown to be under selection for efficient binding of ribosomes to mRNAs and hence can have a weak effect on the frequency of codon usage at those sites [50]. A more complex model could include variation in codon bias between genes due to gene-specific selective pressures such as splice site conservation, mRNA stability, ribosome binding and mRNA abundance. Taken together these factors may account for the ~6 % of missing variation not explained by the selection-mutation model presented here. Incorporation of these factors into the model would be an interesting avenue of future research.

Thermophilic bacteria purine-load their genomic sequences to the extent that amino acid composition is affected [10]. However, this effect is not seen in the mRNA of mesophilic organisms [51, 52] and so would not be expected to feature in the dataset analysed here. For example, the difference observed between $M_N$ and $H_N$ parasites cannot be due to temperature as both groups infect animal hosts with very similar (if not identical) temperatures. Furthermore, some of the $H_N$ (*Mycoplasma crocodyli* and *Leishmania tarentolae*) and $M_N$ (*Trypanosoma grayi*) species in both the Mollicutes and the

kinetoplastids infect cold-blooded reptiles and thus would have host temperatures more similar to plant-infecting $L_N$ parasites than to warm blooded animals. Even though these parasites infect cold-blooded animals, their nitrogen use profiles are consistent with their metabolic group rather than their host temperature. Finally, conducting our analysis on parasites in the same ecological niche revealed that, at the same temperature, in the same microenvironment, the parasites exhibited different nucleotide nitrogen content consistent with their dietary nitrogen availability. These results indicate that while temperature may be important in extreme environments, temperature is not a factor in the comparisons presented here. This is consistent with previous analyses that showed that even at relatively freely evolving sites, mRNA GC content did not appear to be adapted to the thermal environment [52].

## Conclusions

This analysis demonstrates via multiple complementary approaches that differential nitrogen availability, due to differences in host environment and metabolic inputs, contributes to changes in codon bias and genome composition. Specifically, adaptation to low nitrogen availability results in reduced nitrogen content in nucleotide sequences. These results reveal a previously hidden relationship between cellular metabolism and genome evolution and provide new insight into how genome sequence evolution can be influenced by adaptation to different diets.

## Methods

### Data sources

We obtained 17 Mollicutes genomes from the NCBI GenBank. These comprised four plant glycolytic parasite species (*Candidatus Phytoplasma asteris* [53], *Candidatus Phytoplasma australiense* [54], *Candidatus Phytoplasma mali* [55], *Candidatus Phytoplasma solani* [56], *Candidatus Phytoplasma strawberry* [57]), five animal glycolytic parasite species (*Mycoplasma conjunctivae* [58], *Mycoplasma genitalium* [59], *Mycoplasma haemocanis* [60], *Mycoplasma hyopneumoniae* [61], *Mycoplasma parvum* [62]) and seven parasite species known to obtain energy from catabolism of amino acids or amino sugars (*Mycoplasma arthritidis* [63], *Mycoplasma capricolum* [PRJNA16208], *Mycoplasma fermentans* [64], *Mycoplasma hominis* [23], *Mycoplasma penetrans* [65], *Mycoplasma putrefaciens* [66], *Mycoplasma synoviae* [67]). A further four parasite species were used for testing the predictive capacity of the model for synonymous codon use (*Mycoplasma crocodyli* [68], *Mycoplasma hyorhinis* [69], *Mycoplasma leachii* [70], *Mycoplasma mycoides* [70]).

15 kinetoplastid genomes were obtained online from TriTrypDB [71], NCBI genbank or the European Nucleotide Archive. These comprised four plant glycolytic parasite species (*Phytomonas EM1* [GCA_000582765] [72], *Phytomonas françai* [PRJNA343003], *Phytomonas HART1* [GCA_000982615] [72], *Phytomonas serpens* [PRJNA80957], 5 animal glycolytic parasite species (*Trypanosoma brucei* [PRJNA15565], *Trypanosoma congolense* [PRJNA12958], *Trypanosoma cruzi* [PRJNA15540/PRJNA11755], *Trypanosoma grayi* [PRJNA258390], *Trypanosoma vivax* [PRJNA12957]) and six parasite species who obtain energy primarily from catabolism of amino acids (*Leishmania braziliensis* [PRJNA19185], *Leishmania donovani* [PRJNA171503], *Leishmania infantum* [PRJNA19187], *Leishmania major* [PRJNA10724], *Leishmania mexicana* [PRJNA172192], *Leishmania tarentolae* [PRJNA15734]).

### Inference of orthogroups and construction of multiple sequence alignments

The predicted amino acid sequences for each species were subject to orthogroup inference using OrthoFinder [73] using the default program parameters. Single copy genes were selected for analysis to ensure orthology and so that paired comparisons could be made; i.e. a single-copy orthologous gene that is present in two different species can be treated as a paired observation. Single copy gene orthogroups were further filtered to retain those that had representation from at least three species per group ($L_N$, $M_N$ and $H_N$). Protein sequences for these orthogroups were aligned using MergeAlign [74]. The corresponding coding sequences were re-threaded back through the aligned amino acid sequences using custom Perl scripts. These multiple sequence alignments were then filtered so that only un-gapped columns that obtained a MergeAlign column score of >0.75 were retained for further analysis. These stringent filtration criteria ensured that only high accuracy, unambiguously aligned orthologous positions were used for all analyses. The accession numbers for the full set of orthogroups used in this analysis are provided in Additional file 3: Table S1.

### Evaluation of nitrogen content of nucleotide sequences

The filtered multiple sequence alignments above were used to calculate the number of nitrogen atoms used per codon, per gene per species. The number of nitrogen atoms per codon per gene was evaluated as the arithmetic mean of the number of nitrogen atoms in the filtered aligned codons for that gene described above. The average number of nitrogen atoms contained within the mRNA and the dsDNA were recorded for each gene. These data were plotted as probability density functions using the R density distribution plot function with the total area under each curve equal to one.

### Analysis of rRNA

A database of representative rRNA sequences was generated and blasted against the genomes of all the parasites in this study to find the locations of the rRNAs. In the event of no or partial blast hits, sequences were

downloaded from NCBI and the accession numbers noted in Additional file 3: Table S1. Sequences were then aligned using MAFFT [75] to identify the true start and end of the rRNA molecules. The nitrogen content of these sequences was calculated. Due to difficulties in sequencing and assembling repetitive rDNA loci, some species did not have complete sequences to include in this analysis. Those were labelled NF (not found).

### Statistical tests

Given that single copy orthologous genes present in different species can be treated as paired observations, Wilcoxon signed-rank tests were used to compare nitrogen content between different parasite groups. In each case the null hypothesis was that the difference between the two groups was due to chance (symmetric around zero). The alternative hypothesis was that the difference in nitrogen content between each group was not due to chance. In all cases, the test used was two-tailed so that either a greater or lesser nitrogen content difference would reject the null hypothesis. Pairing of samples is justified as the paired observations (genes) are orthologous and descended from the same common ancestor under different environmental and metabolic conditions.

Goodness of fit and the statistical significance of the inclusion of additional parameters to the model were assessed by comparison of AIC values and by using a permutation test, respectively. For the permutation test, the log likelihood values obtained by the model when run with real values were compared with the log likelihood values obtained by the model when it was run with shuffled/randomised values. To analyse the significance of the inclusion of nitrogen selection to the model, the codon nitrogen contents were calculated and then shuffled to randomly assign the values to each codon. The model was then fit to the data using these randomised values and the log likelihood compared to the log-likelihood obtained using the real values. This was repeated for 1000 independently shuffled sets. The same principle was applied to significance testing of the tAI values. An example of the distributions generated when codon nitrogen content was shuffled is provided in Additional file 1: Figure S10.

### A model for synonymous codon use under the joint pressures of selection and mutation bias

To determine whether nitrogen availability influences interspecies variation in codon use and nucleotide content, a model for synonymous codon use was constructed. This model considers the selection bias acting to modulate a codon's nitrogen content and an organism's mutation bias. The system of equations describing the model are as follows.

#### *Synonymous codon use considering selection acting on mRNA nitrogen content*

Here we consider that selection acts to bias synonymous codon use in proportion to the number of nitrogen atoms contained within each codon, i.e.:

$$S(\mathcal{C}_i) = sN_{mRNA} \tag{1}$$

where $S(\mathcal{C}_i)$ is a measure of the relative fitness of codon $\mathcal{C}_i$, with $N_{mRNA}$ being the number of nitrogen atoms in codon $\mathcal{C}_i$ and $s$ being the selection coefficient. Following previous published work [41–43], we model the selection bias towards codon $\mathcal{C}_i$ as:

$$\alpha(\mathcal{C}_i) = e^{2N_g S(\mathcal{C}_i)} \tag{2}$$

where $\alpha(\mathcal{C}_i)$ is the selection bias towards codon $\mathcal{C}_i$ and $N_g$ is the effective number of genes at a locus. Only considering this selection bias, we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as:

$$p(\mathcal{C}_i \mid \theta) = \frac{\alpha(\mathcal{C}_i)}{\sum_{\theta}\alpha(\mathcal{C})} \tag{3}$$

That is, the probability of observing codon $\mathcal{C}_i$ is the selection bias towards codon $\mathcal{C}_i$ divided by the sum of selection biases for all codons encoding amino acid θ. Equation 3 satisfies the law of total probability such that the sum of the probabilities of observing of all the codons that encode the same amino acid sum to one.

#### *Synonymous codon use considering mutation bias only*

Mutation bias is known to be influenced by a range of factors including but not limited to the bias of an organism's polymerase-α subunit [44], gene conversion [45] and differences in repair efficiency [32]. We propose that nitrogen-mediated changes in nucleotide pools also contribute to this mutation bias, as changes in nucleotide pools result in changes in mutation bias [20]. For example, the amount of biologically available nitrogen within a cell could alter the relative abundance of nucleotides via enzymes such as CTP synthase that catalyse nitrogen-dependent nucleotide interconversion of UTP and CTP. Here we have amalgamated these factors into the single variable $m$.

$$\delta = \frac{m}{m+1} \tag{4}$$

where $\delta$ is the probability that a particular site is A or T given a mutation bias towards AT of $m$ as previously described [46]. Due to base pairing, the probability of A or T is equivalent. This equation assumes that the nucleotide composition of the genome is at equilibrium and that the mutation rate per site is independent of the status of neighbouring sites [46]. For example, if there is

no mutation bias towards AT or GC, $m$ will be 1 and $\delta$ will be 0.5 and thus there is an equal likelihood of any site being AT or GC. We model the mutation bias towards codon $\mathcal{C}_i$ as:

$$\beta(\mathcal{C}_i) = \delta^{AT}(1-\delta)^{GC} \quad (5)$$

where $\beta(\mathcal{C}_i)$ is the mutation bias towards codon $\mathcal{C}_i$, $AT$ is the number of A or T nucleotides in codon $\mathcal{C}_i$ and $GC$ is the number of G or C nucleotides in codon $\mathcal{C}_i$. Considering only mutation bias we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as:

$$p(\mathcal{C}_i \mid \theta) = \frac{\beta(\mathcal{C}_i)}{\sum_\theta \beta(\mathcal{C})} \quad (6)$$

That is, the probability of observing codon $\mathcal{C}_i$ is the mutation bias towards codon $\mathcal{C}_i$ divided by the sum of mutation biases for all codons encoding amino acid θ. Equation 6 also satisfies the law of total probability such that the sum of the probabilities of observing all the codons that encode the same amino acid sum to one. For example, if $m = 3$ then $\delta = 0.75$ and we consider amino acid C (encoded by codons TGC and TGT), then the mutation bias towards codon TGC = $\beta(TGC) = 0.75^1(1 - 0.75)^2 = 0.047$ and the mutation bias towards codon TGT = $\beta(TGT) = 0.75^2(1 - 0.75)^1 = 0.141$. Thus, the genome-wide probability of observing codon TGC = $\frac{0.047}{0.047+0.141} = 0.25$ and the genome-wide probability of observing codon TGT = 0.75.

#### *A model for synonymous codon use under the joint pressures of selection and mutation bias*

We model the bias towards codon $\mathcal{C}_i$ under the joint pressures of selection and mutation as the product of Eqs. 2 and 5.

$$\gamma(\mathcal{C}_i) = \alpha(\mathcal{C}_i)\beta(\mathcal{C}_i) \quad (7)$$

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as:

$$p(\mathcal{C}_i \mid \theta) = \frac{\gamma(\mathcal{C}_i)}{\sum_\theta \gamma(\mathcal{C})} \quad (8)$$

It should be noted that selection in this model only considers the nitrogen content of a codon and does not consider other factors such as biased gene conversion [46]. However, kinetoplastids primarily reproduce by clonal expansion and the prokaryotic genomes are haploid; thus, gene conversion may have limited impact in these organisms.

#### *Calculation of codon tRNA adaptation index values*

The tRNA adaptation index (tAI) [76] of a codon takes into account both the abundance of iso-accepting tRNAs and wobble-base pairing to evaluate the efficiency of translation of a given codon. Using the equation developed by dos Reis et al. [77] below and the optimised $s_{ij}$ values obtained by Tuller et al. [78], tAI values for each codon were evaluated:

$$\omega(C_i) = \sum_{j=1}^{n_i} \left(1 - s_{ij}\right) tGCN_{ij} \quad (9)$$

where $\omega(C_i)$ is the absolute adaptiveness value for each codon $C_i$ (referred to in the rest of the text as the tAI value), $n_i$ is the number of tRNA isoacceptors that recognise codon $C_i$, $tGCN_{ij}$ is the gene copy number of the $j^{\text{th}}$ tRNA that recognises codon $C_i$, and $s_{ij}$ is the selective constraint on the efficiency of codon-anticodon coupling.

We model the translational selection bias towards codon $\mathcal{C}_i$ as:

$$\eta(\mathcal{C}_i) = e^{2N_g\sigma\omega(\mathcal{C}_i)} \quad (10)$$

where $\omega(\mathcal{C}_i)$ is the translational selection bias towards codon $\mathcal{C}_i$, $\sigma$ is the selection coefficient and $N_g$ is the effective number of genes at a locus.

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as:

$$p(\mathcal{C}_i \mid \theta) = \frac{\eta(\mathcal{C}_i)}{\sum_\theta \eta(\mathcal{C})} \quad (11)$$

When considering all three parameters (mutation bias, selection acting on the nitrogen content of coding sequences and translational selection) we model the bias towards codon $\mathcal{C}_i$ as the product of Eqs. 2, 5 and 10.

$$\varepsilon(\mathcal{C}_i) = \alpha(\mathcal{C}_i)\beta(\mathcal{C}_i)\eta(\mathcal{C}_i) \quad (12)$$

As above we evaluate the genome-wide probability of observing codon $\mathcal{C}_i$ for amino acid θ as:

$$p(\mathcal{C}_i \mid \theta) = \frac{\varepsilon(\mathcal{C}_i)}{\sum_\theta \varepsilon(\mathcal{C})} \quad (13)$$

#### Model fitting and implementation

Using the system of equations in the model, the parameters ($2N_gs$, $m$ and tAI) were estimated for each of the parasite groups using a maximum likelihood approach. The models for both the Mollicute and kinetoplastid parasites each contain a maximum of three free parameters (selection acting on nitrogen content, mutation bias and translational efficiency) and thus a brute-force parameter search was conducted to find their optimal values. Here, the likelihood of observing the set of sequences contained within each parasite group was evaluated given the

model for synonymous codon use and the values of the parameters. It was evaluated as follows:

$$\mathscr{L}(s, m|X) = \prod_{\mathcal{C}_i} p(\mathcal{C}_i|\theta)^{N_{\mathcal{C}_i}}$$

where $X$ is the set of coding sequences for a given species and $N_{\mathcal{C}_i}$ is the number of times that codon $\mathcal{C}_i$ occurs in the set of sequences $X$. The optimal parameter values were determined as those with the maximum likelihood. This was applied to look at both orthologous genes (the same set as those described in the "Inference of orthogroups and construction of multiple sequence alignments" section) and the full set of coding sequences. Source code and data files for this analysis are available from the Zenodo research data repository (https://doi.org/10.5281/zenodo.154493).

### Classification of additional species using the metabolic model for synonymous codon use

Four additional Mollicutes genomes not included in the initial analysis were downloaded from NCBI to test the ability of the model for synonymous codon use to predict the metabolic properties of these organisms from analysis of codon use. These species were *M. crocodyli*, *Mycoplasma hyorhinis*, *Mycoplasma leachii* and *Mycoplasma mycoides*. Based on literature evidence and phylogeny (Fig. 6), it was expected that some of the additional species would be classified as $H_N$ and some as $M_N$ parasites. Using the system of equations described above and the values obtained for the dependency parameters ($2N_e s$ and $m$) for each of the $L_N$, $M_N$ and $H_N$ Mollicutes parasite groups, a likelihood that each species belonged to each group was calculated (Additional file 3: Table S1). The model with the highest likelihood was determined to be $H_N$ in all instances. This classification was confirmed using a Wilcoxon signed-rank test on the nitrogen cost of the mRNA. Each of the additional species was significantly different ($p < 0.001$) from the $L_N$ and $M_N$ groups and not significantly different ($p > 0.05$) from the $H_N$ group. The one exception to this was *M. crocodyli*. This species had the highest mRNA nitrogen cost of any Mollicutes species in this analysis and was significantly higher than the other species in the $H_N$ group. This may indicate increased dependence on nitrogen liberating metabolic pathways or an increased availability of nitrogen in the host environment.

## Additional files

**Additional file 1:** Supplemental figures. (PDF 3133 kb)

**Additional file 2:** Supplemental information on parasite metabolism. (PDF 425 kb)

**Additional file 3:** Sheet 1: Accession numbers and corresponding orthogroups for all kinetoplastid species used in this analysis. Sheet 2: Accession numbers for the genes required for arginine and amino sugar metabolism in the Mollicutes. Sheet 3: The model was run on the dataset of orthologous genes (as described in the "Inference of orthogroups and construction of multiple sequence alignments" section in the "Methods") and it was also run separately on all complete coding sequences. A complete coding sequence was considered as any coding sequence with start and stop codons whose length was divisible by 3 and longer than 30 nucleotides. This sheet shows the parameters obtained for each of the $L_N$, $M_N$ and $H_N$ groups for both the Mollicutes and kinetoplastid parasites. Highlighted in *yellow* are the parameters obtained for the original two-parameter model that only considers selection acting on nitrogen content in coding sequences and mutation bias. Highlighted in *green* are the values obtained for the three-parameter model, which also considers selection acting on translational efficiency (tAI). The model parameters that produce the highest log likelihood values and the lowest AIC values are the best fit to the observed data. Sheet 4: Genome locations, nitrogen use and aligned sequences for Mollicutes 5S, 16S and 23S rRNA. Total nitrogen used in rRNA per 70S ribosome is given in the top left corner. Sheet 5: Genome locations, nitrogen use and aligned sequences for kinetoplastid 5.8S, 18S, 23S alpha and 23S beta rRNA. Total nitrogen used in rRNA per 80S ribosome is given in the top left corner. Sheet 6: Accession numbers for the genes required for arginine and amino sugar metabolism in the Mollicutes. (XLSX 639 kb)

#### Acknowledgements
The authors would like to thank Michael Lynch, Michael Bulmer, Paul Higgs and the anonymous reviewers for their comments and advice on the manuscript.

#### Funding
EAS is supported by a BBSRC studentship through BB/J014427/1. SK is a Royal Society University Research Fellow. Work in SK's lab is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement number 637765.

#### Availability of data and materials
Data required to interpret and replicate this analysis are available in the supplementary table and supplementary files provided. Source code and data files are also available from the Zenodo research data repository (https://doi.org/10.5281/zenodo.154493).

#### Authors' contributions
SK conceived the study, EAS conducted the analysis, SK and EAS wrote the manuscript. Both authors read and approved the final manuscript.

#### Competing interests
The authors declare that they have no competing interests.

#### Ethics approval and consent to participate
Not applicable.

Received: 21 July 2016 Accepted: 12 October 2016
Published online: 15 November 2016

#### References
1. Francino MP, Ochman H. Isochores result from mutation not selection. Nature. 1999;400:30–1.
2. Eyre-Walker AC. An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol. 1991;33:442–9.
3. Rao Y, Wu G, Wang Z, Chai X, Nie Q, Zhang X. Mutation bias is the driving force of codon usage in the Gallus gallus genome. DNA Res. 2011;18:499–512.
4. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A. 2004;101:12588–91.
5. Sørensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in Escherichia coli. J Mol Biol. 1989;207:365–77.
6. Hu H, Gao J, He J, Yu B, Zheng P, Huang Z, et al. Codon optimization significantly improves the expression level of a keratinase gene in Pichia pastoris. PLoS One. 2013;8(3):e58393.
7. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994;136:927–35.

8. Shah P, Gilchrist MA. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. Proc Natl Acad Sci U S A. 2011;108:10231–6.
9. Novoa EM, Ribas de Pouplana L. Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet. 2012;28:574–81.
10. Lao PJ, Forsdyke DR. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res. 2000;10:228–36.
11. Paz A, Mester D, Baca I, Nevo E, Korol A. Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. Proc Natl Acad Sci U S A. 2004;101:2951–6.
12. Subramanian S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. Genetics. 2008;178:2429–32.
13. Rocha EPC, Feil EJ. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? PLoS Genet. 2010;6:1–4.
14. McEwan CE, Gatherer D, McEwan NR. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. 1998;128:173–8.
15. Elser JJ, Acquisti C, Kumar S. Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition. Trends Ecol Evol. 2011;26:38–44.
16. Baudouin-Cornu P, Surdin-Kerjan Y, Marliere P, Thomas D. Molecular evolution of protein function. Science. 2001;293:297–300.
17. Acquisti C, Kumar S, Elser JJ. Signatures of nitrogen limitation in the elemental composition of the proteins involved in the metabolic apparatus. Proc Biol Sci. 2009;276:2605–10.
18. Li N, Lv J, Niu DK. Low contents of carbon and nitrogen in highly abundant proteins: Evidence of selection for the economy of atomic composition. J Mol Evol. 2009;68:248–55.
19. Rocha EPC, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18:291–4.
20. Buckland RJ, Watt DL, Chittoor B, Nilsson AK, Kunkel TA, Chabes A. Increased and imbalanced dNTP pools symmetrically promote both leading and lagging strand replication infidelity. PLoS Genet. 2014;10:e1004846.
21. Acquisti C, Elser JJ, Kumar S. Ecological nitrogen limitation shapes the DNA composition of plant genomes. Mol Biol Evol. 2009;26:953–6.
22. Bragg JG, Quigg A, Raven JA, Wagner A. Protein elemental sparing and codon usage bias are correlated among bacteria. Mol Ecol. 2012;21:2480–7.
23. Pereyre S, Sirand-Pugnet P, Beven L, Charron A, Renaudin H, Barré A, et al. Life on arginine for Mycoplasma hominis: clues from its minimal genome and comparison with other human urogenital mycoplasmas. PLoS Genet. 2009;5(10):e1000677.
24. Creek DJ, Nijagal B, Kim DH, Rojas F, Matthews KR, Barrett MP. Metabolomics guides rational development of a simplified cell culture medium for drug screening against trypanosoma brucei. Antimicrob Agents Chemother. 2013;57:2768–79.
25. Razin S, Knight BC. A partially defined medium for the growth of Mycoplasma. J Gen Microbiol. 1960;22:492–503.
26. Jaskowska E, Butler C, Preston G, Kelly S. Phytomonas: trypanosomatids adapted to plant environments. PLoS Pathog. 2015;11:e1004484.
27. Ginger M, Fairlamb A, Opperdoes F. Comparative genomics of trypanosome metabolism. Trypanosomes: after the genome. 2007;373-417
28. Arraes FBM, de Carvalho MJA, Maranhão AQ, Brígido MM, Pedrosa FO, Felipe MSS. Differential metabolism of Mycoplasma species as revealed by their genomes. Genet Mol Biol. 2007;30:182–9.
29. Kube M, Mitrovic J, Duduk B, Rabus R, Seemüller E. Current view on phytoplasma genomes and encoded metabolism. Sci World J. 2012;2012:1–25.
30. Kolev NG, Franklin JB, Carmi S, Shi H, Michaeli S, Tschudi C. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. PLoS Pathog. 2010;6:1–15.
31. Pollack JD, Williams MV, McElhaney RN. The comparative metabolism of the mollicutes (Mycoplasmas): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. Crit Rev Microbiol. 1997;23:269–354.
32. Williams MV, Pollack JD. A mollicute (Mycoplasma) DNA repair enzyme: purification and characterization of uracil-DNA glycosylase. J Bacteriol. 1990;172:2979–85.
33. Fiebig M, Kelly S, Gluenz E. Comparative life cycle transcriptomics revises Leishmania mexicana genome annotation and links a chromosome duplication with parasitism of vertebrates. PLoS Pathog. 2015;11:e1005186.
34. Wanasen N, Soong L. L-arginine metabolism and its impact on host immunity against Leishmania infection. Immunol Res. 2008;41:15–25.
35. Elser JJ, Fagan WF, Subramanian S, Kumar S. Signatures of ecological resource availability in the animal and plant proteomes. Mol Biol Evol. 2006;23:1946–51.
36. Stoletzki N, Eyre-Walker A. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol Biol Evol. 2007;24:374–81.
37. Ran W, Higgs PG. Contributions of speed and accuracy to translational selection in bacteria. PLoS One. 2012;7(12):e51652.
38. Ran W, Higgs PG. The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. Mol Biol Evol. 2010;27:2129–40.
39. Higgs PG, Ran W. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol Biol Evol. 2008;25:2279–91.
40. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. Nat Rev Genet. 2009;10:715–24.
41. Shields DC. Switches in species-specific codon preferences: the influence of mutation biases. J Mol Evol. 1990;31:71–80.
42. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129:897–907.
43. Li WH. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 1987;24:337–45.
44. Worning P, Jensen LJ, Hallin PF, Stærfeldt H, Ussery DW. Environmental microbiology. Environ Microbiol. 2006;8:2912.
45. Galtier N. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 2003;19:65–8.
46. Lynch M. The origins of genome architecture. 1st ed. Sunderland: Sinauer Associates, Inc. Publishers; 2007.
47. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 2005;33:1141–53.
48. Supek F, Škunca N, Repar J, Vlahoviček K, Šmuc T. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 2010;6:1–13.
49. Krisko A, Copic T, Gabaldón T, Lehner B, Supek F. Inferring gene function from evolutionary change in signatures of translation efficiency. Genome Biol. 2014;15:R44.
50. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010;107:3645–50.
51. Lambros RJ, Mortimer JR, Forsdyke DR. Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles. 2003;7:443–50.
52. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci. 2001;268:493–7.
53. Oshima K, Kakizawa S, Nishigawa H, Jung H-Y, Wei W, Suzuki S, et al. Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat Genet. 2004;36:27–9.
54. Tran-Nguyen LTT, Kube M, Schneider B, Reinhardt R, Gibb KS. Comparative genome analysis of "Candidatus Phytoplasma australiense" (subgroup tuf-Australia I; rp-A) and "Ca. phytoplasma asteris" strains OY-M and AY-WB. J Bacteriol. 2008;190:3979–91.
55. Kube M, Schneider B, Kuhl H, Dandekar T, Heitmann K, Migdoll AM, et al. The linear chromosome of the plant-pathogenic mycoplasma "Candidatus Phytoplasma mali". BMC Genomics. 2008;9:306.
56. Mitrović J, Siewert C, Duduk B, Hecht J, Mölling K, Broecker F, et al. Generation and analysis of draft sequences of "stolbur" phytoplasma from multiple displacement amplification templates. J Mol Microbiol Biotechnol. 2014;24:1–11.
57. Andersen MT, Liefting LW, Havukkala I, Beever RE. Comparison of the complete genome sequence of two closely related isolates of "Candidatus Phytoplasma australiense" reveals genome plasticity. BMC Genomics. 2013;14:529.
58. Calderon-Copete SP, Wigger G, Wunderlin C, Schmidheini T, Frey J, Quail MA, et al. The Mycoplasma conjunctivae genome sequencing, annotation and analysis. BMC Bioinf. 2009;10 Suppl 6:S7.
59. McGowin CL, Ma L, Jensen JS, Mancuso MM, Hamasuna R, Adegboye D, et al. Draft genome sequences of four axenic Mycoplasma genitalium strains isolated from Denmark, Japan, and Australia. J Bacteriol. 2012;194:6010–1.
60. Do Nascimento NC, Guimaraes AMS, Santos AP, SanMiguel PJ, Messick JB. Complete genome sequence of Mycoplasma haemocanis strain Illinois. J Bacteriol. 2012;194:1605–6.

61. Liu W, Xiao S, Li M, Guo S, Li S, Luo R. Comparative genomic analyses of Mycoplasma hyopneumoniae pathogenic 168 strain and its high-passaged attenuated strain. BMC Genomics. 2013;14:80.
62. do Nascimento NC, Dos Santos AP, Chu Y, Guimaraes AMS, Pagliaro A, Messick JB. Genome sequence of Mycoplasma parvum (formerly Eperythrozoon parvum), a diminutive hemoplasma of the pig. Genome Announc. 2013;1:1–2.
63. Dybvig K, Zuhua C, Lao P, Jordan DS, French CT, Tu AHT, et al. Genome of Mycoplasma arthritidis. Infect Immun. 2008;76:4000–8.
64. Shu HW, Liu TT, Chan HI, Liu YM, Wu KM, Shu HY, et al. Genome sequence of the repetitive-sequence-rich Mycoplasma fermentans strain M64. J Bacteriol. 2011;193:4302–3.
65. Sasaki Y, Ishikawa J, Yamashita A, Oshima K, Kenri T, Furuya K, et al. The complete genomic sequence of Mycoplasma penetrans, an intracellular bacterial pathogen in humans. Nucleic Acids Res. 2002;30:5293–300.
66. Calcutt MJ, Foecking MF. Genome sequence of mycoplasma putrefaciens type strain KS1. J Bacteriol. 2011;193:6094.
67. Vasconcelos ATR, Vasconcelos ATR, Ferreira HB, Ferreira HB, Bizarro CV, Bizarro CV, et al. Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. Microbiology. 2005;187:5568–77.
68. Brown DR, Farmerie WG, May M, Benders GA, Durkin AS, Hlavinka K, et al. Genome sequences of Mycoplasma alligatoris A21JP2T and Mycoplasma crocodyli MP145T. J Bacteriol. 2011;193:2892–3.
69. Dabrazhynetskaya A, Soika V, Volokhov D, Simonyan V, Chizhikov V. Genome sequence of Mycoplasma hyorhinis strain DBS 1050. Genome Announc. 2014;2(2):e00127–14.
70. Wise KS, Calcutt MJ, Foecking MF, Madupu R, DeBoy RT, Röske K, et al. Complete genome sequences of Mycoplasma leachii strain PG50T and the pathogenic Mycoplasma mycoides subsp. mycoides small colony biotype strain Gladysdale. J Bacteriol. 2012;194:4448–9.
71. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2009;38:457–62.
72. Porcel BM, Denoeud F, Opperdoes F, Noel B, Madoui M-A, Hammarton TC, et al. The streamlined genome of Phytomonas spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. PLoS Genet. 2014;10:e1004007.
73. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
74. Collingridge PW, Kelly S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinf. 2012;13:117.
75. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
76. dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res. 2003;31:6976–85.
77. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32:5036–44.
78. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141:344–54.



Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

