



University of Huddersfield Repository

Ginés Rivas, Juan José

Codon Usage Evolution in Opisthokonta

Original Citation

Ginés Rivas, Juan José (2021) Codon Usage Evolution in Opisthokonta. Doctoral thesis, University of Huddersfield.

This version is available at <https://eprints.hud.ac.uk/id/eprint/35791/>

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

<http://eprints.hud.ac.uk/>

Doctoral Thesis

Codon Usage Evolution in Opisthokonta

University of
HUDDERSFIELD
Inspiring global professionals

Juan José Ginés Rivas

Supervised by Dr Martin Carr

Co-supervised by Dr Christopher Cooper and Dr Roy Chaudhuri

Department of Biological and Geographical Sciences

School of Applied Sciences

University of Huddersfield

*A thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy*

October 2021

COPYRIGHT STATEMENT

- The author of this thesis (including any appendices and/ or schedules to this thesis) owns any copyright in it (the “Copyright”) and he has given The University of Huddersfield the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes
- Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.
- The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

DECLARATION OF AUTHORSHIP

I, Juan José Ginés Rivas, declare that this thesis titled, “Codon Usage Evolution in Opisthokonta” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- I have acknowledged all main sources of help.

Signed:



Date: 17/10/2021

ABSTRACT

The emergence of life on Earth developed a system of DNA containing the information inherited in the reproduction of organisms, directing also the variability of life as this changes. The language using for deciphering the DNA is the genetic code, which is degenerated as 20 amino acids are codified by 64 codons, or triplets of nitrogenous bases (adenine, guanine, cytosine and thymine) in the nucleotides that form the DNA. 59 codons have other/s synonymous codons, and those are not used randomly but are a product of different phenomena. The drivers of codon usage bias can be mutations, genetic drift and natural selection processes. The study of these codons can provide evolutionary information, like gene family evolution and organisation, can offer an insight in population size, and also can be used to create software to predict ORFs and coding regions. Codon study also can be important in biotechnology for heterologous expression, as the codon used have a control over the efficiency and accuracy during translation, resulting in determinant for protein folding, growth rate and avoidance of defective proteins.

Opisthokonta is a group of organisms, close to Apusomonadida, and gathering Holozoa (Ichthyosporea, Filasterea, Choanoflagellata and Metazoa) and Holomycota. The present study uses the transcriptome of 20 choanoflagellates species, *Sphaeroforma arctica* (Class Ichthyosporea), *Thecamonas trahens* (class Apusomonadea) and five early branching metazoans (*Amphimedon queenslandica*, *Mnemiopsis leidyi*, *Nematostella vectensis*, *Thelohanellus kitauei* and *Trichoplax adhaerens*) to study the codon usage bias and anticodon related phenomena in the mentioned species.

Analysis revealed optimal codons for each organism, with a GC ended codons enrichment in *T. trahens*, unicellular Opisthokonta, and retained after the transition to pluricellularity (with divergences in *M. leidyi* and *N. vectensis*). All the unicellular species showed evidence for natural selection upon translational accuracy and efficiency, even the less biased ones. In metazoan only translational efficiency was evidenced, with lower strength of selection, and poor support for translational accuracy in ctenophoran and cnidarian species. For other strategies giving advantages on protein translation, indirect evidence of deamination in high degenerated amino acids was

found in all the species studied, while direct evidence was found in *S. arctica*, *A. queenslandica*, *M. leidy* and *N. vectensis*. Additionally, tRNA presenting isotype-anticodon disagreements were found in apusomonadean, ichthyosporean, metazoan species and in *S. rosetta*. Deamination, and possibly chimeric tRNA appeared early in evolution, before the divergence of Apusomonadea and Opisthokonta, benefiting the use of optimal codons through evolution until the apparition of the latest metazoan organisms.

ACKNOWLEDGEMENTS

First, with no doubt I would like to thank Dr Martin Carr, who gave me this opportunity to grow as a researcher and as a person. I will always be really thankful for that and all the support received through the years. I also would like to thank my co-supervisors Christopher Cooper and Roy Chaudhuri for offering their advice in the methodologies employed in this project, and Daniel Richter for his kindness on letting us work with the choanoflagellates data months before its publishing.

Next, I would like to acknowledge the help I have received from great researchers like Cooper Grace, a person with an extraordinary human quality and invaluable guidance. Also Jarek Bryk, Maria Luisa Martin and Rohan Raval for their teachings in bioinformatics and their always ready to help attitude. And to Bernardo de Haro, for his contribution on establishing the choanoflagellates screening methodology.

And last but not less important, I want to thank my family, although far they are always really close to me, and my amazing friends. I could have not asked for better people to support me through this journey.

CONTENTS

Copyright Statement	III
Declaration of Authorship	V
Abstract	VII
Acknowledgements	IX
Contents	XI
List of Figures	XV
List of Tables	XVII
List of Abbreviations	XXIII

Chapter 1 Introduction

1.1 Origin and evolution of eukaryotes	1
1.1.1 The origin of life on Earth	1
1.1.2 The primal RNA world	2
1.1.3 Evolution of the protocell	3
1.1.4 Evolution of the Last Common Ancestor	5
1.2 Taxonomy of the group Opisthokonta	6
1.2.1 Framing the group Opisthokonta	6
1.2.2 Holomycota and Holozoa	9
1.2.3 The choanoflagellates	11
1.2.4 Evolution in Metazoa	14
1.3 Codon usage and evolutionary forces	18
1.3.1 The genetic code and the codon usage of genes	18
1.3.2 The origin of the genetic code	20
1.3.3 Evolutionary forces upon the codon bias	23
1.3.4 Consequences of codon bias	25
1.4 Codon bias in evolution and the transition to pluricellularity	27
1.4.1 Codon bias in microscopic organisms	27
1.4.2 Codon bias in holozoan protists	29

1.4.3 The transition to pluricellularity	30
1.4.4 Codon bias in pluricellular organisms	33
1.4.5 Beyond codon usage rules, the tRNA modifications	38
1.5 Codon usage and tRNA analyses	41
1.5.1 Relevant concepts in codon usage analysis	41
1.5.2 Characteristics of the software employed	43
1.5.3 Statistical techniques employed for evolutionary analyses	46
1.5.4 Difficulties in the sequencing of choanoflagellates for codon usage studies	48
1.6 Objectives of this project	49
Chapter 2 Materials and Methods	
2.1 Sources and treatment of data	51
2.1.1 Genetic file sources	51
2.1.2 Coding sequences	54
2.1.3 Screening of choanoflagellates CDS	54
2.2 Coding regions analysis	57
2.2.1 Mapping of reads for relative expression	57
2.2.2 Determination of optimal codons	58
2.2.3 Codon usage	59
2.2.4 Strength of selection	59
2.3 Anticodon and tRNA analysis	60
2.3.1 Indirect evidence for tRNA deamination through selective advantage	60
2.3.2 Direct evidence for tRNA deamination through transcriptome screening	61
2.3.3 Differential anticodon abundance based on tRNA copy number	62
2.3.4 Chimeric tRNA	62
Chapter 3 Codon usage and tRNA results in choanoflagellates	
3.1 Treated data from choanoflagellates	63
3.2 Coding region analysis	66

3.2.1 Mapping of reads for relative expression	66
3.2.2 Determination of optimal codons	66
3.2.3 Codon usage statistics	88
3.2.3.1 General codon statistic values	88
3.2.3.2 Effective number of codons and GC3 plots	101
3.2.3.3 Codon usage in protein domain and non-domain regions	113
3.2.4 Strength of selection	122
3.3 tRNA analysis	142
3.3.1 Indirect evidence for tRNA deamination through selective pressure	142
Chapter 4 Codon usage and tRNA results in unicellular eukaryotes and metazoan species	
4.1 Genomic and transcriptomic data from unicellular and metazoan species	147
4.2 Coding region analysis	148
4.2.1 Mapping of reads for relative expression	148
4.2.2 Determination of optimal codons	149
4.2.3 Codon usage statistics	154
4.2.3.1 General codon statistic values.	154
4.2.3.2 Effective number of codons and GC3 plots	159
4.2.3.3 Codon usage in protein domain and non-domain regions	165
4.2.4 Strength of selection	170
4.3 tRNA and anticodon analysis	174
4.3.1 Indirect evidence for tRNA deamination through selective pressure	174
4.3.2 Direct evidence for tRNA deamination through transcriptome screening	177
4.3.3 Differential anticodon abundance based on tRNA copy number	179
4.3.4 Chimeric tRNA	185
4.4 Final results overview	194
Chapter 5 Discussion and Conclusions	
5.1 Data sources for eukaryote organisms and treated data of choanoflagellates	197
5.2 Coding regions analysis	200

5.3 Anticodon and tRNA analysis	203
5.4 Impact of this project in future research	204
5.5 Conclusions.	207
 Bibliography	 209

LIST OF FIGURES

Figure 01. Phylogenetic trees of Eukarya with root between Opimoda and Diphoda	8
Figure 02. Consensus phylogenetic tree of the main Opisthokonta groups	10
Figure 03. Maximum-likelihood tree based on a six-gene phylogeny of choanoflagellates	12
Figure 04. Maximum-likelihood tree of the early branching metazoan	17
Figure 05. Genetic code emphasising its degeneracy	19
Figure 06. Comparison of the level of bias in choanoflagellates using four genes	29
Figure 07. Steps in the procedure for choanoflagellates transcriptome screening	56
Figure 08. Nc and GC3 plot for <i>A. spectabilis</i> genes	102
Figure 09. Nc and GC3 plot for <i>C. perplexa</i> genes	102
Figure 10. Nc and GC3 plot for <i>C. hollandica</i> genes	103
Figure 11. Nc and GC3 plot for <i>D. grandis</i> genes	103
Figure 12. Nc and GC3 plot for <i>D. costata</i> genes	104
Figure 13. Nc and GC3 plot for <i>H. balthica</i> genes	104
Figure 14. Nc and GC3 plot for <i>H. gracilis</i> genes	105
Figure 15. Nc and GC3 plot for <i>H. nana</i> genes	106
Figure 16. Nc and GC3 plot for <i>M. roanoka</i> genes	106
Figure 17. Nc and GC3 plot for <i>M. fluctuans</i>	107
Figure 18. Nc and GC3 plot for <i>S. dolichothecata</i> genes	107
Figure 19. Nc and GC3 plot for <i>S. helianthica</i> genes	108
Figure 20. Nc and GC3 plot for <i>S. infusionum</i> genes	108
Figure 21. Nc and GC3 plot for <i>S. kjevrii</i>	109
Figure 22. Nc and GC3 plot for <i>S. macrocollata</i>	109
Figure 23. Nc and GC3 plot for <i>S. punica</i> genes	110
Figure 24. Nc and GC3 plot for <i>S. urceolata</i> genes	110
Figure 25. Nc and GC3 plot for <i>S. parva</i> genes	111
Figure 26. Nc and GC3 plot for <i>S. diplocostata</i> (Australia isolated) genes	112
Figure 27. Nc and GC3 plot for <i>S. diplocostata</i> (France isolated) genes	112
Figure 28. Nc and GC3 plot for <i>S. rosetta</i> genes	159

Figure 29. Nc and GC3 plot for <i>S. arctica</i> genes	160
Figure 30. Nc and GC3 plot for <i>T. trahens</i> genes	160
Figure 31. Nc and GC3 plot for <i>A. queenslandica</i> genes	161
Figure 32. Nc and GC3 plot for <i>M. leidy</i> genes	162
Figure 33. Nc and GC3 plot for <i>N. vectensis</i> genes	162
Figure 34. Nc and GC3 plot for <i>T. kitauei</i> genes	163
Figure 35. Nc and GC3 plot for <i>T. adhaerens</i> genes	164
Figure 36. Final phylogeny with main results	194

LIST OF TABLES

Table 01. Accession numbers for choanoflagellates transcriptome files	52
Table 02. Accession numbers for unicellular eukaryotes transcriptome files	52
Table 03. Accession numbers for metazoan transcriptome files	53
Table 04. Accession numbers for additional files for tRNA deamination screening	53
Table 05. SAMtools commands for treatment of data	57
Table 06. Codon distribution in a codon frequency file	58
Table 07. Transcriptomic data for choanoflagellates	64
Table 08. Comparison of optimal codons detected in <i>A. spectabilis</i>	67
Table 09. Comparison of optimal codons detected in <i>C. perplexa</i>	68
Table 10. Comparison of optimal codons detected in <i>C. hollandica</i>	69
Table 11. Comparison of optimal codons detected in <i>D. grandis</i>	70
Table 12. Comparison of optimal codons detected in <i>D. costata</i>	71
Table 13. Comparison of optimal codons detected in <i>H. balthica</i>	72
Table 14. Comparison of optimal codons detected in <i>H. gracilis</i>	73
Table 15. Comparison of optimal codons detected in <i>H. nana</i>	74
Table 16. Comparison of optimal codons detected in <i>M. roanoka</i>	75
Table 17. Comparison of optimal codons detected in <i>M. fluctuans</i>	76
Table 18. Comparison of optimal codons detected in <i>S. dolichothecata</i>	77
Table 19. Comparison of optimal codons detected in <i>S. helianthica</i>	78
Table 20. Comparison of optimal codons detected in <i>S. infusionum</i>	79
Table 21. Comparison of optimal codons detected in <i>S. kvevrii</i>	80
Table 22. Comparison of optimal codons detected in <i>S. macrocollata</i>	81
Table 23. Comparison of optimal codons detected in <i>S. punica</i>	82
Table 24. Comparison of optimal codons detected in <i>S. urceolata</i>	83
Table 25. Comparison of optimal codons detected in <i>S. parva</i>	84
Table 26. Comparison of optimal codons detected in <i>S. diplocostata</i> (Australia isolated)	85
Table 27. Comparison of optimal codons detected in <i>S. diplocostata</i> (France isolated)	86
Table 28. GC content mean and standard deviation for original CDS in choanoflagellates	89

Table 29. GC content mean and standard deviation for screened CDS in choanoflagellates	90
Table 30. GC3 mean and standard deviation for original CDS in choanoflagellates	91
Table 31. GC3 mean and standard deviation for screened CDS in choanoflagellates	92
Table 32. CBI mean and standard deviation for original CDS in choanoflagellates	94
Table 33. CBI mean and standard deviation for screened CDS in choanoflagellates	95
Table 34. Nc mean and standard deviation for original CDS in choanoflagellates	96
Table 35. Nc mean and standard deviation for screened CDS in choanoflagellates	97
Table 36. Fop mean and standard deviation for original CDS in choanoflagellates	99
Table 37. Fop mean and standard deviation for screened CDS in choanoflagellates	100
Table 38. GC content mean and standard deviation in protein domain and non-domain regions of choanoflagellates	114
Table 39. GC3 mean and standard deviation in protein domain and non-domain regions of choanoflagellates	116
Table 40. CBI mean and standard deviation in protein domain and non-domain regions of choanoflagellates	117
Table 41. Nc mean and standard deviation in protein domain and non-domain regions of choanoflagellates	119
Table 42. Fop mean and standard deviation in protein domain and non-domain regions of choanoflagellates	120
Table 43. Strength of selection values in <i>A. spectabilis</i> sets of coding genes	122
Table 44. Strength of selection values in <i>C. perplexa</i> sets of coding genes	123
Table 45. Strength of selection values in <i>C. hollandica</i> sets of coding genes	124
Table 46. Strength of selection values in <i>D. grandis</i> sets of coding genes	125
Table 47. Strength of selection values in <i>D. costata</i> sets of coding genes	126
Table 48. Strength of selection values in <i>H. balthica</i> sets of coding genes	127
Table 49. Strength of selection values in <i>H. gracilis</i> sets of coding genes	128
Table 50. Strength of selection values in <i>H. nana</i> sets of coding genes	129
Table 51. Strength of selection values in <i>M. roanoka</i> sets of coding genes	130
Table 52. Strength of selection values in <i>M. fluctuans</i> sets of coding genes	131
Table 53. Strength of selection values in <i>S. dolichothecata</i> sets of coding genes	132
Table 54. Strength of selection values in <i>S. helianthica</i> sets of coding genes	133

Table 55. Strength of selection values in <i>S. infusionum</i> sets of coding genes	134
Table 56. Strength of selection values in <i>S. kvevrii</i> sets of coding genes	135
Table 57. Strength of selection values in <i>S. macrocollata</i> sets of coding genes	136
Table 58. Strength of selection values in <i>S. punica</i> sets of coding genes	137
Table 59. Strength of selection values in <i>S. urceolata</i> sets of coding genes	138
Table 60. Strength of selection values in <i>S. parva</i> sets of coding genes	139
Table 61. Strength of selection values in <i>S. diplocostata</i> (Australian isolate) coding genes	140
Table 62. Strength of selection values in <i>S. diplocostata</i> (France isolate) coding genes	141
Table 63. Overrepresented cytosine ended codons in protein domain regions of <i>A. spectabilis</i> , <i>C. perplexa</i> , <i>C. hollandica</i> , <i>D. grandis</i> and <i>D. costata</i>	143
Table 64. Overrepresented cytosine ended codons in protein domain regions of <i>H. balthica</i> , <i>H. gracilis</i> , <i>H. nana</i> , <i>M. roanoka</i> and <i>M. fluctuans</i>	143
Table 65. Overrepresented cytosine ended codons in protein domain regions of <i>S. dolichothecata</i> , <i>S. helianthica</i> , <i>S. infusionum</i> , <i>S. kvevrii</i> and <i>S. macrocollata</i>	144
Table 66. Overrepresented cytosine ended codons in protein domain regions of <i>S. punica</i> , <i>S. urceolata</i> , <i>S. parva</i> and <i>S. diplocostata</i> (France isolated)	144
Table 67. Genomic and transcriptomic data for unicellular eukaryotes species	148
Table 68. Genomic and transcriptomic data for metazoan species	148
Table 69. Comparison of optimal codons determined for the unicellular species <i>S. rosetta</i> , <i>S. arctica</i> and <i>T. trahens</i>	150
Table 70. Comparison of optimal codons determined for the metazoan species <i>A. queenslandica</i> and <i>M. leidy</i>	151
Table 71. Comparison of optimal codons determined for the metazoan species <i>N. vectensis</i> , <i>T. kitauei</i> and <i>T. adhaerens</i>	152
Table 72. GC content mean and standard deviation for unicellular and metazoan species	154
Table 73. GC3 mean and standard deviation for unicellular and metazoan species	155
Table 74. CBI mean and standard deviation for unicellular and metazoan species	156
Table 75. Nc mean and standard deviation for unicellular and metazoan species	157
Table 76. Fop mean and standard deviation for unicellular and metazoan species	158
Table 77. GC content mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species	166

Table 78. GC3content mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species	166
Table 79. CBI mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species	167
Table 80. Nc mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species	167
Table 81. Fop mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species	169
Table 82. Strength of selection values for <i>S. rosetta</i> , <i>S. arctica</i> and <i>T. trahens</i>	171
Table 83. Strength of selection values for <i>A. queenslandica</i> and <i>M. leidy</i>	172
Table 84. Strength of selection values for <i>N. vectensis</i> , <i>T. kitauei</i> and <i>T. adhaerens</i>	173
Table 85. Overrepresented cytosine ended codons in protein domain regions of unicellular species	175
Table 86. Overrepresented cytosine ended codons in protein domain regions of metazoan species	175
Table 87. Detection of direct deamination in transcripts of tRNA genes in unicellular species	177
Table 88. Detection of direct deamination in transcripts of tRNA genes in metazoan species	178
Table 89. Anticodons identified in DNA toplevel files for unicellular species	180
Table 90. Anticodons identified in DNA toplevel files for <i>A. queenslandica</i> and <i>M. leidy</i>	182
Table 91. Anticodons identified in DNA toplevel files for <i>N. vectensis</i> , <i>T. kitauei</i> and <i>T. adhaerens</i>	183
Table 92. Types of tRNA detected in unicellular and metazoan species	186
Table 93. Isotype-anticodon mismatches identified in the species <i>S. rosetta</i> , <i>S. arctica</i> and <i>T. trahens</i>	187
Table 94. Isotype-anticodon mismatches identified in the species <i>A. queenslandica</i>	188
Table 95. Isotype-anticodon mismatches identified in the species <i>M. leidy</i>	189
Table 96. Isotype-anticodon mismatches identified in the species <i>N. vectensis</i> for TAPSILVR amino acids	190

Table 97. Isotype-anticodon mismatches identified in the species <i>N. vectensis</i> for non-TAPSILVR amino acids	191
Table 98. Isotype-anticodon mismatches identified in the species <i>T. kitauei</i>	192
Table 99. Isotype-anticodon mismatches identified in the species <i>T. adhaerens</i>	192

LIST OF ABBREVIATIONS

Abbreviation	Complete word/expression
A	adenine
acetyl-CoA	acetyl coenzyme A
ADAR	adenosine deaminase acting on RNA
Ala	alanine
Arg	arginine
Asn	asparagine
Asp	aspartate
AT	adenine or thymine
BLAST	basic local alignment search tool
C	cytosine
CBI	codon bias index
CDD	conserved domain database
CDS-O	original set of CDS
CDS-S	screened set of CDS
CEG	Conserved eukaryotic genes
COA	correspondence analysis
Cys	cysteine
DELTA	domain enhanced lookup time accelerated
DNA	deoxyribonucleic acid
DR	protein domain regions
DRIP	<i>Dermocystidium</i> , rosette agent, <i>Ichthyophonus</i> and <i>Psorospermium</i>
EF-1 α ,	elongation factor 1-alpha
EFL	elongation factor like
Fop	frequency of optimal codons
G	guanine
GC	guanine or cytosine

GC3	guanine or cytosine at third codon position
Gln	glutamine
Glu	glutamate
Gly	glycine
GPCR	G protein coupled receptors
HE	highest expressed genes
His	histidine
HMM	Hidden Markov Models
I34	deamination of adenosine in tRNA position 34
Ile	isoleucine
iMet	initiator methionine
LE	lowest expressed genes
Leu	leucine
LUCA	last universal common ancestor
Lys	lysine
MBRTK1	<i>Monosiga brevicolis</i> receptor tyrosine kinase 1
Met	methionine
mRNA	messenger ribonucleic acid
MYA	million years ago
NAC/U codons	codon starting with any nitrogenous base, followed by A, and finished in C or U
Nc	effective number of codons
NCBI	national centre for biotechnology information
NDR	protein non-domain regions
NGS	next generation sequencing
ORFs	open reading frames
PANTHER	protein analysis through evolutionary relationships
Phe	phenylalanine
Pro	proline
RNA	ribonucleic acid
RNA-Seq	sequencing of ribonucleic acid

RSCU	random synonymous codon usage
SelCys	selenocysteine
Ser	serine
T	thymine
Thr	threonine
tRNA	transfer ribonucleic acid
Trp	tryptophan
Tyr	tyrosine
U	uracil
Val	valine

CHAPTER 1. INTRODUCTION

The first chapter aims to give a background on the fundamental questions that leads this research. To understand evolution we first answer to the question of how this process started on earth (1.1 Origin and evolution of eukaryotes), with a brief review on how life was originated and developed into eukaryotes. The second question we address is how evolution diversified the eukaryotic lineages, developing the animal tree of life (1.2 Taxonomy of the group Opisthokonta). The third section responded to what forces drive evolution itself and the elements it acts upon, with a special focus on codon usage (1.3 Codon usage and evolutionary forces). In the fourth part we explain how evolutionary forces shaped organisms and increased its complexity due to the development of pluricellularity (1.4 Codon bias in evolution and the transition to pluricellularity). The fifth section aims to give an explanation to software and concepts that we use for the study of evolution through codon bias and tRNA modifications (1.5 Codon usage and tRNA analyses). At the end, the objectives of this project are summarised (1.6 Objectives of this project).

1.1 ORIGIN AND EVOLUTION OF EUKARYOTES.

1.1.1 The origin of life on Earth.

The origin of life on Earth from inorganic compounds, process also known as abiogenesis, rose when the planet was approximately 340 million of years old, between 3.77 and 4.28 billion years ago, soon after the formation of the oceans (Dodd et al., 2017), with the most solid evidence dated on 3.5 billion years ago (Schopf, 2006; Schopf et al., 2007). This possibly happened in a chemically reactive environment similar to hydrothermal vents, as it resemble primordial conditions (Baross and Hoffman, 1985; Weiss et al., 2016), and has been discovered harbouring rich ecosystems, utilising the energy released in the oceanic volcanism (Corlis et al., 1979; Spiess et al. 1980).

The most extended theory for the origin of this organic compounds is the "Prebiotic soup theory". In this model was proposed that gases and primitive environmental conditions in the water gave origin to proteins as the first source of genetic information. This protein containing information formed from inorganic molecules would be the source of all life (Oparin, 1924). The formation of organic molecules on primitive environments was supported by later experiments (Miller, 1953).

Whether information or metabolism was first is still in discussion, although in the first case it is believed that formaldehyde condensations formed tetrose nucleic acids, and Hydrogen-Carbon-Nitrogen condensations turned into bases, that became RNA as self-replicating entities. If this is the case, at some point the lacking of preformed organic compounds on the environment would have pushed the biosynthesis directed by RNA, before the appearance of the more stable DNA (Bada, 2004; Leslie, 2004).

Alternatives to this hypothesis involves the arrival of the organic compounds on comets (De Duve, 1995), or the development of organically catalysed reactions from other reactions, that were spontaneous or inorganically catalysed, and ended up suffering a constriction by thermodynamic convenience (Morowitz et al., 2000). Energy sources could have come from hydrothermal-vents settings, through carbon reduction (Proskurowski et al., 2008) or by a primitive path to acetyl-CoA. Acetil-CoA reaction can take place in acetogenesis as in methanogenesis, and its presence could have support the synthesis of biomolecules (Martin and Rusell, 2007).

1.1.2 The primal RNA world.

The problem of the origin of life was changed elegantly with the proposal of the theory called "RNA World" (Gilbert, 1986). This theory proposed that the RNA used the minus strand for replication meanwhile the plus strand acted as ribozyme (a ribonucleic acid enzyme), rendering it like a self-replicated system. The catalytic activity of the RNA is justified with the discovery of phosphodiester bonds that are cut on the maturation of tRNA of *Escherichia coli* (Guerrier-Takada et al., 1983; Guerrier-Takada and Altman, 1984), and a self-splicing exon in the ciliated protozoan *Tetrahymena*, that can use a molecule of RNA as a RNA polymerase (Kruger et al., 1982; Zaug and Cech, 1986). This examples support the idea of a world where just RNA is necessary to develop a range of

enzymatic activities that would end up developing the synthesis of proteins. If this theory is correct, prebiotic chemistry cannot be studied through nowadays protein enzyme reactions, even if some resemblance could exist. RNA world translate the origin of life into the origin of RNA, rendering the development of DNA and proteins a subsequent natural selection (Leslie, 2004).

Some authors disagree about the fact that RNA appeared *de novo* due to its complexity, instead they propose the evolution of a system developed on clay minerals (Cairns-Smith, 1982) or a simpler linear polymer (Cairns-Smith and Davies, 1977) that underwent a "genetic takeover" to displace the primitive mineral for RNA. Although there is no experimental support for this hypothesis and no proposal for that mineral, the existence of an earlier world due to primitive extreme conditions has been discussed. This debate continue with the success of synthetic biology experiments: creating DNA hybrids (Groebke et al., 1998; Pitsch et al., 2003); hybrid double helices with antisense and phosphate-free analogues of nucleic acid, keeping the ability to transmit chiral information (Egholm et al., 1993; Koppitz et al., 1998; Kozlov et al., 2000); stable heteroduplexes with RNA (a duplex formed of strands from different sources) (Schöning et al., 2000); formation of stable antiparallel double helices with alternation of D and L amino acids (Diederichsen, 1996). These results keep open the possibility that simpler analogues of RNA may have existed previously, even if their first function was not information related.

Over time RNA was relegated to be an intermediate in favour of the more stable DNA, although some RNA reactions and the structure of exons-and-introns for recombination could be relics of this stage in evolution.

1.1.3 Evolution of the protocell.

The next important step in evolution was the encapsulation of the RNA and other biomolecules in a lipid membrane, forming a protocell. Although if this happened late in the RNA World or early in the RNA/Protein World is unknown. This system let molecules evolve together, split the ribozymes in different compartments and enforce selection against lower level of activity or possible parasitic sequences (Szathmáry and Smith, 1995). As an isolating membrane would keep environmental resources out, is more possible that it was originated as a permeable membrane, evolving as the

enzymatic synthesis took over the acquisition of small organic molecules from the media. Initially RNA monomers and small metabolites could diffuse through the membrane, in an obligate heterotrophic lifestyle lacking macromolecular transport machinery (Chen et al., 2005; Mansy et al., 2008).

Another possibility for the membrane development state that the first replicators were attached to the surface of anion-exchanging minerals, supporting the process, and reaching the independence in a late stage (Schwartz & Orgel, 1985; Hill et al., 1998; Orgel, 1998). Although this is a possibility, many studies suggest that the formation of membranous vesicles was possible in a prebiotic world (Deamer and Pashley, 1989; Dworkin et al., 2001; Deamer et al., 2002) and tend to agree that it was important from an early stage. Theoretically both approaches could be partially true in a case where the RNA is adsorbed to mineral particles inside the membranes, helping RNA polymerisation and vesicle formation (Ferris et al., 1996; Ferris, 2002; Hanczyc et al., 2003).

The formation of a protocell present numerous advantages: increasing of replication rate and fidelity (Matsuura et al., 2002; Ichihashi et al., 2013), promote selection (Biaconi et al., 2013), creates high concentrations of polymers and small molecules (Saha et al., 2014), allows the apparition of transmembrane gradients, that can be used to store energy (Chen and Szostak, 2004), and generates osmotic pressure that could benefit the growth of active cells over the inactive ones (Chen et al., 2004). In addition, lipids dehydration could induce RNA polymerization (Rajamani et al., 2008), while lipids were present in prebiotic syntheses and meteorite samples (Allen and Ponnampereuma, 1967; Yuen and Kvenvolden, 1973; Yuen et al., 1981; Deamer, 1985; McCollom et al., 1999; Loison et al., 2010). This discoveries elucidates the path that lead to the formation of LUCA (the Last Universal Common Ancestor) of all living beings, linking the abiotic phase with microbial life (Fox et al., 1980; Arndt and Nisbet, 2012).

1.1.4 Evolution of the Last Common Ancestor.

The progenote or LUCA to all extant life was a pre-prokaryotic organism, still evolving the connection between phenotype and genotype, but presenting a genetic code and biochemical pathways (Woese and Fox, 1977; Fox et al., 1980). This organism was estimated to have only 500-600 genes (Koonin, 2003), but reduced to 355 genes after removing possible horizontal transfer (Weiss et al., 2016). Genes categorised as ancestral revealed that LUCA may have been anaerobic, CO₂ and N₂ fixing, dependent on H₂ and transition metals, turned from thermophilic to mesophilic, and presented nucleoside modifications and methylations (Glansdorff et al., 2008). If those genes were codified by RNA or DNA is unclear, as there is evidence of proteins related to replication and tRNA emerging more than once. The fact is that bacterial and archaeal/eukaryotic replication system shares no similarities, implying that both evolved from RNA or RNA/DNA mixed LUCA (Leipe et al., 1999; Forterre, 2002; Poole and Logan, 2005; Di Giulio, 2006; Forterre, 2006). A theory suggest that this transition was made initially in viruses and transmitted to RNA cells, creating the plasmid, and differentiating the prokaryote/eukaryote traits by the nature of different DNA viruses (Forterre, 2002; 2006; Glansdorff et al., 2008).

LUCA was the origin of both Bacteria and Archaea, possibly by reductive evolution (Glansdorff et al., 2008). Those first cellular organisms, as they lack of nucleus formation are called prokaryotes, but presented core divergences to found different domains (the higher taxonomy rank). From the symbiosis of both described organisms Eukarya would have emerge, constituting a third domain. Current data points out to an archaeal symbiont member of the Asgard Archaea, possible from the group Heimdallarchaeota (Zaremba-Niedzwiedzka et al., 2017; Liu et al., 2018), and a bacterial endosymbiont belonging to Alphaproteobacteria, which evolved into the mitochondrion (Gray et al., 1999). Additional bacterial genes were acquired from chloroplast endosymbionts at various times, like those related to photosynthesis, from a cyanobacteria similar to *Gleomargarita lithophora* (Gray and Doolittle, 1982; Ponce-Toledo et al., 2017; Sato, 2020).

The endosymbiont theory was proposed by Mereschkowsky in 1905 (Martin and Kowallik, 1999), and rescued in 1967 (Margulis, 1967) to explain how mitochondria and chloroplast had an origin in free living primitive microbes that established multiple relations of endosymbiosis. She also includes the basal bodies (present at the base of cilia and flagella) in the theory, adding genes that

gave origin to centromeres and centrioles, but due to the lack of genetic information in this element there is still no proof for it. The genome reduction of endosymbionts is a process studied mainly in insects, suggesting that this process occur simply by genetic drift, and also demonstrating that the horizontal transfer from the symbiont to the nucleus of the cell is possible (Sloan et al., 2014; Wernegreen, 2015; Latorre and Manzano-Marín, 2017).

Eukaryotic cells were defined mainly by the presence of membrane-bound organelles and a nucleus containing the genetic material. Microfossils and molecular clock studies tend to agree that this group arose and begun its diversification in the Paleoproterozoic era, around 1800-1300 MYA (Javaux et al., 2003; Porter, 2004; Knoll et al., 2006; Parfrey et al., 2011), not without the possibility of its existence before any evidence (Brocks et al., 1999). This diversification led to the evolution of the major groups in eukaryotes.

1.2 TAXONOMY OF THE GROUP OPISTHOKONTA.

1.2.1 Framing the group Opisthokonta.

Vischer was the first one to use the term “opisthokonten” to describe unicellular organisms characterised by a radially symmetrical protoplast plus a posterior flagellum that is used for propulsion (Vischer, 1945). This term was used later in two opposite ways, for one author it must be expanded to include Metazoa (Gams, 1947) and for the other should be restricted to name the phylum of chytrids, spelling it as “Opisthokonta” (Copeland, 1956).

Opisthokonta understood in the current definition is a taxonomic supergroup set by Thomas Cavalier-Smith to gather the kingdoms Metazoa and Fungi (Cavalier-Smith, 1987). Initially this classification attended to the presence of posterior flagellum on reproductive cells, and although numerous exceptions weaken the morphological support (like the flagellum losing events through Fungi), its ultrastructural complexity suggest that this was the ancestral trait (James et al., 2006). Other common traits are the flattened mitochondrial cristae (Patterson, 1988), and an insertion of 9-17 amino acids in EF-1 α , a highly conserved translation elongation factor protein (Baldauf and Palmer, 1993, Steenkamp et al., 2006). Although quite widespread these characteristics have

exceptions, as the mentioned flagellum lost, demonstrated to happen from 4 to 6 times in different branches of Fungi (James et al., 2006), and also in Nuclearioid amoebae, Filasterea and Ichthyosporea (Torruella et al., 2015); the lamellar, tubular and discoidal cristae showed in the group Nucleariidae (Zettler et al., 2001; Ragan et al., 1996); and the lack of EF-1 α insertion found in holozoan protists and early branching fungal species (both codify for a related protein called EFL instead) (Keeling and Inagaki, 2004; Gile et al., 2009). Despite the lack of complete homogeneity of traits, numerous phylogenies gave strong support to the described classification (Cavalier-Smith and Chao, 1995; Lang et al., 2002; Medina et al., 2003; Ruiz-Trillo et al., 2004, 2006, 2008; Steenkamp et al., 2006).

To understand the place of Opisthokonta in evolution a discussion on the root of eukaryotes is necessary, which has been object of diverse hypothesis. Initially Opisthokonta was clustered with its closest sister group identified, Amoebozoa, to form a monophyletic group named Unikonta (Cavalier-Smith, 2002), setting the root between monoflagellated organisms and biflagellated (Bikonta) (as an early hypothesis of what is also well supported by Derelle et al, 2015). After that the root has also been changed to Plantae (Rogozin et al., 2009), and Excavata (Cavalier-Smith, 2010) but always concluding a single origin (except for Katz et al., 2002 based on the polyphyletic distribution of some eukaryotic genes). Amoebozoa as a sister group of Opisthokonta was strongly supported by phylogenies and synapomorphies (Stechmann and Cavalier-Smith, 2002; Richards and Cavalier-Smith, 2005; Ruiz-Trillo et al., 2006, 2008; Rodriguez-Ezpeleta et al., 2007; Burki et al., 2007; Rogozin et al., 2009), but two groups were found closer to Opisthokonta, Apusomonadida and Breviatea.

Apusomonadida is the closest Opisthokonta relative (Cavalier-Smith and Chao, 2010; Paps et al., 2013), and order containing unicellular flagellated that predates prokaryotes. Within this group we can find *Thecamonas trahens* (Larsen and Patterson, 1900), an amoeboid organism living in marine surfaces. While Breviatea (Walker et al., 2006) contains another kind of amoeboid flagellates, without mitochondrion (Edgcomb et al., 2002), and form with Opithokonta and Apusomonadida a clade named Obazoa (Brown et al., 2013). Afterwards the necessity of the division of eukaryotes in unikont and bikont was questioned due to the fact that the most parsimonious trees would consider all eukaryotes coming from a bikont ancestor, invalidating this trait to set groups (Stechmann and Cavalier-Smith, 2002; Kim et al., 2006; Roger and Simpson, 2009). This fact lead to

replace the confusing term based on synapomorphy Unikonta by the phylogenetic term Amorphea, and Bikonta by Diaphoretickes (Adl et al., 2012). Later results, as the polyphyletic origin of Excavata, led to rename again both groups as Opimoda and Diphoda respectively, setting the Eukaryota root between both (Derelle et al., 2015) with a strong support (as it is shown in Figure 01).

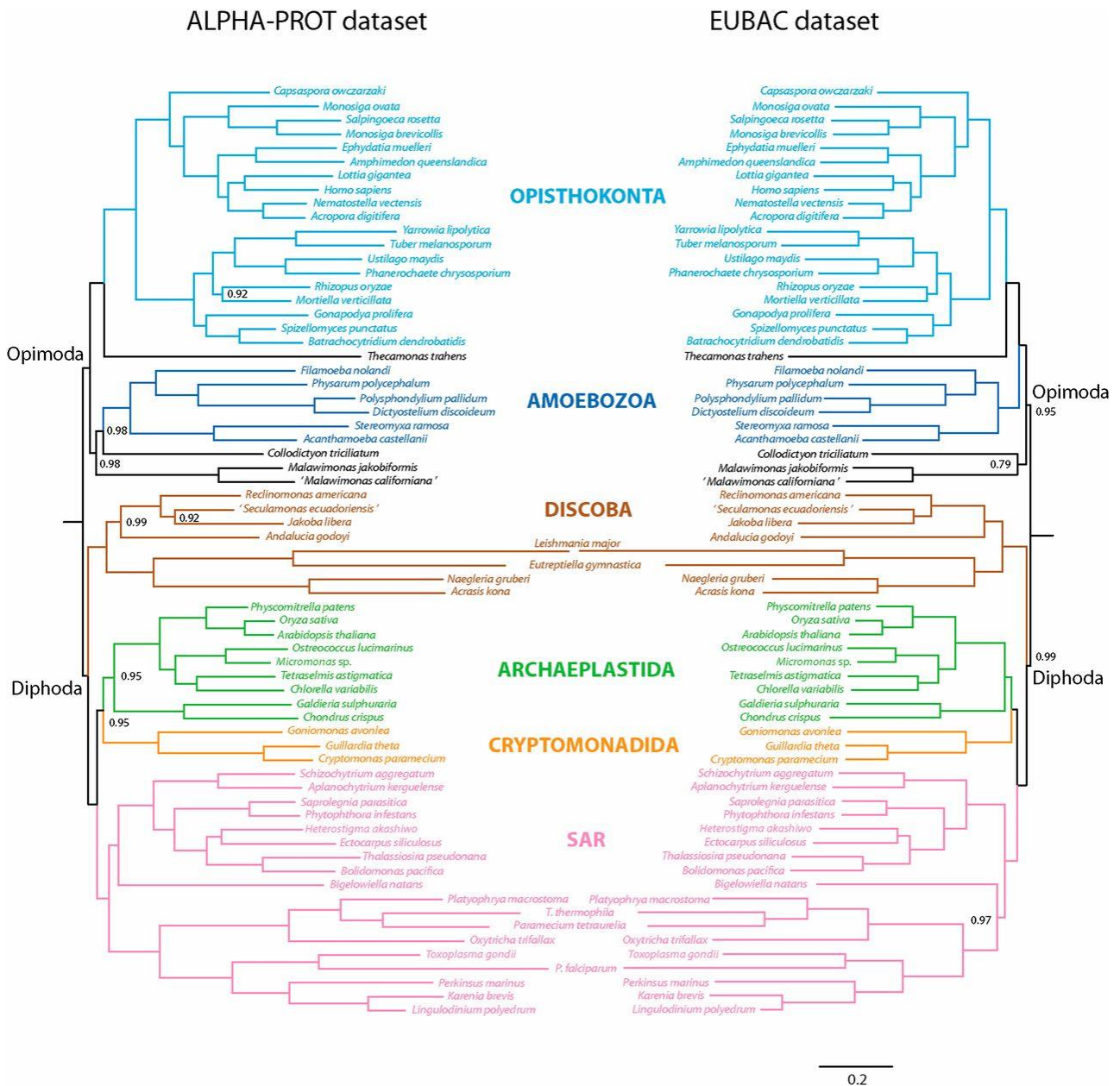


Figure 01. Phylogenetic trees of Eukarya with root between Opimoda and Diphoda. These trees also show the relation of Opisthokonta with higher ranked groups. Both are Bayesian consensus trees obtained from two datasets (ALPHA-PROT for the left and EUBAC for the right). Probabilities are equal to 1 except when specified. Figure 01 is reproduced under the terms of PNAS Copyright license from Figure 1 of Derelle et al., 2015.

1.2.2 Holomycota and Holozoa.

Inside the Opisthokonta group we can differentiate two major lines: Holomycota and Holozoa (as we can appreciate in Figure 02). Holomycota was proposed to comprise the closely related Fungi and Nucleariida, an amoeba group with filose pseudopodia (Liu et al., 2009). Holomycota was also called Nucletmycea (Brown et al., 2009) to gather the genus *Nuclearia* and *Fonticula* with Fungi (which can be seen as the earliest branches of Holomycota in Figure 02), setting it as a group that includes the unicellular fungi relatives, and in accordance with later analysis (Lang et al., 2002; Ruiz-Trillo et al., 2006 and Steenkamp et al., 2006).

Holozoa, on the other hand, is a taxonomic group created to comprise Ichthyosporea, Choanoflagellata and Metazoa as a consequence of discovering the phylogenetic position of Ichthyosporea (Lang et al., 2002). Ichthyosporea is a group of mostly parasites, with amoeboid and uniflagellated stages. What we know as Ichthyosporea born as DRIP clade, as an acronym of the first four organisms identified (Ragan et al., 1996), but changed to Ichthyosporea to reference them as fish parasites (Cavalier-Smith, 1998), and renamed again as Mesomycetozoa to include more species after a phylogenetic study of 24 organisms of the DRIP clade with 18S small-subunit ribosomal DNA (Herr et al., 1999). Mesomycetozoa was revised to include most unicellular opisthokont protists, except for the choanoflagellates (Adl et al., 2005). Afterwards, the class Filasterea was added to include the clade *Ministeria* (amoebae with radiating pseudopodia)/*Capsaspora*, both clades a monotypic genus containing *M. vibrans* and *C. owczarzaki* respectively (which can be seen in Figure 02 branching in Holozoa earlier than Choanoflagellata). In last years, the species *Pigoraptor vietnamica* and *P. chileana* have been added to Filasterea with strong support, and a new Class Pluriformea was set as sister to Filasterea, Choanoflagellata and Metazoa (containing *Syssomonas multiformis* and *Corallochytrium limacisporum*) (Hehenberger et al., 2017). Although the position of Pluriformea has been contentious, it was previously proposed to form group with ichthyosporeans (Torruella et al., 2015), as its closeness to Filasterea is moderately supported (Hehenberger et al., 2017), and also it has been placed as sibling to Filozoa (Tikhonenkov et al., 2020). The class Filozoa was set to gather those organism that could have develop filose tentacles before their aggregation (Metazoa, Choanoflagellata and Filasterea). As result, Holozoa was defined as Filozoa plus Ichthyosporea and Pluriformea (Shalchian-Tabrizi et al., 2008; Hehenberger et al., 2017).

Both Holomycota and Holozoa were proposed to have an origin in two different choanoflagellates ancestors, coming from salpingoecid and codosigid respectively (Moore, 1987) but later molecular evidence contradicted this hypothesis (Medina et al., 2003). Despite the contradiction of mentioned hypothesis with later phylogenetic analysis, indeed pluricellularity has evolved more than once in Opisthokonta, separating the evolution of animals and fungi (Lang et al., 2002; Medina et al., 2003). Multicellularity as an evolutionary strategy means an increase in body size, allowing specialisation and aerobic respiration, and improving their success in competition for limited sources (Cox and Bonner, 2001).

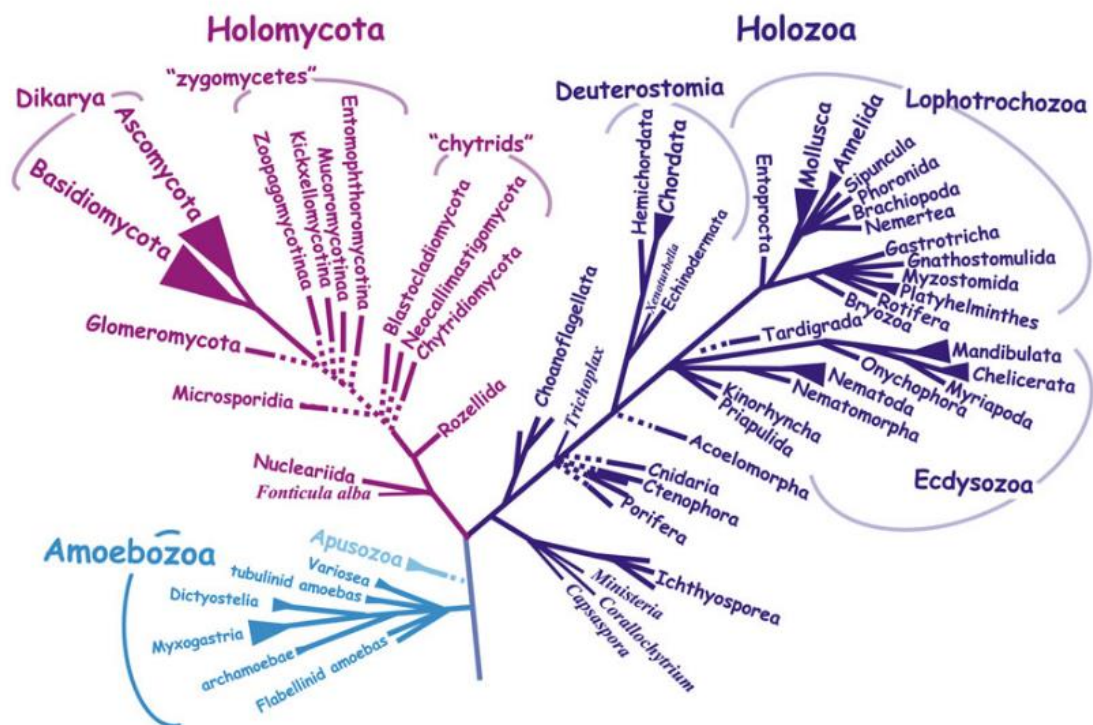


Figure 02. Consensus phylogenetic tree of the main Opisthokonta groups. Filled triangles highlight multicellular groups. The colour code separates Holomycota (purple) from Holozoa (dark blue) from the outgroup Amoebozoa (light blue). Figure 02 is reproduced under the terms of Copyright Clearance Center's license number 5302511348923 from Figure 7.3 of Baldauf et al., 2013.

1.2.3 The choanoflagellates.

The choanoflagellates have their origin in the theorised Urchoanozoan, the common ancestor between the mentioned group and metazoan. After the divergence of both groups, choanoflagellates retained at least 372 gene families that were previously considered metazoan specific, including genes involved in animal development and immunity. The evolution of choanoflagellates also implied that the retention of ancestral genes is not homogeneous, showing *Monosiga brevicollis* and *Salpingoeca rosetta* the lowest level keeping these traits (Richter et al., 2018).

The class Choanoflagellata, commonly known as choanoflagellates, was established when Kent described all known (at that time) flagellate, ciliate and tentaculiferous protozoa, plus sponges organisation in its manual of Infusoria (term used for microorganisms able of movement in liquids). Kent set the order Choanoflagellata up with the families Codonosigidae and Salpingoecidae, plus the family Phalansteriidae, which contained the genus *Protospongia* that was later renamed *Proterospongia* and moved into Amoebozoa (Kent, 1880). Although James-Clark did the first unequivocal choanoflagellate description fourteen years before when discussed about the consideration of sponges as animals instead of plants (James-Clark, 1866). The family Acanthoecidae was described much later (Norris, 1965). In the final morphological classification we can find the three aforementioned families: Codonosigidae, which includes the choanoflagellates “naked” but with a covering of glycocalyx (a coat of polysaccharides), Salpingoecidae, comprising species with a theca (microfibril-based), and Acanthoecidae, with those species with a complex basket-like lorica (composed by silica strips called costae).

Phylogenies based on the small-subunit of ribosomal DNA revealed that Codonosigidae is a polyphyletic group included in Salpingoecidae (Cavalier-Smith and Chao, 2003; Medina et al., 2003; Carr et al., 2008a; Nitsche et al., 2011). In the revised taxonomy (Nitsche et al., 2011) we find the order Craspedida and Acanthoecida (proposed by Cavalier-Smith, 1997), as it is shown in the phylogeny of Figure 03. Meanwhile Craspedida includes the choanoflagellates with organic cell coat and contain the family Salpingoecidae (term from Kent, 1880); Acanthoecida includes the species with siliceous lorica and present the families Acanthoecidae (based on Acanthoecaceae proposed by Norris, 1965) and Stephanoecidae (taxon by Leadbeater, 2011, description found in Nitsche et

al., 2011). The family Acanthoecidae includes the nudiform taxa (“naked” species with a brief stage without loricae, which is characterised by horizontal and helical costae), while the family Stephanoecidae contains the tectiform taxa (species with no naked stage, which loricae is characterised by the formation of rings).

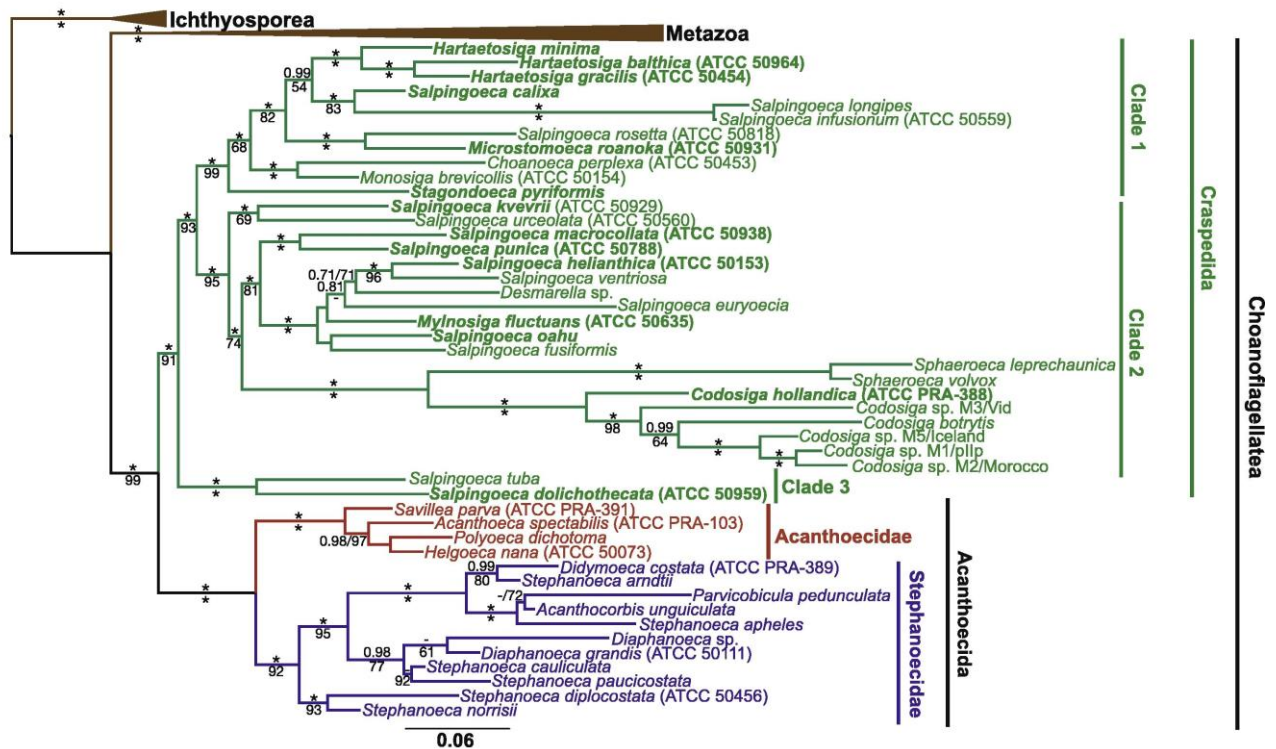


Figure 03. Maximum-likelihood tree based on a six-gene phylogeny of choanoflagellates. 47 species of choanoflagellates revealed the major clades of the group, with colour code for Craspedida (green), and within Acanthoecida it is differentiated Acanthoecidae (red) and Stephanoecidae (blue). Figure 03 is reproduced under the terms of CC BY 4.0 license from Figure 2 of Carr et al., 2017.

The choanoflagellates' morphology through species is quite consistent, as a spherical or ellipsoidal protoplast, with a single anterior flagellum surrounded by a collar of actin based microvilli (Karpov and Leadbeater, 1997, 1998). The flagellum is used both for movement and feeding, creating a flow of water to trap mostly bacteria. Furthermore, the flagellum is helped by pseudopodia that arise from the base of the microvilli, when they are activated by signals from the prey particles (Leadbeater, 2015). Choanoflagellates show an extracellular matrix or periplast that present a huge diversity, from simple ones to complex nets of more than 100 μm (Leadbeater, 1981). They are a widespread group with importance in ecology as the main bacteria predator of different aquatic ecosystems (Fenchel, 1982; Leadbeater, 2008). Most of the choanoflagellates can be found in marine water, although there are a significant amount of species, over 50, living in fresh water (Leadbeater and Karpov, 2000). In fact choanoflagellates seems to have an origin on marine water and invaded fresh water (Carr et al., 2008a). Furthermore, in adverse conditions some species can form cyst, like those described in the *Codonosiga* genus (Kent, 1880).

James-Clark (James-Clark, 1866, 1868, 1871) was also the first one noticing the similarities between the choanoflagellates and the choanocytes of the sponge *Leucosolenia botryoides*, starting a debate about the metazoan evolution. Choanoflagellates are usually free living cells, although many species are able to be seen forming colonies, what possibly could be a conserved ancestral trait (Kent, 1880; Carr et al., 2008a). The most widely studied colony of choanoflagellates belong to the species *S. rosetta*, which is able to present two kind of colonial forms: a linear chain colony, and a characteristic rosette shape with tightly packed cells, resembling a morula stage from an embryo, with the flagella facing out. The colonies were demonstrated to be produced by cellular division, not by aggregation (Fairclough et al., 2010). Thus, of the three cell types that can be distinguished in *S. rosetta* (thecate, slow swimmer and fast swimmer), only slow swimmers able to replicate can form these colonies (Dayel et al., 2011). Both chains and rosettes are connected by intercellular bridges, extracellular matrix, and filopodia, and both are free swimming (Dayel et al., 2011). While other species also have been found to be connected by intercellular bridges, like *Codosiga botrytis* and *Desmarella* Kent (Hibberd, 1975; Karpov and Coupe, 1998), some others appear attached by pairing of the collar of microvilli (Leadbeater, 1983).

The formation of colonies in choanoflagellates could resemble an early stage of pluricellular evolution. It is also relevant the fact that numerous molecular functions considered unique to Metazoa have been discovered in choanoflagellates, like tyrosine kinases, tyrosine kinase receptors as GPCR and MBRTK1, cadherins and C-type lectins (King and Carroll, 2001; King et al., 2003). To these discoveries we need add the fact that choanoflagellates were identified and well supported as the closest group to Metazoa (Medina et al., 2003; Lang et al., 2002; Ruiz-Trillo et al., 2006, 2008; Carr et al., 2008a; Torruella et al., 2012). All of this makes choanoflagellates an organism with the potential key to highlight the complex evolution to Metazoa.

Numerous choanoflagellates' sequencing have been made to study evolution and the transition of life to pluricellularity. To the first publication of genomes for *S. rosetta* and *M. brevicolis* (King et al., 2008; Fairclough et al., 2013), multiple protein-coding genes of choanoflagellates were added publicly (Carr et al., 2008b; Leadbeater et al., 2008; Richter, 2013). Choanoflagellates' database has been expanded even more with the publication of sequences for 18 new species inside the order Craspedida, which was gathered in a taxonomic analysis of 47 species of choanoflagellates for a comprehensive six-gene phylogeny (shown in Figure 03) (Carr et al. 2017). In the present study we use the transcriptome sequencing of the 19 species published by Daniel Richter (Richter et al., 2018), that represent each major lineage of the choanoflagellates tree (Carr et al., 2017) (shown also in Figure 03).

1.2.4 Evolution in Metazoa.

As the Urchoanozoan gave rise to choanoflagellates, the Urmetazoan has been proposed as the key organism to represent the first animal or the last common ancestor of animals, signalling the transition of the ancestor of choanoflagellates to pluricellularity around 600 MYA (Müller, 2001). The features of this organism have been of great interest, its genome has been trying to be reconstructed, and its evolution resulted in a comprising of old, new, rearranged and repurposed genes (Richter et al., 2018).

Metazoa is a monophyletic group that comprises pluricellular organisms, and the earliest metazoan to branch out were traditionally considered the poriferans, but that statement is now debated, with some authors supporting the earliest branching of the phylum Ctenophora (Whelan et al., 2015; Simion et al., 2017) (a phylogeny showing this non-traditional approach can be seen in Figure 04). Ctenophora are marine invertebrates with eight characteristic rows of comb plates (Ryan et al., 2013), and includes the Classes Tentaculata, organisms with a pair of contractile tentacles (Eschscholtz, 1825) and Nuda, organisms without tentacles (Chun, 1879). Ctenophora contains the commonly known as comb jellies, like the Tentaculata *Mnemiopsis leidyi* (the sea walnut or warty comb jelly, included in the phylogeny of Figure 04), whose genome was sequenced to give an insight of early animal evolution (Ryan et al., 2013). This phylum has presented a strong support as sister group to the rest of animals, challenging the traditional view and recognising comb jellies having an independent path in evolution for neurogenetic, immune and developmental genes (Dunn et al., 2008; Moroz et al., 2014; Whelan et al., 2015). Although this discussion is still ongoing, as more phylogenies with strong support for Sponges as the sister group of all animals are still recovered (Simion et al., 2017; Redmond and McLysaght, 2021).

At the base of the metazoan tree of life the next organisms we can find are the sponges (phylum Porifera), still supported by some authors as the earliest branching Metazoa (Feuda et al., 2017), followed by its diversification into the four main lineages, the classes Hexactinellida, Calcarea, Demospongiae and Homoscleromorpha (Lake, 1990; Kobayashi, 1996; Kruse et al., 1998; Müller, 1998; Zrzavý et al., 1998; Borchellini, 2000; Medina et al., 2001; Gazave et al., 2011). Although later studies suggest that Porifera is paraphyletic, with Calcarea more related to Eumetazoa than to other sponges (Zrzavý et al., 1998; Borchellini, 2000, Borchellini, 2001), phylogenetic studies agrees with the morphological classification in a monophyletic origin (Philippe et al., 2009), with incongruences in Calcarea yet to be resolved. Thus, the poriferan morphology, based on choanocytes and aquiferous system, has been speculated to resemble a first stage on pluricellular evolution. So, with the aim to unveil animal complexity development, the genome of *Amphimedon queenslandica* was sequenced, a sponge belonging to the class Demospongiae (Srivastava et al., 2010) (species included in the phylogeny of Figure 04).

From this point, organisms are classified as Eumetazoa, as they are able to form true tissues derived from germ layers that appear during the gastrula stage of the embryo. A consideration is necessary here, as if Porifera branched earlier than Ctenophora, this latter would be included in Eumetazoa, and not evolving eumetazoan traits independently. The first group branching out at this stage is Cnidaria (Dunn et al., 2008), with the exception of Placozoa, whose uncertainty will be discussed later (although this exception is kept showing both groups in the phylogeny of Figure 04). This group has nematocytes as main trait (also known as cnidocytes or cnidoblasts), a cell type used for predation, adhesion, locomotion and defence (Lenhoff and Lenhoff, 1988; Beckmann and Özbek, 2012). It includes the subphylum Anthozoa (corals, sea anemones and sea pens), Medusozoa (sea wasps, jellyfish and *Hydra*) and Endocnidozoa (jellyfish-like obligate parasites of the groups Myxozoa and Polypodiozoa) (Kayal et al., 2018). Although there is still some discussion about the groups inside the major clades, and the parasitic species needed a long history of research, this classification is now well supported. As representatives of the phylum Cnidaria, the genome of the species *Nematostella vectensis* (the starlet sea anemone, subphylum Anthozoa, included in Figure 04) and *Thelohanellus kitauei* (an endoparasite of the common carp, group Myxozoa) have been sequenced to study early animal evolution (Putnam et al., 2007; Yang et al., 2014).

The phylum Placozoa is a controversial one. As it contains a simple layered organism, formed by a lower and an upper epithelium plus a fibre cell net-layer joining them, with just six different cells (Rassat and Ruthmann, 1979; Guidi et al., 2011). This species was believed to be at the base of the tree as the simplest animal. Phylogenetic studies reveal a strong support to consider it a sister group to Cnidaria, but if Cnidaria is closer to Bilateria or Placozoa is still unknown (Laumer et al., 2018). For over a century Placozoa contained a single organism, *Trichoplax adhaerens*, that can be found in diverse ecosystems of coastal waters, from tropical to moderately cool waters (Tomassetti et al., 2005; Pearse and Voigt, 2007; Eitel and Schierwater, 2010). In a search for diversity within this phylum, *Trichoplax* was divided into haplotypes (eight initially) (Voigt et al., 2004). In the last years, one of the haplotypes was differentiated in the species *Hoilungia hongkongensis* (Eitel et al., 2018), and also the species *Polyplacotoma mediterranea* was described (Osigus et al., 2019).

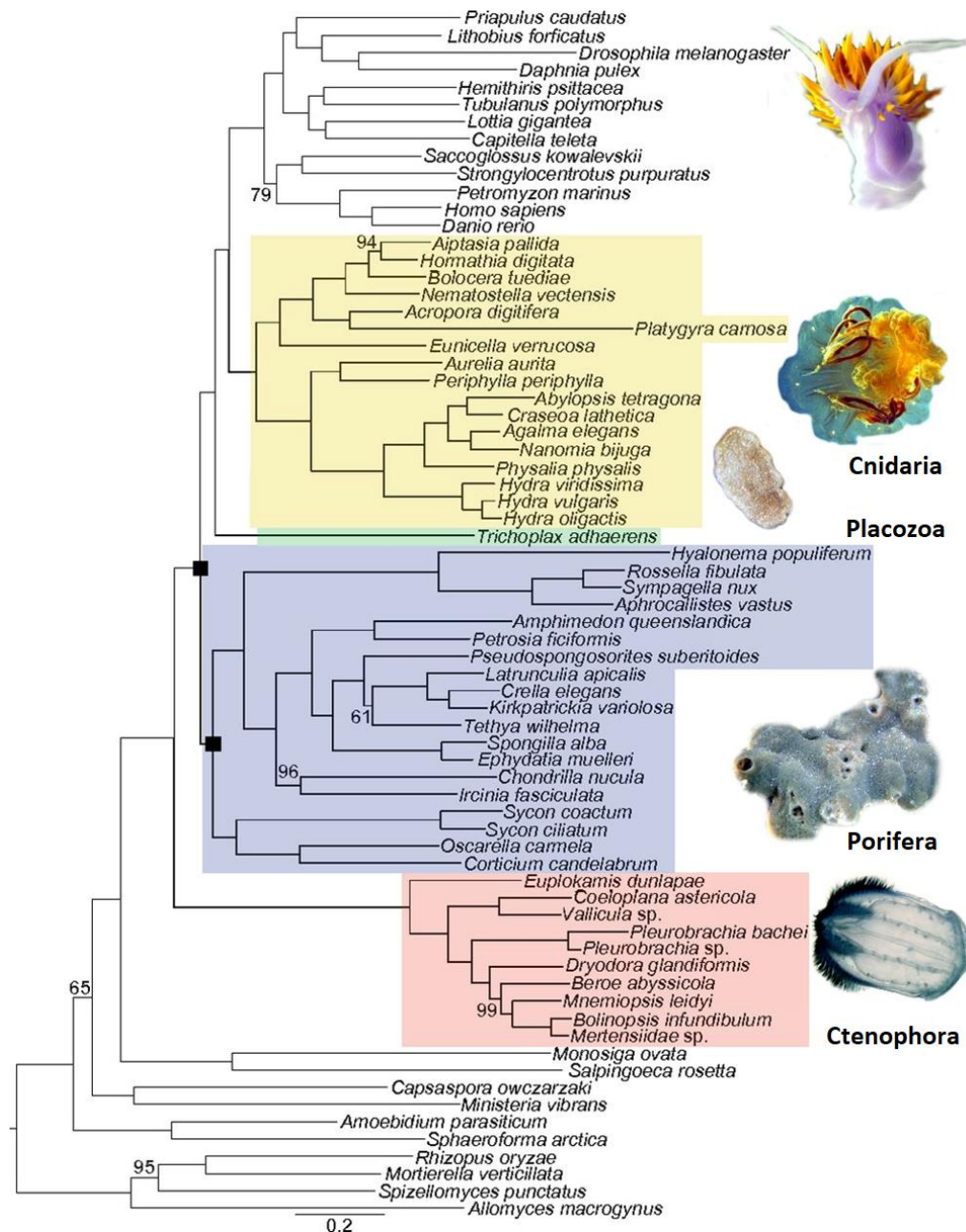


Figure 04. Maximum-likelihood tree of the early branching metazoan. Nodes are supported with 100% bootstrap except when specified. Different datasets with maximum-likelihood and Bayesian analysis gave the same topology. This phylogeny highlight Ctenophora (species framed in a pink square) as sister group to the rest of animals, and also emphasise the position of Porifera (species with blue background), Placozoa (species framed in green), and Cnidaria (species framed in yellow). The image has been modified to include the name of the mentioned groups at the right side of the figure. Figure 04 is reproduced under the terms of PNAS Copyright license from Figure 3 of Whelan et al., 2015.

The next major clade classify the rest of organisms as Bilateria, animals presenting bilateral symmetry during the embryo development (Haeckel, 1874). This clade emerged in the Ediacaran period (635-541 MYA) and it is considered to establish the rules for ecology and evolution of pluricellular life before the diversification of Metazoa in the Cambrian Explosion (Peterson et al., 2008). This group include Deuterostomia, and Protostomia (Grobbs, 1908), both are differentiated attending to the embryonic blastopore (first opening), becoming the mouth (protostomes) or the anus (deuterostomes), among other developmental features.

For the rest of the tree of life, Protostomia were hugely diversified a wide array of organisms (Mollusca, Annelida, Phoronida, Brachiopoda, Nemertea, Platyhelminthes, Priapulida, Kinorhyncha, Nematoda, Nematoporida, Tardigrada, Onychophora, Tetraconata, Myriapoda, Chelicerata), and Deuterostomia gave rise to the most complex extant organisms (from the simpler Xenotubellida, Echinodermata and Hemichordata to the variety of Chordata where *Homo sapiens* can be found) (Dunn et al., 2008).

This short review through the eukaryotic tree, from the very beginning to the most recent and complex species, still does not reply many evolutionary questions. As we have already seen diversification of species, we will discuss next some of the forces influencing evolution.

1.3 CODON USAGE AND EVOLUTIONARY FORCES.

1.3.1 The genetic code and the codon usage of genes.

The central dogma of molecular biology states the directional flow of information, that can go from DNA to RNA and finally to proteins, but not from protein to protein or amino acids (Crick 1958), a statement that is still valid today. The information contained in the DNA is copied in a complementary messenger RNA during the process of transcription, and deciphered in the process of translation carried out in the ribosomes. The rules to decipher this information are what we know as the genetic code (the genetic code is shown in Figure 05).

Considering that for both DNA and RNA the units of information are the nitrogenous bases contained in the nucleotides, George Gamow was the first one proposing sets of three nitrogenous bases encoding for each amino acid (Gamow, 1954). This was demonstrated experimentally seven years later in an experiment inducing mutations (Crick et al., 1961). The triplets of nitrogenous bases are known as codons, containing adenine (A), guanine (G), cytosine (C) and thymine (T) in the case of DNA, and uracil (U) replacing thymine for the RNA. The combination of 4 possible bases in triplets gives 64 possible sets to codify for the 20 amino acids, meaning that all amino acids (except for methionine and tryptophan) are codified for more than one codon, which is commonly known as the degeneracy of the genetic code (this degeneracy is shown in Figure 05).

		SECOND									
		U		C		A		G			
FIRST	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U	THIRD
		UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C	
		UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A	
		UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G	
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U	
		CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C	
		CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A	
		CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G	
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U	
		AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C	
		AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A	
		AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G	
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U	
		GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C	
		GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A	
		GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G	

Figure 05. Genetic code emphasising its degeneracy. Codon in grey background are two fold degenerate, the one in green is threefold degenerate, the yellow are fourfold degenerate, and in blue the sixfold degenerate. Those non degenerate or stop codon are in white and red respectively. Modified image deleting non-relevant information relative to non-canonical codon usage. Figure 05 is reproduced under the terms of Copyright Clearance Center's license number 5302520108640 from Figure 1 of Agris et al., 2007.

Different codons that codify for the same amino acid are called synonymous codons, and theoretically they should appear in the same proportion, but in some cases the use of codons have been reported not randomly distributed (see Chapter 1.4 for a deep review). The discussion about codon usage is even older than DNA sequencing itself, the first method for sequencing appeared in 1970, but previous research let us find references about relative codon usage discussions in 1966 as an inexplicable phenomenon (Woese et al., 1966). More than a decade later, in 1979, the preferential codon usage on ribosomal protein genes of *Escherichia coli* was correlated with the most abundant tRNA extant, revealing its impact in translation efficiency (Post et al., 1979). An early explanation for the origin of codon bias in viral genomes involved thermodynamic considerations in the primitive DNA (Rowe and Trainor, 1983). Since these early days, the study of codon usage has been widely developed with the apparition of whole genome sequencing (Duret, 2002).

A clarification here is necessary, as we use both the terms “preferred codons” and “optimal codons”. For Ikemura, both were similar terms but with nuances (Ikemura, 1982). While preferred codons reference to the highly used codons in an organism, and it is the codon choice, optimal implies an efficiency for the translation process in that organism. This nuance sometimes disappeared from literature, and just remained the definition of optimal codons as any codon used significantly higher than other codons in highly expressed genes (Duret, 2000). We will use preferred when necessary to keep the nomenclature of the original authors and make easier the intertextuality, although optimal codons is the definition that implies a consideration for expression levels.

1.3.2 The origin of the genetic code.

The genetic code is generally considered universal for all life forms (Abel 1964; Woese et al., 1964; Vetsigian et al., 2006), as its structure remains the same through all the species (shown in Figure 05), presenting deviations that involves only a few codons. The variations of the genetic code were discovered for the first time in human mitochondria, as it uses UGA and AUA for tryptophan and methionine, instead of termination and isoleucine respectively (Barrel et al., 1979). Followed by the discover of other variations in mitochondria and chloroplasts, beside the fact that green plant mitochondria uses tRNAs acquired from chloroplasts (Jukes and Osawa, 1990); variations in

bacteria, like the CGG as an unassigned codon in *Mycoplasma capricolum* (Oba et al., 1991); variations in fungi, like the identification of one clade using CTG for serine (like *Candida albicans*), in opposition to other major clade using CTG for leucine (like *C. lusitanae*) (Fitzpatrick et al., 2006); and even with virus showing in some cases the same variation that their hosts, like *Scheffersomyces segobiensis*, a fungal RNA virus (Taylor et al., 2013). This array of exceptions demonstrate that the genetic code can still evolve.

How the amino acids were assigned to codons is one of the major questions in biology and gave rise to numerous theories. The key-lock theory address the issue proposing a direct stereochemical affinity between amino acids and oligonucleotides (Orgel, 1968; Yarus, 1998), while the coevolution theory claim that the code could have co-evolved reflecting pathways of amino acid biosynthesis (Wong, 1975). In the same way, the error minimisation theory, states that the main selective force for the code was the reduction of translational misreading (Woese, 1965; Goldberg and Wittes, 1966), and differs from the frozen accident, which assumes a random assignment of codons and that the universality of the genetic code is due to the fact that any change on it would be highly deleterious (Crick, 1968).

Evolution does not have foresight, it works in the present without considering future advantages, so the code must have appear before translation attending to necessities of the primitive RNA world. With that premise, a theory set the translation system as a by-product of the selection of amino acid binding ribozymes for reactions catalysis (Wolf and Koonin, 2007). All the aforementioned possibilities for the origin of the genetic code are not exclusive, for instance the universality of the code could have been fixed in the common ancestor, where could have evolve a few amino acids thanks to chemical affinity, that even could have been ambiguous for class of amino acids (Woese, 1965). Afterwards, more amino acids were assigned through modification of existing tRNA, and selected against nonsense codons. The fact that the actual code is not the most stable possible (Koonin and Novozhilov, 2009) suggest that it appeared randomly and was optimised by codon relocation, until reaching a stage where any reassignment could be highly deleterious.

The hypothesis of “collective evolution in primordial replicators” frames the universality of the code both as an outcome and a necessity for the horizontal transfer of genetic information in primitive communities. The horizontal transfer could have caused competition among groups of organisms with similar codes, competing for niches. The genetic code would have appeared by a frozen accident, while the advantage of the flow of genetic information among communities of organisms could have acted as a selective factor, due to an advantage of harmonising the code to get benefit from foreign sequences. The theory has proven to lead to robust and universal codes (Anderson, 1970; Syvanen, 1985; Martin and Russell, 2003; Vetsigian et al., 2006; Goldenfeld and Woese, 2007).

The origin of tRNA is proposed by the ligation of a couple of RNA hairpins, from minigenes used for protein synthesis that helped to establish the main phyletic lines, rendering its origin polyphyletic (Di Giulio, 1992; 1999). In this duplicated structure, the anticodon loop is equivalent to the acceptor stem (Di Giulio, 2004). This theory could be seen as complementary to the exon theory, that explains the formation of RNA by splicing of minigenes mediated by introns, as in fact tRNA introns appear mainly in the anticodon loop joining the two parts of the molecule (Gilbert, 1987; Di Giulio, 1995; Roy, 2003).

To address the apparition of the translation system, a point of view more tRNA centric is considered through two key interactions: amino acids-tRNA and tRNA-mRNA. The first one is carried out by an enzyme called aminoacyl-tRNA synthetase, but in the primitive world there should have appeared a mean to capture amino acids selectively. Primitive aminoacylation could have been carried out by ribozymes (Bessho et al., 2002; Murakami et al., 2003), so this interaction could have evolved before the enzyme apparition (Wong, 1991; Noller, 2004), taking advantage of the easy and precise recognition of anticodons by complementarity.

The second interaction, tRNA-mRNA, is done by an adaptor tRNA thanks to an anticodon and the amino acid attached to the 3' end. This interaction could have been originated from the participation of triplets in amino acid binding (Yarus et al., 2005) or from oligomers that stabilized early tRNA (Grosjean et al., 1976). The mRNA could have emerged as random sequences coding small favourable peptides, while the ribosome evolved from a simple peptide ribozyme, acting as an entropy trap, developing the ordering later by itself or by the action of aminoacyl-RNA (Sun et al., 2002; Tamura and Schimmel, 2003; Sievers et al., 2004; Bernhardt and Tate, 2010). These

favourable peptides may have increase competitiveness through RNA stabilisation, RNA catalysis for replication or ligation, or even vesicle stability and permeability (Gilbert, 1986; Zaher and Unrau, 2007; Mayer et al., 2018; Frenkel-Pinter et al., 2020). Over time, enzymes competed against ribozymes, erasing the trace of their origin during their replacement (Benner et al., 1989).

1.3.3 Evolutionary forces upon the codon bias.

The preferential use of certain codons during translation is what we know as “codon bias”. As the genetic code is degenerate (as was shown in Figure 05), most amino acids are encoded by multiple codons, so each organism has the possibility to use mainly certain codons over others through its genome. Different codons coding for the same amino acid is what we know as synonymous codons, and isoaccepting tRNAs are the tRNA for synonymous codons, meaning that they are attached to the same amino acid.

Early in the study of codon usage something was noticed, codon bias was a general phenomenon, in each species genes have a similar bias trend, with more similarity to related species than to those distant in the phylogenetic tree (Grantham et al., 1980; Ikemura, 1985; Naotake, 1985). But genes in each species also showed bias in a subset of codons, correlated with highly expressed genes and tRNA isoacceptor, and in mammalian it was also related with GC richness (Bennetzen and Hall, 1982; Ikemura 1985). The first ones addressing this question with a specific model were Kimura (Kimura, 1981; 1983), proposing an optimal state when the amount of codon frequency matched the amount of isoaccepting tRNA available, and We Li (Li, 1987), who propose bias as a balance between two categories of codons to achieve fitness. Although both theories are outdated, this led to formulate the selection-mutation-drift theory (Sharp and Li, 1986; Bulmer, 1987; 1991), the most accepted nowadays and that we will discuss below.

The forces shaping the codon bias have been defined in two major categories: mutation and selection, while drift is a random process that is difficult to test. The mutation explanation states that codon bias is due to an inherent bias in the mutation patterns, thus the difference in codon patterns among organisms would be due to different mutational biases (Eyre-Walker, 1991). An example of this can be seen when organisms present GC bias in a genome level, instead of in specific

sequences (Kanaya et al., 2001a; Chen et al., 2004). Selectionists attribute codon bias to a selective pressure for an advantage in efficiency or accuracy during translation. This phenomenon could explain why the most abundant codons have been correlated with the most abundant tRNA in the cell, and also how abundant codons also could correlate with highly expressed genes (Ikemura, 1985; Kanaya et al., 1999; 2001b), generating different biases within genomes.

As neither mutation nor selection can explain the bias attributed to the other category, both were reconciled in the currently most accepted model, the selection-mutation-drift balance, also called the major codon preference model (Sharp and Li, 1986; Bulmer, 1987; 1991). This model proposes that while mutation pressure and genetic drift introduce changes and bias in the genome, selective pressure act upon genes to keep those mutation increasing the preferred codons for the cell, and decreasing the number of mutations introducing unpreferred codons. This model propose a weak selection acting constantly over time, although other studies, like the performed on *E. coli*, showed continuing strong selection for highly expressed genes, and continuing weak selection for genes with low level of expression (Sharp et al., 2010).

The selective pressure acting in the selection-mutation-drift model could be the product of the advantage of using codons that increase the accuracy or the efficiency of the translation. Efficiency was the first explanation proposed and the traditional view (Ehrenberg and Kurland, 1984). Under efficiency pressure the organisms with preferred codons would have the advantage of reducing the translation time, increasing the speed of protein production, while other codons retard the expression. Another effect of efficiency is that highly expressed genes does not hoard all ribosomes, allowing mRNA from weakly expressed genes to be translated. Efficiency would maximize the growth rate of organisms, giving them competitive advantage (Andersson and Kurland, 1990; Sharp et al., 2010). Efficiency data are also related with the amount of tRNA available in the cell, as its availability is the direct responsible for this increase in translation speed (Kanaya et al., 2001b). The selective pressure for accuracy does not favour directly the preferred codons, but as unpreferred codons could produce mistranslated proteins, that can be toxic and decrease their fitness (Drummond and Wilke, 2008). Higher expressed genes have more opportunities to be mistranslated, so they are under more pressure to reduce toxicity, but an impact of mistranslation upon efficiency cannot be completely ruled out (Bulmer, 1991).

In the circumstances of weak selection, mutation bias is sensitive to the population size, showing a positive correlation with unpreferred mutations (Akashi, 1995). Although in other studies have been found no correlation of selection with effective population size in numerous animals, including insects (like the mosquito *Culex pipiens*) molluscs (like *Ostrea edulis*), crustaceans (like *Necora puber*), echinoderms (like *Abatus cordatus*), annelids (like *Allolobophora chlorotica*), nematodes (like *Caenorhabditis brenneri*) and a cnidarian (*Eunicella cavolinii*) (Galtier et al., 2018), and has not been found correlation with 41 mammals like humans, dog (*Canis familiaris*), chimpanzee (*Pan troglodites*), rabbit (*Oryctolagus cuniculus*), mouse (*Mus musculus*, *M. domesticus* and *M. castaneus*) or dolphin (*Tursiops truncatus*) (Kessler and Dean, 2014). Nonetheless this is indeed important in other organisms with large effective population sizes, as it has an impact on the efficiency of selection (Snoke et al., 2006). A large population size increases the efficiency of selective pressure upon synonymous sites, favouring the fixation of beneficial mutations and lowering the mildly deleterious ones (Ohta, 1973; 2002). This effect has been seen in species like ciliate eukaryotes of the genus *Paramecium* (Snoke et al., 2006), that showed a high level of synonymous codon polymorphism that is attributed to its big population over time.

The selection coefficient measures the fitness difference of the mutation among the individuals of a population, and its outcome is the potential fixation of favoured mutations or alleles (Bersaglieri et al., 2004). This statistic has been widely used to describe evolution quantitatively as it reflects selection through a change in genotype frequencies. Overall, high mutation rates, a small population size or small selection coefficients are negative factors for the selection for codon bias (Hartl, 1994; Akashi, 1995).

1.3.4 Consequences of codon bias.

The codon bias importance branches from affecting directly the protein synthesis. During the flow of genetic information, codon bias could affect the secondary structure of the RNA. A strong secondary structure in the 5' end of mRNA can inhibit translation, and other folding could change the protein levels, having a control over translational efficiency (Kudla et al., 2009; Novoa et al., 2012). Codon bias control the speed of the translation process, as preferred codons by the organisms reduces the time needed to incorporate amino acids, meanwhile other codons

contribute to reduce the errors in translation (Andersson and Kurland, 1990; Drummond and Wilke, 2008; Sharp et al., 2010; Novoa et al., 2012). It has been reported that codon optimisation could increase protein production up to 1000 fold (Gustafsson et al., 2012). The next advantage derives from the previous one, as controlling the speed of translation can impact the protein folding and function (Zhao et al., 2017; Lyu and Liu, 2020). It has been demonstrated that pauses occur during translation of some proteins, and also that synonymous mutations have an impact in the folding of the protein, so it has been hypothesized that controlling synonymous codons can improve protein folding (Novoa et al., 2012). As an example, the de-optimisation of an antibody resulted in an increase of its expression (Magistrelli et al., 2017).

The potential of codon bias was recognised quite early, due to the correlation of bias with level of expression it was used to develop an early software to identify ORFs (open reading frames) (Staden and McLachlan, 1982), and also it was used to identify coding regions (Warrick and Spudich, 1988). Population biology also could benefit from codon studies, as variations in the strength of selection give insights, for instance, in the changes of effective population size in the species (Ohta, 1973; 2002; Bersaglieri et al., 2004). With the development of research in codon usage, biotechnology could take advantage of the explained traits, but in addition it could be used during heterologous expression. Knowledge of codon usage can be used to optimise a transgene for a specific host, helping to avoid the phenomena of tRNA starvation, that could happen when a transgene uses codons very differently from the host tRNA abundance. An improved heterologous gene expression could be translated in an improvement to design gene therapy and vaccines, and this could be achieved using libraries of randomized genes that determine the parameters for efficiency (Plotkin and Kudla, 2011), although the other effects of codon usage must also be considered.

The evolutionary information that we can extract from the codon bias, important to study gene family evolution and gene organisation, together with its potential impact in biotechnology, makes codon usage an interesting field to study evolution both with basic and applied impact. In addition, the focus in choanoflagellates, other Opisthokonta and early branching metazoan could be a step for unveiling one of the most important steps in evolution history, the transition of animals to pluricellularity.

1.4 CODON BIAS IN EVOLUTION AND THE TRANSITION TO PLURICELLULARITY.

1.4.1 Codon bias in microscopic organisms.

One of the first studies in codon bias can be dated in 1972, showing a clear synonymous codon bias of nuclear genes in yeast and seven bacteria (including the model organism *E. coli*) that was attributed to a pressure to keep ratios in synonymous codons (Goel et al., 1972). Codon usage was further discussed in *E. coli*, that showed a high bias using genes for ribosomal proteins, and also showed preferred codons align with the most abundant tRNA. In this case it was attributed more accurately to a selection by a need for efficiency or accuracy (Post et al., 1979; Moriyama and Powel, 1998). In bacteria, the leading strand is the one with higher GT bias (Lobry, 1996), like in *E. coli*, *Haemophilus influenzae* and *Bacillus subtilis*. In opposition, examples of bacteria with different GC content but no evidence for selection are spirochaetes (like *Borrelia burgdorferi* and *Treponema pallidum*) (Lafay et al., 1999) and the proteobacteria *Helicobacter pylori* (Lafay et al., 2000). Sharp et al. analysed a diverse array of 80 bacterial species and concluded that 30% of them were no under selection (Sharp et al., 2005), and also showed that strong codon bias is highly correlated with bacterial growth rate (Sharp et al., 2010), pointing out towards translational efficiency causes.

The model organism *Saccharomyces cerevisiae* is also a species thoroughly studied for codon usage bias. An early study of nine genes revealed its strong bias in highly expressed genes and *vice versa*, with preferred codons correlated to isoaccepting tRNA abundance, indicating natural selection (Bennetzen and Hall, 1982; Ikemura, 1982). Results were consistent with those obtained with a more comprehensive gene sample (Moriyama and Powel, 1998). *S. cerevisiae* possess a great variability of gene GC content among regions, ranging from 35 to over 50%, attributed in this work to mutational biases (Sharp and Lloyd, 1993). The codon usage knowledge in this species has been applied to improve the production of L-Arabinose transgenes (Wiedemann and Boles, 2008). The study of codon bias in yeasts have also a biotechnology potential in the production of terpene and other lipids. In this context, red yeasts species were characterised for codon bias, determining mutational bias playing an important role in the genome composition, but also showing evidence for selection upon efficiency (Camiolo et al., 2019).

Considering other entities with genomic information, RNA viruses have shown codon bias, like the influenza A virus, evolving through mutational bias and natural selection, and AT rich (uncorrelated with its hosts) (Goñi et al., 2012). Another example are exogenous retroviruses like the human immunodeficiency virus, strongly biased towards A, although with restrictions of the host to highest expressed genes (Hemert and Berkhout, 1995). DNA viruses, despite a lower level of nucleotide substitution, still show codon bias, like in papillomavirus, whose bias is correlated with the tRNA availability (Zhou et al., 1999).

Codon bias could appear in all kind of organisms. In the domain Archaea we have the example of the methanogenic *Methanococcus maripaludis*, whose genome is strongly AT biased and also contain optimal codons correlated with expression level (Emery and Sharp, 2011). But in this case, the bias was stronger in twofold degenerate amino acids, and weaker in fourfold ones.

In the domain Eukarya, after the model organism *S. cerevisiae*, numerous species across the phylogenetic tree have shown codon bias. Trypanosomatids (belonging to Excavata) like *Trypanosoma cruzi*, *T. brucei* and *Leishmania major* showed a codon bias conserved in syntenic genes, and a bias correlated with gene expression and gene copy number. These species showed translational selection (Horn, 2008), while another excavate, *Trichomonas vaginalis*, showed a codon bias shaped by a combination of mutational pressure plus translational selection (McInerney, 1997). In *Entamoeba histolytica*, an amoebozoan with extremely AT rich genome of 75%, the third codon position is also biased towards AT, but it keeps a subset of codons ending in C despite the trend (Romero et al., 2000). *Dictyostelium discoideum* (Amoebozoa) and *Plasmodium falciparum* (Alveolata) are also AT rich species, and as *E. histolytica*, appeared to be biased by a combination of compositional pressure and selection. In particular, *D. discoideum*, with AT of 78%, also showed a correlation of preferred codons with gene expression (Sharp and Devine, 1989). *Tetrahymena thermophila* and *Paramecium tetraurelia*, both organisms also classified as Alveolata, are also AT rich (72.3% and 69.9%), but with higher GC content in highly expressed genes. Also, both presented reassigned stop codons with different efficiency among codons and species (Salim et al., 2008).

This array of unicellular organisms (plus viruses) with codon bias lead us to discuss holozoan protists as the next step. Nonetheless, due to their importance for this project we will discuss them more extensively in the next section.

1.4.2 Codon bias in holozoan protists.

Choanoflagellates, as cosmopolitan eukaryotes with intercontinental distributions (Carr et al., 2008; 2017; Nitsche et al., 2011), may have a large effective population size, so they are strong candidates to study process of natural selection. Carr et al. studied for the first time the codon usage in the group Choanoflagellatea (Carr et al., 2017), in their phylogenetic study using six genes in 47 diverse species (phylogeny previously presented in Figure 03 of section 1.2.3). Although the study is mainly focused on phylogeny, codon usage were studied in elongation factor 1A and elongation factor like, calculating the effective number of codons (see 1.5.2 for a deeper explanation) and finding data consistent with selection (with no definitive evidence). In a Carr unpublished work, he also analysed the level of codon bias in choanoflagellates using four genes in 21 species, giving the first sorting by the level of bias (presented in Figure 06).

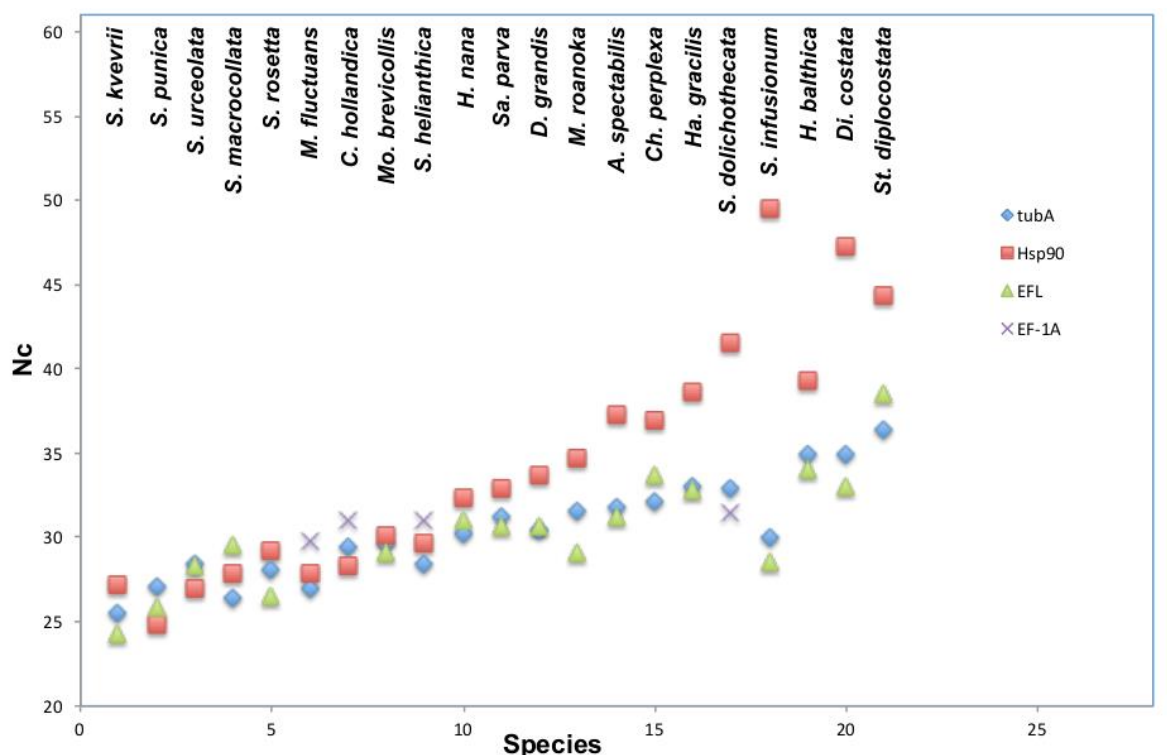


Figure 06. Comparison of the level of bias in choanoflagellates using four genes. Bias is measured as Nc or effective number of codons, an overall codon bias value (see 1.5.2 for an extended explanation) (Carr et al., pers. commun.). The lower the Nc value, the higher is the codon usage bias, thus we could point out *Salpingoeca kvevrii* as the more biased species and *Stephanoecca diplocostata* as the less biased one.

Southworth studied the choanoflagellates *S. rosetta* and *M. brevicollis*, and the filasterean *C. owczarzaki*, all showing patterns of codon bias. There is a preferential use of codons ending in G or C, clearly distinguishable for lower, medium and higher expressed genes in the species *S. rosetta* and *C. owczarzaki*. Optimal codons also showed a high correlation with the higher tRNA gene copy number in these two species, revealing selection for translational efficiency as the major bias driver in these species. In addition, protein domains display higher codon bias than protein non-domains, a result compatible with selection for translational accuracy. Thus, *S. rosetta* and *C. owczarzaki* revealed a strong evidence for selection driven by translational efficiency, and the three of them also showed to be driven by translational accuracy, which is reflected even in the weakly biased genes. At the same time, noncoding regions did not show the same GC pattern, discarding mutational pressure as a possible major bias driver. In this deep study of codon usage, the strength of selected codon usage bias was calculated for the genome and the two fold degenerate amino acids (Southworth et al., 2018), which will need further discussion in this work (see Chapter 5.2). The similarities in these three species are not just a matter of similar evolutionary forces acting upon their genomes, despite the divergence of choanoflagellates and *C. owczarzaki* more than a billion years ago (Parfrey et al., 2011), codon usage is highly conserved among them. Beside the similarities in GC ending codons, we need to highlight the similar set of optimal codons except for seven amino acids, which at least shares one optimal codon. The main difference among them is that *C. owczarzaki* may have the amino acid selenocysteine identified as stop codon, due to the fact that tRNA for this amino acid with UGA anticodons were identified (Southworth et al., 2018).

1.4.3 The transition to pluricellularity.

The development of pluricellular organisms meant a historical step on the complexity of life. These organisms are constituted by differentiated cells, able to communicate among themselves and sharing an overall coordination, which made them lose their ability to survive as free living cells, but gained a division of labour. Nonetheless, pluricellularity is not a single lucky chance in evolution, this is a trait that has appeared at least 25 times, including multiple times in plants, fungi and bacteria (Medina et al., 2003; Grosberg and Strathmann, 2007; Parfrey and Lahr, 2013). Although in 3.5 billion years not all transitions developed the same level of diversity and complexity, emerging the whole Metazoan group from a single transition (Schopf, 1993; King, 2004).

Among the main reasons for the development of pluricellularity, the first and most intuitive advantage is the possibility to escape phagotrophic predators. This selective pressure was demonstrated successfully in lab conditions using a unicellular algae (*Chlorella vulgaris*) with a phagotrophic flagellated protist (*Ochromonas vallescia*) (Boraass et al., 1998). The second characteristic is that it increases the efficiency of nourishment and growth rate, and it suppose an advantage especially in environments rich in nutrients (Koufopanou and Bell, 1993). In bacteria, aggregations have demonstrated to increase their resistance to environmental stress both for physical and chemical conditions (Justice et al., 2008). Additionally, multicellularity allows different cells to be diversified and specialised, increasing the performance of different processes that can be divided by location and time (Flores and Herrero, 2010), contrasting with unicellular organisms, that have the limitation of dividing labour exclusively by time.

The Urmetazoan would have lived 800 MYA, an enigmatic pluricellular organism ancestor of all Metazoan, and also the simplest one (Hirabayashi and Kasai, 1993; Müller, 2001; Müller et al., 2004). Although unique, its development could have been promoted by a set of genetic tools present in the unicellular eukaryotes. The genes promoting this pivotal event are related with cell adhesion, cell signalling, and cell differentiation. For instance, choanoflagellates showed protein families for cell interactions previously assigned to animal cells, like cadherins, C type lectins, G protein coupled receptor, tyrosine kinases, and multidomain polypeptides for signalling and adhesion (like the complement control protein, Src homology 2, epidermal growth factor and tumour necrosis factor receptor) (King et al., 2003). The choanoflagellate *M. brevicollis* was used to study the origin of metazoans (King et al., 2008), showing cell adhesion domains previously attributed just to animals (like 23 genes encoding cadherin domains, 12 genes for C type lectins, including 2 transmembrane lectins, integrin alfa and immunoglobulins), but also revealing differences in the protein domains linkages with metazoans. It is speculated that over this genetic bases new genetic information came through three different methods that lead to pluricellularity: the domain shuffling, gene duplication and divergence, and evolution of gene regulation.

New genes can appear from domain shuffling with new functions (Long, 2001). We can see evidence of this phenomena in the tyrosine kinase family, kinases that phosphorylate proteins, important for cell communication, differentiation and proliferation (Hunter and Sefton, 1980; Law et al., 2017). At least 15 of the 30 subfamilies of tyrosine kinase receptor were already present in the common ancestor of poriferan with the rest of metazoan, all showing different combinations of ligand binding domains (Suga et al., 2001). This fact points to an Urmetazoan rich in tyrosine kinase, although the little we still know about choanoflagellates tyrosine kinases did not show homology to metazoan (King et al., 2003), requiring further research to prove this connection.

Gene duplication, with its later divergence, also can generate important variability. This is what it is called a paralog gene, either coming from total or partial gene duplication, and it allows different drift and selection act upon both, generating new functions without sacrificing the ancestral one (Lundin, 1999). Tyrosine kinases, tyrosine phosphatases, Ga proteins and collagens have shown evidence of using this method before the diversification of animals (Ono et al., 1999; Suga et al., 1999; Exposito et al., 2002).

The evolution of gene regulation is another major issue in the transition to pluricellularity. It has been shown that the complexity of this transcriptional regulation correlates with metazoan complexity. By comparison, metazoan cis-regulatory DNAs uses enhancers and a more efficient spacing that allows an expression multiple times more complex than in yeasts like *S. cerevisiae* (Levine and Tjian, 2003). Also, the small RNA gene regulation is quite widespread among eukaryotes (protozoan, plants and animals), suggesting that their common ancestor could use this method for gene and transcriptional regulation (Ngo et al., 1998; Carrington and Ambros 2003; Bartel and Chen 2004). Nonetheless, little is known how the transcriptional and posttranscriptional regulation could have contribute to the apparition of pluricellularity, further research of the link with close related species to Metazoa is still necessary.

1.4.4 Codon bias in pluricellular organisms.

The interest to translate codon usage bias studies to pluricellular organisms appeared early in comparative studies (Sharp et al., 1988). Pluricellular Fungi also stimulated early interest for codon analyses in species like *Aspergillus nidulans*, a filamentous ascomycete, showing genes from lowly to very highly bias for optimal codons. Despite the fact that GC content in this species is 50%, most of the optimal codons are GC ending (Lloyd and Sharp, 1991). Fungal xylanase, a protein important for its industrial applications, also rise a great interest. Among other parameters, GC ending and optimal codons were analysed in 12 species, showing evidence for mutation and selection forces, and establishing as highly biased the species *A. nidulans*, *Chaetomium globosum*, *A. terreus* and *A. clavatus* (Shrivastava et al., 2009). The fungal model organism *Neurospora crassa* was screened at large scale for codon mutations. It was found that optimal codons segregate more frequently than non-optimal codons, a result consistent with natural selection, which with some level of genetic interference shapes the genome of this species (Whittle et al., 2012). Other fungal organisms have been analysed revealing optimal codons and bias, like the basidiomycete *Volvariella volvacea* (Jiang et al., 2014). In another relevant study, optimal codons were determinant in fungal pathogen parasites for host colonisation. This shows codon usage as a necessary part of the genome evolution for the success of parasitism (Badet et al., 2017).

The kingdom Plantae also rose an early interest for codon usage studies, like in a gene study to differentiate monocotyledons and dicotyledons synonymous codon bias, finding differences between groups, mainly in GC ending codons (while in monocotyledons 16 of 18 amino acids are coded preferably by codons ending in GC, dicotyledons only present 7 of 18). Additionally maize (*Zea mays*) and soybean (*Glycine max*) were examined, with a correlation of codon bias with highly expressed genes (Murray et al., 1989). In further analysis of these groups, GC ending codons were correlated with higher bias, and general GC content was found higher in monocotyledons, but the effective number of codons (a general measure of codon bias, see Chapter 1.5) was similar in both groups, which shared many preferred codons (Kawabe and Miyashita, 2003). An updated analyses of this division, with 15 eudicotyledons (like *Brassica rapa*, *G. max* and *Solanum lycopersicum*), 4 monocotyledons (including *Sorghum bicolor* and *Z. mays*) and 2 mosses (*Physcomitrella patens* and *Selaginella moellendorffii*), showed a constant baseline of mutational pressure, but with an overall high codon bias with evidence for translational efficiency and mRNA folding (Camiolo et al., 2015).

Soybean also received a more thorough analysis, showing GC content and GC ending codons correlated with expression levels and codon bias. Both GC content and GC ending codons were found higher in specific genes for seeds than in those for leaves (Zhang et al., 2011). In other studies focused on species like the Asian rice (*Oryza sativa*) and the thale cress (*Arabidopsis thaliana*), synonymous codons were screened, revealing optimal codons related to the most important and conserved proteins in plants. These data are consistent with the relation of optimal codons with protein folding, and also with selection for translational accuracy (Porceddu et al., 2013).

Metazoan are not exempt of presenting numerous genomes and genes with codon bias, although the codon usage in early branching animals is mostly unknown and previous studies have not been focused on determining bias. We can see a bias study in the poriferan *A. queenslandica* to identify putative horizontal gene transfer (Conaco et al., 2016). The analysis included a correspondence analysis on codon usage (see Chapter 1.5), but instead of going deep in codon usage, a general value for codon bias was plotted against GC content to separate genes in two groups, speculated to be formed by host and foreign genes. In the phyla Ctenophora and Cnidaria, codon usage has been used just to study mitochondrial genome, in the species *Pleurobrachia bachei* (Kohn et al., 2018) and *Bolocera* sp. respectively (Zhang et al., 2017). For neither phylum codon bias was the aim of the study, instead a high level of polymorphism was revealed for *P. bachei*, while in *Bolocera* sp. the mitogenome was characterised, revealing a preference for codons ending in AT.

The codon usage in the phylum Nematoda was studied in *Caenorhabditis elegans* first (Marais and Duret, 2001), and expanded on a comprehensive study of 32 species, including *C. elegans* and *C. briggsae* (Mitrevva et al., 2006). *C. elegans* showed a correlation between bias and gene length, and evidence for selection upon translational accuracy. In the mentioned comprehensive study, GC compositional pressure appears to drive the codon usage bias, varying in a content range of 32% to 51%, with the filarial parasites showing the lowest GC levels (*Brugia malayi*, *Onchocerca volvulus* and *Dirofilaria immitis*) and the cyst nematodes the highest levels (*Globodera pallida*, *G. rostochi* and *Haemonchus contortus*). The GC at third position appeared very consistently around 52.1% with a deviation of 1.5%, with similar distribution to the species named in the general GC content, with an addition of whipworms to the species with endings preferred in GC (*Trichuris muris* and *T. vulpis*). Overall, the author acknowledge that the species studied with similar GC content were also showing similar codon usage, even those distantly related.

In the class Cestoda, different species of tapeworms of the genus *Taenia* have been studied. The species *T. pisiformis* showed a GC content of 49.48%, but it was correlated with codon bias, and of the 24 optimal codons identified just CGU were not ended in GC (Chen et al., 2013). In *T. saginata* 23 optimal codons were identified, all of them ending in GC, and also the major forces shaping this genome were mutation and selection (Yang et al., 2014). Similarly, *T. solium* presented 26 optimal codons, all ended in GC, with also evidence for mutation and selection driving forces (Yang et al., 2015). In contrast, *T. multiceps* showed a weak codon bias with 51% of GC content (Huang et al., 2017), but with GC enriched optimal codons, just CGU differing of the 21 identified. Among the influential forces, selection was the major driver for essential genes. Very differently, in the phylum Annelida, *Hirudinaria manillensis* (the Asian medicinal leech) showed 4 optimal codons ending in GC, although there is a general significant preference for C ending codons (of the 30 codons with highest RSCU, see 1.5.1 for details about this acronym, 13 were ended in C while just 5 in G, a close number to the 6 found in A and U). This leech appears also under natural selection, but function seems to play a role, as blood-diet genes with no homology in the host species are less biased (Guan et al., 2018).

Within the phylum Arthropoda, the crustacean *Alvinocaris longirostris*, a shrimp inhabiting deep sea, presented 43.75% of GC content. Although correspondence analysis was performed for this species (see Chapter 1.5), no optimal codons or evolutionary forces were identified (Hui et al., 2018). Codon usage was also studied in three mitochondrially encoded cytochrome C oxidase genes across different species of Arthropoda, revealing AT richness and low bias, but with influence of natural selection and mutation pressure (Barbhuiya et al., 2020).

The genus *Drosophila* is the most widely studied in the phylum Arthropoda, and a referent for codon bias overall. From the early descriptive and comparative studies of *D. melanogaster* with other species like *H. sapiens* and *E. coli* (Sharp et al., 1988), many studies have followed. Codon bias was evidenced in *D. melanogaster*, *D. pseudoobscura* and *D. virilis* nuclear genes, correlating with level of expression, and in opposition to non-coding regions (Moriyama and Hartl, 1993). The next step in *Drosophila* was demonstrating codon selection upon translational accuracy (Akashi, 1994). The codon usage was determined to lack changes for long periods of time in evolution, but at the same time codon bias was correlated with tRNA abundance, showed preference for GC ending codons, and presented intraspecific polymorphism (Powel and Moriyama, 1997).

The species *D. melanogaster*, *D. pseudoobscura* and *D. virilis* were analysed, underlying the correspondence between codon usage and isoaccepting tRNA abundance, but considering all codons (except for Asp, whose isoaccepting tRNA changes during stages) contributing to the general bias (Moriyama and Powel, 1997). The codon bias during the different life stages of *D. melanogaster* and *D. pseudoobscura* was more deeply studied, revealing that larval stage genes are the most highly biased in comparison with embryos, pupae and adults, probably due to higher pressure for rapid growth in that stage (Vicario et al., 2008). In another study, *D. virilis* species was used to analyse polymorphism and divergence of 18 genes, supporting the hypothesis of selection plus mutation driving forces, but also determining *D. americana* genome base composition and codon usage in almost equilibrium (Maside et al., 2004).

In further analysis, *D. melanogaster* was screened for codon bias of sex biased gene expression. A negative correlation of codon bias and expression for male biased genes was found, in opposition to female biased and non sex biased, determining both categories affected by a different selection (Hambuch and Parsch, 2005). In a comprehensive study with 12 different species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananasse*, *D. pseudocabra*, *D. persimilis*, *D. willistoni*, *D. majovensis*, *D. virilis* and *D. grimshawi*), all shown a quite consistent codon usage, with GC ending optimal codons. The only divergence in GC ending codons was for serine, with half of the species favouring the codon TCC while the other half preferred AGC. This generalisation is true except for *D. willistoni*, which presented 7 differing codons ended in AT. In this study evidence for natural selection was found, but also for genetic drift and mutation (Vicario et al., 2007). *D. melanogaster* also showed that genes with introns, needing splicing, keep optimal codons biased at the centre of exons, hinting an interference in this mechanism (Warnecke et al., 2007). Codon bias was demonstrated to be key for the protein structure and function of the gene dPER, a key element of the circadian oscillator. The increase of optimal codons altered the protein folding, phosphorylation and stability, giving an abnormal circadian behaviour (Fu et al., 2006), possibly as consequence of optimal codons speeding up the protein translation, an effect that has been demonstrated in other proteins (Zhao et al., 2017). Finally, codon usage in *D. melanogaster* was studied in populations, detecting and quantifying selection, with the highest biased genes affecting 10-20% of optimal codons. Codon usage bias was associated with higher polymorphism of synonymous sites inside the populations, and it contributes to splicing and transcription factor binding, highlighting the importance of codon usage bias (Machado et al., 2020).

Larger animals of subphylum Vertebrata present limitations for codons bias analyses due to smaller effective population sizes, which may render natural selection less efficient, attributing bias mainly to mutation and genetic drift. Research on chicken (*Gallus gallus*), rat (*Rattus rattus*), mouse (*M. musculus*), bovine (*Bos taurus*) and human (*H. sapiens*) found a correlation of codon bias with C ending codon preference in these species, except for the G ending codons leucine and valine (Martin et al., 1989). A study in humans using single nucleotide polymorphism revealed pattern of natural selection bias, although also affecting non coding areas of the DNA, either by proximity to selected genes or by effect of biased gene conversion (Smith and Eyre-Walker, 2001).

Coding genes studies, like the one performed over 38 vertebrates including chicken, bovine, dog (*C. familiaris*), finch (*Fringilla coelebs*), panda (*Ailuropoda melanoleuca*), pig (*Sus scrofa*) and platypus (*Ornithorhynchus anatinus*), revealed that although mutational bias was the major bias force in these species, translational selection is also an extended phenomenon in vertebrates (Doherty and McInerney, 2013). In an even more comprehensive study genomes were characterised for GC content, involving 44 species of vertebrates representing mammals, birds and reptiles. Organisms characterised included species like the American alligator (*Alligator mississippiensis*), the red tailed boa (*Boa constrictor*), the loggerhead sea turtle (*Caretta caretta*), the grey short-tailed opossum (*Monodelphis domestica*) and the wild duck (*Anas platyrhynchos*). This study revealed a similar GC ending codons distribution in all the categories, and that GC content in vertebrates is directed by karyotypic elements (Figueroa et al., 2015). A specific study in species of the genus *Bungarus*, containing venomous snakes, showed that these species contain relatively low codon bias, but with natural selection driving the codon usage pattern in their genomes over mutation pressure (Chakraborty et al., 2017).

Transcriptomic data of populations in 30 animal species showed translational selection in species with large effective population sizes (like the nematode *C. brenneri*, or the common vole *Microtus arvalis*), but with no evidence in small populations (like the cuttlefish *Sepia officinalis*). Also C and T ending codons were preferred over G and A ones, although an explanation was not found. This study showed that the common GC gene bias in Metazoa is not correlated with the effective population size (Galtier et al., 2018).

Across the studies described in pluricellular organisms, it is patent that despite their smaller population size in comparison with unicellular species, a consistent evidence for codon bias driven by mutation and natural selection forces is shown, favouring mainly those codons ending in GC. These evidences emphasise the value of codon usage bias as a hugely widespread feature to research evolutionary and translational features through the complete evolutionary tree. Nonetheless, it is manifest that many of the mechanisms and patterns driving the composition of genomes are still unknown.

1.4.5 Beyond codon usage rules, the tRNA modifications.

The discussion about codon usage would not be complete without considering the wobble hypothesis proposed by Crick, as an answer to the fact that exist fewer tRNA than codons (Crick, 1966). This states that a same tRNA can bind to multiple codons, not following the base pairing rules, giving to all the research previously seen a potential layer of exceptions and nuances. The mechanism of this phenomena is the ambiguity of recognition in the third pair of bases, in the tRNA the most modify nucleosides are in position 34 (the third position) and the purine 37 (3' adjacent to the anticodon) (Agris, 1996). Depending on the species, this type of recognition could reach 30-40% of tRNA interactions (Nakamura et al., 2000; Agris et al., 2007).

The ambiguity of recognition in this hypothesis is a consequence of nucleoside modification. Nucleoside modification has different functionalities, like improving aminoacylation kinetics, helping the codon-anticodon pairing, keeping translation in frame (Stahl et al., 2002; Rodriguez-Hernandez et al., 2013), improving accuracy and efficiency of translation (Hagervall and Björk, 1984; Stahl et al., 2002; Rodriguez-Hernandez et al., 2013), and preparing the anticodon stem loop with the rigidity needed for translocation (Agris, 2008; Zhang et al., 2014). Crick predicted in his theory that although U in the anticodon would pair with A, inosine would pair with U and C preferably, but also with A. In his model, pairing purine-purine is acceptable despite the widening of the double helix, but the pyrimidine-pyrimidine pairing structure could not be stable (Crick, 1966).

Inosine is a modification of adenosine (as it is called the adenine joined to the ribose ring) carried out in eukaryotes by the enzymes adenosine deaminase acting on RNA, also known as ADAR (Gerber and Keller, 1999). More specifically, for deamination in position 34 is necessary the complex ADAT2/ADAT3, the first one focused on the enzyme activity, while the second one is involved in tRNA binding (Elias and Huang, 2005; Ramos-Morales et al., 2021). The deamination of adenosine in tRNA position 34 (I34) is common in eukaryotes and its complementarity is enriched in highly expressed genes (Rafels-Ybern et al., 2018). In bacteria I34 is also quite widespread, catalysed by the tRNA adenosine deaminase, an enzyme that acts exclusively in arginine tRNA (Wolf et al., 2002; Rafels-Ybern et al., 2018). Both enzymes for adenosine deamination have been determined essential for different species like *E. coli* (Wolf et al., 2002), *S. cerevisiae* (Gerber and Keller, 1999), *A. thaliana* (Karcher and Bock, 2009) and *H. sapiens* (Shoshan et al., 2015). In contrast, I34 does not appear to be present in archaea (Marck and Grosjean, 2002).

After the predictions of the Crick theory, over 80 tRNA modifications have been discovered (Boccaletto et al., 2018). For instance, threonylcarbamoyladenosine is a modification conserved in all domains of life, essential in *E. coli* and most prokaryotes, but with a role that may differ radically among different species (Thiaville et al., 2015). Also in the three domains (Björk et al., 2001), the 1-methylguanosine in the position 37 of the tRNA is the base for other multiple modifications, like the wybutosine in Eukarya (Noma and Suzuki, 2006), with an important function keeping the reading frame (Atkins and Bjork, 2009). The queuosine is also part of all eukaryotic cells, but its synthesis is exclusively carried out by bacteria (Reader et al., 2004), although this modification is necessary to translate NAC/U codons, presenting correlations with tumour growth (in deficiency) and stress tolerance (Katze and Beck, 1980; Reisser et al., 1994). The 5-methylcytosine in position 38 is another modification that can be found in a diverse array of eukaryotes (Jurkowski and Jeltsch, 2011), contributing to tRNA stability, to translational accuracy (Tuorto et al., 2015) and affecting the early development of organisms (Zhao and Chen, 2013).

In humans, the inosine modification in position 34 has a function of expanding the base pairing (allowing pairings like IA and IC) (Novoa et al., 2012), its hypomodification has been correlated with myositis (Becker et al., 1999), and the missense of this modification is associated with intellectual disability and strabismus (Alazami et al., 2013). Another example in humans is the 2'O-methylation of G or C 34, a common modification that is suggested to improve the codon-anticodon pairing (Liu

et al., 2013), although not all the identified modifications have known functions. The uridine thiolation in the position 34 in eukaryotes has been proved to enhance its binding to the ribosome A site, improving the translational accuracy and efficiency (Rezgui et al., 2013), coordinate metabolism and cell growth (Laxman et al., 2013), while its absence causes ribosome pausing and proteotoxic stress (seen in *S. cerevisiae* and *C. elegans*) (Nedialkova et al., 2015). In *Salmonella enterica*, the uridine in the anticodon UGG has shown to be able to pair with all four Pro codons, and is important for the pairing of Pro codons ended in U/C (Näsvall et al., 2004). Uridine methyltransferases and a heterodimeric form of adenosine deaminase have been shown to contribute to diversification, increasing the divergence between Eukarya and Bacteria from Archaea (Novoa et al., 2012). This project will be focused on inosine for its prevalence and importance in both prokaryotic and eukaryotic translation process.

Wobble positions were discovered in choanoflagellates in highly degenerate amino acids like the TAPSILVR (abbreviation for Thr, Ala, Pro, Ser, Ile, Leu, Val and Thr). The most abundant tRNA genes in *M. brevicolis* were found containing adenosine at the wobble position, pairing frequently with C, although in the filasterean *C. owczarzaki* the anticodon for serine is more frequently paired with G. On the contrary, tRNA genes with G at the wobble position were not found. Due to the presence of inosine commonly among metazoan RNA (Porath et al., 2017; Rafels-Ybern et al., 2017), and its presence in choanoflagellates, large scale deamination appeared to have evolved early in holozoan species (Southworth et al., 2018).

At least 32 codons are necessary to decode the 61 codons with Crick wobble rules (Crick, 1966), but some codes include less than the minimum tRNAs. These codes can be found in mitochondria, plastids, and parasitic bacteria like mycoplasmas (Osawa et al., 1992). This could be explained with two very similar theories, one is the theory of “two out of three decoding”, in which just the two first nucleotides are informative, and the third is completely wobble, pairing with any third nucleotide. Here we should note that in case of codon misreading risk, it is proposed that the organisation of the genetic code would prevent the use of this method (Lagerkvist, 1978). The alternative is the “superwobble position” or “four way wobble” theory, in which the third position is an unmodified U, but also reading the 4 bases, so each codon family needs just one tRNA type (Rogalski et al., 2008). Finally, modifications of non-transfer RNA exist, but representing a reduced percentage, less studied and with less impact than the tRNA modifications (Cantara et al., 2010).

An additional variation in tRNA interaction was discovered with the publication of a software for tRNA recognition (Chan et al., 2019; 2021). In this case, the mismatched recognition was not in the anticodon interaction, but in a tRNA with a stable anticodon but carrying the wrong amino acid. This would suppose more exceptions to the rules that operate the translation process, but before considering its real impact it is necessary the determination of the origin of this phenomena. It has been speculated that some of the isotype mismatch elements contain traits of pseudogenes, but further research in this identified elements is necessary. In choanoflagellates, *M. brevicolis* have shown five possible tRNA pseudogenes for lysine with CUU anticodons (Southworth et al., 2018).

With the aim to evaluate our results in codon usage bias, tRNA wobble position and this tRNA isotype mismatch recognition, an analysis of the concepts and limitations of the software applied is also required.

1.5 CODON USAGE AND tRNA ANALYSES.

1.5.1 Relevant concepts in codon usage analysis.

A series of indices are presented reflecting codon usage bias and correlated with gene expression. Despite sharing a similar goal, each of them show a different aspect and nuance, resulting in a different information provided. Definitions are presented in alphabetical order.

Codon bias index (CBI) measures the codon bias towards the use of optimal codons, but differing of Fop (see later in this section), it considers the expected usage a scaling factor. The numerator contains the difference between times optimal codons are used and what are expected, while the denominator is the difference between number of amino acids present with synonymous codons, and the expected random usage. The level of codon bias inferred in this value can help to differentiate foreign genes in a set. Its range goes from 1, maximum optimal codon usage, to 0, completely random codon usage. If the value is below zero indicates that there are fewer optimal codons than those expected randomly (Bennetzen and Hall, 1982).

Correspondence analysis (COA) is a technique belonging to multivariate statistics, the part of statistics that analyses at the same time multiple dependent factors that can influence a single outcome. This technique uses a model that allows different chi-square values for its variables, so it avoids treating data with *a priori* assumptions. The correspondence analysis of synonymous codons is based on the distribution of genes along continuous axes, one axis per synonymous codon codifying for an amino acid, in accordance with the major trends of the data (Benzecri, 1992; Peden, 2005). A variant of COA is performed on **relative synonymous codon usage (RSCU)**. COA on RSCU results from the division between current codon usage frequency and unbiased synonymous codon usage (Sharp et al., 1986). It sacrifices some degrees of freedom to reduce the deviation produced by the fact that all amino acids are not used equally. The software CodonW (Peden, 1997; 2005) uses this method to calculate optimal codons and assumes that the most biased genes are also the highly expressed, which give an accurate estimation if bias is produced by natural selection. An RSCU of 1.0 indicates that the times a codon is observed is similar to its frequency in a random usage, while a higher number means that a codon is used more frequently than expected, and a lower number that the codon is used less frequently than in a random situation.

Effective number of codons (Nc) measures the totalled codon bias as the difference between the actual codon usage and the complete absence of bias (all synonymous codons are used equally). This value reflects the variety of codons contained in the sequence. Its range goes from 61, random codon usage, to 20, maximum bias (Wright, 1990).

Frequency of optimal codons (Fop) measures the codon bias towards the use of optimal codons. It is a ratio calculated as optimal codons divided into synonymous codons. Its range goes from 1, meaning that optimal codons are used exclusively, to 0 in the case that optimal codons are not used at all. Besides showing the level of codon bias, this value helps to predict protein expression levels (Ikemura, 1981). Nonetheless, prediction of expression levels by codon usage are accurate in case of the bias as a result of natural selection, but not in bias by other origins (Henry and Sharp, 2007).

GC3 represents the fraction of synonymous codons of guanine or cytosine at third codon position. GC at this site is correlated with optimal codons, and if those also are originated by natural selection, creates the circumstances for GC3 to be used for determination of expression levels. Its range goes from 0 (no content), to 1 (maximum content in the sequence) (Stenico et al., 1994).

1.5.2 Characteristics of the software employed.

The software used for choanoflagellates' transcriptome screening (see 1.5.4 for the causes and Chapter 2.1.3 for a description of the process) and analysis with its characteristics is summarised below, organised by its appearance in Chapter 2.

Protein Analysis Through Evolutionary Relationships (PANTHER) classification system is a database part of the Gene Ontology Resource, containing the biggest gene function information worldwide. It offers different online services oriented toward high-throughput analysis, making use of Hidden Markov Models (HMM), Multiple Sequences Alignments and phylogenetic trees, with the aim to study relationships among proteins and predict their function. For a clarification in the functioning of PANTHER, HMM is a statistical model based on predicting unknown values from others that can be observed. PANTHER is divided into two components: an index, gathering information function-related of proteins, and a library, containing all the family relationships among proteins like an evolutionary tree. The HMM is used to refine alignments, for a more accurate classification in subtrees (Thomas et al., 2003a; 2003b).

Basic Local Alignment Search Tool (BLAST), is a program created to compare query sequences of nucleotides or amino acids with an extant database. The outcome are the most similar sequences based on a heuristic algorithm, which means prioritising the efficiency of the analysis over the most optimal answer. The optimisation of alignments is achieved with a score called maximal segment pair (MSP), and the algorithm is fast and with great flexibility for different analysis (Altschul et al., 1990). This program is hosted in the National Centre for Biotechnology Information. The modification of this tool, the **Domain Enhanced Lookup Time Accelerated (DELTA) BLAST** uses an additional database of pre-constructed position-specific scoring matrix, that is the Conserved Domain Database (CDD), aligning the query with conserved domains before the protein database searching step, which improves the homology detection. This advantage could be reduced if multiple iterations are applied, due to a saturation of results and detection of too many families that reduces the sensitivity. The performance also depends on the completeness of the CDD, as this previous similarity is key to improve the full sequences detection, also improving if the same query contains more than one CDD. If the sequence has no CD homology, BLAST for proteins is performed (Boratyn et al., 2012).

OBITools 1.01 is an open source software designed for NGS data analysis and typically used for DNA metabarcoding. One of its main features allows sorting and filtering processes based on taxonomic annotations (Boyer et al., 2016). Using this database, the script **ecofind** is a utility developed from the same team of OBITools as a complement able to retrieve specific taxonomic information, commonly using a *taxid* (a taxon identification number).

The tools used to calculate both to map reads for relative expression (see 2.2.1) and tRNA gene copy number (see 2.3.3) were Samtools and SMALT. **SAMtools** is a collection of programs created to manage high-throughput sequencing (Li et al., 2009). SAMtools uses the SAM format, consisting on headers starting with “@”, lines delimited by TAB, and a specific structure of 11 mandatory fields plus optional ones. This format is simple and flexible, it will be used for the utilities for different functions, like indexing, mapping and sorting. Sorting has a function of saving memory in future analysis, as the optimisation prevent extra alignments. Indexing is necessary for a fast access to the sequences required. And mapping reorganises the alignments in accordance with its putative positions in the genome (this last options with the help of integrated software like SMALT). SAMtools also allows the user to transform SAM format into BAM (Binary Alignment/Map), a compressed binary transformation that is equivalent, but lighter and fast to access at the same time. The next stage of the sequencing analysis could use BAM for different software of result visualisation, like the use of the graphical viewer Tablet (Milne et al., 2013) for alignments and assemblies.

SMALT (SMALT is Copyright © 2010-2015 Genome Research Ltd.) is a program for mapping sequences over genomic references. It is compatible with different sequencing platforms and with simple and paired-end reads. During the alignment, the potential matches are identified and tested with a Smith-Waterman algorithm (Ponstingl and Ning, 2010). This algorithm is specific to align sub sequences (sequences derived from other sequences, through a possible partial deletion) of nucleic acids or amino acids by similarity. The method compare all possible sequence fragments and finds the best score considering a substitution matrix plus gap-scoring (Smith and Waterman, 1981). In SMALT the balance speed and sensibility can be customised, as in general it is slower than the mappers BWA (Li and Durbin, 2009) and BOWTIE (Langmead et al., 2009) (which are fast software commonly used in human genomes), but also SMALT tend to be much more accurate and in most cases is able to map more reads (Ponstingl and Ning, 2010).

Tablet is a graphical viewer for alignments and assemblies (Milne et al., 2013). Visualising assemblies and mapping can be an important scientific tool, like offering a visual screening of the quality, allowing confirmation of polymorphisms, and even revealing splicing proteins that would not be apparent otherwise (Miller et al., 2010). This software supports both SAM and BAM files (Li et al., 2009), different alignment formats, supplementary information, and many quality of life traits like a very visual interface. Tablet can work with multi-core processor, and the output format is FASTA or as copied clipboard, which also makes it quite versatile (Milne et al., 2013).

CodonW (Peden, 1997) version 1.4.2 (Peden, 2005) is a widespread used package, a tool to analyse codon usage bias and amino acid composition through COA (correspondence analysis), with an optional relative synonymous codon usage or RSCU, a variant of COA (see 1.5.1 for a description of these concepts). It is created to simplify these operations for the user and to offer different codon usage indices like CAI (codon adaptation index), CBI (codon bias index), Nc (effective number of codons) and Fop (frequency of optimal codons) (see 1.5.1 for an extended explanation). For the study of amino acids the indices Hydropathicity and Aromaticity are also available.

tRNAscan-SE 2.0 (Chan et al., 2019) is a program aimed to predict tRNA genes, improved in a 2.0 version for a more accurate search and predict gene functions. It offers both a UNIX based software and an online version, both based on covariance model search software, Infernal v1.1 in this update (Nawrocki and Eddy, 2013), and tested in thousands of genomes (Chan and Lowe, 2016). The covariance models consider the sequence and secondary structure of tRNA in a high discriminating search for probable tRNA related sequences. For this purpose tRNAscan contains models for very diverse tRNA genes. With the covariance search software, a HMM (Hidden Markov Model) filter can be applied to improve the processing time but reducing slightly the sensitivity of the search (Chan et al., 2019).

1.5.3 Statistical techniques employed for evolutionary analyses.

The main statistical procedure used is the traditional **Pearson's chi-square test** (Pearson, 1900; Pandis, 2016) that is used to know if there is association among categorical variables. Thus, it is widely used in bioinformatics to compare genes belonging to different categories. Similar to the procedure with other test for null hypothesis, the chi-square can evaluate if codons present bias based on its distribution. The statistical formula is showed below.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Formula for chi-squared. "Σ" references the sum of the followed procedure for all the variables. "O" stands for Observed and "E" for Expected values.

The result of the chi-squared formula correspond to a determinate **P value** in the distribution table, for this we need to calculate the degrees of freedom as (nº of rows -1)*(nº of columns - 1). For a confidence of 5%, if $P > 0.05$ we accept the null hypothesis, meaning that would not be a significant difference among the categories studied, otherwise if $P < 0.05$ would mean that the variables are significantly different and we can find that difference at least in 95% of the samples observed.

The **Yates' correction** for continuity (Yates, 1934; Adler, 1951) is a modification of the chi-squared distribution to reduce error in approximation, making the requirements to determine a significant correlation more restrictive. The modification consist in a fixed subtraction of 0.5 in the numerator of the equation as the example below.

$$\chi^2_{yates} = \sum \frac{(O-E-0.5)^2}{E}$$

Formula for chi-squared modified with Yates' correction. "Σ" references the sum of the followed procedure for all the variables. "O" stands for Observed and "E" for Expected values.

Yates' approximation is considered to be excessively conservative for some samples, even recommending its avoidance (Haviland, 1990), but in datasets with small numbers this test can be useful to avoid the overestimation of significance.

Odds ratio is a statistical test to measure the association between two elements. These two elements are an exposure and an outcome, and the objective is to calculate the probability of the outcome in the presence and absence of the exposure (Szumilas, 2010), although it could be used in other cases like pairs of synonymous codons in different conditions (Eyre-Walker and Bulmer, 1995). This technique was used by Eyre-Walker and Bulmer to demonstrate that the reduced synonymous substitution rate of higher expressed genes of *E. coli* and *S. typhimurium* was not due to selection, but a consequence of a decreased mutation rate. The odds ratio will be addressed in 2.2.5 applied to codons, but the general formula is as follows:

$$Oddsratio = \frac{(ExposedOutcome/NotexposedOutcome)}{(ExposedNotoutcome/NotexposedNotoutcome)}$$

Odds ratio formula of a contingency square for the frequencies for presence and absence of an outcome, in the presence and absence of an exposure.

The interpretation of the results of the odds ratio are taken as how many times is more probable to happen the “exposed outcome” than the “exposed not outcome”. Thus, if the result is higher than 1, the exposure is correlated with the outcome, if is lower and 1 the correlation is inverted, with a 1 meaning that there is no correlation (Szumilas, 2010). The natural logarithm of the odds ratio is used to normalise the data for symmetric representations, allowing ratio comparisons (Eyre-Walker and Bulmer, 1995; Bland and Altman, 2000).

Statistic S ($\hat{S}$) is a statistical value to measure the strength of selection on codon usage. Initially, the model was used to evaluate the strength of codon bias in an array of 80 diverse bacteria (Sharp et al., 2005), but additional research with this method analysed 10 species of eukaryotes, including fungi like *Cryptococcus neoformans*, protozoan representation in *P. falciparum*, plants like *A. thaliana*, and metazoan species like *D. melanogaster* and *M. musculus* (dos Reis and Wernisch, 2009). Holozoan protist also have been evaluated with the Statistic $\hat{S}$ in the choanoflagellates *S. rosetta* and *M. brevicollis*, and the filasterean *C. owczarzaki*. (Southworth et al., 2018). *S. rosetta* and *C. owczarzaki* gave similar results that Dikarya fungi, and *M. brevicollis* a lower value similar to those in ecdysozoans (species studied in dos Reis and Wernisch, 2009). The statistic $\hat{S}$ will be addressed in 2.2.5.

1.5.4 Difficulties in the sequencing of choanoflagellates for codon usage studies.

With the aim to study codon usage bias in choanoflagellates species, either genomic or transcriptomic sequencing is crucial. In many cases, like in this research, transcriptomic data offer the advantage of considering the level of expression of genes or its abundance of reads (remarking again that optimal codons are those presenting more frequency in highly expressed genes). The first choanoflagellates sequencing was carried out in *M. brevicollis* and *S. rosetta* (King et al., 2008; Fairclough et al., 2013), relying on a CsCl gradient to separate bacterial DNA (with low GC content of 33%) from genomic DNA, previous generation of libraries.

The choanoflagellates transcriptomes published by Daniel Richter in 2018 (in which the present research is based on) received physical and bioinformatic treatment to avoid any contamination. The physical methods included the isolation of each choanoflagellate, using a micromanipulator and a microinjector in sterile conditions, and rounds of application of 10 different antibiotics before the cell multiplication in cultures. Once the cells were multiplied and harvested, RNA was extracted and separated from bacterial sequences by a test of polyA. This method is based on using poly-dT beads during the preparation of the standard library to separate polyadenylated mRNA of those who are not, like the bacterial sequences. Bioinformatically, transcriptomes were screened to remove all contigs assigned by an incorrect index, based on a blastn of query contigs from all the species against the rest of transcriptomes. Sequences nearly identical were removed (96% of identity or more), as it is a likely cross contamination, while sequences with 80% of sequence similarity were considered conserved genes. Then they identified the sources of cross contamination, generating mapping ratios between species sharing the cross contamination. It was removed from 1.7 to 8.8% of contigs in each species, expecting to erase also any trace of bacterial contamination (Richter et al., 2018).

Bacterial contamination in choanoflagellates sequencing is particularly difficult to avoid. As natural prey of choanoflagellates, bacteria can be found either attached to their surfaces or even inside the predator cell body, even if bacteria are removed completely from the environment. The size of both organisms are also relevant, as bacteria can multiply exponentially their representation during sequencing, rendering the choanoflagellates genome or transcriptome just a fraction of the sample. Methods for cleaning based on GC content could not be reliable because for many choanoflagellates this values are unknown, so the difference in GC-content for discriminating genes cannot be assumed.

Due to the difficulty of the process, there are reasons enough to speculate about the capacity of the bioinformatic screening performed to filter all the bacterial sequences. As a precedent, in Carr et al., 2017, during the alignment for the six-gene revised phylogeny, bacterial genes were found in the transcriptome sequences of choanoflagellates (pers. Comm.). In this project, the contamination theory can be reflected for instance in the Nc-GC3 plots, showing disperse genes and in some cases a double dot cloud that can be attributed to a completely different set of genes (this discussion will be addressed in Chapter 5). The results of the database Ecofind also revealed the presence of sequences previously described exclusively in bacteria. Thus we decided to further filtering this transcriptomes as we describe in the section 2.1.3. of the next chapter.

1.6 OBJECTIVES OF THIS PROJECT.

The aim of this project is to study evolution in organisms surrounding the metazoan transition to pluricellularity, more specifically, addressing elements involved in the translation process. For this purpose, the goals of this research are directed towards the study of codon usage bias and anticodons in Opisthokonta organisms, mainly in choanoflagellates, although other unicellular species and metazoan are also included, and a species belonging to Apusomonadida (the closest sibling to Opisthokonta), *T. trahens*, is added as outgroup. This knowledge can help to clarify to its possible extent the influence of codon usage and anticodons during this key event in the evolution of life on Earth.

The goals are addressed analysing transcriptomic sequencing for 19 choanoflagellates species published by Daniel Richter et al., 2018, 3 other unicellular species and 5 metazoan. For Daniel Richter transcriptomes, and additional objective is the creation of coding regions database for genuine eukaryotic sequences, avoiding all bacterial contamination that could bias the statistics. For the rest of species, transcriptome sequencing are downloaded from NCBI and the genomic data from EnsemblProtist and EnsemblMetazoa, to perform a similar analysis on codon usage and anticodons but without the previous coding regions screening.

The codon usage bias study included four main objectives: the determination of optimal codons, the analysis of codon usage statistic values, the estimation of the strength of selection, and the study of processes affecting the codon usage. These processes related to codon usage are deamination, which is studied through indirect and direct evidence, the study of anticodon abundance, and the search for chimeric tRNAs (these three last elements are only performed on the second set of species, those in Chapter 4 due to limitations of the data available).

CHAPTER 2. MATERIALS AND METHODS

This chapter contains the details of the data sources, software and procedures used for the analysis of all the species of this research. Starting with the obtaining of genetic information, including the published sequencing data, assembled contigs, predicted proteins and ncRNA data (2.1.1). Following a description for the process of coding regions extraction (2.1.2) and eukaryotic sequences filtering (2.1.3). On the next section the analysis of coding regions is explained, from the mapping to the obtaining of relative expression (2.2.1), determination of optimal codons (2.2.2), codon usage statistics (2.2.3), and the strength of selection using Odds ratio and statistic $\hat{S}$ (2.2.4). The final section includes analyses relative to tRNA and anticodons, covering direct and indirect evidence for the deamination process (2.3.1 and 2.3.2), anticodon abundance based on copy number (2.3.3) and searching for chimeric tRNA (2.3.4). Brief explanations for the software and concepts mentioned in “Materials and Methods” can be found in Chapter 1 section 5.

2.1 SOURCES AND TREATMENT OF DATA.

2.1.1 Genetic files sources.

Sequenced transcriptomes for 19 species of choanoflagellates (described in Table 01) were provided prior to publication by Daniel Richter (Richter et al., 2018). These were obtained by RNA-Seq with polyA mRNA selection with Illumina HiSeq 2000. Assembled transcriptomes in contigs and predicted proteins were also provided by Daniel Richter and are available publicly (<https://dx.doi.org/10.6084/m9.figshare.5686984.v2>) in the online open access repository Figshare. The RNA-Seq files can be found on NCBI database (Sayers et al., 2010) in the Sequence Read Archive (SRA) (Leinonen et al., 2010), by the accession numbers specified in Table 01.

Table 01. Accession numbers for choanoflagellates transcriptome files. These were obtained through RNA-Seq by Daniel Richter (Richter et al. 2018) and downloaded from NCBI.

Chaonoflagellate spp.	Accession nº	Chaonoflagellate spp.	Accession nº
<i>Acanthoecca spectabilis</i>	SRR6344971	<i>Salpingoecca dolichothecata</i>	SRR6344972
<i>Choanoeca perplexa</i>	SRR6344967	<i>Salpingoecca helianthica</i>	SRR6344974
<i>Codosiga hollandica</i>	SRR6344973	<i>Salpingoecca infusionum</i>	SRR6344968
<i>Diaphanoeca grandis</i>	SRR6344970	<i>Salpingoecca kjevrii</i>	SRR6344978
<i>Didymoeca costata</i>	SRR6344966	<i>Salpingoecca macrocollata</i>	SRR6344984
<i>Hartaetosiga balthica</i>	SRR6344985	<i>Salpingoecca punica</i>	SRR6344976
<i>Hartaetosiga gracilis</i>	SRR6344977	<i>Salpingoecca urceolata</i>	SRR6344979
<i>Helgoeca nana</i>	SRR6344981	<i>Savillea parva</i>	SRR6344983
<i>Microstomoea roanoka</i>	SRR6344969	<i>Stephanoeca diplocostata</i> - Australia	SRR6344980
<i>Mylnosiga fluctuans</i>	SRR6344975	<i>Stephanoeca diplocostata</i> - France	SRR6344982

Transcriptome files from RNA-Seq, in FASTQ format, were also downloaded from NCBI-SRA for the early-branching metazoan species and unicellular used in this study (species names and accession number codes can be found in Table 02 and Table 03 for unicellular and metazoan organisms respectively).

Table 02. Accession numbers for unicellular eukaryotes transcriptome files. These were obtained through RNA-Seq and downloaded from NCBI, complementing the sequenced species by Daniel Richter.

Unicellular species	Class	Accession number
<i>Salpingoecca rosetta</i>	Choanoflagellatea	SRR6333396
<i>Sphaeroforma arctica</i>	Ichthyosporea	SRR350206, SRR350207, SRR350209
<i>Thecamonas trahens</i>	Apusomonadea	SRR343042

Table 03. Accession numbers for metazoan transcriptome files. These were obtained through RNA-Seq and downloaded from NCBI.

Metazoan species	Phylum	Accession number
<i>Amphimedon queenslandica</i>	Porifera	SRR8951910, SRR8951911
<i>Mnemiopsis leidyi</i>	Ctenophora	SRR6828928
<i>Nematostella vectensis</i>	Cnidaria	SRR9916742
<i>Thelohanellus kitauei</i>	Cnidaria	SRR1103279
<i>Trichoplax adhaerens</i>	Placozoa	SRR6768521

Additionally, ncRNA (non-coding RNA) sequencing files in FASTA format were also downloaded from EnsemblProtists (*S. rosetta*, *S. arctica*, *T. trahens*) and EnsemblMetazoa (*A. queenslandica*, *M. leidyi*, *N. vectensis*, *T. kitauei*, *T. adhaerens*) on the websites <https://protists.ensembl.org/> and <https://metazoa.ensembl.org> respectively. Also, a copy of ncRNA files used can be found in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>. Some extra transcriptomes were downloaded with the purpose of increasing the sources to find evidence for tRNA deamination as explained in 2.3.2, those extra RNA-Seq transcriptomes are listed below in Table 04.

Table 04. Accession numbers for additional files for tRNA deamination screening. These were obtained through RNA-Seq and downloaded from NCBI.

Species	Accession number
<i>Amphimedon queenslandica</i>	SRR7734360, SRR7734361, SRR7734362, SRR7734363
<i>Mnemiopsis leidyi</i>	SRR538486, SRR538487
<i>Nematostella vectensis</i>	SRR7244441
<i>Sphaeroforma arctica</i>	ERR2662459, ERR2662460

DNA toplevel files were obtained from EnsemblProtists and EnsemblMetazoa in a similar way to ncRNA. In the same way, a copy of DNA toplevel files used can be found in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>. These files provided genomic DNA with no masking and containing haplotype information.

2.1.2 Coding sequences.

Coding regions for each gene (CDS) were extracted for choanoflagellates using a script in Python 3.0 (all scripts are available in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>), matching the predicted proteins by Daniel Richter with the contigs file containing the untranslated regions (available online as described in 2.1.1). Those contigs were trimmed in frame for all the sequences and the reverse complement was generated when the coding frame was in the negative strand, thus a translation of this file would match the predicted proteins file. Furthermore, for proteins codified in the same contig, “REP” codes were added to avoid conflict with software unable to process different sequences with the same name.

In addition, for the protozoan and early-branching metazoan species, CDS files and DNA contigs were downloaded (DNA toplevel, genomic and no masked, as in 2.1.1) from EnsemblProtists (*S. rosetta*, *S. arctica*, *T. trahens*) and EnsemblMetazoa (*A. queenslandica*, *M. leidy*, *N. vectensis*, *T. kitauei*, *T. adhaerens*) on the websites <https://protists.ensembl.org/> and <https://metazoa.ensembl.org> respectively, both in FASTA format. A copy of all CDS files described in this section can be found in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>.

2.1.3 Screening of choanoflagellates CDS.

The complete set of coding genes found in all choanoflagellates transcriptomes were screened to create a catalogue of high reliable sequences for eukaryote organisms. This means sequences that have been proven by homology to have a eukaryotic origin, considering contaminant those genes that are only described in Bacteria and Archaea domains, those whose family of protein were impossible to classify and those unable to detect homology reciprocally. The stages of this filter are depicted in Figure 07 and described below:

Protein files of genes for choanoflagellates were used for a PANTHER v.14.0 screening (Thomas et al., 2003a; 2003b) with the purpose of assigning sequences to protein families. This was carried out locally with default parameters as it was published in 2013 (Mi et al., 2013) and with the benefits of the update in 2017 (Mi et al., 2017), using PANTHER HMM scoring tool (which integrates

HMMER3) with PANTHER HMM library as reference. Sequences with no possible protein family were discarded, as not all the query proteins were able to be mapped.

The amino acid sequences that passed the previous filter were used for a DELTA-BLAST analysis. As these proteins were assigned to families, the probability of hypothetical proteins to be genuine increases. This analysis was made with default parameters, setting a minimum e-value of 0.05, custom output format and 10 as maximum target sequences (meaning that just the top 10 results were evaluated). The databases used from NCBI were nr for general proteins and cdd_delta for improving protein domain search. The aim of this analysis was to assign functions and species in which homologous proteins have been previously described. This will allow a later classification, at the same time that would increase the certainty of the origin of these sequences, as the contamination in choanoflagellates sequencing is extremely difficult to prevent.

DELTA-BLAST results were screened with Ecofind (OBITools scripts 1.01). The role of this taxonomic database was classifying the previous results into different files with sequences belonging to Archaea, Bacteria, and Eukarya or uncertain for each one. This was made with a script taking the taxid (taxonomic identifier) for each sequence and comparing them with the own database for its classification. Only those proteins marked as eukaryotes were directly used to create the new high reliable genes files.

DELTA-BLAST output files were also used to extract the GI (GenInfo identifier) for the best hit of each protein. And the list of GI codes were used to recover sequences in Batch Entrez from NCBI, creating a custom database.

GI sequences from custom files were used as input for a BLAST analysis, using the eukaryote obtained from the Ecofind filter as the reference database (transformed into blastdb format). This reciprocal BLAST aimed to test if the previous results obtained could be also recovered using the opposite search. Files with the best eukaryote sequence for each GI identifier were retrieved, discarding repeated sequences and eukaryote proteins unable to give a top match in this reverse search for similarity.

Finally, CDS in fasta format were matched with the correspondent output protein sequences using a script in Python 3.0, creating similar filtered files but containing nucleotides. All scripts and filtered CDS are available in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>).

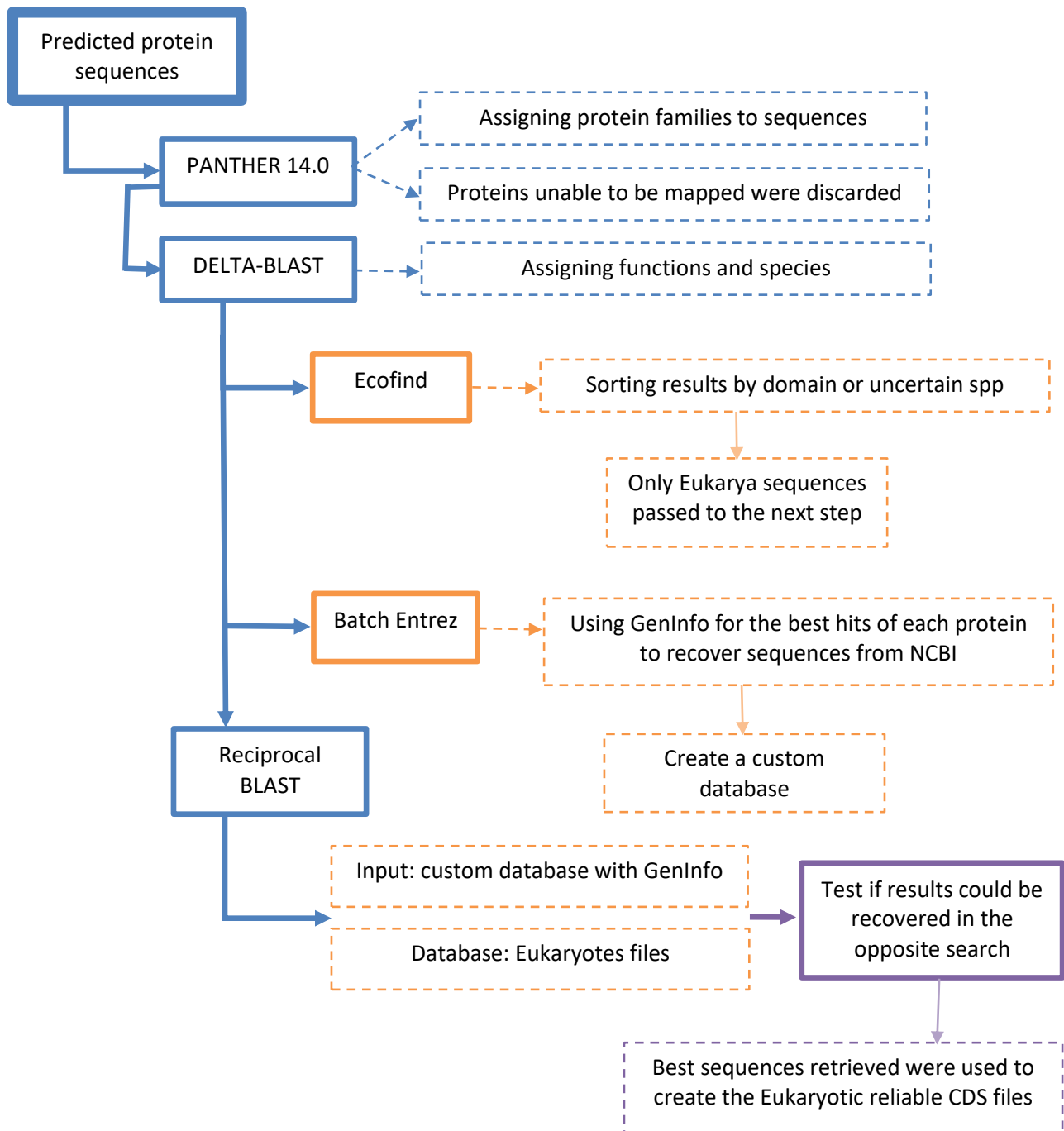


Figure 07. Steps in the procedure for choanoflagellates transcriptome screening. From the initial predicted protein files, to the creation of the screened CDS files, and including all the software used.

2.2 CODING REGIONS ANALYSIS.

2.2.1 Mapping of reads for relative expression.

Sequencing reads mapping to CDS for the species of this study was performed with the software SMALT (Ponstingl and Ning, 2010) (SMALT is Copyright © 2010-2015 Genome Research Ltd.). Previously to the mapping process, CDS data were prepared with SAMtools (Li et al., 2009), as it was necessary an indexing, mapping, format transformation and sorting of the query file (using the commands shown below in Table 05). The mapping was performed using the CDS files of each species at the time, and the raw data from the respective sequencing in fastq format (obtained as explained in 2.1.2 and 2.1.1 respectively).

Table 05. SAMtools commands for treatment of data. The first column shows the previous steps to the mapping of reads using SMALT, and the second column specify the command for these actions.

SAMtools utility	Command
Indexing	smalt index -k 20 -s 1 /path_output /path_CDS_file
Mapping	smalt map -f samsoft -o /path_output.sam /path_indexed_CDS /path_sequencing_file
Format transformation	samtools view -S -b /path_file.sam > /path_file.bam
Sorting of the query file	samtools sort /path_file.bam -o /path_file.sorted.bam

The mapping process also generated the length and number of reads associated with each gene ID. These parameters were used to calculate the relative expression as number of reads divided into length for each gene, which was used for the sorting of the genes IDs attending to their overrepresentation during the transcriptome sequencing. Then, the 5% of the CDS sequences for each species was calculated, and the genes IDs with this 5% highest and lowest relative expression were copied into separate files. The genes IDs were matched with their respective sequences using a Python 3.0 script to create the files used in the codon analysis (all scripts are available in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>).

2.2.2 Determination of optimal codons.

Initially, optimal codons were calculated by correspondence analysis on relative synonymous codon usage (COA on RSCU), employing CodonW 1.4.2 (Peden, 2005) software acquired from SourceForge.net. This was performed on the CDS sequences for each species to obtain frequencies of optimal codons (Fop) files, which would be used as input for later CodonW analysis without performing COA.

Due to the fact that correspondence analyses may generate some false positives if selection is not the primary driver of the bias, as it would not meet the requirement of bias directly correlated with expression (Peden, 2005), optimal codons were also identified manually using relative expression from SMALT analyses (as described in 2.2.1). The codon data from the analysis of the genes with 5% highest and lowest relative expression were used to create a contingency table, first considering that relative codon use needed to be higher in genes with highest relative expression than in genes with lowest relative expression (like it was shown in Sharp and Devine, 1989). Then, the table was prepared as genes with highest and lowest relative expression *versus* the number of other codons used for that amino acid for those genes with highest and lowest relative expression, to calculate P value and Chi Squared with Yates' correction. Codons with p-value lower than the standard value for optimal codons 0.01 were included in custom Fop files attending to the specific order required for the software (example of the format is shown in Table 06).

Table 06. Codon distribution in a codon frequency file. The input format does not presents spaces and changes each respective codon for the number 2 (for rare and average frequency codons), or 3 (for optimal codons).

UUU, UCU, UAU, UGU, UUC, UCC, UAC, UGC, UUA, UCA, UAA, UGA, UUG, UCG, UAG, UGG
CUU, CCU, CAU, CGU, CUC, CCC, CAC, CGC, CUA, CCA, CAA, CGA, CUG, CCG, CAG, CGG
AUU, ACU, AAU, AGU, AUC, ACC, AAC, AGC, AUA, ACA, AAA, AGA, AUG, ACG, AAG, AGG
GUU, GCU, GAU, GGU, GUC, GCC, GAC, GGC, GUA, GCA, GAA, GGA, GUG, GCG, GAG, GGG

2.2.3 Codon usage statistics.

Codon usage statistics were determined with CodonW 1.4.2. Among the information this software can provide, we focus our attention in the more relevant data for our study: Fop (frequency of optimal codons), Nc (effective number of codons) and GC3s (guanine or cytosine at third position of the codon), beside a consideration of CBI (codon bias index) values (see Chapter 1 section 5 for an explanation of this concepts and references).

The different categories for coding genes files of each species were analysed with default parameters and custom files for frequency of optimal codons (Fop) based on expression level analyses (as shown in Table 06). Sets of genes with the 5% highest and lowest relative expression (level of relative expression and creation of this files is explained in 2.2.1) were separated and analysed also with default parameters and custom Fop.

The same SMALT data were used to select the 100 genes presenting highest and lowest relative expression for each species. For these selected genes functional domains were identified using DELTA-BLAST on NCBI website, and separated manually into genetic regions encoding the functional domains and regions encoding non-domains. In case any gene did not present both kind of regions, functional domains and non-domains, this was replaced by an additional gene situated next in the respective relative expression sorted list. The statistics aforementioned were calculated using CodonW for both sets of regions independently, using default parameters and custom Fop.

2.2.4 Strength of selection.

Two methodologies were used to measure the strength of selection for the transcriptomic data. The first one was the odds ratio, developed through a study on substitution rate in *Escherichia coli* and *Salmonella typhimurium* (Eyre-Walker and Bulmer, 1995) and the second one was the statistic $\hat{S}$, a population genetics based model that has been tested in 80 different bacterial species (Sharp et al., 2005).

Odds ratio was calculated from a contingency square showing codon data from genes with highest and lowest relative expression, plus the number of other codon used instead for each amino acid. From this table the values for the odds ratio formula were selected: codon frequency of optimal codons / codon frequency of suboptimal codons, both in high expressed genes, multiplied by the codon frequency of suboptimal codons / codon frequency of optimal codons, both in low expressed genes. Optimal codons are here defined as those used preferentially in highly expressed genes.

$$Oddsratio = \frac{f_{optimalcodonsH}}{f_{suboptimalcodonsH}} \times \frac{f_{suboptimalcodonsL}}{f_{optimalcodonsL}}$$

In this Odds ratio formula f stands for frequency, H for high expressed genes and L for low expressed genes.

Statistic $\hat{S}$ was calculated (based on Sharp et al., 2005, as in Southworth et al., 2018) using the natural logarithm of the odds ratio above mentioned, multiplied by the relative usage of each codon in highest expressed genes (meaning the frequency of optimal codons in highest expressed genes / the sum of all frequencies of optimal codons in highest expressed genes).

$$\hat{S} = \ln(oddsratio) \times \frac{f_{optimalcodonsH}}{\sum f_{optimalcodonsH}}$$

For this statistic $\hat{S}$ formula f stands for frequency, H for high expressed genes, and $\sum f$ for the sum of all frequencies in the category.

2.3 ANTICODON AND tRNA ANALYSIS.

2.3.1 Indirect evidence for tRNA deamination through selective advantage.

Deamination events occur when tRNA genes have adenosine at the wobble position, which is the first base in anticodons, and this is deaminated by the enzymes ADAT2 and ADAT3 in this position into inosine (for an extended explanation see 1.4.5). Inosine may form pairs with adenine, uracil or cytosine, but preferentially with cytosine (Bass and Weintraub, 1988). Evidence for deamination was searched indirectly analysing optimal codons with cytosine at the synonymous third position, meaning that the candidate codons were CUC (Leu), AUC (Ile), GUC (Val), UCC (Ser), CCC (Pro), ACC

(Thr), GCC (Ala) and CGC (Arg) (with a similar methodology as in Southworth et al., 2018). This approach required also a lack of tRNA genes with G at the wobble position.

The aforementioned codons were used to create a contingency square of genetic regions containing functional domains and regions encoding non-domains (as explained in 2.2.3) *versus* the number of other codons used for the same amino acid. Both set of genetic regions were tested separately in the case of genes with highest relative expression and lowest relative expression. For each case Chi-squared and p-value were calculated, considering significant only p-values under 0.01. The aim of this contingency square was to detect an overrepresentation of these codons as a possible product of a selective advantage on using deaminated tRNA.

2.3.2 Direct evidence for tRNA deamination through transcriptome screening.

Evidence for deamination events can be found directly at the wobble position of tRNA molecules in transcriptomes. In this case, if inosine is found in the sequence, this is replaced by guanine in cDNA molecules (Suzuki et al., 2015), but guanine is absent at this sites from genes in the genome. As an approach to this evidence, data from RNA sequencing were mapped into tRNA found in ncRNA files (obtained as it is explained in 2.1.1) using SMALT. Thus, just species described in tables 02 and 03 were searched for this method.

For each tRNA anticodon wobble position, presence or absence of guanine and adenine mapped was screened manually. This was carried out for all TAPSILVR amino acids (Ala, Arg, Val, Ile, Leu, Ser, Pro and Thr), all highly-degenerated amino acids except for Gly, as its deamination is absent in eukaryotes due to the lack of stability of inosine in the wobble position of its tRNA (Yokoyama and Nishimura, 1995; Saint-Léger et al., 2016). These amino acids were studied in all the ncRNA files obtained in 2.1.1, for the transcriptomes of Tables 02, 03 and 04.

2.3.3 Differential anticodon abundance based on tRNA copy number.

Anticodons were identified in tRNA sequences from DNA toplevel files (obtained as in 2.1.1) using tRNAscan-SE 2.0 (Chan et al., 2019) online with default parameters. Downloaded data were available for the species listed in Tables 02 and 03. tRNA-scan also provide a summary of the total number of each anticodon identified per species, giving an insight of the difference of anticodon abundance, as anticodon copy number has showed previously a close correlation with tRNA molecular abundance (Kanaya et al, 1999).

Optimal codons identified previously (see 2.2.3) were compared with the respective anticodons. This comparison was made counting the number of anticodons corresponding to optimal codons, and then dividing this between the total numbers of anticodons for each amino acid. This method offered the percentage of anticodons corresponding to optimal codons, showing the correlation between optimal codons and tRNA copy number for each amino acid and the total per species.

2.3.4 Chimeric tRNA.

Chimeric tRNA was searched in DNA toplevel files (obtained as explained in 2.1.1) with a tRNAscan-SE 2.0 online version exclusively with default parameters. This detection is based on the fact that is possible to use two approaches to predict the encoded amino acid: while one is based on identifying directly the anticodon in predicted tRNA genes, the other one is based on an isotype specific covariance model. Normally this is coincident in over 95% of the samples (Chan et al., 2019), but in cases of isotype-anticodon disagreement this could mean a tRNA with identity elements belonging to a certain type of tRNA synthetase, but with an anticodon altered, pairing with a codon for a different amino acid, a chimeric tRNA.

Isotype-anticodon disagreement events were screened for the species with DNA toplevel sequences available, summarised in Tables 02 and 03.

CHAPTER 3. CODON USAGE AND tRNA RESULTS IN CHOANOFLAGELLATES

In this chapter results obtained for the analysis of choanoflagellates transcriptomes are presented. Initially the results of the new data generated from the contigs and predicted proteins will be addressed (3.1). The second subsection is focused on the different data relative to codon usage bias (3.2), including the calculation of relative expression (3.2.1), the determination of optimal codons (3.2.2), the codon usage statistics (3.2.3), and a determination of the strength of selection (3.2.4). In the last subsection tRNA related analysis are presented, searching for tRNA deamination through codon quantification (3.3.1).

3.1 TREATED DATA FROM CHOANOFLAGELLATES.

Coding sequences (CDS) for each gene were obtained from choanoflagellates transcriptomes as explained in 2.1.2. All files generated, both for nucleotide and amino acid versions, are available in FASTA format in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>. Richter et al., 2018 showed that this sequencing contained complete conserved eukaryotic genes represented above 88% for all the species (see Table 07), through the use of software BUSCO (Simao et al., 2015).

The coding sequences obtained from choanoflagellates were screened to create a custom library reliable for eukaryotic genes as it was described step by step in 2.1.3. All files in FASTA format, with the selected set of genes for each species, are available also in the aforementioned repository. The comparison of CDS number before and after the screening is listed in Table 07.

Table 07. Transcriptomic data for choanoflagellates. Number of contigs (N° contigs) and average length of contigs (contig N50) are extracted from supplementary file 2 of Richter et al., 2018. These data are the result of sequences decontamination, redundancy removal and elimination of noise. Transcriptomes size (Transcript. size) are calculated by multiplication of both previous parameters. Conserved eukaryotic genes (CEG) detected in Richter et al. 2018 are shown, although the analysis was not performed on *S. diplocostata* Australian culture due to the equivalence of predicted proteins between isolates. CDS total describes the full number of CDS extracted, while CDS filtered results from the screening for eukaryotic reliable genes (not performed on *S. diplocostata* from Australia, as in the CEG, and due to the similarities between isolates shown in 3.3.2 and 3.3.3).

Species	N° contigs	Contig N50 (nt)	Transcript. size (Mb)	CEG (%)	CDS total	CDS filtered
<i>Acanthoecca spectabilis</i>	44,002	2,916	128.31	95	40467	2958
<i>Choanoeca perplexa</i>	25,369	2,548	64.64	93.4	23213	3544
<i>Codosiga hollandica</i>	70,401	1,785	125.67	91.1	61053	2845
<i>Diaphanoeca grandis</i>	64,955	1,531	99.45	93.8	44901	3474
<i>Didymoeca costata</i>	36,575	2,379	87.01	96	33705	3333
<i>Hartaetosiga balthica</i>	19,524	2,428	47.4	92.5	18816	3058
<i>Hartaetosiga gracilis</i>	24,580	2,342	57.57	91.1	23114	3148
<i>Helgoeca nana</i>	58,806	2,258	132.78	95	38250	3183
<i>Microstomoea roanoka</i>	43,523	1,864	81.13	94.1	30393	3647
<i>Mylnosiga fluctuans</i>	31,671	1,974	62.52	95.1	25158	3504
<i>Salpingoeca dolichothecata</i>	52,403	1,446	75.77	95.1	39355	4079
<i>Salpingoeca helianthica</i>	32,829	1,780	58.44	93.1	25209	3445
<i>Salpingoeca infusionum</i>	45,247	2,670	120.81	91.4	28123	3000
<i>Salpingoeca kvevrii</i>	25,813	2,462	63.55	94.7	23252	3548
<i>Salpingoeca macrocollata</i>	44,734	1,373	61.42	93.1	36525	3756
<i>Salpingoeca punica</i>	40,846	1,482	60.53	93.7	28067	3255
<i>Salpingoeca urceolata</i>	57,706	1,878	108.37	94.7	35982	3735
<i>Savillea parva</i>	47,054	1,716	80.74	92.8	32055	3161
<i>Stephanoeca diplocostata</i> (Australia)	76,811	2,404	184.65	---	36167	---
<i>Stephanoeca diplocostata</i> (France)	80,355	2,701	217.04	88.1	43380	3259

The choanoflagellates studied presented a wide range in the estimation of transcriptome size, as showed in Table 07, from 47.4 in *H. balthica* to 217.04 Mb in *S. diplocostata* (France culture). As this is the first sequencing for this species its size cannot be compared to previous results, although the genome has been studied for the species *S. rosetta* (Fairclough et al., 2013) and *M. brevicolis* (King et al., 2008), showing 55.4 Mb and 41.6 Mb respectively. In a situation where all contaminating sequences have been previously removed, the transcriptomes of these 19 species would be bigger than the genome size of *M. brevicolis*, and most of them also bigger than the genome size of *S. rosetta*.

A comparison among choanoflagellates revealed a high variability in the sequencing. *S. diplocostata* has a transcriptome around fourfold larger than *H. balthica*, but the number of CDS for this species is just 2.3 times larger. This proportion is the clearest example for a lack of correlation between transcriptome size and number of genes in these data. Another discrepancy in the transcriptomes is the difference between both isolates of *S. diplocostata*, with 184.65 Mb for the Australian one (17.5% smaller), although both species were deemed equivalent in protein content. After the CDS screening, the differences among the species was reduced, showing a number of genes ranging from 2845 (in *C. hollandica*) to 4079 (in *S. dolichothecata*).

3.2 CODING REGION ANALYSIS.

3.2.1 Mapping of reads for relative expression.

Mapping of reads from RNA sequencing over the CDS allowed the calculation of relative expression of these fragments (as explained in 2.2.1). This was done as number of reads divided into gene length. The results for relative expression were used for sorting the CDS by these values and determining the 5% higher and lower sequences in the ranking, both for the original and filtered transcriptome independently. Files containing relative expression are available in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>.

3.2.2 Determination of optimal codons.

Optimal codons for the 20 choanoflagellates transcriptomes were determined as explained in 2.2.2, for both sets of CDS (from Richter et al., 2018 predicted proteins and screened for high eukaryotic reliable sequences) and with two methodologies (correspondence analysis predictions based on codon bias, and using the 5% highest and lowest expressed genes). Results for all the choanoflagellates species are showed below in Tables 08 to 27 comparing each category, highlighting in bold the differences found between the two methodologies. To describe each category the abbreviations CDS-O (original CDS) and CDS-S (screened CDS) will be used.

Table 08. Comparison of optimal codons detected in *A. spectabilis*. Both the whole set of CDS and the screened high reliable eukaryotic CDS are included. Results are shown comparing the two methods for detecting optimal codons, based on the expression level and by correspondence analysis in CodonW (differences between both are marked in bold).

<i>A. spectabilis</i> original CDS			<i>A. spectabilis</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUC, CUG	CUC, CUG	CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUC, GUG	GUC, GUG	GUC
Ser	UCC, UCG, AGC	UCC, UCG	UCC, UCG, AGC	UCC, UCG
Pro	CCC , CCG	CCG	CCC , CCG	CCG
Thr	ACC, ACG	ACC, ACG	ACC, ACG	ACC
Ala	GCC, GCG	GCC, GCG	GCC, GCG	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC, CGG	CGC, CGA , CGG	CGC, CGG	CGU , CGC
Gly	GGC, GGG	GGC	GGC, GGG	GGU , GGC

A. spectabilis presented 3 false positives in the CDS-O, and 8 in CDS-S with CodonW, failing to avoid false negatives in both cases. By expression levels, 25 optimal codons were detected in CDS-O and 21 in CDS-S, with 19 of them coincident between both data sets. Compared to CDS-S, CDS-O showed 6 different optimal codons and considered not significant CGU (Arg) and GGU (Gly), codons also not detected in the predictions. For both expression level-based results, most optimal codons were ended in GC, all except 1 in CDS-O and 2 in CDS-S. All optimal codons are specified in Table 08.

Table 09. Comparison of optimal codons detected in *C. perplexa*. The formatting is presented in a similar style to Table 08.

<i>C. perplexa</i> original CDS			<i>C. perplexa</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	UUG, CUC	UUG, CUC	UUG, CUC	UUG, CUC
Ile	AUU, AUC	AUU, AUC	AUU , AUC	AUC
Val	GUC, GUG	GUC, GUG	GUC, GUG	GUC
Ser	UCU, UCC, UCG	UCU, UCC	UCU, UCC	UCU, UCC
Pro	CCU, CCC	CCU, CCC	CCU, CCC	CCU, CCC
Thr	ACU, ACC	ACU, ACC	ACU, ACC	ACU, ACC
Ala	GCU, GCC	GCU, GCC	GCU, GCC	GCU, GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAA	CAA	CAA	CAA
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGU, CGC	CGU, CGC	CGU, CGC	CGU, CGC
Gly	GGU, GGC	GGU, GGC	GGU, GGC	GGU

C. perplexa showed consistent result through methodologies for the analysis of optimal codons, as it is specified for each amino acid in Table 09. CodonW predicted only 1 false positive in CDS-O (on 28 codons), and 3 in CDS-S, but with no false negatives in both cases. 27 optimal codons were determined in CDS-O, and 24 in CDS-S, with no different codons in the second case. This species presented 8 codons ended in AU in CDS-O and 7 in CDS-S, giving lower percentage of GC ended optimal codons than in other species like *A. spectabilis*, but still showing values over 66%.

Table 10. Comparison of optimal codons detected in *C. hollandica*. The formatting is presented in a similar style to Table 08.

<i>C. hollandica</i> original CDS			<i>C. hollandica</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUG	CUU, CUC	CUC, CUG	CUC, CUG
Ile	AUC	AUU	AUC	AUC
Val	GUC, GUG	GUU	GUC	GUC
Ser	UCC, UCG, AGC	UCU, UCC	UCC, UCG, AGC	UCC, AGC
Pro	CCC, CCG	CCU, CCA	CCC	CCC
Thr	ACC, ACG	ACU, ACA	ACC	ACC
Ala	GCC, GCG	GCU, GCA	GCC	GCU, GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAA	CAG	CAG
Asn	AAC	-	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAU	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC	CGU, CGA, AGA	CGU, CGC	CGU, CGC
Gly	GGC	GGU, GGA	GGU, GGC	GGU, GGC

C. hollandica presented divergence between codon usage results, as it is detailed in Table 10. In CDS-O 17 optimal codons were wrongly predicted and 6 coincident with expression levels results, while the CodonW prediction in CDS-S showed only 1 false positive and 1 false negative. CDS-O presented 25 optimal codons, with no result for Asn and 11 coincident with CDS-S. CDS-S showed 23 optimal codons, with 20 of them ended in GC, a result that was consistent with the high GC3 content presented across choanoflagellates optimal codons, in opposition to the 8 GC3 optimal codons determined in CDS-O of *C. hollandica*. Thus, in this species the screening process improved the ability of optimal codons prediction and also increased the optimal codons GC3 value.

Table 11. Comparison of optimal codons detected in *D. grandis*. The formatting is presented in a similar style to Table 08.

<i>D. grandis</i> original CDS			<i>D. grandis</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUU	UUC	UUC	UUC
Leu	UUA, CUA	CUU, CUC	CUC	CUU, CUC
Ile	AUA	AUC	AUC	AUC
Val	GUA	GUC	GUC	GUC
Ser	UCA, AGC	UCU, UCC	UCC, UCG	UCC
Pro	CCA	CCA	CCC, CCA	CCA
Thr	ACA	ACC	ACC	ACC
Ala	GCA	GCC	GCC	GCC
Tyr	UAU	UAC	UAC	UAC
His	-	CAC	CAC	CAC
Gln	CAA	CAA	CAA	CAA
Asn	AAU	AAC	AAC	AAC
Lys	AAA	AAG	AAG	AAG
Asp	-	GAC	GAC	GAC
Glu	GAA	GAA	GAG	GAA
Cys	UGU	UGC	UGC	UGC
Arg	CGA, AGA, AGG	CGU	CGU, CGC	CGU
Gly	GGA, GGG	GGU, GGC	GGU, GGC	GGU

D. grandis exhibited a change in the CodonW accuracy post-screening, similarly to *C. hollandica*. While the prediction in CDS-O misidentified 18 optimal codons and showed 18 false negatives, CDS-S presented 5 false positives and 2 false negatives. The expression level analysis determined 19 optimal codons in CDS-S, shared completely with CDS-O, which presented two additional optimal codons, UCU (Ser) and GGC (Gly). The results in CDS-S changed from 7 to 6 codons ended in AU, keeping a result of GC over 66% in a similar way to *C. perplexa*. All optimal codons for *D. grandis* can be found in Table 11, while *C. hollandica* and *C. perplexa* were shown in Tables 10 and 09 respectively.

Table 12. Comparison of optimal codons detected in *D. costata*. The formatting is presented in a similar style to Table 08.

<i>D. costata</i> original CDS			<i>D. costata</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUU	UUC	UUC	UUC
Leu	UUA, CUA	UUG, CUU, CUC	UUG, CUU, CUC	UUG, CUU, CUC
Ile	AUA	AUU, AUC	AUU, AUC	AUU, AUC
Val	GUU, GUA	GUU, GUC	GUU , GUC	GUC
Ser	UCA, AGU	UCU, UCC, AGC	UCU , UCC, AGC	UCC, AGC
Pro	CCA	CCU , CCA	CCU , CCA	CCA
Thr	ACA	ACC	ACU , ACC	ACC
Ala	GCA	GCU, GCC	GCU , GCC	GCC
Tyr	UAU	UAC	UAC	UAC
His	CAU	CAC	CAC	CAC
Gln	CAA	CAA	CAA	CAA
Asn	AAU	AAC	AAC	AAC
Lys	AAA	AAG	AAG	AAG
Asp	GAU	GAC	GAC	GAC
Glu	GAA	-	-	-
Cys	UGU	UGC	UGC	UGC
Arg	AGA, AGG	CGU, CGA	CGU, CGC, CGA	CGU, CGC
Gly	GGA, GGG	GGA	GGU, GGC , GGA	GGU, GGA

D. costata, which optimal codons are detailed in Table 12, presented 18 false positives and 22 false negatives in CDS-O, although the set CDS-S presented only 7 false positives and 0 false negative. This effect has also been seen in species like *C. hollandica* and *D. grandis*. The expression analysis revealed similar results in CDS-O and CDS-S, with no significant results for Glu and sharing 21 optimal codons. The GC3 content value also increased post-screening, identifying 11 AU3 codons in CDS-O and 7 in CDS-S, which keeps the consistency of high GC3 optimal codons across choanoflagellates over 66%. *D. costata* shared 18 optimal codons with its sister species *D. grandis* in CDS-S (codons in Table 11), including all the codons detected in this species except for GAG.

Table 13. Comparison of optimal codons detected in *H. balthica*. The formatting is presented in a similar style to Table 08.

<i>H. balthica</i> original CDS			<i>H. balthica</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUU, CUC	CUU, CUC	CUU, CUC	CUU, CUC
Ile	AUC	AUU , AUC	AUC	AUC
Val	GUU, GUC	GUU, GUC	GUC	GUU , GUC
Ser	UCU, UCC, AGC	UCU, UCC	UCU, UCC, AGC	UCU, UCC
Pro	CCU	CCA	CCA	CCA
Thr	ACU, ACC	ACU, ACA	ACC	ACU , ACC
Ala	GCU, GCC	GCU, GCC	GCU, GCC	GCU, GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	-	CAG	-
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	-	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGU, CGC	CGU, CGC	CGU, CGC	CGU, CGC
Gly	GGU, GGC	GGU, GGA	GGU, GGC	AGC , GGU

H. balthica predictions were consistent in both optimal codons analysis, which results are presented in Table 13. While CodonW detected 6 false positives and 4 false negatives in CDS-O, it also gave 3 false positives and 3 false negatives in CDS-S. Considering expression levels, both datasets revealed 24 optimal codons, sharing 21 of them. In the case of CDS-O no optimal codons were detected for Gln or Asp, but in CDS-S only Gln lacked of optimal codons. In this species, like in *D. costata* and *C. perplexa* (Tables 12 and 09 respectively), GC3 content increased post-screening, changing from 11 to 8 AU3 codons, and keeping the GC3 value in optimal codons slightly over 66%.

Table 14. Comparison of optimal codons detected in *H. gracilis*. The formatting is presented in a similar style to Table 08.

<i>H. gracilis</i> original CDS			<i>H. gracilis</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUU	UUC	UUC	UUC
Leu	UUA, UUG, CUA	CUU, CUC	CUU, CUC	CUU, CUC
Ile	AUA	AUC	AUC	AUC
Val	GUU, GUA	GUC	GUC	GUU, GUC
Ser	UCA, AGU	UCC, AGC	UCU, UCC, AGC	UCU, UCC, AGC
Pro	CCA	-	CCC	CCA
Thr	ACA	-	ACU, ACC	ACC
Ala	GCA	GCU, GCC	GCU, GCC	GCU, GCC
Tyr	UAU	UAC	UAC	UAC
His	CAU	CAC	CAC	CAC
Gln	CAA	CAG	CAG	CAG
Asn	AAU	AAC	AAC	AAC
Lys	AAA	AAG	AAG	AAG
Asp	GAU	GAC	GAC	GAC
Glu	GAA	GAG	GAG	GAG
Cys	UGU	UGC	UGC	UGC
Arg	CGA, AGA, AGG	CGU, CGC	CGU, CGC	CGU, CGC
Gly	GGA, GGG	GGA	GGU, GGC	GGU

H. gracilis had a great increase of CodonW accuracy post-screening, as can be seen in the results of Table 14, and in species like in *D. grandis* and *D. costata*. The predictions in CDS-O differ in 19 of the 20 optimal codons identified by expression, while in CDS-S presented only 3 false positives and 2 false negatives. Analysis by expression levels shared 19 optimal codons, presenting in CDS-O the codon GGA (Gly) and no results for Pro and Thr. CDS-S showed 24 optimal codons, and despite the tendency in previously seen species, AU3 rose from 4 to 7, but still with a high GC3. Comparing CDS-S results in both *Hartaetosiga* species, *H. gracilis* included all the optimal codons identified in *H. balthica* except for AGC (Gly) and ACU (Thr), presenting instead CAG (Gln) and AGC (Ser).

Table 15. Comparison of optimal codons detected in *H. nana*. The formatting is presented in a similar style to Table 08.

<i>H. nana</i> original CDS			<i>H. nana</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUC	CUC, CUG	CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUC	GUC	GUC
Ser	UCC, UCG, AGC	UCC, UCG	UCC, UCG	UCC, UCG
Pro	CCC , CCG	CCG	CCG	CCG
Thr	ACC, ACG	ACC, ACG	ACC, ACG	ACC
Ala	GCC, GCG	GCC, GCG	GCC, GCG	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC , CGG	CGG	CGC, CGG	CGU , CGC, CGG
Gly	GGC, GGG	GGC	GGC	GGU , GGC

H. nana presented few optimal codon changes among datasets, as can be seen in Table 15. CodonW misidentified 6 false positives in CDS-O and 3 in CDS-S, with 2 false negatives in the last case. The expression levels determined 21 optimal codons in CDS-O and 22 in CDS-S, sharing 19 of them. While the not shared codons belonged to Thr and Ala in the first case, the CDS-S presented different optimal codons for Arg and Gly. In this species the GC3 content for optimal codons was 100% in CDS-O, and 90.9% in CDS-S (20 codons). This species was very similar in optimal codons for CDS-S to its closest relative *A. spectabilis*, with an identical set of results plus the optimal codon CGG (Arg), and also shared 17 optimal codons with its second closest related species, *S. parva* (Table 25).

Table 16. Comparison of optimal codons detected in *M. roanoka*. The formatting is presented in a similar style to Table 08.

<i>M. roanoka</i> original CDS			<i>M. roanoka</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUU	UUC	UUC
Leu	CUC, CUG	UUG , CUU, CUC	CUU, CUC	CUU, CUC
Ile	AUC	AUU	AUC	AUC
Val	GUC , GUG	GUU	GUU, GUC	GUU, GUC
Ser	UCC, UCG , AGC	UCU , UCC	UCU, UCC, AGC	UCU, UCC, AGC
Pro	CCC , CCG	CCU , CCA	CCU, CCC	CCU, CCC
Thr	ACC , ACG	ACU , ACA	ACU, ACC	ACU, ACC
Ala	GCC , GCG	GCU , GCA	GCU, GCC	GCU, GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	-	CAC	CAC
Gln	CAG	CAA	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAU	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC , CGG	CGU , AGA , AGG	CGU, CGC	CGU, CGC
Gly	GGC , GGG	GGU , GGA	GGU, GGC	GGU

M. roanoka gave big differences in the optimal codons pre and post-screening, as it is shown in the results of Table 16. CDS-O was poorly predicted by CodonW, with 20 false positives and 19 false negatives, showing also a high percentage of optimal codons ended in AU (17 of 26 codons). CDS-S was predicted almost perfectly, showing only 1 false positive, and the AU3 codons were reduced to 8 (from 26). In this CDS set the GC3 content was consistent with the high values found in the rest of choanoflagellates.

Table 17. Comparison of optimal codons detected in *M. fluctuans*. The formatting is presented in a similar style to Table 08.

<i>M. fluctuans</i> original CDS			<i>M. fluctuans</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUU , CUC	CUC, CUG	CUU , CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUU , GUC, GUA	GUC	GUU , GUC
Ser	UCC, UCG , AGC	UCU , UCC	UCC, UCG	UCU , UCC
Pro	CCC, CCG	CCC	CCC	CCU , CCC
Thr	ACC, ACG	ACC	ACC, ACG	ACU , ACC
Ala	GCC, GCG	GCU , GCC	GCC	GCU , GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	-	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	-	GAC	-
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC	CGU , CGC	CGC	CGU , CGC
Gly	GGC	GGU , GGC	GGC	GGU , GGC

M. fluctuans exhibited a poor optimal codon prediction for both datasets, as it can be seen in the results of Table 17. CDS-O presented 9 false positives and 7 false negatives, while CDS-S showed 4 false positives and 8 false negatives. The expression levels determined 23 optimal codons in CDS-O and 25 in CDS-S, sharing 22 codons between them. The only different codon detected for the first case was GUA (Val), while CDS-S presented CCU (Pro), ACU (Thr) and CAC (His). CDS-O and CDS-S showed 7 and 8 codons ended in AU respectively, keeping the common GC3 value across choanoflagellates over 66%.

Table 18. Comparison of optimal codons detected in *S. dolichothecata*. The formatting is presented in a similar style to Table 08.

<i>S. dolichothecata</i> original CDS			<i>S. dolichothecata</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	UUG, CUU , CUC	UUG, CUC, CUG	UUG, CUU , CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC	GUU , GUC	GUC	GUC
Ser	UCU, UCC, UCG	UCU, UCC	UCU, UCC, AGC	UCU, UCC
Pro	CCU, CCC, CCG	CCU, CCC	CCU, CCC	CCU, CCC
Thr	ACU , ACC	ACC, ACA	ACU , ACC	ACC
Ala	GCU, GCC	GCU, GCC	GCU , GCC	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAA	-	-
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	-	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGU , CGC, CGG	CGC, CGA	CGU, CGC	CGU, CGC
Gly	GGU, GGC	GGU, GGC	GGU, GGC	GGU, GGC

S. dolichothecata, in a similar way to species like *H. nana*, increased the optimal codon prediction accuracy post-screening (results shown in Tables 18 and 15 respectively). CDS-O presented 8 false positives and 6 false negatives, while CDS-S only 4 false positives and 1 false negative. In CDS-O 26 optimal codons were determined, 21 shared with CDS-S and 9 ended in AU. CDS-S showed 23 optimal codons, reducing also the number of AU3 ones to 5, and increasing the GC3 value to 78.2%.

Table 19. Comparison of optimal codons detected in *S. helianthica*. The formatting is presented in a similar style to Table 08.

<i>S. helianthica</i> original CDS			<i>S. helianthica</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUG	CUC , CUG	CUC, CUG	CUC, CUG
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUC	GUC, GUG	GUU , GUC
Ser	UCC, UCG, AGC	UCC, UCG, AGC	UCC, UCG, AGC	UCC, UCG
Pro	CCC, CCG	CCC	CCC, CCG	CCC
Thr	ACC, ACG	ACC	ACC, ACG	ACC
Ala	GCC, GCG	GCU , GCC	GCC	GCU , GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	-	GAC	-
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC, CGG	CGC	CGC	CGU , CGC
Gly	GGC	GGC	GGU, GGC	GGU, GGC

S. helianthica also presented optimal codon differences among categories, as specified in Table 19. Predictions in CDS-O showed 6 false positives and 2 false negatives, while in CDS-S were 5 and 2 respectively. CDS-O showed 21 optimal codons, 20 shared with CDS-S and one different (AGC for Ser). CDS-S showed 23 optimal codons, neither with results for Asp and showing exclusively GUU (Val), CGU (Arg) and GGU (Gly). As in the case of *H. gracilis* (Table 14) and *H. nana* (Table 15), post-screening increased the AU3 content, from 1 to 4 in this species but still showing a high percentage of optimal codons ended in GC (82.6%). *S. helianthica* shared 21 optimal codons with its sibling species *M. fluctuans* in CDS-S (results in Table 17), with both species lacking optimal codon for Asp.

Table 20. Comparison of optimal codons detected in *S. infusioinum*. The formatting is presented in a similar style to Table 08.

<i>S. infusioinum</i> original CDS			<i>S. infusioinum</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	UUA , CUU, CUA	CUU, CUC , CUG	CUU, CUC	CUU, CUC
Ile	AUA	AUU , AUC	AUC	AUC
Val	GUC, GUG	GUU , GUC	GUC	GUC
Ser	UCA , UCG , AGC	UCU , UCC , AGU	UCU, UCC, AGC	UCU, UCC
Pro	CCG	CCA	CCA	CCA
Thr	ACU , ACG	ACC , ACA	ACU, ACC	ACU, ACC
Ala	GCU , GCC	GCG	GCU, GCC	GCU, GCC
Tyr	UAU	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAA	CAG	CAG	CAG
Asn	AAU	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAU	GAC	GAC
Glu	-	GAG	GAG	GAG
Cys	UGU	UGC	UGC	UGC
Arg	CGC, CGA , AGA	CGU , CGC	CGU, CGC	CGU, CGC
Gly	GGU, GGG	GGU, GGC	GGU, GGC	GGU, GGC

S. infusioinum evidenced a higher accuracy of CodonW prediction post-screening (results shown in Table 20) like in the species *H. gracilis* and *M. roanoka* (Tables 14 and 16 respectively). In CDS-O CodonW determined 20 false positives and 20 false negatives, while in CDS-S only 1 false positive. The expression levels analysis revealed 27 optimal codons in CDS-O and 24 in CDS-S, with 20 in common. Those only detected in CDS-S were ACU (Thr), GCU, GCC (Ala) and GAC (Asp). The AU ended optimal codons were 8 in CDS-O and 7 in CDS-S, showing a GC3 content over 70% in both cases.

Table 21. Comparison of optimal codons detected in *S. kvevrii*. The formatting is presented in a similar style to Table 08.

<i>S. kvevrii</i> original CDS			<i>S. kvevrii</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUC	CUC	CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC	GUC	GUC	GUC
Ser	UCC, UCG, AGC	UCC, UCG, AGC	UCG	UCC , UCG
Pro	CCC	CCC	CCC	CCC
Thr	ACC, ACG	ACC, ACG	ACC, ACG	ACC
Ala	GCC	GCC	GCC	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC	CGC	CGC	CGC
Gly	GGC	GGC	GGC	GGC

S. kvevrii showed consistent results for optimal codons through the different datasets and methods, as it is described in Table 21. CodonW detected 1 false positive in CDS-O, with no false negatives, and 1 false positive and 1 false negative in CDS-S. The expression levels revealed 21 optimal codons in CDS-O and 19 in CDS-S, all completely coincident. The codons only detected in CDS-O were AGC (Ser) and ACG (Thr). In both datasets the optimal codons were 100% ended in GC.

Table 22. Comparison of optimal codons detected in *S. macrocollata*. The formatting is presented in a similar style to Table 08.

<i>S. macrocollata</i> original CDS			<i>S. macrocollata</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	-	UUC	UUC
Leu	CUG	CUU	CUG	CUG
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUU , GUC	GUG	GUC
Ser	UCC , UCG, AGC	UCU , UCA , UCG	UCC , UCG, AGC	UCU , UCC, UCG
Pro	CCC, CCG	CCU , CCC	CCC, CCG	CCU , CCC
Thr	ACC , ACG	ACA	ACC , ACG	-
Ala	GCC , GCG	GCU , GCA	GCC , GCG	GCU , GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	-	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	-
Glu	GAG	-	GAG	GAG
Cys	UGC	UGU	UGC	UGC
Arg	CGC , CGG	CGU , CGA , AGA , AGG	CGC	CGU , CGC
Gly	GGC	GGC, GGA	GGC	GGU , GGC

S. macrocollata, as seen in *D. costata* and *S. dolichothecata*, presented low accuracy in optimal codon predictions for both datasets (results in Tables 22, 12 and 18). CodonW determined 15 false positives and 14 false negatives for CDS-O, and 6 false positives and 5 false negatives in CDS-S. The expression level revealed a very different set of optimal codons in both datasets, CDS-O showed 24 optimal codons with no result for Phe, Gln and Glu, while CDS-S showed 22 optimal codons with no results for Thr and Asp. Both datasets shared 13 codons, and the screening reduced the amount of AU3 from 13 to 5, keeping the consistent high GC3 value across choanoflagellates only in CDS-S.

Table 23. Comparison of optimal codons detected in *S. punica*. The formatting is presented in a similar style to Table 08.

<i>S. punica</i> original CDS			<i>S. punica</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUU , CUC	CUC, CUG	CUC
Ile	AUC	AUC	AUC	AUC
Val	GUC, GUG	GUC	GUC	GUC
Ser	UCC, UCG , AGC	UCC	UCC, UCG	UCU , UCC
Pro	CCC, CCG	CCC	CCC, CCG	CCC
Thr	ACC, ACG	ACC	ACC	ACC
Ala	GCC	GCC	GCC	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	-	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC	CGU , CGC	CGC	CGC
Gly	GGC	GGC	GGC	GGU , GGC

S. punica had a low amount of misidentified codons in both datasets in a similar way to *H. balthica* and *H. nana* (results in Tables 23, 13 and 15 respectively). CodonW produced 7 false negatives and 2 false positives in CDS-O, and 3 false positives and 2 false negatives in CDS-S. Expression levels revealed 19 optimal codons in CDS-O and 20 in CDS-S, sharing 17 between them. While in CDS-O was detected CUU (Leu) and CGU (Arg), in CDS-S was detected GAC (Asp), UCU (Ser) and GGU (Gly). Both datasets contained 2 optimal codons ended in U, and 0 ended in A, keeping a high GC3 level. *S. punica* shared 17 optimal codons in CDS-S with *S. macrocollata* (Table 22), its close relative, but showing differences like a change of CUG for CUC (Leu) and optimal codons for Gln and Glu.

Table 24. Comparison of optimal codons detected in *S. urceolata*. The formatting is presented in a similar style to Table 08.

<i>S. urceolata</i> original CDS			<i>S. urceolata</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUU	UUU	-	UUC
Leu	CUC, CUG	CUU , CUC	CUC, CUG	CUC
Ile	AUC	AUU	AUC	AUC
Val	GUC, GUG	GUC	GUC	GUC
Ser	UCC, UCG, AGC	UCU , UCC, UCG	UCC, UCG	UCU , UCC, UCG
Pro	CCC, CCG	CCC	CCC, CCG	CCC
Thr	ACC, ACG	ACC	ACC, ACG	ACC
Ala	GCC, GCG	GCC	GCC	GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC, CGG	CGC	CGC	CGU , CGC
Gly	GGC	GGC	GGC	GGU , GGC

S. urceolata had some misidentified optimal codons in both datasets, as in *S. punica* (results in Tables 24 and 23 respectively). CodonW predicted 8 false positives and 3 false negatives in CDS-O, and 3 false positives and 4 false negatives in CDS-S. The expression levels revealed 21 optimal codons in CDS-O and 22 in CDS-S, sharing 18 codons. Only CDS-O presented UUU (Phe), CUU (Leu) and AUU (Ile), and only CDS-S showed results for the codons UUC (Phe), AUC (Ile), CGU (Arg) and GGU (Gly). While CDS-O contained 4 optimal codons ended in U, CDS-S showed 3, with no A3 result in any case, thus both presenting GC3 optimal codons over 80%. *S. urceolata*, as close species to *S. kevrrii* (Table 21), included all the results for this species for CDS-S, plus 3 additional codons.

Table 25. Comparison of optimal codons detected in *S. parva*. The formatting is presented in a similar style to Table 08.

<i>S. parva</i> original CDS			<i>S. parva</i> filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUU	-	UUC
Leu	UUG , CUC, CUG	CUU , CUC	CUC, CUG	CUC
Ile	AUC	AUU	AUU	AUC
Val	GUC, GUG	GUC	GUC, GUG	GUC
Ser	UCC , UCG	UCU , UCG	UCC , UCG	UCG
Pro	CCC , CCG	CCA	CCC , CCG	-
Thr	ACC , ACG	ACG	ACC , ACG	ACG
Ala	GCC , GCG	GCU	GCC, GCG	GCU , GCC
Tyr	UAC	UAC	UAC	-
His	CAC	CAC	CAC	CAC
Gln	CAG	CAA	CAG	CAG
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAC
Glu	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	UGC
Arg	CGC , CGG	CGA	CGU, CGC, CGG	CGU, CGC
Gly	GGC , GGG	GGA	GGU, GGC	GGU, GGC

S. parva presented a high number of misidentified codons in both sets of CDS, as in *M. fluctuans* (results in Tables 25 and 17 respectively). CodonW detected 16 false positives and 9 false negatives in CDS-O, reduced to 10 and 3 in CDS-S respectively. Expression level revealed 20 optimal codons in CDS-O, with 9 ended in AU, and 19 in CDS-S, with AU3 reduced to 3. Both data sets shared just 11 codons, and while no result was found for Phe in CDS-O, CDS-S showed no result for Pro and Tyr. Only CDS-S presented a high GC3 content in optimal codons, as is the trend seen in the rest of choanoflagellates, reaching 84.2% in this species.

Table 26. Comparison of optimal codons detected in *S. diplocostata* (Australia isolated). Only CDS with no screening process is shown in this species. The formatting is presented in a similar style to Table 08.

<i>S. diplocostata</i> AU original CDS		
Amino Acid	CodonW	Expression level
Phe	UUC	UUC
Leu	CUC, CUG	CUU , CUC, CUG
Ile	AUC	AUU , AUC
Val	GUC, GUG	GUU , GUC
Ser	UCC , UCG, AGC	UCG
Pro	CCC , CCG	-
Thr	ACC, ACG	ACU , ACC, ACG
Ala	GCC, GCG	GCU , GCC
Tyr	UAC	UAC
His	CAC	CAC
Gln	CAG	CAG
Asn	AAC	AAC
Lys	AAG	AAG
Asp	GAC	-
Glu	GAG	-
Cys	UGC	UGC
Arg	CGC, CGG	CGU , CGC
Gly	GGC , GGG	GGA

S. diplocostata Australian isolate optimal codons were analysed exclusively without CDS screening (results shown in Table 26) due to the fact that it was considered identical to the France isolated at protein level (Richter et al., 2018). CodonW prediction resulted unreliable, as showed 11 false positives and 7 false negatives. The expression level revealed 23 optimal codons, with no result for Pro, Asp and Glu. Although 7 codons were AU3, the GC3 level remained over 66%, similar to species like *D. grandis* and *C. perplexa* (Tables 11 and 09 respectively).

Table 27. Comparison of optimal codons detected in *S. diplocostata* (France isolated). The formatting is presented in a similar style to Table 08.

<i>S. diplocostata</i> FR original CDS			<i>S. diplocostata</i> FR filtered CDS	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUC	UUC	UUC	UUC
Leu	CUC, CUG	CUU , CUC, CUG	CUC, CUG	UUG , CUU , CUC
Ile	AUC	AUU , AUC	AUC	AUC
Val	GUC, GUG	GUC	GUC, GUG	GUU , GUC
Ser	UCC , UCG, AGC	UCG	UCG, AGC	UCC , UCG
Pro	CCC , CCG	-	CCC , CCG	CCA
Thr	ACC, ACG	ACC, ACG	ACC , ACG	ACU
Ala	GCC, GCG	GCC	GCC, GCG	GCU , GCC
Tyr	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	-
Asn	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	-
Glu	GAG	-	GAG	GAA
Cys	UGC	UGC	UGC	UGC
Arg	CGC, CGG	CGC	CGC, CGG	CGU , CGC
Gly	GGC , GGG	GGA	GGC , GGG	GGU , GGA

For the *S. diplocostata* France isolate optimal codons predictions (results in Table 27) were as unreliable as for the Australian isolate and species like *M. fluctuans* and *S. parva* (Tables 17 and 25 respectively). CodonW showed 10 false positives and 3 false negatives in CDS-O, increased to 14 and 11 in CDS-S, an effect only seen in *A. spectabilis* and *C. perplexa* (Tables 01 and 09 respectively). Expression level revealed 20 optimal codons in CDS-O and 23 in CDS-S, with no result for Pro and Glu in the first case, and no result for Gln and Asp in the second one. In this species the AU3 content increased post-screening, from 3 to 9 codons, the lowest GC3 content in optimal codons for all the choanoflagellates studied (with 60.9%).

Comparing the optimal codon for both isolates of *S. diplocostata* (shown in Tables 26 and 27) in CDS-O, CodonW results were completely coincident but the expression levels gave results slightly different. Only in the Australian isolate results were not significant for Asp, while the France isolate lacked the optimal codons GUU (Val), ACU (Thr), GCU (Ala) and CGU (Arg). The GC3 content in the Australian isolate was closer to the lower value of the France isolate CDS-S than to the higher in CDS-O.

CodonW analysis for prediction of optimal codons presented a variable number of false positives and negatives across the species. Although it predicted with high accuracy some data sets like *C. hollandica*, *M. roanoka* and *S. infusionum* (Tables 10, 16 and 20 respectively), the high number of misidentification in other species like in *M. fluctuans* and *S. diplocostata* (Tables 17, 26 and 27) rendered the prediction in choanoflagellates not reliable. Interestingly, the number of misidentifications by CodonW decreased considerably after the screening in many species, like in *M. roanoka* and *S. infusionum*, suggesting that the previous false results could not be due to the method, but to data sets biased due to contamination. The presence of false positive codons by this software revealed some level of mutation or genetic drift in all the choanoflagellates studied.

Optimal codons results were consistent across choanoflagellates, as the codons UUC (Phe), GUC (Val), AAG (Lys), AUC (Ile), CAC (His), UGC (Cys), AAC (Asn) and GCC (Ala) were present in all the species studied. Also the amino acids UAC (Tyr), UCC (Ser), CUC (Leu), GGU (Gly) and CGC (Arg) were shared in 18 species, while CGU (Arg) appeared in 17 species, ACC (Thr) and GAG (Glu) in 16 species, GAC (Asp) in 15 species, CAG (Gln) in 13, GGC (Gly) in 12, and GCU (Ala) in 11. The amino acid less shared across the species were Pro, with only 10 species sharing CCC. Also, as detailed previously, all the species presented GC3 content for optimal codons from 66% to 100% (as in *S. kvevrii*), except for *S. diplocostata* France isolated, which presented 60.9%.

3.2.3 Codon usage statistics.

3.2.3.1 General codon statistic values.

Codon usage statistic values were calculated using CodonW over the different sets of CDS for each species (as explained in 2.2.3). The values for GC content, GC at third position (GC3), codon bias index (CBI), effective number of codons (Nc) and frequency of optimal codons (Fop) are listed below (Tables 28 to 37), separated by datasets and species, and comparing the results obtained by different expression levels. The abbreviations used here are LE (lowest expressed genes) and HE (highest expressed genes).

The GC content calculation will allow us to compare it with GC3 and study a correlation with optimal codons for choanoflagellates (results shown in Table 28 for original CDS and 29 for screened CDS). This value did not show a clear tendency post-screening, neither was consistently higher in HE genes. Although this correlation was not expected, if this is found in GC3 it would be attributable to selection and not the general composition. Nonetheless, many species showing high general GC content in HE like *D. costata*, *H. balthica* o *M. roanoka* could be due to the influence of the high number of optimal codons rich in GC3, but this influence was not so strong over the general value in species like *S. marcocollata*. Species showing higher GC content post-screening were *M. fluctuans* and *S. punica*, while the lowest content was found in *D. costata* and *H. balthica*.

Table 28. GC content mean and standard deviation for original CDS in choanoflagellates. The general data (All) is compared with the 5% of lowest (LE) and highest expressed genes (HE).

GC content – Original CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.518 ± 0.083	0.563 ± 0.065	0.569 ± 0.046
<i>C. perplexa</i>	0.507 ± 0.042	0.509 ± 0.029	0.518 ± 0.029
<i>C. hollandica</i>	0.611 ± 0.084	0.567 ± 0.056	0.579 ± 0.041
<i>D. grandis</i>	0.483 ± 0.048	0.480 ± 0.044	0.478 ± 0.052
<i>D. costata</i>	0.433 ± 0.065	0.432 ± 0.039	0.447 ± 0.039
<i>H. balthica</i>	0.441 ± 0.051	0.429 ± 0.037	0.453 ± 0.039
<i>H. gracilis</i>	0.439 ± 0.046	0.445 ± 0.040	0.468 ± 0.041
<i>H. nana</i>	0.560 ± 0.062	0.587 ± 0.044	0.585 ± 0.036
<i>M. roanoka</i>	0.552 ± 0.098	0.502 ± 0.065	0.488 ± 0.060
<i>M. fluctuans</i>	0.618 ± 0.049	0.627 ± 0.040	0.613 ± 0.033
<i>S. dolichothecata</i>	0.490 ± 0.075	0.478 ± 0.044	0.485 ± 0.051
<i>S. helianthica</i>	0.572 ± 0.062	0.584 ± 0.055	0.571 ± 0.059
<i>S. infusionum</i>	0.467 ± 0.034	0.473 ± 0.028	0.489 ± 0.039
<i>S. kjevrii</i>	0.558 ± 0.067	0.592 ± 0.045	0.607 ± 0.044
<i>S. macrocollata</i>	0.585 ± 0.074	0.586 ± 0.063	0.560 ± 0.080
<i>S. punica</i>	0.620 ± 0.053	0.623 ± 0.046	0.623 ± 0.043
<i>S. urceolata</i>	0.574 ± 0.059	0.577 ± 0.048	0.575 ± 0.051
<i>S. parva</i>	0.578 ± 0.051	0.573 ± 0.044	0.569 ± 0.036
<i>S. diplocostata</i> (Au)	0.526 ± 0.060	0.536 ± 0.052	0.525 ± 0.044
<i>S. diplocostata</i> (Fr)	0.520 ± 0.059	0.532 ± 0.052	0.526 ± 0.045

Table 29. GC content mean and standard deviation for screened CDS in choanoflagellates. The formatting is similar in style as Table 28.

GC content – Screened CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.574 ± 0.057	0.565 ± 0.047	0.575 ± 0.034
<i>C. perplexa</i>	0.512 ± 0.024	0.509 ± 0.023	0.526 ± 0.027
<i>C. hollandica</i>	0.566 ± 0.033	0.564 ± 0.027	0.595 ± 0.029
<i>D. grandis</i>	0.486 ± 0.030	0.482 ± 0.027	0.495 ± 0.031
<i>D. costata</i>	0.437 ± 0.027	0.440 ± 0.022	0.464 ± 0.027
<i>H. balthica</i>	0.433 ± 0.028	0.435 ± 0.025	0.468 ± 0.030
<i>H. gracilis</i>	0.455 ± 0.028	0.462 ± 0.026	0.486 ± 0.028
<i>H. nana</i>	0.603 ± 0.039	0.580 ± 0.031	0.583 ± 0.028
<i>M. roanoka</i>	0.507 ± 0.025	0.499 ± 0.023	0.525 ± 0.032
<i>M. fluctuans</i>	0.635 ± 0.036	0.627 ± 0.029	0.605 ± 0.025
<i>S. dolichothecata</i>	0.479 ± 0.027	0.480 ± 0.025	0.505 ± 0.034
<i>S. helianthica</i>	0.593 ± 0.035	0.588 ± 0.033	0.589 ± 0.031
<i>S. infusionum</i>	0.497 ± 0.024	0.498 ± 0.023	0.539 ± 0.030
<i>S. kjevrii</i>	0.588 ± 0.031	0.593 ± 0.027	0.618 ± 0.027
<i>S. macrocollata</i>	0.607 ± 0.042	0.583 ± 0.034	0.590 ± 0.038
<i>S. punica</i>	0.625 ± 0.035	0.618 ± 0.029	0.626 ± 0.031
<i>S. urceolata</i>	0.585 ± 0.033	0.574 ± 0.030	0.587 ± 0.034
<i>S. parva</i>	0.569 ± 0.036	0.562 ± 0.031	0.568 ± 0.028
<i>S. diplocostata</i> (Au)	-	-	-
<i>S. diplocostata</i> (Fr)	0.539 ± 0.048	0.532 ± 0.041	0.521 ± 0.035

Table 30. GC3 mean and standard deviation for original CDS in choanoflagellates. The formatting is similar in style as Table 28.

GC3 content – Original CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.508 ± 0.137	0.592 ± 0.117	0.643 ± 0.102
<i>C. perplexa</i>	0.497 ± 0.084	0.496 ± 0.061	0.534 ± 0.072
<i>C. hollandica</i>	0.727 ± 0.175	0.626 ± 0.116	0.671 ± 0.100
<i>D. grandis</i>	0.474 ± 0.092	0.472 ± 0.082	0.479 ± 0.102
<i>D. costata</i>	0.366 ± 0.095	0.366 ± 0.059	0.390 ± 0.059
<i>H. balthica</i>	0.393 ± 0.091	0.379 ± 0.066	0.420 ± 0.075
<i>H. gracilis</i>	0.418 ± 0.084	0.429 ± 0.074	0.474 ± 0.088
<i>H. nana</i>	0.587 ± 0.121	0.640 ± 0.088	0.686 ± 0.080
<i>M. roanoka</i>	0.573 ± 0.171	0.493 ± 0.113	0.493 ± 0.105
<i>M. fluctuans</i>	0.734 ± 0.086	0.745 ± 0.081	0.765 ± 0.091
<i>S. dolichothecata</i>	0.483 ± 0.141	0.460 ± 0.078	0.487 ± 0.095
<i>S. helianthica</i>	0.634 ± 0.107	0.638 ± 0.093	0.670 ± 0.123
<i>S. infusionum</i>	0.464 ± 0.064	0.462 ± 0.049	0.495 ± 0.068
<i>S. kjevrii</i>	0.584 ± 0.115	0.656 ± 0.089	0.727 ± 0.115
<i>S. macrocollata</i>	0.674 ± 0.132	0.662 ± 0.106	0.656 ± 0.144
<i>S. punica</i>	0.773 ± 0.114	0.755 ± 0.106	0.786 ± 0.111
<i>S. urceolata</i>	0.631 ± 0.111	0.626 ± 0.085	0.655 ± 0.097
<i>S. parva</i>	0.630 ± 0.093	0.610 ± 0.080	0.636 ± 0.072
<i>S. diplocostata</i> (Au)	0.542 ± 0.101	0.554 ± 0.087	0.546 ± 0.087
<i>S. diplocostata</i> (Fr)	0.527 ± 0.098	0.546 ± 0.088	0.541 ± 0.086

Table 31. GC3 mean and standard deviation for screened CDS in choanoflagellates. The formatting is similar in style as Table 28.

GC3 content – Screened CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.631 ± 0.109	0.629 ± 0.102	0.726 ± 0.074
<i>C. perplexa</i>	0.503 ± 0.057	0.510 ± 0.054	0.587 ± 0.077
<i>C. hollandica</i>	0.652 ± 0.092	0.659 ± 0.070	0.769 ± 0.081
<i>D. grandis</i>	0.485 ± 0.067	0.484 ± 0.057	0.514 ± 0.068
<i>D. costata</i>	0.372 ± 0.050	0.378 ± 0.037	0.431 ± 0.057
<i>H. balthica</i>	0.388 ± 0.059	0.385 ± 0.054	0.461 ± 0.071
<i>H. gracilis</i>	0.444 ± 0.065	0.461 ± 0.061	0.515 ± 0.067
<i>H. nana</i>	0.675 ± 0.070	0.669 ± 0.062	0.741 ± 0.062
<i>M. roanoka</i>	0.501 ± 0.060	0.493 ± 0.055	0.581 ± 0.082
<i>M. fluctuans</i>	0.772 ± 0.057	0.791 ± 0.056	0.797 ± 0.067
<i>S. dolichothecata</i>	0.460 ± 0.050	0.467 ± 0.051	0.546 ± 0.081
<i>S. helianthica</i>	0.664 ± 0.066	0.685 ± 0.068	0.755 ± 0.067
<i>S. infusionum</i>	0.469 ± 0.065	0.475 ± 0.061	0.604 ± 0.085
<i>S. kvevrui</i>	0.668 ± 0.064	0.703 ± 0.067	0.812 ± 0.068
<i>S. macrocollata</i>	0.718 ± 0.091	0.700 ± 0.067	0.762 ± 0.079
<i>S. punica</i>	0.813 ± 0.082	0.799 ± 0.068	0.851 ± 0.059
<i>S. urceolata</i>	0.662 ± 0.063	0.649 ± 0.057	0.732 ± 0.073
<i>S. parva</i>	0.616 ± 0.066	0.619 ± 0.055	0.687 ± 0.065
<i>S. diplocostata</i> (Au)	-	-	-
<i>S. diplocostata</i> (Fr)	0.569 ± 0.096	0.566 ± 0.080	0.572 ± 0.075

The study of the GC3 content revealed inconsistent results for the data sets pre-screening (detailed in Table 30), showing in some cases a higher value for LE than in any other set, like in *C. hollandica*, *M. roanoka* and *S. macrocollata*. In other cases like in both *S. diplocostata* isolates, the content was higher in the All CDS set than in any other. Unlike GC3 pre-screening and GC content, post-screening results (shown in Table 31) presented a clear pattern with higher GC3 content in HE for all the species without exceptions. This value was not consistently and clearly determined higher or lower in LE genes than in the general dataset, but the difference with the HE genes could reach over 0.1, like in *C. hollandica* and *S. infusionum*. The species showing higher GC3 value in HE were *S. punica*, *S. kvevrii*, *M. fluctuans* and *S. macrocollata*, while the lowest values were found in *D. costata*, *H. balthica*, *D. grandis* and *H. gracilis*. These results were consistent with the high number of GC3 optimal codons determined for all the choanoflagellates, including the species with the smallest increment in HE genes, *S. diplocostata*, which presented the lowest GC3 level in optimal codons. The other species with the smallest difference of GC3 with the general data set was *M. fluctuans*, which GC3 on optimal codons was slightly over 66%.

As explained in 1.5.1, CBI index measures the codon bias towards the use of optimal codons. In the CBI analysis, as in the results in GC3 content, pre-screening values were not consistent across species (results shown in Table 32). While some organisms showed higher CBI in HE, like *A. spectabilis* and *H. balthica*, many other showed higher values in LE, like in *C. hollandica*, *D. costata* and *M. roanoka*, or in the All data set, like in *D. costata* and *S. infusionun*. The post-screening analysis (detailed in Table 33) changed the erratic CBI values across the species, giving a consistently higher result for HE in all the choanoflagellates, even in species previously detected with higher values in LE like *S. infusionum*, this value was now almost 5.8 times higher in HE than in LE. CBI showed a high bias of highly expressed genes towards use of optimal codons, consistent with the high level of GC3 content both in optimal codons and HE dataset. The species presenting higher values in this biased index were *S. punica*, *S. kvevrii*, *S. helianthica* and *M. fluctuans*, while the lowest results were found in *S. diplocostata*, which correlates with the species showing less GC3 optimal codons and less GC3 content in HE in comparison with CDS-S.

Table 32. CBI mean and standard deviation for original CDS in choanoflagellates. The formatting is similar in style as Table 28.

Codon Bias Index – Original CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.026 ± 0.260	0.190 ± 0.228	0.294 ± 0.202
<i>C. perplexa</i>	0.095 ± 0.139	0.123 ± 0.124	0.265 ± 0.194
<i>C. hollandica</i>	0.420 ± 0.329	0.229 ± 0.215	0.319 ± 0.188
<i>D. grandis</i>	0.013 ± 0.140	0.023 ± 0.137	-0.002 ± 0.180
<i>D. costata</i>	0.130 ± 0.150	0.141 ± 0.102	0.074 ± 0.117
<i>H. balthica</i>	-0.054 ± 0.150	-0.080 ± 0.139	0.101 ± 0.199
<i>H. gracilis</i>	0.094 ± 0.145	0.083 ± 0.135	-0.023 ± 0.173
<i>H. nana</i>	0.184 ± 0.229	0.285 ± 0.169	0.381 ± 0.156
<i>M. roanoka</i>	0.164 ± 0.339	0.001 ± 0.225	0.001 ± 0.212
<i>M. fluctuans</i>	0.462 ± 0.161	0.482 ± 0.153	0.542 ± 0.175
<i>S. dolichothecata</i>	0.041 ± 0.204	0.002 ± 0.146	0.069 ± 0.214
<i>S. helianthica</i>	0.241 ± 0.205	0.247 ± 0.178	0.312 ± 0.243
<i>S. infusionum</i>	0.115 ± 0.204	0.113 ± 0.191	0.101 ± 0.189
<i>S. kjevrii</i>	0.170 ± 0.178	0.295 ± 0.159	0.453 ± 0.221
<i>S. macrocollata</i>	0.326 ± 0.243	0.306 ± 0.201	0.300 ± 0.284
<i>S. punica</i>	0.511 ± 0.211	0.478 ± 0.194	0.559 ± 0.213
<i>S. urceolata</i>	0.275 ± 0.213	0.270 ± 0.162	0.340 ± 0.189
<i>S. parva</i>	0.270 ± 0.177	0.233 ± 0.152	0.285 ± 0.139
<i>S. diplocostata</i> (Au)	0.090 ± 0.196	0.115 ± 0.169	0.103 ± 0.170
<i>S. diplocostata</i> (Fr)	0.063 ± 0.188	0.100 ± 0.171	0.093 ± 0.170

Table 33. CBI mean and standard deviation for screened CDS in choanoflagellates. The formatting is similar in style as Table 28.

Codon Bias Index – Screened CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.268 ± 0.212	0.266 ± 0.200	0.461 ± 0.144
<i>C. perplexa</i>	0.114 ± 0.100	0.174 ± 0.124	0.435 ± 0.204
<i>C. hollandica</i>	0.252 ± 0.163	0.273 ± 0.132	0.546 ± 0.173
<i>D. grandis</i>	0.027 ± 0.079	0.043 ± 0.098	0.256 ± 0.149
<i>D. costata</i>	0.019 ± 0.090	0.039 ± 0.116	0.320 ± 0.185
<i>H. balthica</i>	-0.085 ± 0.100	-0.057 ± 0.129	0.241 ± 0.188
<i>H. gracilis</i>	-0.032 ± 0.102	-0.005 ± 0.122	0.209 ± 0.174
<i>H. nana</i>	0.282 ± 0.114	0.284 ± 0.110	0.466 ± 0.124
<i>M. roanoka</i>	0.090 ± 0.114	0.113 ± 0.138	0.422 ± 0.220
<i>M. fluctuans</i>	0.395 ± 0.095	0.449 ± 0.100	0.552 ± 0.137
<i>S. dolichothecata</i>	0.060 ± 0.102	0.075 ± 0.118	0.311 ± 0.203
<i>S. helianthica</i>	0.331 ± 0.120	0.370 ± 0.130	0.557 ± 0.141
<i>S. infusionum</i>	0.083 ± 0.135	0.115 ± 0.147	0.481 ± 0.195
<i>S. kjevrii</i>	0.270 ± 0.106	0.336 ± 0.126	0.570 ± 0.140
<i>S. macrocollata</i>	0.394 ± 0.163	0.369 ± 0.124	0.503 ± 0.149
<i>S. punica</i>	0.494 ± 0.158	0.480 ± 0.133	0.631 ± 0.131
<i>S. urceolata</i>	0.272 ± 0.107	0.261 ± 0.109	0.452 ± 0.146
<i>S. parva</i>	0.278 ± 0.122	0.288 ± 0.111	0.448 ± 0.142
<i>S. diplocostata</i> (Au)	-	-	-
<i>S. diplocostata</i> (Fr)	0.148 ± 0.184	0.143 ± 0.153	0.152 ± 0.145

Table 34. Nc mean and standard deviation for original CDS in choanoflagellates. The formatting is similar in style as Table 28.

Effective number of codons – Original CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	49.333 ± 8.161	51.170 ± 6.884	49.073 ± 6.797
<i>C. perplexa</i>	50.771 ± 7.873	53.790 ± 5.904	49.731 ± 7.157
<i>C. hollandica</i>	40.439 ± 11.316	48.260 ± 8.535	44.940 ± 7.367
<i>D. grandis</i>	50.107 ± 8.218	50.950 ± 7.914	46.037 ± 8.623
<i>D. costata</i>	48.641 ± 8.139	52.460 ± 6.166	51.384 ± 6.049
<i>H. balthica</i>	48.015 ± 8.083	49.960 ± 6.483	46.652 ± 6.398
<i>H. gracilis</i>	48.48 ± 8.036	50.500 ± 6.635	48.625 ± 6.873
<i>H. nana</i>	47.627 ± 8.356	49.360 ± 6.520	46.478 ± 6.408
<i>M. roanoka</i>	44.657 ± 9.624	49.630 ± 8.430	45.787 ± 7.756
<i>M. fluctuans</i>	42.007 ± 7.542	42.580 ± 6.171	39.109 ± 6.185
<i>S. dolichothecata</i>	47.471 ± 9.183	51.850 ± 7.564	48.059 ± 8.172
<i>S. helianthica</i>	44.899 ± 7.889	46.730 ± 6.453	41.216 ± 6.785
<i>S. infusionum</i>	53.710 ± 8.108	55.050 ± 6.282	52.223 ± 8.288
<i>S. kjevrii</i>	46.034 ± 8.313	45.430 ± 6.792	38.677 ± 8.175
<i>S. macrocollata</i>	43.159 ± 8.556	44.750 ± 6.941	41.634 ± 7.400
<i>S. punica</i>	38.199 ± 8.575	39.750 ± 7.358	36.841 ± 6.792
<i>S. urceolata</i>	45.512 ± 8.082	48.050 ± 6.625	44.775 ± 7.082
<i>S. parva</i>	45.781 ± 8.448	49.660 ± 6.325	48.349 ± 5.911
<i>S. diplocostata</i> (Au)	49.182 ± 8.369	52.650 ± 6.539	52.372 ± 6.426
<i>S. diplocostata</i> (Fr)	49.594 ± 8.045	52.780 ± 6.555	52.527 ± 6.347

Table 35. Nc mean and standard deviation for screened CDS in choanoflagellates. The formatting is similar in style as table 28.

Effective number of codons – Screened CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	51.261 ± 5.675	51.351 ± 6.085	44.126 ± 6.214
<i>C. perplexa</i>	55.207 ± 4.481	54.313 ± 4.524	45.314 ± 7.106
<i>C. hollandica</i>	49.891 ± 6.993	49.062 ± 5.784	37.36 ± 6.716
<i>D. grandis</i>	54.912 ± 3.727	54.542 ± 4.040	45.420 ± 6.261
<i>D. costata</i>	54.476 ± 4.261	54.734 ± 3.368	49.572 ± 5.740
<i>H. balthica</i>	53.067 ± 3.088	51.44 ± 3.798	44.273 ± 5.544
<i>H. gracilis</i>	53.140 ± 4.411	52.802 ± 3.730	47.449 ± 5.792
<i>H. nana</i>	48.807 ± 5.457	49.556 ± 4.724	42.439 ± 5.979
<i>M. roanoka</i>	54.532 ± 4.291	53.964 ± 4.288	45.092 ± 7.025
<i>M. fluctuans</i>	41.890 ± 5.126	39.949 ± 4.533	35.779 ± 5.219
<i>S. dolichothecata</i>	54.927 ± 4.838	55.082 ± 4.021	47.457 ± 6.576
<i>S. helianthica</i>	47.724 ± 5.011	46.012 ± 5.375	37.483 ± 5.723
<i>S. infusionum</i>	53.057 ± 4.863	52.122 ± 4.922	40.469 ± 6.862
<i>S. kjevrii</i>	46.114 ± 5.308	43.002 ± 5.738	33.563 ± 5.580
<i>S. macrocollata</i>	43.502 ± 6.717	44.944 ± 5.294	39.260 ± 6.004
<i>S. punica</i>	37.203 ± 6.545	38.348 ± 5.427	32.962 ± 5.155
<i>S. urceolata</i>	48.028 ± 4.684	48.900 ± 4.789	41.357 ± 6.499
<i>S. parva</i>	51.080 ± 4.424	50.967 ± 4.072	45.408 ± 5.505
<i>S. diplocostata</i> (Au)	-	-	-
<i>S. diplocostata</i> (Fr)	53.780 ± 5.285	54.213 ± 4.297	50.815 ± 5.869

The Nc values range from 20 (maximum bias) to 61 (minimum bias). Considering this format, we can see that as in the CBI value, bias in HE was not consistent pre-screening (results shown in Table 34). While Nc values were higher in LE, as expected in some species like *A. spectabilis*, *H. nana* or *M. fluctuans*, the species *C. hollandica*, *D. costata*, *H. gracilis*, *M. roanoka*, *S. dolichothecata* and both *S. diplocostata* gave higher results in HE than in LE. Post-screening (detailed in Table 35), all species gave lower Nc values consistently in the HE datasets, and also lower than both LE and All sets. The second group of results obtained are consistent with selection favouring optimal codons and with the results obtained for CBI and GC3 content. The species showing higher level of general bias were *S. punica*, *S. kvevrii*, *M. fluctuans* and *C. hollandica*, while the species showing less bias were *S. diplocostata* and *D. costata*. The lowest biased species, *S. diplocostata*, also gave the lowest results in CBI and GC3 in optimal codons, and was one of the species with less increment of GC3 in the HE dataset.

The Fop results followed the same tendency as seen for previous statistics. In the pre-screening set, optimal codons were generally more frequent in HE (as shown in Table 36), but the species *S. infusionum* showed the All dataset even higher in those codons, and species like *S. punica* and *S. parva* gave higher results in LE than in the general set. A post-screening analysis (which results can be seen in Table 37) revealed LE genes showing less optimal codons than the average for the whole CDS, and HE presented consistently a higher value. All results for HE were higher post-screening, except for *S. dolichothecata*, which frequency became more apparent by a decreasing of optimal codons in the other gene sets.

The species showing higher frequencies for optimal codons in HE were *M. fluctuans*, *S. helianthica*, *C. hollandica*, *S. punica* and *S. kvevrii*, all close related choanoflagellates (belonging to Craspedida Clade 2), while the species with lowest Fop values were *D. grandis*, *S. diplocostata*, *H. gracilis* and *D. costata*, three of them (except *H. gracilis*) also close related (the three species of the family Stephanoecidae).

Table 36. Fop mean and standard deviation for original CDS in choanoflagellates. All values were obtained using custom Fop files. The formatting is similar in style as table 28.

Frequency of optimal codons – Original CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.449 ± 0.129	0.530 ± 0.110	0.592 ± 0.104
<i>C. perplexa</i>	0.518 ± 0.073	0.534 ± 0.066	0.610 ± 0.104
<i>C. hollandica</i>	0.369 ± 0.131	0.450 ± 0.086	0.465 ± 0.059
<i>D. grandis</i>	0.399 ± 0.074	0.405 ± 0.075	0.456 ± 0.117
<i>D. costata</i>	0.473 ± 0.075	0.475 ± 0.058	0.534 ± 0.083
<i>H. balthica</i>	0.463 ± 0.081	0.481 ± 0.075	0.587 ± 0.112
<i>H. gracilis</i>	0.377 ± 0.082	0.391 ± 0.080	0.463 ± 0.110
<i>H. nana</i>	0.445 ± 0.108	0.494 ± 0.087	0.566 ± 0.092
<i>M. roanoka</i>	0.424 ± 0.140	0.516 ± 0.098	0.563 ± 0.075
<i>M. fluctuans</i>	0.564 ± 0.084	0.581 ± 0.076	0.674 ± 0.098
<i>S. dolichothecata</i>	0.512 ± 0.087	0.538 ± 0.071	0.590 ± 0.097
<i>S. helianthica</i>	0.528 ± 0.104	0.531 ± 0.093	0.587 ± 0.147
<i>S. infusionum</i>	0.448 ± 0.069	0.507 ± 0.065	0.500 ± 0.109
<i>S. kjevrii</i>	0.453 ± 0.105	0.528 ± 0.104	0.636 ± 0.147
<i>S. macrocollata</i>	0.428 ± 0.083	0.449 ± 0.072	0.490 ± 0.077
<i>S. punica</i>	0.534 ± 0.123	0.529 ± 0.110	0.606 ± 0.124
<i>S. urceolata</i>	0.469 ± 0.095	0.470 ± 0.081	0.529 ± 0.110
<i>S. parva</i>	0.433 ± 0.061	0.385 ± 0.060	0.433 ± 0.061
<i>S. diplocostata</i> (Au)	0.485 ± 0.080	0.496 ± 0.066	0.528 ± 0.080
<i>S. diplocostata</i> (Fr)	0.413 ± 0.084	0.429 ± 0.072	0.453 ± 0.083

Table 37. Fop mean and standard deviation for screened CDS in choanoflagellates. All values were obtained using custom Fop files. The formatting is similar in style as Table 28.

Frequency of optimal codons – Screened CDS			
Species	LE	All	HE
<i>A. spectabilis</i>	0.455 ± 0.071	0.467 ± 0.081	0.611 ± 0.088
<i>C. perplexa</i>	0.465 ± 0.050	0.495 ± 0.063	0.634 ± 0.111
<i>C. hollandica</i>	0.550 ± 0.097	0.564 ± 0.082	0.739 ± 0.111
<i>D. grandis</i>	0.374 ± 0.040	0.393 ± 0.060	0.548 ± 0.091
<i>D. costata</i>	0.432 ± 0.044	0.442 ± 0.056	0.579 ± 0.086
<i>H. balthica</i>	0.409 ± 0.056	0.431 ± 0.078	0.619 ± 0.113
<i>H. gracilis</i>	0.433 ± 0.056	0.450 ± 0.066	0.573 ± 0.101
<i>H. nana</i>	0.510 ± 0.068	0.523 ± 0.067	0.666 ± 0.089
<i>M. roanoka</i>	0.492 ± 0.059	0.503 ± 0.072	0.665 ± 0.117
<i>M. fluctuans</i>	0.595 ± 0.056	0.636 ± 0.067	0.779 ± 0.096
<i>S. dolichothecata</i>	0.445 ± 0.055	0.457 ± 0.061	0.584 ± 0.106
<i>S. helianthica</i>	0.569 ± 0.064	0.602 ± 0.075	0.753 ± 0.096
<i>S. infusionum</i>	0.468 ± 0.076	0.487 ± 0.083	0.694 ± 0.112
<i>S. kjevrii</i>	0.507 ± 0.074	0.555 ± 0.086	0.710 ± 0.095
<i>S. macrocollata</i>	0.561 ± 0.067	0.570 ± 0.064	0.678 ± 0.090
<i>S. punica</i>	0.597 ± 0.100	0.598 ± 0.088	0.718 ± 0.091
<i>S. urceolata</i>	0.495 ± 0.061	0.498 ± 0.069	0.636 ± 0.097
<i>S. parva</i>	0.458 ± 0.055	0.474 ± 0.059	0.596 ± 0.091
<i>S. diplocostata</i> (Au)	-	-	-
<i>S. diplocostata</i> (Fr)	0.452 ± 0.052	0.465 ± 0.048	0.570 ± 0.084

In general, results pre-screening were inconsistent across the parameters studied in this section, presenting results that are difficult to explain without considering bacterial contamination biasing the highly expressed genes identified, as the presence of low Fop in HE of *S. parva* and *S. infusionum* (Table 36), or the higher Nc value in HE than in LE of 7 species (Table 34). The screened dataset provided consistent high GC3 content in HE while this was not the case in general GC (Tables 31 and 29), high CAI and Fop values (Tables 33 and 37), plus lowest Nc values in HE of all the choanoflagellates studied in this chapter (Table 35). These results were consistent with the optimal codons identified in 3.2.2, where it was shown the preference of choanoflagellates for certain optimal codons and mainly ended in GC, with sibling species sharing similar results. This distribution was found also in the frequencies of these optimal codons, although neither completely fitting the phylogenetic tree, as for instance *H. gracilis* presented similar Fop as the Stephanoecidae species, but belongs to a distant place in Craspedida Clade 1, while its sibling *H. balthica* presented very different values. Despite this nuances, the screened dataset provided, for all the choanoflagellates, evidence compatible with natural selection upon transcriptional efficiency.

3.2.3.2 Effective number of codons and GC3 plots.

The effective number of codons (Nc) and GC3 data for each gene (calculated in 2.2.3) were plotted to study their tendency over a line representing the expected position under a model for neutral mutation evolution (Wright, 1990). The plots resulted allow a visualisation of the level of bias towards GC3. While whole datasets for CDS could result unclear with the numerous dots/genes covering a big part of the graphic, the screened transcriptome offers a slenderised version that also clarify a double cloud of dots in some cases. Below is listed a comparison of the whole set of CDS and the screened transcriptome, marked by “(S)” for screened (shown in Figures 08 to 27).

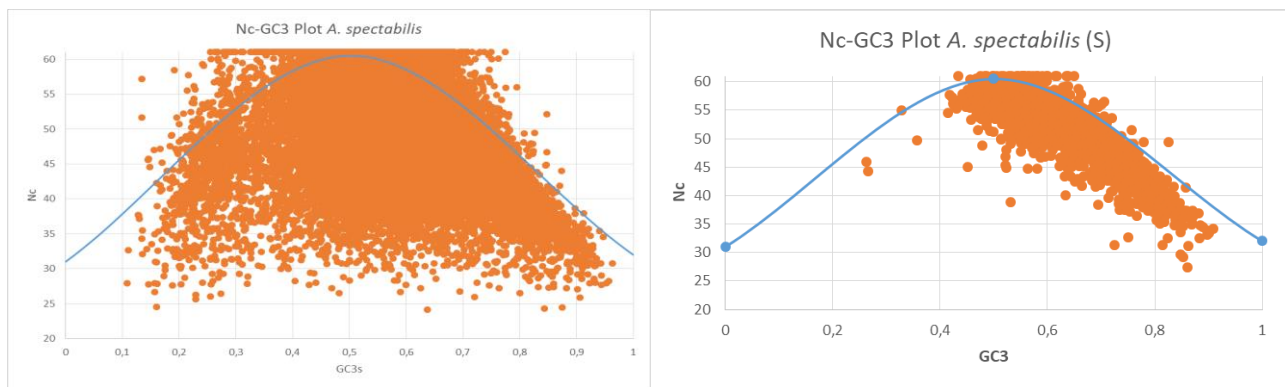


Figure 08. Nc and GC3 plot for *A. spectabilis* genes. The graphs present these values for genes in the complete CDS dataset (left side) and screened (S) transcriptome (right side). Genes are represented as orange dots and the neutral mutation line is showed in blue.

A. spectabilis presented in Figure 08 an unclear tendency pre-screening, with biased genes showing both high and low GC3 use, and most of them under the line for bias expected in the neutral mutation model. The fact that most genes are not in the area predicted in a neutral mutation model implies that those genes may be linked to bias by natural selection and/or genetic drift. In this graph, the double cloud of dots represents a subset of genes with different values to the rest of CDS, possibly showing genes with bacterial origin. Post-screening, bacterial genes were removed and the plot presented a clear correlation between general codon bias and GC3 content. These results correlates with its 90.5% of GC3 content in optimal codons, as highly expressed genes are more biased and enriched on these codons.

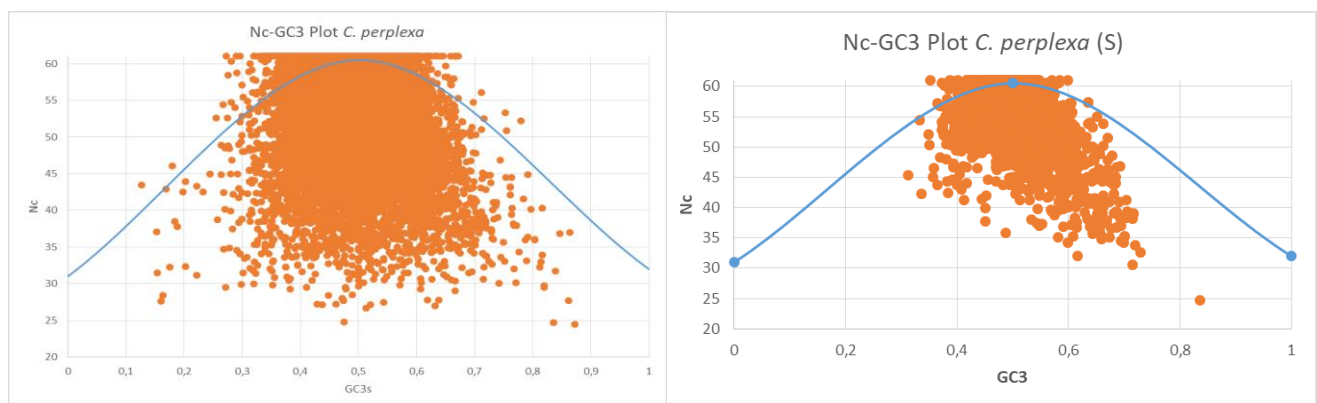


Figure 09. Nc and GC3 plot for *C. perplexa* genes. The formatting is similar in style as Figure 08.

C. perplexa exhibited in Figure 09 a single cloud of dots under the line, suggesting a set of genes that had non-neutral mutations through evolution. Although genes appeared in the centre of the graph for the original CDS in a symmetrical way, the screened set evidenced some correlation of biased genes towards GC3. This correlation is consistent with the lower GC3 content in optimal codons in comparison with *A. spectabilis* (Figure 08), slightly over 66%.

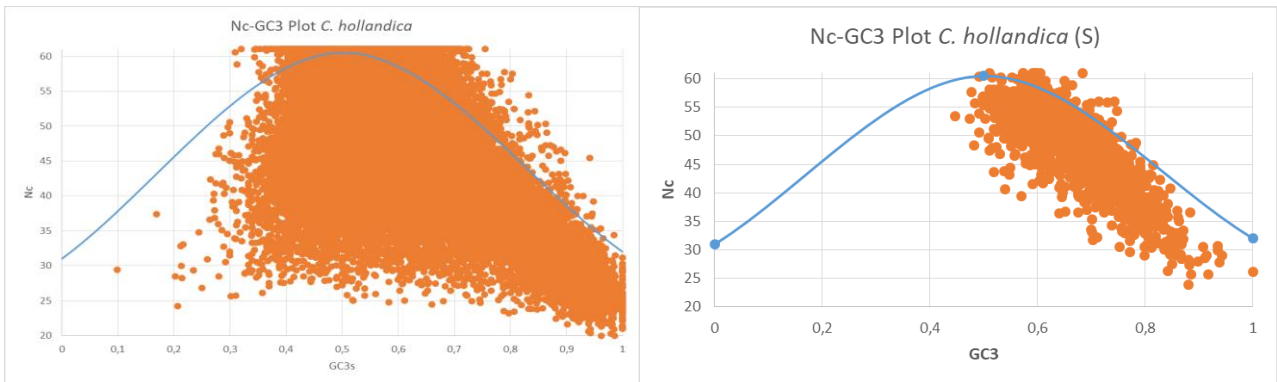


Figure 10. Nc and GC3 plot for *C. hollandica* genes. The formatting is similar in style as Figure 08.

C. hollandica presented in Figure 10 numerous genes pre-screening, distributed around the centre and right side of the graph, mainly under the line for neutral mutation genes. This set did not show a clear trend, although highly biased genes high in AU3 content were absent. Post-screening analysis revealed a strong clear correlation of the level of bias and GC3, consistent with the high level of GC3 found in optimal codons and high Fop in this species (Tables 31 and 37), as most of the Craspedida Clade 2 relatives.

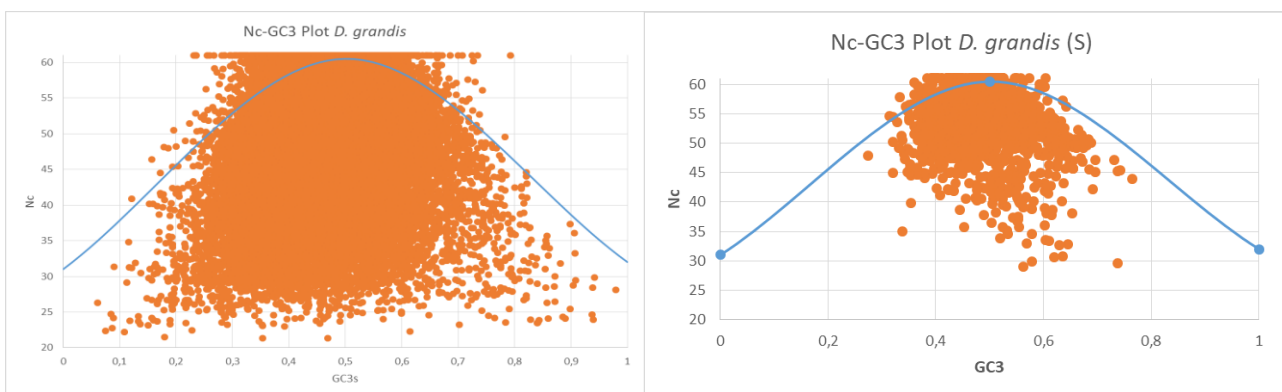


Figure 11. Nc and GC3 plot for *D. grandis* genes. The formatting is similar in style as Figure 08.

D. grandis presented in Figure 11 genes surrounding the middle of the graph in both pre and post-screening, with genes far from the neutral mutation line. Although disperse genes may present some high GC3 content, the tendency was generally towards the centre, meaning a dataset with highly biased genes not correlated with GC3 content. This result was consistent with the lowest level of Fop found in HE for the species of the family Stephanoecidae, although the results obtained in GC3, CBI, Nc and Fop in HE, and optimal codon GC3 (Tables 31, 33, 35 and 37), contradict the total absence of GC3 and general bias in HE genes.

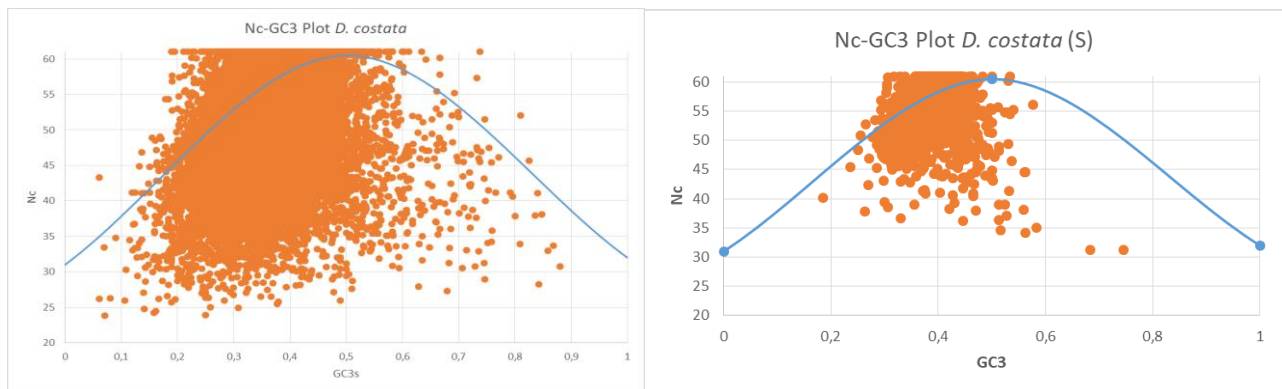


Figure 12. Nc and GC3 plot for *D. costata* genes. The formatting is similar in style as Figure 08.

D. costata showed in both pre and post-screening most of genes biased towards the left side of the graph and under the neutral mutation line (shown in Figure 12). This result is consistent with the fact that this species presented the lowest GC3 content in all the choanoflagellates studied and was one of the family Stephanoecidae presenting low Fop values in comparison with other species. Although the plot showed a general low GC3 content and no correlation with highly biased genes, these HE separately showed a clear bias in GC3, CBI, Nc and Fop in previous results.

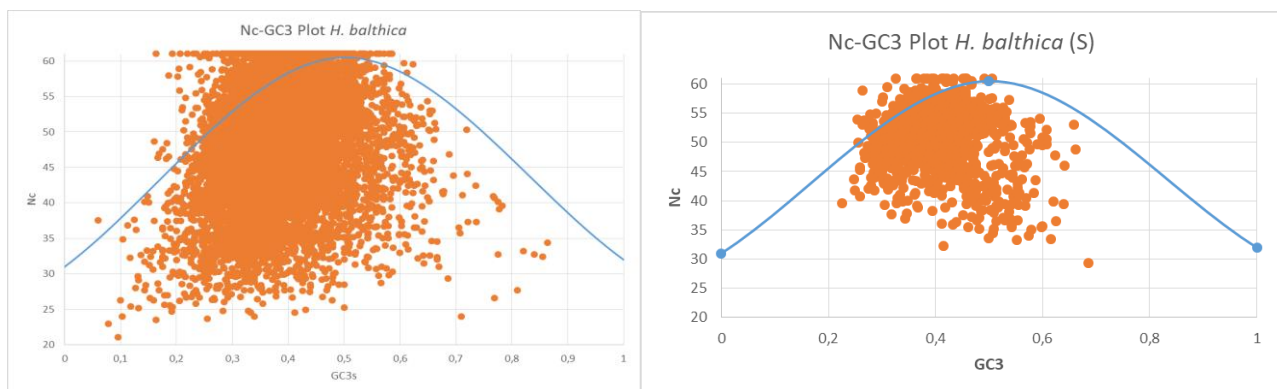


Figure 13. Nc and GC3 plot for *H. balthica* genes. The formatting is similar in style as Figure 08.

H. balthica presented in Figure 13 a cluster of genes around the centre of the plot, slightly towards the left side in both pre and post-screening, and far from the neutral mutation predictions. Although this species erased most of the AU3 biased genes during the screening, the low GC3 resemble the data seen in the Stephanoecidae family and the plot that will be show next for its sibling *H. gracilis* (Figure 14). Nonetheless, the Fop values in *H. balthica* were higher than in the previous mentioned species and the HE genes showed previously a clear bias in the statistics GC3, CBI, Nc and Fop. From the data obtained can be inferred that this species is low in GC3 content but biased for optimal codons and GC3 increased in HE genes.

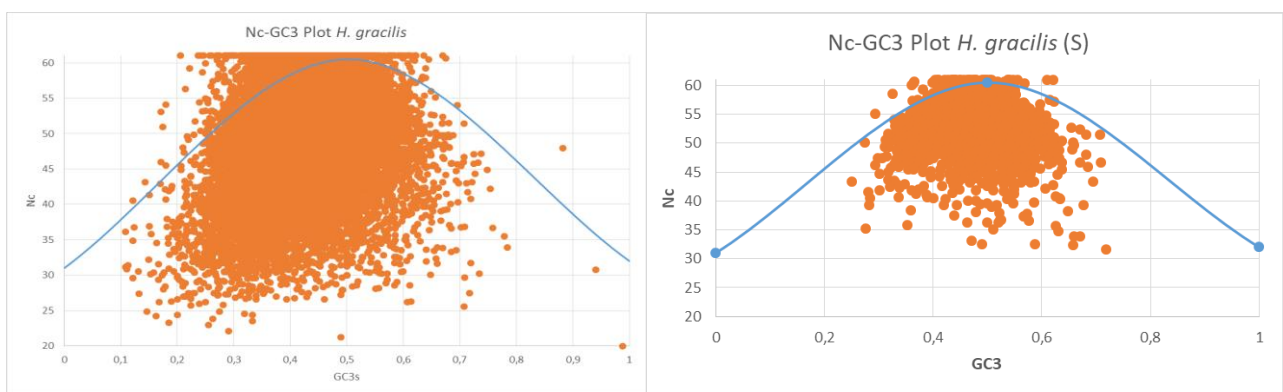


Figure 14. Nc and GC3 plot for *H. gracilis* genes. The formatting is similar in style as Figure 08.

H. gracilis exhibited genes mainly in the centre of the plot for both pre and post-screening, mainly below the neutral mutation line (shown in Figure 14). Although for the first dataset more biased genes with AU3 tendency appeared, these were deleted during the screening, giving a symmetrical plot. The pre-screening plot can be linked to the high CBI found in LE and All genes (Table 32), and the similarities of LE to HE in Nc values (Table 34), that revealed a possible contamination bias. Results post-screening were consistent with the values obtained for its sibling *H. balthica* (Figure 13) and the low Fop level found in comparison with other species (Table 37), as was seen in the Stephanoecidae studied. Nonetheless, HE genes revealed previously bias in the statistic GC3, Nc, CBI and Fop, as it happened for its sibling species, not reflected in this plot where no clear correlation can be seen.

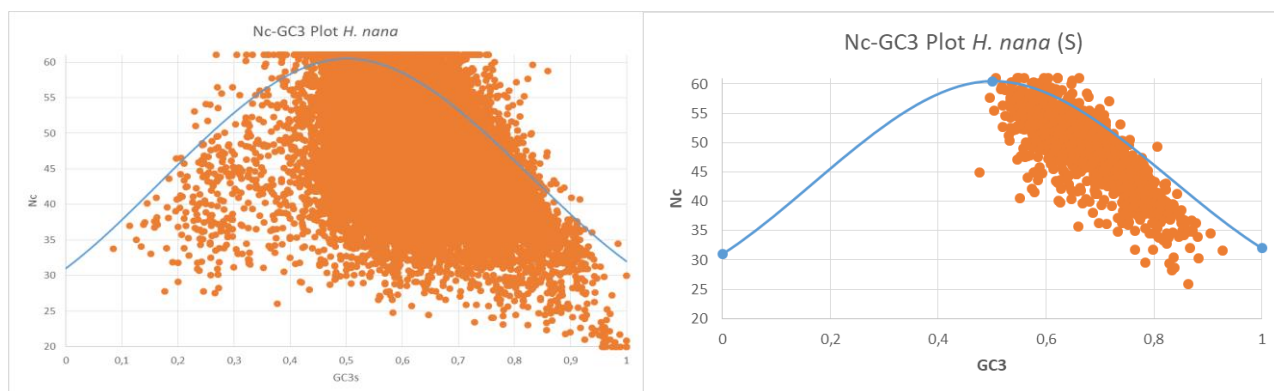


Figure 15. Nc and GC3 plot for *H. nana* genes. The formatting is similar in style as Figure 08.

H. nana showed in Figure 15 a main distribution of genes around the centre and right side of the plot, plus a cloud of scattered genes on the left side (possibly a different set of genes with bacterial origin), and mainly below the neutral mutation line. Post-screening, only a delimited cloud of dots remained showing a clear correlation of general bias and GC3 in genes. This plot was similar to the result seen in its closest related species *A. spectabilis* (Figure 08), and consistent with the results obtained in the codon usage statistics.

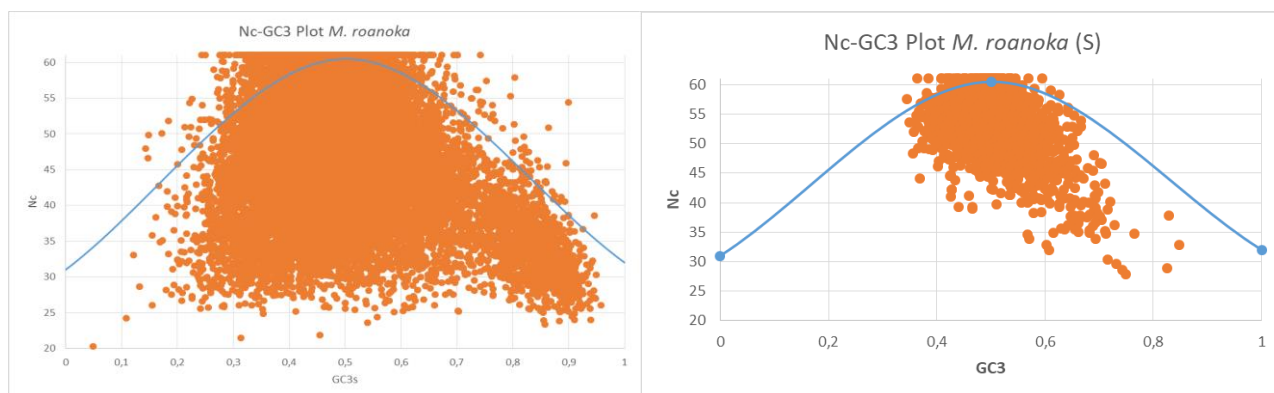


Figure 16. Nc and GC3 plot for *M. roanoka* genes. The formatting is similar in style as Figure 08.

M. roanoka presented two clouds of dots pre-screening, one in the centre of the plot and another one of genes at the right side, mainly below the neutral mutation line (as can be seen in Figure 16). The double cloud of dots, possibly showing a highly biased set of genes from bacterial origin, were deleted in the screened dataset, which showed a correlation of overall bias and GC3. The contamination theory in this species is in agreement with the high level of GC3 content found in LE and the great increase of this value in HE post-screening (Table 30 compared with 31), as it also happened with the value CBI and a reduction in Nc (Tables 32 vs 33, and 34 vs 35).

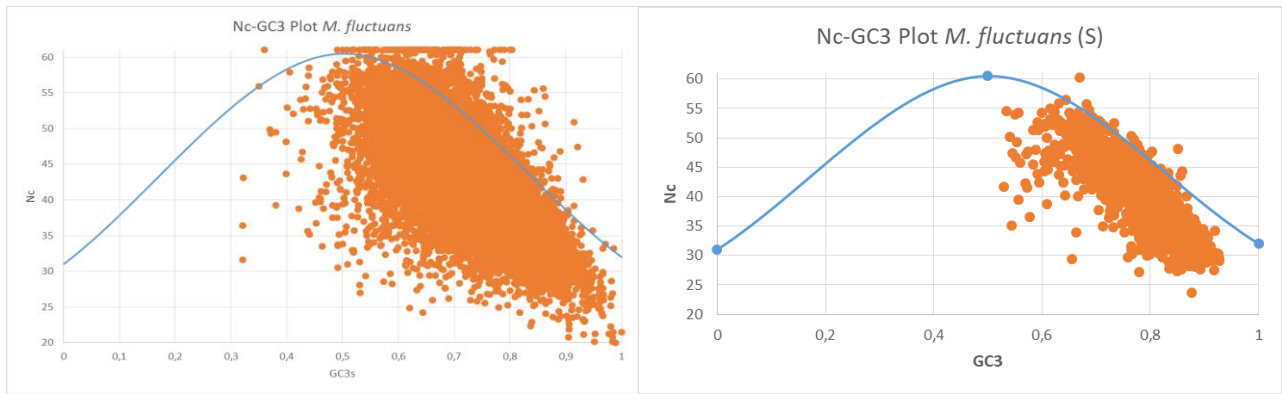


Figure 17. Nc and GC3 plot for *M. fluctuans* genes. The formatting is similar in style as Figure 08.

M. fluctuans, in Figure 17, showed pre-screening genes occupying the right side of the graph mainly below the neutral mutation line, showing biased genes with increasingly higher GC3 content. Post-screening, results remained similar, erasing many of the less biased genes and showing a clear correlation overall bias and GC3. This result was consistent with the fact that this species presented the highest Fop value in HE (Table 37), with the CBI, Nc and GC3 statistics showed before (Tables 33, 35 and 31 respectively) and with the related Craspedida Clade 2 species, which showed the most biased HE genes.

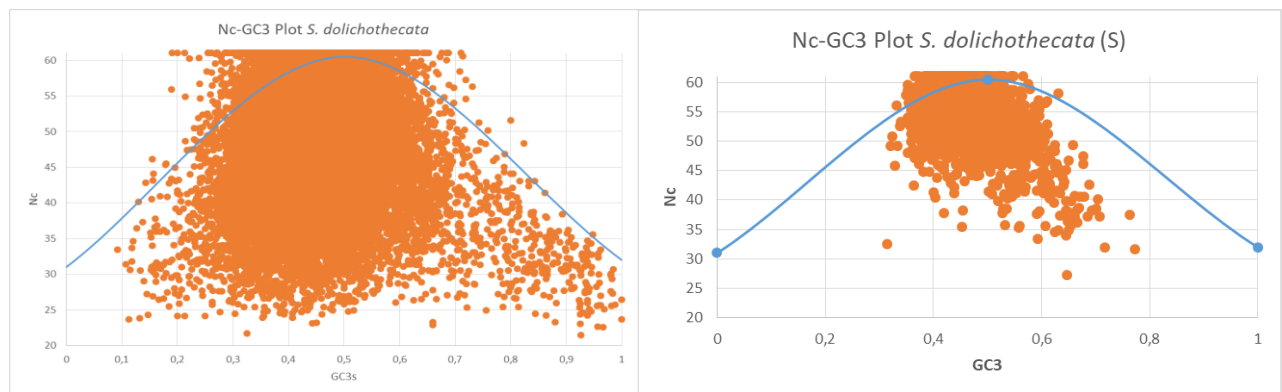


Figure 18. Nc and GC3 plot for *S. dolichothecata* genes. The formatting is similar in style as Figure 08.

S. dolichothecata exhibited genes grouped as a cloud in the middle of the graph, and also mainly below the neutral mutation line (shown in Figure 18). The pre-screening set showed a diversity of genes scattered with high and low GC3 use, but post-screening only remained genes in the centre and progressively increasing its Nc value and GC3, revealing a correlation. This correlation was not shown as biased towards the right side of the plot as in the Craspedida Clade 2 species, and as the

only representative of Craspedida Clade 3, *S. dolichothecata* presented a Fop level in a middle point between those calculated for the families Stephanoecidae and Acanthoecidae. This fact was also consistent with the statistics and the graph presented, but not phylogenetically, as Craspedida Clade 2 is closest related. An explanation for this phenomena is a higher selective pressure or its efficiency in Craspedida Clade 2, while the Clade 3 retained more similarities with the ancestral state.

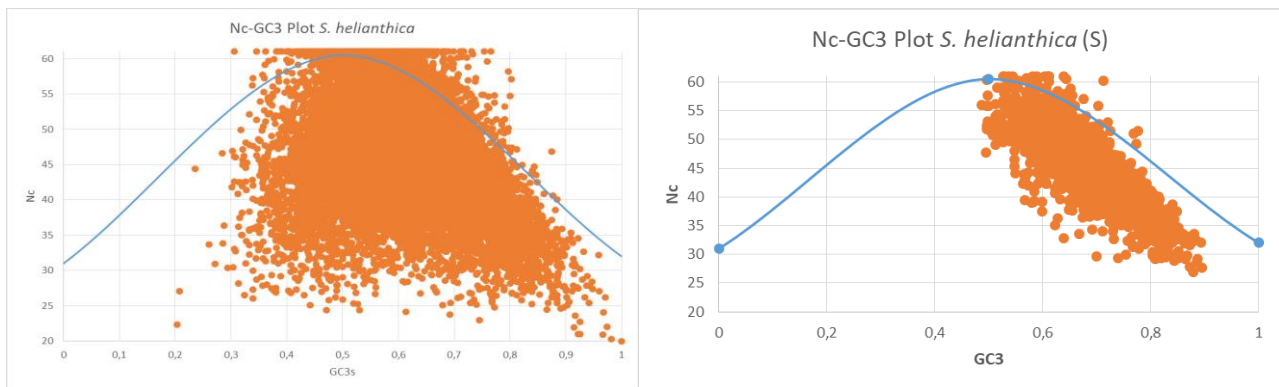


Figure 19. Nc and GC3 plot for *S. helianthica* genes. The formatting is similar in style as Figure 08.

S. helianthica manifested a single cloud of dots around the centre and the right side of the plot, which post-screening was reduced to a set of genes increasing progressively both in bias and GC3 content (as can be seen in Figure 19). The positive correlation of both parameters in this species was clear, and was consistent with the statistics showed previously, as the second with highest Fop in HE among all the species studied, and with its phylogenetic situation among the highly biased Craspedida Clade 2 choanoflagellates (like *M. fluctuans* and *S. punica*, which Nc-GC3 correlation is shown in Figures 17 and 23 respectively).

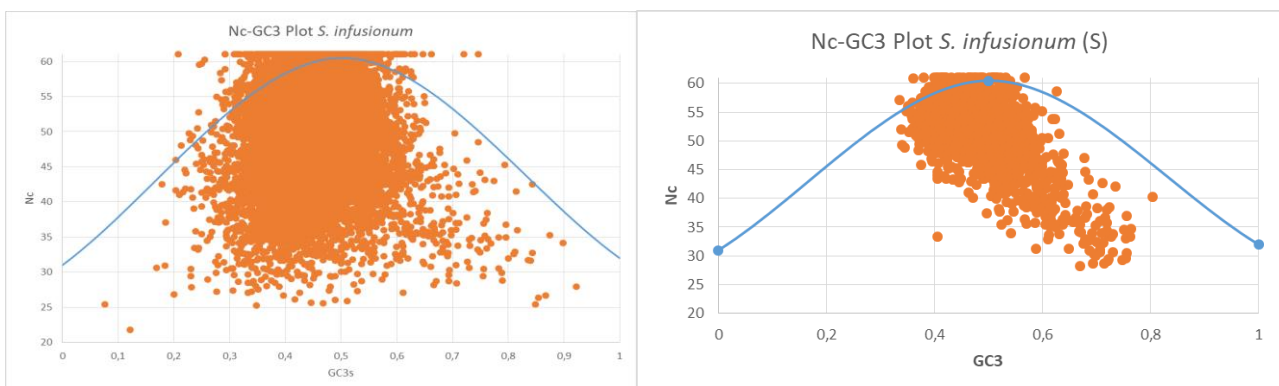


Figure 20. Nc and GC3 plot for *S. infusionum* genes. The formatting is similar in style as Figure 08.

S. infusionum presented in the pre-screening set a cloud of dots around the centre of the plot, below the neutral mutation line. Post-screening, the almost symmetrical cloud revealed a set of genes increasing in overall bias and GC3 at the same time (as it is shown in Figure 20). This clear correlation was similar to that found in *M. roanoka* (in Figure 16), the closest related species studied in this chapter (belonging to Craspedida Clade 1). Results were also consistent with the statistics presented in the previous section, which showed HE genes slightly more biased than in its mentioned related species, and GC3 in optimal codons over 70%.

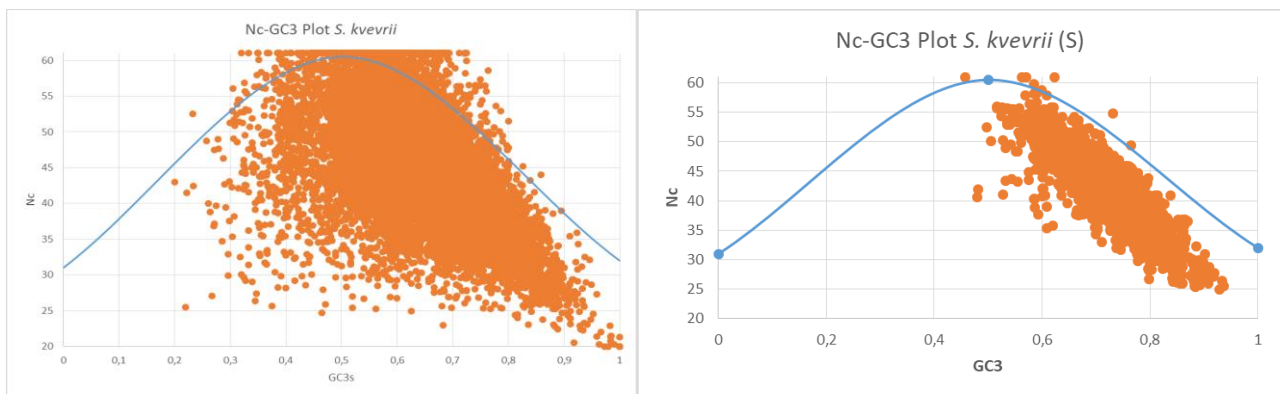


Figure 21. Nc and GC3 plot for *S. kvevrii* genes. The formatting is similar in style as Figure 08.

S. kvevrii showed in both pre and post-screening graphs a cloud of dots skewed towards the right side, although more clear in the screened set, and almost completely under the neutral mutation line (shown in Figure 21). The skewed distribution of genes revealed a clear correlation of overall bias and GC3, as it is expected in a species with the statistic presented in the previous section, included in the species with highest Fop in HE, exhibiting 100% of optimal codons ended in GC (Table 21), and among the previously seen highly biased Craspedida Clade 2 species.

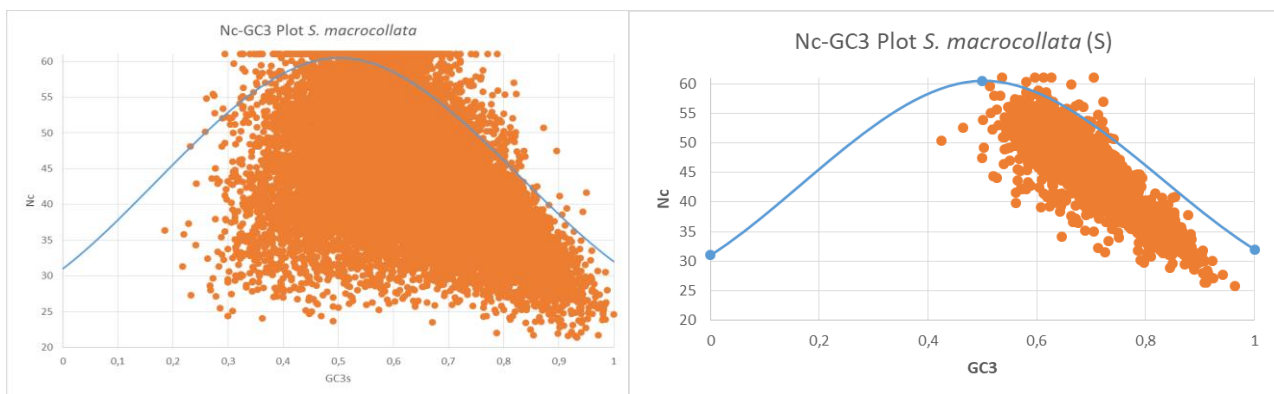


Figure 22. Nc and GC3 plot for *S. macrocollata* genes. The formatting is similar in style as Figure 08.

S. macrocollata exhibited in Figure 22, similar to other Craspedida Clade 2 species, a cloud representing genes mainly in the centre and right side of the plot that became very clear post-screening. A correlation between overall bias and GC3 content is evident in this species, and it is consistent with the previous statistics showed, with a GC3 content in optimal codons of 81% (shown in Table 22), and with its phylogenetic position close to species presenting the highest biased HE genes (like *S. punica*, *S. kvevrii* and *S. urceolata*).

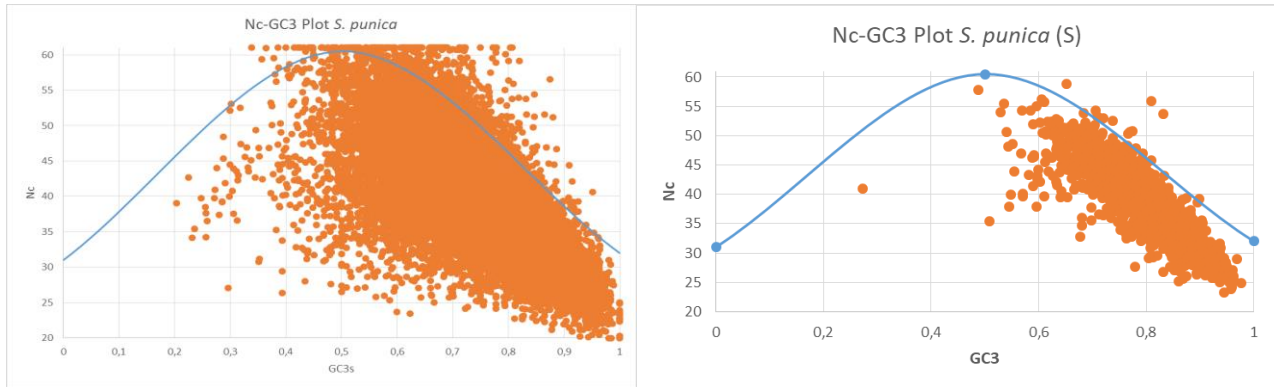


Figure 23. Nc and GC3 plot for *S. punica* genes. The formatting is similar in style as Figure 08.

S. punica presented, in Figure 23, very similar results to its sibling *S. macrocollata* (Figure 22). Both datasets, but more clear in post-screening genes, showed a correlation of overall bias with GC3. This, beside the support of the phylogenetic position in Craspedida Clade 2 and its close related species, is supported by HE genes with highest level of bias than its sibling, situated as the 4th species with highest Fop values, and with GC3 content in optimal codons of 90% (seen in Table 23).

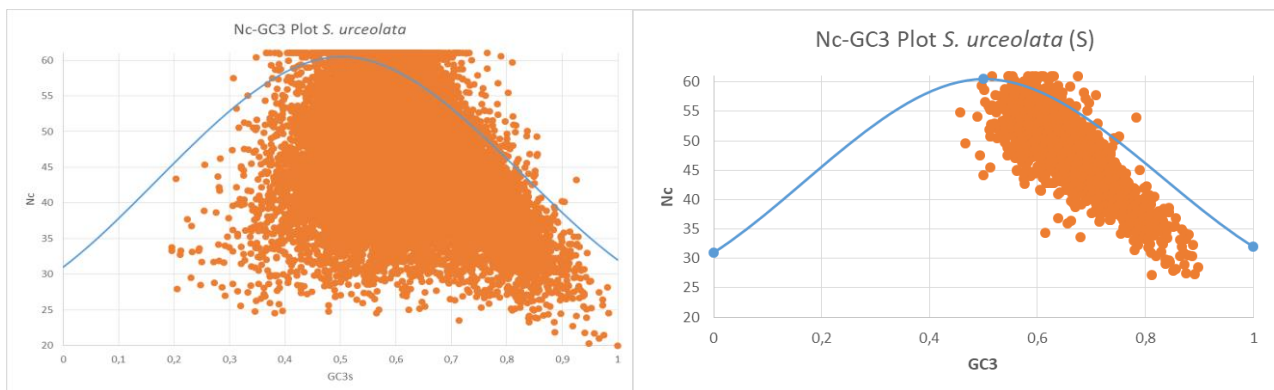


Figure 24. Nc and GC3 plot for *S. urceolata* genes. The formatting is similar in style as Figure 08.

S. urceolata manifested biased genes both around the middle and right side of the plot, mainly under the neutral mutation line (as detailed in Figure 24). The screening process revealed a clear correlation between overall bias and GC3 content, a result very similar to its closest sibling species *S. kvevrii* (shown in Figure 21), although showing it presented HE statistics with lower levels of bias. The correlation is also consistent with the bias presented in other Craspedida Clade 2 species, and the lower values like Fop in comparison with its sibling species are also reflected in the GC3 content in optimal codons, which decreased from 100% in *S. kvevrii* to 86% (as shown in Tables 21 and 24).

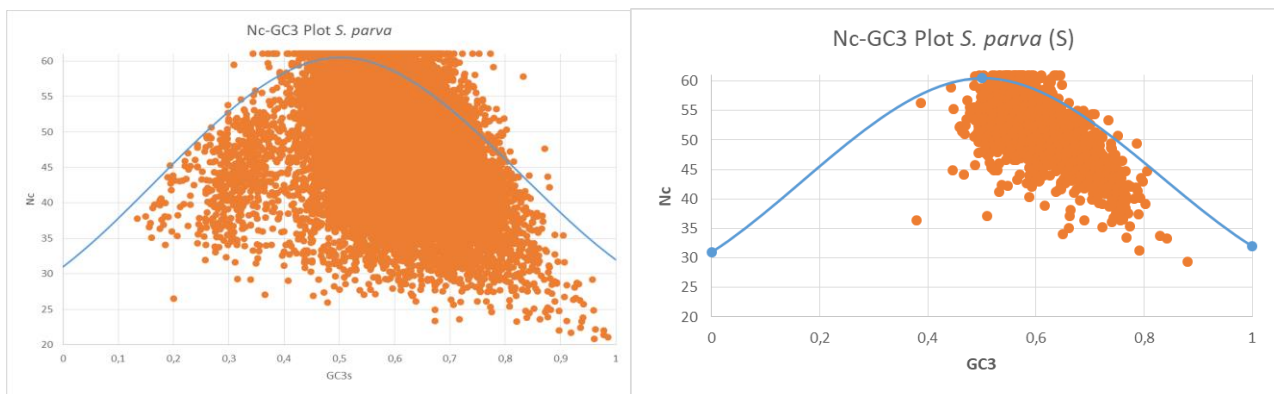


Figure 25. Nc and GC3 plot for *S. parva* genes. The formatting is similar in style as Figure 08.

S. parva showed a double cloud of dots for the pre-screening genes, a main one skewed towards the right side, and a secondary one at the left side (as can be seen in Figure 25). The double cloud of dots was mainly showed in the other Acanthoecidae species, *A. spectabilis* and *H. nana* (in Figures 08 and 15), revealing a subset of genes in this family that could have a bacterial origin, possibly by contamination. As in the mentioned species, the bacterial origin is also supported by the statistics calculated in the previous section, giving abnormalities like LE showing the same Fop as HE (while this statistics were calculated confronting HE and LE genes as explained in 2.2.2), or showing LE with lower Nc value than in HE. These statistical results changed post-screening to show a correlation of overall bias and GC3, as it also showed a clear codon bias in HE genes. The post-screening change was similar to the statistical values and plots of the other Acanthoecidae species, although of the three species studied in this family this was presenting lowest codon bias values, which is reflected also in the Figure 25 plot.

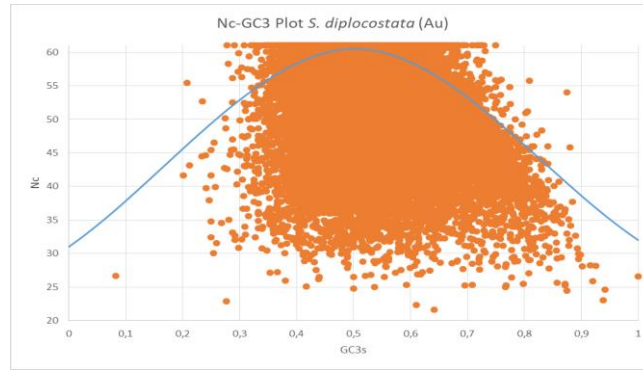


Figure 26. Nc and GC3 plot for *S. diplocostata* (Australia isolated) genes. The graph includes the results for all CDS. The formatting is similar in style as Figure 08.

S. diplocostata (Australia isolated) presented a full set of genes around the centre and toward the left side of the graph, under the neutral mutation line (shown in Figure 26). Nonetheless, the high amount of dispersed genes make unable to extract conclusions, and as this result was identical to the France isolate, *S. diplocostata* will be discussed more thoroughly based on Figure 27.

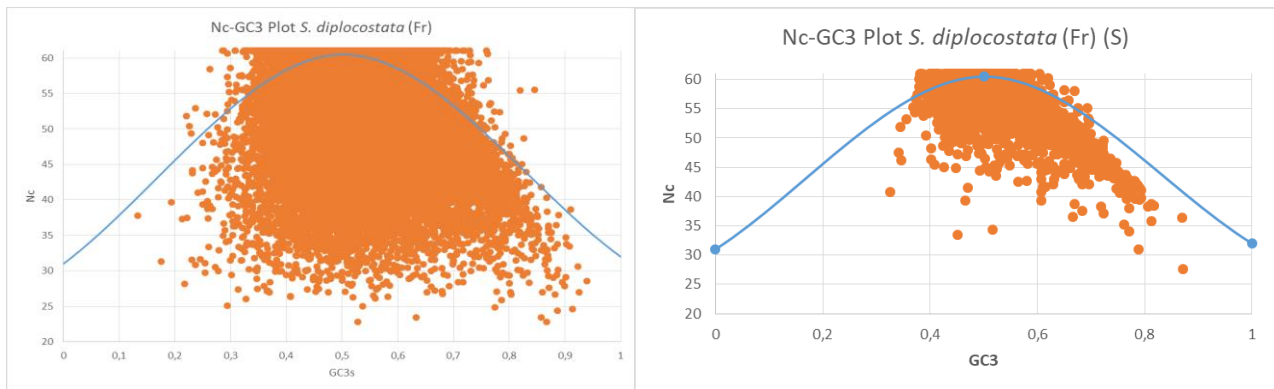


Figure 27. Nc and GC3 plot for *S. diplocostata* (France isolated) genes. The formatting is similar in style as Figure 08.

S. diplocostata (France isolate, shown in Figure 27) was very similar to the one from Australia (shown in Figure 26), with genes forming a cloud symmetrically around the centre with a slight skew toward the right side. Post-screening, the dataset was reduced to show a clear correlation of overall codon bias and GC3 content in genes. This species, together with the others belonging to the family Stephanoecidae (*D. costata* and *D. grandis*) was determined with one of the lowest Fop value in the HE genes. *S. diplocostata* also showed one of the lowest increment of GC3 in HE, the lowest GC3 in optimal codons, and the lowest bias in CBI and Nc values among all the choanoflagellates.

As a summary, correlation between overall codon bias and GC3 was evidenced for most choanoflagellates, showing consistency with the statistic values presented in the previous section, with the GC3 content of the optimal codons determined, and also related with the position in the phylogenetic tree. The plots showing the highest correlations between the named parameters were those of the species included in the highly biased Craspedida Clade 2, while the lowest correlation was found in the species of the family Stephanoecidae. The species representative of Craspedida Clade 3 (*S. dolichothecata*) resulted more similar to Acanthoecidae species than to Craspedida Clade 2. The family Acanthoecidae gathered most of the species presenting a double cloud of dots pre-screening, possibly belonging to a subset of genes with bacterial origin that were considered horizontal transfer in the initial bioinformatic filtering. The Craspedida Clade 1 included very different species, like the highly biased *S. infusionum*, species presenting low GC3 and showing the most unclear correlation Nc-GC3 like both *Hartaetosiga*, and *M. roanoka*, the only non-Acanthoecidae showing a double cloud of dots.

3.2.3.3 Codon usage in protein domain and non-domain regions.

The 100 genes with highest and lowest expressed genes for each choanoflagellate species were separated into domain and non-domains regions manually as explained in 2.2.3. The purpose of this separation was to analyse codon bias in predicted protein domain regions, areas where incorporate the wrong amino acid is more critic and could reveal selection upon translational accuracy. Both categories were analysed for codon usage statistics using CodonW, and the comparison of results for each statistics is shown below in Tables 38 to 42. Abbreviations used were NDR (protein non-domain regions), DR (protein domain regions), LE (lowest expressed genes), HE (highest expressed genes).

Table 38. GC content mean and standard deviation in protein domain and non-domain regions of choanoflagellates. The analysis was performed on the 100 HE and LE of the screened CDS (S).

GC content				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>A. spectabilis</i> (S)	0.578 ± 0.082	0.597 ± 0.068	0.555 ± 0.062	0.578 ± 0.035
<i>C. perplexa</i> (S)	0.514 ± 0.044	0.534 ± 0.075	0.506 ± 0.023	0.530 ± 0.025
<i>C. hollandica</i> (S)	0.582 ± 0.073	0.595 ± 0.076	0.554 ± 0.050	0.598 ± 0.028
<i>D. grandis</i> (S)	0.492 ± 0.049	0.501 ± 0.076	0.482 ± 0.039	0.498 ± 0.029
<i>D. costata</i> (S)	0.445 ± 0.070	0.472 ± 0.070	0.440 ± 0.034	0.466 ± 0.026
<i>H. balthica</i> (S)	0.440 ± 0.036	0.432 ± 0.092	0.435 ± 0.041	0.484 ± 0.030
<i>H. gracilis</i> (S)	0.464 ± 0.077	0.447 ± 0.092	0.455 ± 0.031	0.493 ± 0.030
<i>H. nana</i> (S)	0.603 ± 0.058	0.569 ± 0.071	0.584 ± 0.037	0.584 ± 0.027
<i>M. roanoka</i> (S)	0.477 ± 0.053	0.546 ± 0.061	0.501 ± 0.032	0.534 ± 0.032
<i>M. fluctuans</i> (S)	0.648 ± 0.060	0.602 ± 0.066	0.625 ± 0.041	0.604 ± 0.025
<i>S. dolichothecata</i> (S)	0.508 ± 0.090	0.513 ± 0.074	0.479 ± 0.033	0.523 ± 0.027
<i>S. helianthica</i> (S)	0.600 ± 0.049	0.568 ± 0.116	0.587 ± 0.039	0.593 ± 0.028
<i>S. infusionum</i> (S)	0.509 ± 0.032	0.547 ± 0.055	0.492 ± 0.03	0.541 ± 0.029
<i>S. kvevrii</i> (S)	0.600 ± 0.074	0.648 ± 0.064	0.577 ± 0.029	0.614 ± 0.029
<i>S. macrocollata</i> (S)	0.618 ± 0.054	0.571 ± 0.105	0.610 ± 0.049	0.595 ± 0.035
<i>S. punica</i> (S)	0.647 ± 0.081	0.609 ± 0.076	0.621 ± 0.040	0.626 ± 0.029
<i>S. urceolata</i> (S)	0.595 ± 0.052	0.593 ± 0.068	0.566 ± 0.034	0.592 ± 0.030
<i>S. parva</i> (S)	0.585 ± 0.074	0.590 ± 0.067	0.555 ± 0.051	0.567 ± 0.026
<i>S. diplocostata</i> (Fr) (S)	0.544 ± 0.070	0.542 ± 0.069	0.540 ± 0.058	0.519 ± 0.031

The GC content was calculated, as in 3.2.3.1, to differentiate the values of general GC and GC3, among HE, LE, DR and NDR in this case. For this parameter, results are shown in Table 38. If the mutation pressure is not strong enough to overwhelm amino acid usage, therefore is unlikely to overwhelm codon usage, and high GC3 in DR for species with high GC3 in optimal codons can be attributed to natural selection upon translational accuracy. In choanoflagellates no correlation of GC was presented between DR and NDR, or with HE and LE genes. Any comparison among the categories studied in Table 38 would give a list of exceptions that would reject any consistency. Nonetheless, these results would support a later GC3 analysis with a bias in this position, as it can be considered not previously determined by the general GC bias. Species containing the highest GC in DR of HE were *S. punica*, *S. kevrii*, *M. fluctuans*, *C. hollandica*, *S. macrocollata*, *S. helianthica* and *S. urceolata*, which correlate perfectly with the species belonging to Craspedida Clade 2. The species showing lowest value in the cited category were *D. costata*, *H. balthica*, *H. gracilis*, *D. grandis* and *S. diplocostata*, the species of the family Stephanoecidae and the two *Hartaetosiga* species of Craspedida Clade 1. Both described correlations were coincident with the level of bias found in the previous sections of 3.2.3, and this could be due to the fact that although general GC content was not correlated previously with HE, and neither was correlated with DR, GC in HR is correlated with the level of GC3 content, which is a parameter reflecting high codon usage bias.

Analysing the GC3 results, detailed in Table 39, the values presented in DR of HE were considerably higher than any other category, higher than NDR of HE in all the species, and higher than DR of LE (except for *S. diplocostata*). As presented in Table 38 for GC, this difference among categories was not influenced by the general GC bias, but the species showing the highest and lowest GC3 values in HR were similarly those belonging to Craspedida Clade 2, and to Stephanoecidae plus the genus *Hartaetosiga* respectively. An increase of GC3 in HE for species with high GC3 in optimal codons shows evidence of selection upon translational accuracy, due to the fact that those codons are selected in DR due to higher impact of its translation error in the protein function. These results were consistent with the statistics in the previous sections, like the comparison of Fop values in HE across the species (Table 37). Although *H. balthica* was not found previously among the less biased species, it indeed showed resemblance with its sibling species in analysis like in the Nc-GC3 plot presented in 3.2.3.2 (Figures 13 and 14).

Table 39. GC3 mean and standard deviation in protein domain and non-domain regions of choanoflagellates. The formatting is similar in style as Table 38.

GC3 content				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>A. spectabilis</i> (S)	0.620 ± 0.161	0.676 ± 0.150	0.634 ± 0.139	0.764 ± 0.064
<i>C. perplexa</i> (S)	0.511 ± 0.125	0.537 ± 0.168	0.511 ± 0.067	0.633 ± 0.068
<i>C. hollandica</i> (S)	0.678 ± 0.162	0.713 ± 0.156	0.705 ± 0.117	0.815 ± 0.069
<i>D. grandis</i> (S)	0.488 ± 0.113	0.487 ± 0.124	0.490 ± 0.079	0.539 ± 0.066
<i>D. costata</i> (S)	0.367 ± 0.116	0.432 ± 0.187	0.382 ± 0.069	0.465 ± 0.051
<i>H. balthica</i> (S)	0.390 ± 0.098	0.417 ± 0.118	0.404 ± 0.086	0.497 ± 0.076
<i>H. gracilis</i> (S)	0.463 ± 0.135	0.467 ± 0.147	0.452 ± 0.076	0.544 ± 0.068
<i>H. nana</i> (S)	0.654 ± 0.105	0.680 ± 0.147	0.683 ± 0.077	0.774 ± 0.059
<i>M. roanoka</i> (S)	0.512 ± 0.172	0.584 ± 0.144	0.526 ± 0.096	0.626 ± 0.090
<i>M. fluctuans</i> (S)	0.757 ± 0.101	0.725 ± 0.135	0.786 ± 0.071	0.829 ± 0.043
<i>S. dolichothecata</i> (S)	0.511 ± 0.164	0.528 ± 0.122	0.444 ± 0.061	0.604 ± 0.082
<i>S. helianthica</i> (S)	0.647 ± 0.082	0.648 ± 0.191	0.690 ± 0.065	0.802 ± 0.063
<i>S. infusionum</i> (S)	0.459 ± 0.092	0.566 ± 0.146	0.482 ± 0.069	0.640 ± 0.081
<i>S. kvevrii</i> (S)	0.650 ± 0.116	0.757 ± 0.115	0.688 ± 0.086	0.856 ± 0.061
<i>S. macrocollata</i> (S)	0.697 ± 0.131	0.667 ± 0.189	0.749 ± 0.089	0.820 ± 0.075
<i>S. punica</i> (S)	0.801 ± 0.124	0.736 ± 0.142	0.846 ± 0.078	0.895 ± 0.042
<i>S. urceolata</i> (S)	0.637 ± 0.099	0.680 ± 0.132	0.670 ± 0.061	0.786 ± 0.062
<i>S. parva</i> (S)	0.607 ± 0.152	0.656 ± 0.135	0.633 ± 0.091	0.722 ± 0.058
<i>S. diplocostata</i> (Fr) (S)	0.562 ± 0.127	0.564 ± 0.164	0.604 ± 0.117	0.603 ± 0.070

Table 40. CBI mean and standard deviation in protein domain and non-domain regions of choanoflagellates. The formatting is similar in style as Table 38.

Codon Bias Index				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>A. spectabilis</i> (S)	0.243 ± 0.312	0.363 ± 0.283	0.283 ± 0.269	0.535 ± 0.127
<i>C. perplexa</i> (S)	0.130 ± 0.234	0.369 ± 0.321	0.135 ± 0.118	0.546 ± 0.189
<i>C. hollandica</i> (S)	0.325 ± 0.265	0.419 ± 0.260	0.377 ± 0.249	0.651 ± 0.146
<i>D. grandis</i> (S)	0.043 ± 0.151	0.234 ± 0.171	0.015 ± 0.090	0.327 ± 0.155
<i>D. costata</i> (S)	0.047 ± 0.227	0.326 ± 0.303	0.042 ± 0.116	0.417 ± 0.168
<i>H. balthica</i> (S)	-0.093 ± 0.151	0.124 ± 0.232	-0.071 ± 0.141	0.342 ± 0.169
<i>H. gracilis</i> (S)	-0.043 ± 0.259	0.093 ± 0.297	-0.011 ± 0.111	0.288 ± 0.188
<i>H. nana</i> (S)	0.257 ± 0.186	0.371 ± 0.263	0.300 ± 0.130	0.541 ± 0.113
<i>M. roanoka</i> (S)	0.096 ± 0.350	0.360 ± 0.313	0.199 ± 0.218	0.539 ± 0.217
<i>M. fluctuans</i> (S)	0.358 ± 0.170	0.411 ± 0.245	0.428 ± 0.123	0.636 ± 0.097
<i>S. dolichothecata</i> (S)	0.092 ± 0.295	0.308 ± 0.295	0.055 ± 0.115	0.466 ± 0.186
<i>S. helianthica</i> (S)	0.298 ± 0.154	0.358 ± 0.365	0.380 ± 0.119	0.664 ± 0.122
<i>S. infusionum</i> (S)	0.087 ± 0.193	0.400 ± 0.283	0.111 ± 0.166	0.566 ± 0.174
<i>S. kvevrii</i> (S)	0.245 ± 0.164	0.438 ± 0.221	0.315 ± 0.148	0.677 ± 0.113
<i>S. macrocollata</i> (S)	0.366 ± 0.234	0.329 ± 0.345	0.453 ± 0.159	0.610 ± 0.144
<i>S. punica</i> (S)	0.468 ± 0.248	0.436 ± 0.256	0.561 ± 0.154	0.722 ± 0.096
<i>S. urceolata</i> (S)	0.218 ± 0.147	0.362 ± 0.237	0.294 ± 0.111	0.554 ± 0.127
<i>S. parva</i> (S)	0.250 ± 0.274	0.350 ± 0.262	0.307 ± 0.148	0.533 ± 0.121
<i>S. diplocostata</i> (Fr) (S)	0.132 ± 0.252	0.130 ± 0.324	0.213 ± 0.227	0.215 ± 0.133

The CBI data presented in Table 40, as in GC3, showed higher value in DR of HE than in any other category for all the species, followed by the values obtained in NDR for HE, except for three species (*S. macrocollata*, *S. punica* and *S. diplocostata*), which showed higher bias in DR than NDR for both HE and LE genes. Also similarly to the GC3 analysis showed in Table 39, the species showing highest value in DR were included in Craspedida Clade 2, more specifically these 7 species and *S. infusionum* (from Craspedida Clade 1). The species with lowest value in HE also belonged to the families Stephanoecidae and *Hartaetosiga*, a result correlating with that obtained for CBI in HE genes.

The Nc values results are detailed in Table 41. These were consistently lower in HE than in LE genes, meaning a higher level of overall codon bias, but this trend was not seen between DR and NDR. In this analysis, HE genes presented higher Nc value in DR than in NDR for the species *A. spectabilis*, *D. grandis*, *D. costata*, *H. gracilis*, *H. nana*, *S. dolichothecata* and *S. diplocostata*. The species presenting higher overall bias were, as in CBI, those in Craspedida Clade 2 and *S. infusionum*, while the lowest bias was found also in the family Stephanoecidae and *H. gracilis*.

Fop results, described in Table 42, showed the highest values for DR in HE with no exception across all the choanoflagellates studied. The next dataset with highest values were NDR in HE sequences, also without exceptions, followed by the results obtained in LE genes with no trend showing DR or NDR consistently higher. As in CBI and Nc values (Tables 40 and 41), the highest Fop was found in the species of Craspedida Clade 2 and *S. infusionum*. In this case, *H. nana* presented the highest value after the mentioned species, although in the Fop for HE calculated (Table 37) this species showed a higher result than *M. roanoka*. The lowest Fop was found in the family Stephanoecidae and in *H. gracilis*, similarly to the Fop analysis for HE genes.

Table 41. Nc mean and standard deviation in protein domain and non-domain regions of choanoflagellates. The formatting is similar in style as Table 38.

Effective Number of Codons				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>A. spectabilis</i> (S)	50.376 ± 6.401	40.135 ± 8.816	48.887 ± 7.246	41.002 ± 5.988
<i>C. perplexa</i> (S)	51.745 ± 8.385	43.679 ± 9.669	54.067 ± 5.376	41.597 ± 6.843
<i>C. hollandica</i> (S)	46.119 ± 9.934	37.936 ± 10.459	42.854 ± 9.656	33.576 ± 6.004
<i>D. grandis</i> (S)	52.849 ± 7.144	41.914 ± 10.209	54.011 ± 4.691	42.623 ± 6.731
<i>D. costata</i> (S)	51.085 ± 7.657	45.087 ± 10.455	52.739 ± 6.102	46.174 ± 5.576
<i>H. balthica</i> (S)	51.662 ± 5.836	42.459 ± 7.162	53.093 ± 3.998	41.380 ± 5.370
<i>H. gracilis</i> (S)	49.480 ± 6.877	44.846 ± 8.377	51.996 ± 6.016	45.698 ± 7.042
<i>H. nana</i> (S)	49.622 ± 6.784	38.743 ± 7.407	47.707 ± 6.722	39.025 ± 5.505
<i>M. roanoka</i> (S)	52.660 ± 9.660	44.864 ± 8.313	48.547 ± 9.78	40.902 ± 7.581
<i>M. fluctuans</i> (S)	42.166 ± 7.607	37.429 ± 9.546	41.275 ± 6.960	32.333 ± 4.167
<i>S. dolichothecata</i> (S)	52.648 ± 8.574	40.921 ± 9.454	52.927 ± 7.726	42.440 ± 6.123
<i>S. helianthica</i> (S)	47.356 ± 6.747	35.791 ± 8.520	45.843 ± 5.242	34.071 ± 5.451
<i>S. infusionum</i> (S)	50.948 ± 8.038	39.871 ± 8.709	51.866 ± 5.987	37.448 ± 6.564
<i>S. kvevrii</i> (S)	45.726 ± 7.415	36.593 ± 9.096	43.808 ± 8.149	29.472 ± 4.282
<i>S. macrocollata</i> (S)	42.527 ± 7.880	37.909 ± 10.932	41.503 ± 7.218	35.474 ± 6.106
<i>S. punica</i> (S)	37.047 ± 8.751	35.183 ± 8.567	34.613 ± 6.216	29.288 ± 3.721
<i>S. urceolata</i> (S)	47.608 ± 5.679	40.085 ± 8.427	46.984 ± 6.049	38.069 ± 6.391
<i>S. parva</i> (S)	48.526 ± 6.787	45.192 ± 9.475	49.087 ± 6.510	42.316 ± 5.968
<i>S. diplocostata</i> (Fr) (S)	53.640 ± 5.666	44.226 ± 9.151	51.251 ± 7.440	47.888 ± 6.366

Table 42. Fop mean and standard deviation in protein domain and non-domain regions of choanoflagellates. Only custom Fop files were used. The formatting is similar in style as Table 38.

Frequency of Optimal Codons				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>A. spectabilis</i> (S)	0.452 ± 0.132	0.572 ± 0.153	0.474 ± 0.089	0.662 ± 0.080
<i>C. perplexa</i> (S)	0.473 ± 0.118	0.614 ± 0.183	0.472 ± 0.060	0.692 ± 0.106
<i>C. hollandica</i> (S)	0.600 ± 0.171	0.666 ± 0.172	0.630 ± 0.151	0.800 ± 0.091
<i>D. grandis</i> (S)	0.381 ± 0.070	0.537 ± 0.113	0.370 ± 0.047	0.589 ± 0.096
<i>D. costata</i> (S)	0.425 ± 0.102	0.587 ± 0.168	0.448 ± 0.057	0.624 ± 0.078
<i>H. balthica</i> (S)	0.408 ± 0.080	0.550 ± 0.130	0.415 ± 0.071	0.680 ± 0.102
<i>H. gracilis</i> (S)	0.427 ± 0.131	0.513 ± 0.174	0.447 ± 0.066	0.618 ± 0.110
<i>H. nana</i> (S)	0.493 ± 0.113	0.621 ± 0.165	0.528 ± 0.078	0.723 ± 0.080
<i>M. roanoka</i> (S)	0.501 ± 0.195	0.642 ± 0.164	0.549 ± 0.114	0.727 ± 0.117
<i>M. fluctuans</i> (S)	0.572 ± 0.100	0.713 ± 0.166	0.611 ± 0.070	0.841 ± 0.070
<i>S. dolichothecata</i> (S)	0.450 ± 0.162	0.602 ± 0.150	0.442 ± 0.066	0.663 ± 0.103
<i>S. helianthica</i> (S)	0.556 ± 0.089	0.680 ± 0.214	0.593 ± 0.070	0.815 ± 0.083
<i>S. infusionum</i> (S)	0.470 ± 0.110	0.642 ± 0.155	0.485 ± 0.094	0.743 ± 0.102
<i>S. kvevrui</i> (S)	0.486 ± 0.116	0.623 ± 0.145	0.544 ± 0.100	0.779 ± 0.075
<i>S. macrocollata</i> (S)	0.547 ± 0.107	0.634 ± 0.195	0.586 ± 0.069	0.735 ± 0.092
<i>S. punica</i> (S)	0.590 ± 0.158	0.622 ± 0.163	0.640 ± 0.097	0.779 ± 0.068
<i>S. urceolata</i> (S)	0.470 ± 0.080	0.607 ± 0.147	0.506 ± 0.068	0.699 ± 0.083
<i>S. parva</i> (S)	0.509 ± 0.191	0.569 ± 0.143	0.473 ± 0.069	0.643 ± 0.079
<i>S. diplocostata</i> (Fr) (S)	0.440 ± 0.101	0.542 ± 0.164	0.471 ± 0.061	0.616 ± 0.083

The statistical results presented were consistent through the different parameters. A detail to mention is that although *H. balthica* resulted one of the species with lowest GC and GC3 content, it was not included in the lowest biased species, but this was also seen with the contrast of its HE statistics and the GC3-Nc plot. In summary, GC3 analysis revealed a higher content of this parameter for DR than in NDR in HE, and also in DR HE over DR LE (except for *S. diplocostata*). CBI presented also highest values in DR over NDR in HE, and also in DR HE over DR LE (although *S. diplocostata* showed close results in both), which support the results seen for GC3. The Nc value was found lower in DR than in NDR of HE in 12 species, while HE presented lower values consistently than LE. And finally Fop was determined higher in DR of HE for all the species, followed by NDR HE, and LE genes.

Although not reflected in some Nc values, the difference of DR over NDR in GC3, CBI and Fop revealed consistently a selection upon translational accuracy, finding evidence even in the most weakly expressed genes in the genome (except for *C. perplexa*, *D. grandis* and *S. dolichothecata*). The justification of this values and an explanation for this theory is the increased bias for optimal codons (mainly ended in GC) in DR. This phenomena would be a product of the advantage presented by the lowest error rate of these codons for the incorporation of the correct amino acids, as DR fidelity is key in the correct functionality of the proteins. Additionally, the results showing values of higher bias in DR of HE than in LE is consistent with the previous HE genes analysis, which is compatible with natural selection upon translational efficiency, as the reason for this selection would not be the advantage of the minimization of errors in translation, but an advantage of the translation speed in HE genes. At the same time the lowest GC3 and CBI values for *S. diplocostata* in HE over LE is consistent with the fact that was the lowest biased species in the previous section for general statistic codon values (in 3.2.3.1).

3.2.4 Strength of selection.

To measure the strength of selection, odds ratio values for two-degenerated amino acids were calculated using the CDS data (whole set and screened) as described in 2.2.5. For this same purpose, the previous data were used to calculate the statistic $\hat{S}$, also explained in the same section, but as both parameters are correlated and the last one describes selection more accurately (due to its formula, specified in 2.2.5), this will be the focus of descriptions. Expression level based results, as showed more accurate than CodonW predictions in previous analysis, are summarised for each choanoflagellate species in the tables below (from Table 43 to 62).

Table 43. Strength of selection values in *A. spectabilis* sets of coding genes. Odds ratio (OR) and Statistic $\hat{S}$ (St. $\hat{S}$) values are shown for the original CDS (CDS O) and post screening (CDS S). The values for the optimal codon for each of the two-degenerated amino acids are shown, considering only the results obtained for the expression based analysis. The sum and average values for the statistics are showed at the bottom of the columns.

<i>A. spectabilis</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	1.075	0.068	UUC	2.379	0.844
Tyr	UAC	4.210	0.939	UAC	18.557	2.039
His	CAC	5.734	0.982	CAC	27.311	1.567
Gln	CAG	1.360	0.286	CAG	3.634	1.100
Asn	AAC	2.578	0.878	AAC	10.179	2.016
Lys	AAG	2.274	1.136	AAG	13.400	4.546
Asp	GAC	1.911	0.978	GAC	4.473	2.117
Glu	GAG	1.423	0.571	GAG	3.955	2.165
Cys	UGC	3.416	0.581	UGC	9.518	0.885
Sum		23.979	6.417		93.405	17.279
Aver		2.664	0.713		10.378	1.920

A. spectabilis (as shown in Table 43) presented an average for statistic $\hat{S}$ of 0.71 pre-screening, which increased 1.7 times post screening (equal to 1.92). The codons with strongest bias pre-screening were AAG (Lys), CAC (His) and GAC (Asp), while post screening were AAG, GAG (Glu) and GAC. The less biased codon for both dataset were UUC (Phe). In this species, the increment of the strength of selection post-screening is consistent with the increased values seen for the statistics GC3, CBI, Fop and the reduction of Nc, while the overall high strength correlates with the higher mentioned values over the general and LE dataset, showing evidence for natural selection.

Table 44. Strength of selection values in *C. perplexa* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>C. perplexa</i>						
	CDS O			CDS S		
AA	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.621	-0.421	UUC	0.823	-0.191
Tyr	UAC	2.629	0.629	UAC	5.017	1.064
His	CAC	1.154	0.083	CAC	2.277	0.395
Gln	CAA	2.586	1.072	CAA	3.160	1.051
Asn	AAC	5.895	1.667	AAC	15.940	2.487
Lys	AAG	3.950	1.786	AAG	7.924	3.424
Asp	GAC	1.136	0.198	GAC	1.534	0.638
Glu	GAG	1.533	0.635	GAG	2.274	1.270
Cys	UGC	2.117	0.361	UGC	3.658	0.486
Sum		21.621	6.009		42.607	10.623
Aver		2.402	0.668		4.734	1.180

C. perplexa (as detailed in Table 44) showed an increase of 1.7 times in statistic $\hat{S}$ comparing pre and post-screening (from 0.67 to 1.18 of average). In both dataset the codons presenting highest strength of selection were AAC (Asn) and AAG (Lys). Also in both the weakest selected codons were UCC (Phe), followed by CAC (His) and UGC (Cys). In this species, the increment of strength of selection post-screening is compatible with previous results, and resemble the values of CBI for HE.

Table 45. Strength of selection values in *C. hollandica* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>C. hollandica</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	5.366	1.792	UUC	8.774	2.233
Tyr	UAC	8.010	1.313	UAC	13.586	1.656
His	CAC	4.404	0.948	CAC	9.016	1.206
Gln	CAA	0.165	-2.082	CAG	6.750	1.848
Asn	-	-	-	AAC	13.650	2.368
Lys	AAG	20.591	3.430	AAG	59.893	7.069
Asp	GAU	0.258	-1.895	GAC	3.262	1.566
Glu	GAG	5.788	2.441	GAG	11.111	3.492
Cys	UGC	6.521	1.097	UGC	12.568	1.044
Sum		51.103	7.044		138.609	22.483
Aver		6.388	0.881		15.401	2.498

C. hollandica (as disclosed in Table 45) received an increase of 2.8 times in average of statistic $\hat{S}$ post-screening (from 0.881 to 2.498). This screening also showed an optimal codon for Asn and changed the optimal codon CAA for CAG. For both datasets the codons showing strongest selection were AAG (Lys) and GAG (Glu), while the weakest ones were CAA (Gln) and GAU (Asp) pre-screening, and UGC (Cys) and CAC (His) post-screening. Results are consistent with the increment of bias in HE found post-screening in the statistics of 3.2.3.1. These data also are compatible with natural selection, and with the *C. hollandica* phylogenetic position, as this species was determined one of the most codon biased studied, like those included in Craspedida Clade 2.

Table 46. Strength of selection values in *D. grandis* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>D. grandis</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.904	-0.093	UUC	1.390	0.316
Tyr	UAC	3.192	0.765	UAC	5.982	1.155
His	CAC	3.282	0.698	CAC	5.522	0.833
Gln	CAA	8.136	2.001	CAA	10.578	2.193
Asn	AAC	4.865	1.782	AAC	7.090	1.918
Lys	AAG	0.622	-0.688	AAG	1.129	0.210
Asp	GAC	1.084	0.112	GAC	1.233	0.292
Glu	GAA	2.417	1.178	GAA	2.644	1.414
Cys	UGC	0.617	-0.278	UGC	1.135	0.052
Sum		25.120	5.478		36.703	8.384
Aver		2.791	0.609		4.078	0.932

D. grandis strength of selection (as shown in Table 46) increased in 1.53 fold post-screening, from 0.6 to 0.93 on average for statistic $\hat{S}$. The strongest biased codons were CAA (Gln) and AAC (Asn) in both datasets, meanwhile the weakest ones were AAG (Lys), UGC (Cys) and UUC (Phe). These results were compatible with the increment in the statistics post-screening like GC3 and Fop, although CBI presented a bigger increase in this process (from -0.002 to 0.256). The results obtained also can be linked with the data compatible with natural selection obtained in the general codon usage statistics, and with its phylogenetic position. The family Stephanoecidae was consistently presenting the lowest codon bias results for all the previous sections of coding regions analysis, although a ranking for the average strength of selection revealed this species in the seventh position of the lowest values presented across choanoflagellates.

Table 47. Strength of selection values in *D. costata* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>D. costata</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.453	-0.784	UUC	0.731	-0.315
Tyr	UAC	0.726	-0.231	UAC	1.235	0.152
His	CAC	0.580	-0.278	CAC	0.912	-0.049
Gln	CAA	5.743	1.704	CAA	5.890	1.707
Asn	AAC	0.767	-0.289	AAC	1.296	0.244
Lys	AAG	0.501	-1.160	AAG	0.767	-0.524
Asp	GAC	0.396	-1.364	GAC	0.469	-1.088
Glu	-	-	-	-	-	-
Cys	UGC	0.619	-0.270	UGC	0.990	-0.005
Sum		9.785	-2.672		12.288	0.122
Aver		1.223	-0.334		1.536	0.015

D. costata strength of selection (presented in Table 47) had an increment post-screening of 0.349, from -0.334 to 0.015 of statistic $\hat{S}$ average. This species did not present optimal codon for Glu in any case, and it presented strongest selection in the codon CAA (Gln). The weakest biased codons in both cases were UUC (Phe), AAG (Lys) and GAC (Asp). Despite the fact that this is the sibling species of *D. grandis*, the strength of selection was considerably lower (from 0.932 to 0.015), but the codons presenting highest and lowest selection was conserved between both. Although there is previous evidence of natural selection for this species (in 3.2.3.1 and 3.2.3.2), results were consistent with the fact that it presented the lowest Fop values in HE genes and in HE DR of all the species studied. Also it can be linked to its phylogenetic position in the family Stephanoecidae, which has previously shown low codon biased species in comparison with other choanoflagellates.

Table 48. Strength of selection values in *H. balthica* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>H. balthica</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.508	-0.761	UUC	0.589	-0.592
Tyr	UAC	0.925	-0.053	UAC	1.342	0.196
His	CAC	0.657	-0.239	CAC	0.740	-0.142
Gln	-	-	-	-	-	-
Asn	AAC	1.217	0.224	AAC	1.584	0.467
Lys	AAG	0.951	-0.074	AAG	2.269	1.290
Asp	-	-	-	GAC	0.260	-1.845
Glu	GAG	0.516	-1.029	GAG	0.893	-0.161
Cys	UGC	0.998	-0.001	UGC	1.787	0.206
Sum		5.772	-1.933		9.465	-0.580
Aver		0.825	-0.276		1.183	-0.073

H. balthica strength of selection increased from -0.276 to -0.073 on average for statistic $\hat{S}$ (as detailed in Table 48). It did not present optimal codon for Gln in any case, or for Asp pre-screening. The strongest selected codons were AAG (Lys), AAC (Asn), UAC (Tyr), and UGC (Cys), while the weakest ones were GAG (Glu) and UUC (Phe), plus GAC (Asp) post-screening. This species presented the lowest statistic $\hat{S}$ mean value of all the choanoflagellates, which is not consistent with the values presented in 3.2.3.1, as *H. balthica* showed similar Fop values than *A. spectabilis* and the statistics were compatible with natural selection events. Nonetheless, results were consistent with the low levels of GC3 and CBI presented in comparison with other species, and with the low level of codon bias found in its sibling *H. gracilis*.

Table 49. Strength of selection values in *H. gracilis* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>H. gracilis</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.454	-0.757	UUC	0.520	-0.674
Tyr	UAC	1.496	0.230	UAC	2.182	0.531
His	CAC	1.698	0.294	CAC	2.283	0.423
Gln	CAG	0.376	-1.006	CAG	0.720	-0.286
Asn	AAC	2.444	0.958	AAC	3.980	1.370
Lys	AAG	1.600	0.712	AAG	3.166	1.910
Asp	GAC	0.706	-0.453	GAC	0.831	-0.251
Glu	GAG	0.862	-0.236	GAG	1.519	0.640
Cys	UGC	0.671	-0.166	UGC	1.051	0.018
Sum		10.306	-0.424		16.250	3.682
Aver		1.145	-0.047		1.806	0.409

H. gracilis was the species showing the highest difference pre and post-screening, with a change from -0.047 to 0.409 on average for statistic $\hat{S}$ (as disclosed in Table 49). The strongest selected codons were AAC (Asn) and AAG (Lys) for both datasets, while the weakest ones were UUC (Phe), CAG (Gln) and GAC (Asp). The codons presenting strongest and weakest selection were conserved with its sibling species *H. balthica* (shown in Table 48). Although statistics values compatible with natural selection were found in 3.2.3.1, the low result in statistic $\hat{S}$ average in comparison with other species correlates with the low level of Fop, CBI and GC3 found in HE genes, situating both species of the genus *Hartaetosiga* among the choanoflagellates studied with the weakest codon usage bias.

Table 50. Strength of selection values in *H. nana* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>H. nana</i>						
	CDS O			CDS S		
AA	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	1.612	0.445	UUC	2.939	1.034
Tyr	UAC	12.258	1.668	UAC	32.984	2.394
His	CAC	10.150	1.430	CAC	24.648	1.588
Gln	CAG	2.683	0.883	CAG	5.911	1.539
Asn	AAC	7.125	1.782	AAC	18.160	2.356
Lys	AAG	6.299	2.519	AAG	20.887	5.441
Asp	GAC	2.581	1.478	GAC	4.693	2.226
Glu	GAG	2.882	1.638	GAG	7.421	3.128
Cys	UGC	5.635	0.876	UGC	12.480	0.987
Sum		51.225	12.720		130.123	20.693
Aver		5.692	1.413		14.458	2.299

H. nana presented (as detailed in Table 50) an increase of 1.6 times in strength of selection post-screening (from 1.4 to 2.3 of statistic $\hat{S}$ average). The codon presenting highest strength of selection was AAG (Lys), followed by GAG (Glu), UAC (Tyr) and AAC (Asn), while the codons with lowest value were UUC (Phe), CAG (Gln) and UGC (Cys). The bias of these codons presented a correlation with those in the sibling species *A. spectabilis* (shown in Table 43), for both strongest and weakest selected, although the average values in *H. nana* were higher (2.299 on average in comparison with 1.92 for *A. spectabilis*). The high values in this species were consistent with the GC3, CBI, Nc and Fop values presented in 3.2.3.1, which increment post-screening can be reflected specially in the CBI value (as it received an increment from the full set to HE of the same amount, 1.6 times).

Table 51. Strength of selection values in *M. roanoka* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>M. roanoka</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUU	0.820	-0.213	UUC	1.071	0.071
Tyr	UAC	1.387	0.198	UAC	1.837	0.432
His	-	-	-	CAC	0.931	-0.040
Gln	CAA	0.719	-0.354	CAG	0.798	-0.227
Asn	AAC	2.719	0.872	AAC	3.016	1.032
Lys	AAG	2.653	1.175	AAG	6.003	2.506
Asp	GAU	1.314	0.357	GAC	0.768	-0.374
Glu	GAG	1.253	0.314	GAG	2.157	1.149
Cys	UGC	1.741	0.262	UGC	2.675	0.435
Sum		12.606	2.612		19.256	4.986
Aver		1.576	0.326		2.140	0.554

M. roanoka strength of selection had an increase of 1.7 times post-screening, from 0.326 to 0.554 on average for statistic $\hat{S}$ (as presented in Table 51). This species did not present optimal codon for His pre-screening and some optimal codons were changed during this process, as UUU was replaced for UUC, CAA for CAG, and GAU for GAC. The codons with strongest bias were AAG (Lys) and AAC (Asn), plus GAG (Glu) post-screening, while the lowest values were found in UUU (Phe) and CAA (Gln) pre-screening, and CAG (Gln) and GAC (Asp) post-screening. The increment in the strength of selection post-screening was consistent with the one seen in the statistics GC3, CBI, Nc and Fop. The level of bias showed in Fop for HE was similar to that found in *H. nana*, but the difference between both species is reflected in the GC3 content. Although the closest species studied in this chapter is *S. infusionum* (results shown in Table 55), *M. roanoka* showed a strength of selection more similar to its next close relatives, the genus *Hartaetosiga* (Tables 48 and 49).

Table 52. Strength of selection values in *M. fluctuans* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>M. fluctuans</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	6.239	1.769	UUC	10.757	2.532
Tyr	UAC	9.505	1.455	UAC	15.709	1.895
His	-	-	-	CAC	11.157	1.444
Gln	CAG	30.919	3.652	CAG	68.757	4.215
Asn	AAC	8.389	2.032	AAC	17.309	2.780
Lys	AAG	65.775	5.563	AAG	240.257	9.926
Asp	-	-	-	-	-	-
Glu	GAG	22.778	4.886	GAG	81.481	6.515
Cys	UGC	17.679	1.367	UGC	34.414	1.360
Sum		161.284	20.724		479.840	30.668
Aver		23.041	2.961		59.980	3.834

M. fluctuans showed an increase of 1.29 times in strength of selection post-screening, from 2.96 to 3.83 in average for statistic $\hat{S}$ (as disclosed in Table 52). This species did not present optimal codon for His pre-screening, or for Asp in any case. The strongest selected codons were AAG (Lys), GAG (Glu) and CAG (Gln), while the weakest ones were UGC (Cys) and CAC (His) post-screening. The high level in this species for average strength of selection, the second highest across choanoflagellates, is consistent with the phylogenetic position of this species in Craspedida Clade 2, surrounded by other species previously detected with high codon usage bias in this chapter. The other species from Clade 2 described previously in this section, *C. hollandica* (in Table 45), presented similar results for the optimal codons showing high and low levels of selection. These results were also compatible with the relative high values of Fop, CBI and GC3, and low in Nc across choanoflagellates, and the compatibility of these values with natural selection.

Table 53. Strength of selection values in *S. dolichothecata* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. dolichothecata</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.380	-1.032	UUC	0.416	-0.899
Tyr	UAC	1.729	0.320	UAC	2.457	0.585
His	CAC	1.573	0.291	CAC	1.698	0.287
Gln	CAA	2.754	1.328	-	-	-
Asn	AAC	2.816	1.179	AAC	3.347	1.191
Lys	AAG	0.959	-0.058	AAG	1.878	0.982
Asp	GAC	0.871	-0.187	GAC	0.865	-0.197
Glu	-	-	-	GAG	0.572	-0.821
Cys	UGC	0.922	-0.044	UGC	1.492	0.163
Sum		12.004	1.797		12.725	1.292
Aver		1.501	0.225		1.591	0.162

S. dolichothecata was the only species not showing an increase of the strength of selection post-screening, being this reduced in from 0.225 to 0.162 in average for statistic $\hat{S}$ (as presented in Table 53). This species did not present optimal codon for Glu pre-screening, neither for Gln post-screening. The strongest selected codons were CAA (Gln) pre-screening, UGC (Cys) post-screening and AAC (Asn) in both, while the weakest ones were UUC (Phe), GAC (Asp) and GAG (Glu) post-screening. While the reduction in the average strength of selection post-screening is supported by the Fop value in HE and the Nc values in DR, it is not reflected in GC3, CBI and Nc values, where also was revealed consistency with natural selection. This species is the only representative of Craspedida Clade 3 studied, and has shown by the codon usage statistics to be more similar to Stephanoecida and Acanthoecidae than to the Clade 2. In this analysis, *S. dolichothecata* also revealed a more similar strength of selection to species of the family Stephanoecidae and the genus *Hartaetosiga* than to those in Craspedida Clade 2.

Table 54. Strength of selection values in *S. helianthica* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. helianthica</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.742	-0.335	UUC	0.836	-0.195
Tyr	UAC	4.562	0.962	UAC	6.994	1.362
His	CAC	8.471	1.487	CAC	16.712	1.652
Gln	CAG	4.683	1.634	CAG	9.514	2.148
Asn	AAC	6.636	1.957	AAC	12.796	2.468
Lys	AAG	15.394	3.837	AAG	55.377	7.319
Asp	-	-	-	-	-	-
Glu	GAG	14.109	3.988	GAG	32.739	5.196
Cys	UGC	3.777	0.726	UGC	9.180	0.862
Sum		58.374	14.256		144.149	20.812
Aver		7.297	1.782		18.019	2.601

S. helianthica (as detailed in Table 54) increased the strength of selection in 1.46 times post-screening, from 1.78 to 2.6 for statistic $\hat{S}$ average, and did not present optimal codon for Asp in any case. The strongest selected codons were GAG (Glu) and AAG (Lys), while the weakest were UUC (Phe), UGC (Cys) and UAC (Tyr). This species, as its relatives included in Craspedida Clade 2, showed a consistent high strength of selection, the fourth highest across choanoflagellates in this case. Previous results in this species showed the second highest Fop in HE genes, low Nc and high GC3 and CBI values, which also were consistent with natural selection. The closest related species, *M. fluctuans*, also showed the second highest strength of selection across choanoflagellates (shown in Table 52), and presented similar codons with the highest selection.

Table 55. Strength of selection values in *S. infusionum* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. infusionum</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.465	-1.029	UUC	0.543	-0.638
Tyr	UAC	2.556	0.600	UAC	7.796	1.438
His	CAC	2.625	0.629	CAC	2.788	0.492
Gln	CAG	0.451	-0.900	CAG	0.960	-0.041
Asn	AAC	3.116	1.060	AAC	23.681	2.766
Lys	AAG	10.315	2.624	AAG	16.311	4.144
Asp	GAU	0.236	-1.902	GAC	1.258	0.330
Glu	GAG	1.498	0.564	GAG	3.344	1.871
Cys	UGC	0.563	-0.267	UGC	3.347	0.520
Sum		21.825	1.378		60.030	10.883
Aver		2.425	0.153		6.670	1.209

S. infusionum presented an increment from 0.15 to 1.22 in statistic $\hat{S}$ average, over 8 times higher post-screening (as shown in Table 55). During this process the optimal codon for Asp changed from GAU to GAC. The strongest selected codons were AAG (Lys) and AAC (Asn), while the weakest ones were GAU/GAC, UUC (Phe) and CAG (Gln). This species showed higher strength of selection than the closest related species studied in this chapter, *M. roanoka* (Table 51) and the *Hartaetosiga* genus (Tables 48 and 49). Both the increment of strength of selection post-screening and the higher results than in its sibling species were supported by the GC3, CBI, Nc and Fop values, and reflected in the Nc-GC3 plot. *S. infusionum* showed an average statistic $\hat{S}$ closer to the other member of Craspedida Clade 1, *C. perplexa* (Table 44), than to any other species, which also showed resemblance in the strongest and weakest selected codons.

Table 56. Strength of selection values in *S. kvevrii* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. kvevrii</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.541	-0.570	UUC	0.563	-0.551
Tyr	UAC	8.997	1.213	UAC	27.731	2.089
His	CAC	4.270	0.999	CAC	10.737	1.371
Gln	CAG	7.127	1.973	CAG	23.315	2.994
Asn	AAC	6.350	1.627	AAC	15.449	2.330
Lys	AAG	8.704	2.704	AAG	62.640	6.565
Asp	GAC	2.558	1.477	GAC	4.079	1.996
Glu	GAG	7.757	3.202	GAG	33.927	5.627
Cys	UGC	11.250	1.360	UGC	34.016	1.517
Sum		57.554	13.985		212.455	23.940
Aver		6.395	1.554		23.606	2.660

S. kvevrii increased 1.7 times the strength of selection post-screening, from 1.55 to 2.66 of statistic $\hat{S}$ average (as presented in Table 56). The strongest selected codons were GAG (Glu), AAG (Lys) and CAG (Gln), while the weakest ones were UUC (Phe) and CAC (His). This species was the third one showing the highest values for strength of selection, showing a similar level to *S. helianthica* (Table 54), which although it is not its closest related species, it is a close relative within Craspedida Clade 2. Besides its taxonomic position and the previous results showed in related species, these values were consistent with the high values showed in GC3, CBI and Fop, and the low one for the Nc parameter, which also were consistent with natural selection events.

Table 57. Strength of selection values in *S. macrocollata* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. macrocollata</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	-	-	-	UUC	0.720	-0.349
Tyr	UAC	10.297	1.296	UAC	16.921	1.702
His	CAC	9.272	1.834	CAC	16.543	1.647
Gln	-	-	-	CAG	12.185	2.505
Asn	AAC	7.521	1.949	AAC	18.494	2.535
Lys	AAG	4.778	2.531	AAG	18.000	5.464
Asp	GAC	5.485	2.407	-	-	-
Glu	-	-	-	GAG	7.064	3.134
Cys	UGU	0.198	-1.006	UGC	9.431	0.862
Sum		37.552	9.011		99.358	17.499
Aver		6.259	1.502		12.420	2.187

S. macrocollata (detailed in Table 57) showed an increase for statistic $\hat{S}$ average of 1.45 times post-screening, from 1.5 to 2.187. This species did not presented optimal codons for Phe, Gln and Glu pre-screening, but after this process only Asp showed no result and the optimal codon for Cys changed from UGU to UGC. The strongest selected codons were AAG (Lys) and GAC (Asp) pre-screening, while post-screening also GAG (Glu) presented one of the highest values. The weakest selected codons were mainly UGU/UGC (Cys for pre and post-screening respectively) with addition of UUC (Phe) post-screening. As the rest of species of Craspedida Clade 2, high results on strength of selection were consistent with high GC3, CBI, Fop and low Nc, showing also data consistent with natural selection. *S. macrocollata* did not show a similar bias level than its sibling *S. punica* (Table 58), instead it showed closer values to the close related species *S. urceolata* (Table 59), while the described sibling of *S. urceolata*, *S. kvevrii* (Table 56), was more similar to *S. helianthica* (Table 54). These similarities found within Craspedida Clade 2 could be a product of some species retaining ancestral traits, or a similar efficiency in selection due to similarities in the population size.

Table 58. Strength of selection values in *S. punica* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. punica</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	2.649	0.898	UUC	3.191	1.106
Tyr	UAC	21.530	1.777	UAC	47.100	2.417
His	CAC	12.009	2.038	CAC	32.367	2.153
Gln	CAG	28.933	4.390	CAG	73.128	4.304
Asn	AAC	35.227	3.509	AAC	94.688	4.115
Lys	AAG	29.935	4.515	AAG	140.577	8.191
Asp	-	-	-	GAC	29.659	4.401
Glu	GAG	28.337	5.228	GAG	88.286	6.612
Cys	UGC	28.632	1.668	UGC	84.347	2.051
Sum		187.253	24.024		593.343	35.351
Aver		23.407	3.003		65.927	3.928

S. punica increased 1.3 times the statistic $\hat{S}$ average post-screening, from 3 to 3.93, and did not present optimal codon for Asp pre-screening (as disclosed in the results of Table 58). The strongest selected codons were GAG (Glu), AAG (Lys), CAG (Gln) and AAC (Asn), with addition to GAC (Asp) post-screening. The weakest selected codons were UUC (Phe), UAC (Tyr), UGC (Cys) and CAC post-screening. This species presented the highest average statistic $\hat{S}$ across choanoflagellates, which is consistent with its taxonomic position within Craspedida Clade 2, presenting the lowest Nc values in HE, the highest CBI in HE, high values for GC3 and Fop, and consistent data for natural selection. Although the sibling species of *S. punica* is *S. macrocollata*, it presented a more similar strength of selection to its close relative *M. fluctuans* (shown in Table 52), a highly biased species showing the highest Fop values in HE genes, which also shared similar codons with the highest and lowest values.

Table 59. Strength of selection values in *S. urceolata* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. urceolata</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUU	5.769	1.650	UUC	0.106	-2.185
Tyr	UAC	20.383	1.591	UAC	45.511	2.428
His	CAC	9.750	1.774	CAC	15.079	1.789
Gln	CAG	2.332	0.983	CAG	3.819	1.242
Asn	AAC	25.072	2.786	AAC	73.788	3.617
Lys	AAG	4.848	2.001	AAG	21.329	4.907
Asp	GAC	8.533	3.158	GAC	11.008	3.494
Glu	GAG	2.857	1.455	GAG	4.219	2.120
Cys	UGC	4.069	0.843	UGC	10.518	1.017
Sum		83.613	16.239		185.377	18.430
Aver		9.290	1.804		20.597	2.048

S. urceolata was among the species with the lowest increment in statistic $\hat{S}$ average post-screening, only 13% from 1.804 to 2.048 (results shown in Table 59). This species changed the optimal codon for Phe after the screening process, from UUU to UUC. The strongest selected codons were GAC (Asp), AAC (Asn) and AAG (Lys), while the weakest ones were UUC (Phe) post-screening, CAG (Gln) and UGC (Cys). This species, despite being the sibling of *S. kvevrii*, showed lower strength of selection levels, more similar to *S. macrocollata* (Table 57), which shared some of the strongly and weakest selected codons. Despite this fact, *S. urceolata* is one of the species within the highly biased Craspedida Clade 2 and showing high strength of selection among choanoflagellates. These results were consistent with GC3, CBI, Nc and Fop values, which also were compatible with natural selection.

Table 60. Strength of selection values in *S. parva* sets of coding genes. The formatting is presented in a similar style as Table 43.

<i>S. parva</i>						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUU	2.312	0.763	UUC	0.334	-0.980
Tyr	UAC	8.256	1.465	-	-	-
His	CAC	3.763	0.922	CAC	6.298	1.038
Gln	CAA	0.458	-0.744	CAG	1.850	0.535
Asn	AAC	6.022	1.704	AAC	9.630	1.795
Lys	AAG	5.339	2.097	AAG	19.702	4.957
Asp	GAC	2.290	1.308	GAC	3.367	1.655
Glu	GAG	1.350	0.425	GAG	2.047	1.065
Cys	UGC	2.687	0.543	UGC	5.732	0.638
Sum		32.477	8.484		48.959	10.704
Aver		3.609	0.943		6.120	1.338

S. parva exhibited an increment of 1.42 times the strength of selection average post-screening, from 0.943 to 1.338 (as presented in Table 60). After this process no optimal codon for Tyr was found, and the optimal codons changed for Phe (from UUU to UUC) and Gln (from CAA to CAG). The strongest selected codons were AAG (Lys), AAC (Asn) and UAC (Tyr) pre-screening, while weakest ones were CAA (Gln) and UUU/UUC (Phe), GAG (Glu) pre-screening and UGC (Cys). This species, as it is included in the family Acanthoecidae, is close related to *A. spectabilis* (Table 43) and *H. nana* (Table 50), although these results were more similar to the first species than to *H. nana*, which presented higher values. This proportion also correlates with the values obtained for GC3, CBI, Nc and Fop, although *S. parva* also showed data consistent with natural selection.

Table 61. Strength of selection values in *S. diplocostata* (Australian isolate) coding genes. The formatting is presented in a similar style as Table 43.

<i>S. diplocostata</i> (Au)			
AA	CDS O		
	Codon	OR	St. $\hat{S}$
Phe	UUC	1.023	0.024
Tyr	UAC	2.854	0.855
His	CAC	2.967	0.748
Gln	CAG	1.328	0.325
Asn	AAC	2.274	0.954
Lys	AAG	1.571	0.653
Asp	-	-	-
Glu	-	-	-
Cys	UGC	1.854	0.418
Sum		13.870	3.978
Aver		1.981	0.568

S. diplocostata Australia isolate showed a strength of selection of 0.57 in statistic $\hat{S}$ average (as detailed in Table 61). For this isolate no optimal codons for Asp and Glu were detected. The strongest selected codons were AAC (Asn), UAC (Tyr) and CAC (His), while the weakest ones were UUC (Phe), CAG (Gln) and UGC (Cys). The strength of selection in this species was more similar to the France isolate than to any other species, and as both were deemed equivalent in protein content (as it was mentioned in 3.1), this species will be further discussed below with the other isolate (presented in Table 62).

Table 62. Strength of selection values in *S. diplocostata* (France isolate) coding genes. The formatting is presented in a similar style as Table 43.

<i>S. diplocostata</i> (Fr)						
AA	CDS O			CDS S		
	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	1.004	0.004	UUC	1.514	0.406
Tyr	UAC	2.750	0.750	UAC	5.232	1.248
His	CAC	2.913	0.671	CAC	5.109	0.824
Gln	CAG	1.253	0.239	-	-	-
Asn	AAC	2.131	0.803	AAC	3.551	1.270
Lys	AAG	1.410	0.447	AAG	2.914	1.801
Asp	GAC	1.344	0.483	-	-	-
Glu	-	-	-	GAA	1.103	0.156
Cys	UGC	1.644	0.304	UGC	3.037	0.529
Sum		14.449	3.702		22.460	6.235
Aver		1.806	0.463		3.209	0.891

S. diplocostata France isolated (shown in Table 62) almost doubled the strength of selection post-screening (90% increase from 0.46 to 0.89 in statistic $\hat{S}$ average), and while this process revealed an optimal codon for Glu, it also lost the significance for Gln and Asp codons. Although in this dataset Glu was not detected pre-screening, and Asp post-screening, neither were detected in the Australian isolated (Table 61). The strongest selected codons were AAG (Lys) post-screening, AAC (Asn), UAC (Tyr) and CAC (His), while the weakest ones were UUC (Phe), CAG (Gln) pre-screening, and UGC (Cys), both sets identical to those identified in the Australian isolate. *S. diplocostata* is included in Stephanoecidae, a family previously shown containing species with some of the lowest codon bias results, although this species gave a strength of selection close to *D. grandis* (Table 46) and higher than *D. costata* (Table 47).

Looking at the statistical values for the codon analysis, *S. diplocostata* changed from the sixth place in lowest value to share the last position with *D. grandis* in lowest Fop, CBI and highest Nc in HE genes. Although this species did not had the lowest GC3 content, it showed the lowest increment from the whole dataset to HE and the lowest GC3 content in optimal codons in choanoflagellates. Despite these data, and an Nc value for NDR lower than DR in HE, the parameters studied still showed evidences compatible with natural selection based on translational accuracy and efficiency, like the increment of Fop in HE over LE and in DR over NDR.

As a summary, the strength of selection values presented in this section supported the previous evidence for natural selection upon translational accuracy and efficiency. These data were in accord with the GC3 content in optimal codons, the parameters for GC3, CBI, Nc and Fop in codons, with the GC3-Nc plots and the domains and non-domains analysis. The sum of this information has let us compare the strength of selection with additional context and over the phylogenetic tree, which also has been shown important to understand the variations in this selection.

3.3 tRNA ANALYSIS.

3.3.1 Indirect evidence for tRNA deamination through selective pressure.

In choanoflagellates the evidence for the process of deamination was searched indirectly through the TAPSILVR amino acids in codons ended in cytosine. The overrepresentation of deaminated codons of this set of amino acids (threonine, alanine, proline, serine, isoleucine, leucine, valine and arginine), was searched in protein domains regions over non-domains ones. This phenomena would appear as a consequence of a selective advantage on using deaminated codons for the most abundant tRNA molecules, and was calculated comparing domains and non domains regions data for the predicted proteins in the 100 genes with highest and lowest relative expression (with a contingency square as described in 2.3.1, and analysing the highest and lowest expressed genes in separate analysis). The set of genes were chosen from the screened CDS for high reliability in eukaryotic origin, and the process was performed as explained in 2.3.1. Results for each choanoflagellate TAPSILVR amino acid are showed next (in Tables 63 to 66), comparing those codons overrepresented in highest and lowest expressed genes.

Table 63. Overrepresented cytosine ended codons in protein domain regions of *A. spectabilis*, *C. perplexa*, *C. hollandica*, *D. grandis* and *D. costata*. Codons for TAPSVLR amino acids in domains of highest expressed genes (HE) and lowest expressed genes (LE) are compared.

Overrepresented codons in protein domain regions of choanoflagellates (PART 1)										
	<i>A. spectabilis</i>		<i>C. perplexa</i>		<i>C. hollandica</i>		<i>D. grandis</i>		<i>D. costata</i>	
aa	HE	LE	HE	LE	HE	LE	HE	LE	HE	LE
Leu	-	-	CUC	-	CUC	-	CUC	-	-	-
Ile	-	-	-	-	-	-	-	-	-	-
Val	-	-	-	-	-	-	-	-	GUC	-
Ser	-	-	-	-	-	-	UCC	-	-	-
Pro	-	-	-	-	CCC	-	-	-	-	-
Thr	-	-	ACC	-	ACC	ACC	ACC	-	-	-
Ala	GCC	-	GCC	-	GCC	-	GCC	-	GCC	-
Arg	-	-	CGC	-	CGC	-	-	-	-	-

Table 64. Overrepresented cytosine ended codons in protein domain regions of *H. balthica*, *H. gracilis*, *H. nana*, *M. roanoka* and *M. fluctuans*. The formatting is similar to Table 63.

Overrepresented codons in protein domain regions of choanoflagellates (PART 2)										
	<i>H. balthica</i>		<i>H. gracilis</i>		<i>H. nana</i>		<i>M. roanoka</i>		<i>M. fluctuans</i>	
aa	HE	LE	HE	LE	HE	LE	HE	LE	HE	LE
Leu	CUC	-	-	-	-	-	CUC	-	-	-
Ile	-	-	-	-	-	-	AUC	-	-	-
Val	-	-	-	-	-	-	GUC	-	GUC	-
Ser	-	-	UCC	-	-	-	UCC	-	UCC	-
Pro	-	-	-	-	-	-	CCC	-	CCC	-
Thr	-	-	ACC	-	-	ACC	ACC	-	ACC	-
Ala	-	-	GCC	-	-	-	GCC	-	GCC	-
Arg	-	-	-	-	-	-	CGC	-	CGC	-

Table 65. Overrepresented cytosine ended codons in protein domain regions of *S. dolichothecata*, *S. helianthica*, *S. infusionum*, *S. kvevrii* and *S. macrocollata*. The formatting is similar to Table 63.

Overrepresented codons in protein domain regions of choanoflagellates (PART 3)										
	<i>S. dolichothecata</i>		<i>S. helianthica</i>		<i>S. infusionum</i>		<i>S. kvevrii</i>		<i>S. macrocollata</i>	
aa	HE	LE	HE	LE	HE	LE	HE	LE	HE	LE
Leu	-	-	-	-	CUC	-	CUC	CUC	-	-
Ile	-	-	-	-	AUC	-	-	-	-	-
Val	GUC	-	GUC	GUC	GUC	-	GUC	-	-	-
Ser	UCC	-	-	-	UCC	-	-	-	-	-
Pro	CCC	-	CCC	-	-	-	CCC	CCC	CCC	-
Thr	ACC	-	-	-	-	-	-	ACC	-	-
Ala	GCC	-	GCC	-	GCC	GCC	GCC	GCC	GCC	-
Arg	-	-	CGC	-	-	-	CGC	CGC	CGC	-

Table 66. Overrepresented cytosine ended codons in protein domain regions of *S. punica*, *S. urceolata*, *S. parva* and *S. diplocostata* (France isolated). The formatting is similar to Table 63.

Overrepresented codons in protein domain regions of choanoflagellates (PART 4)								
	<i>S. punica</i>		<i>S. urceolata</i>		<i>S. parva</i>		<i>S. diplocostata</i> (Fr)	
aa	HE	LE	HE	LE	HE	LE	HE	LE
Leu	CUC	-	CUC	CUC	-	-	-	CUC
Ile	-	-	-	-	-	-	-	-
Val	GUC	GUC	GUC	-	GUC	-	-	-
Ser	UCC	-	-	-	-	-	-	-
Pro	CCC	-	CCC	-	-	-	-	-
Thr	ACC	-	ACC	-	-	-	-	-
Ala	GCC	GCC	GCC	GCC	GCC	-	GCC	GCC
Arg	CGC	CGC	CGC	CGC	-	-	-	CGC

The analysis of C ending codons in TAPSLVR amino acids, as an effect of selective advantage (shown in Tables 63, 64, 65 and 66), presented 76 overrepresented codons, 18 in lowest expressed genes and 72 in highly expressed genes (as 14 were coincident between both sets). Thus, this codons were 4 times more common in HE than in LE genes. The amino acid with more overrepresented codons was Ala (22 codons), and the less prevalent was Ile (2 codons, in *M. roanoka* and *S. infusionum* HE), the rest were in a range from 7 to 13 codons. Although 11 species did not show any overrepresented codon in LE, only *H. nana* did not show any result for HE genes, in opposition to *M. roanoka*, the only species showing overrepresented codons for all the amino acids. Species with 7 overrepresented codons in HE was only *S. punica*, and with 6 codons *M. fluctuans* and *S. urceolata*. Species showing a single overrepresented codon in HE were *A. spectabilis*, *H. balthica* and *S. diplocostata*, and species showing two results were *D. costata* and *S. parva*.

The amount of overrepresented codons obtained in the evaluation of selective pressure did not correlate with the codon usage bias found in previous sections of this chapter. Although some of the less biased species presented a low number of these codons and *vice versa*, the distribution of the results mismatch any resemblance with the phylogenetic tree or an organisation based on previous parameters like Fop in HE. The analysis was neither limited nor correlated with the presence of C ending optimal codons in choanoflagellates, as the species *C. perplexa*, *C. hollandica*, *M. roanoka*, *M. fluctuans*, *S. dolichothecata*, *S. helianthica*, *S. punica* and *S. urceolata* presented an optimal codon or one of the optimal codons ended in C per amino acid studied. In addition, the species *A. spectabilis*, *D. costata*, *H. balthica*, *H. gracilis*, *H. nana* and *S. infusionum* did not present a C ended optimal codons only for Pro. The species *D. grandis* lacked these codons for Pro and Arg, *S. macrocollata* for Leu and Thr, *S. parva* for Ser, Pro and Thr, and *S. diplocostata* for Pro and Thr. Nonetheless, the optimal codon determined did not conflict with this analysis, as there was not overrepresented codons in amino acids not presenting optimal codon ended in C.

The overrepresented C ending optimal codons showed in this analysis provide an advantage to be selected in domains over non domains, which is compatible with natural selection upon translational accuracy. The selective advantage of using these specific codons is compatible with an advantage of using deaminated tRNA for inosine. This indirect evidence of deamination, although found in some extent in all the choanoflagellates studied, only offers a first insight of the impact of deamination in these species.

CHAPTER 4. CODON USAGE AND tRNA RESULTS IN UNICELLULAR EUKARYOTES AND METAZOAN SPECIES

In this chapter the results obtained for the unicellular eukaryotes and metazoan species of this study will be presented. First the genomic data for the coding genes used in analysis for both sets of organisms will be described (4.1). The next subsection will address the results obtained for codon usage bias analysis (4.2), including the obtaining of relative expression (4.2.1), the determination of optimal codons (4.2.2), codon usage statistics values (4.2.3), and an evaluation of the strength of selection (4.2.4). The last part of the chapter will be dedicated to results relative to tRNA analysis, gathering tRNA deamination through selective pressure (4.3.1) and through transcriptome screening (4.3.2), differential anticodon abundance analysis (4.3.3), and screening for chimeric tRNA (4.3.4).

4.1 GENOMIC AND TRANSCRIPTOMIC DATA FROM UNICELLULAR AND METAZOAN SPECIES.

Coding sequences (CDS) for unicellular and metazoan species were downloaded from EnsemblProtists and EnsemblMetazoa as explained in 2.1.2. A copy of these files also can be found in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>. Relevant data for the genome and transcriptome of these species are showed below in Tables 67 (for unicellular organisms) and 68 (for metazoan).

Table 67. Genomic and transcriptomic data for unicellular eukaryotes species. Number of genome scaffold, genome size in Mb and coding regions (CDS) obtained from EnsemblProtist. The last column shows the size in Mb for the RNA sequencing files listed in Table 02 (in 2.1.1) in respective order.

Unicellular species	Genome scaffold	Genome size (Mb)	CDS	RNA seq. (Mb)
<i>Salpingoeca rosetta</i>	154	55.44	11731	859.1
<i>Sphaeroforma arctica</i>	15619	121.63	18730	1100, 1100, 997.2
<i>Thecamonas trahens</i>	131	28.68	10626	2000

Table 68. Genomic and transcriptomic data for metazoan species. Number of genome scaffold, genome size in Mb and coding regions (CDS) obtained from EnsemblMetazoa. The last column shows the size in Mb for the RNA sequencing files listed in Table 03 (in 2.1.1) in respective order.

Metazoan species	Genome scaffold	Genome size (Mb)	CDS	RNA seq. (Mb)
<i>Amphimedon queenslandica</i>	13397	166.68	43614	242.9, 241.1
<i>Mnemiopsis leidyi</i>	5101	155.87	16058	191
<i>Nematostella vectensis</i>	10804	356.61	24780	461.8
<i>Thelohanellus kitauei</i>	5757	150.35	15020	5000
<i>Trichoplax adhaerens</i>	1414	105.63	11520	19500

4.2 CODING REGION ANALYSIS.

4.2.1 Mapping of reads for relative expression.

Mapping of reads was also performed in the unicellular and metazoan species of this study, in a similar way that was performed for the choanoflagellates in 3.2.1 and as explained in 2.2.1. This was done mapping the reads from RNA sequencing over the coding regions of the organisms, to calculate relative expression as number of reads divided by gene length. Values obtained from this process were used to sort CDS by their relative expression to determine the 5% of higher and lower expressed sequences. Files containing relative expression are available in the repository <https://doi.org/10.6084/m9.figshare.19361888.v1>.

4.2.2 Determination of optimal codons.

Optimal codons for unicellular and metazoan species (which procedure was described in 2.1.1 and the data were specified in 4.1) were determined using two methodologies, optimal codons predicted by CodonW analysis, and results based on a comparison of the 5% highest and lowest expressed genes (as explained in 2.2.2). Results are gathered in the tables below for both methodologies (Tables 69 to 71), comparing at the same time the results for the unicellular species (*S. rosetta*, *S. arctica* and *T. trahens*), the poriferan species (*A. queenslandica*) with the ctenophoran (*M. leidy*), and the cnidarians (*N. vectensis* and *T. kitauei*) with placozoa (*T. adhaerens*). This distribution, in addition to attend to the organisation by complexity and alphabetical order that has been presented until now, allows a comparison of the species with a debate on their taxonomic location (as was presented in 1.2.4).

Optimal codons predictions using CodonW resulted unreliable for *S. rosetta* (showing 7 false positives and 5 false negatives) and *S. arctica* (with 12 false positives and 10 false negatives). Nonetheless, results in *T. trahens* were closer to the expression levels results with no false positives and only 4 false negatives. Optimal codons for these species are detailed in Table 69.

The expression levels revealed 22 optimal codons in *S. rosetta*, 17 ended in GC. This species is included in the Craspedida Clade 1 taxon for choanoflagellates and is the closest sibling to *M. roanoka*, closest that *S. infusionum*, which was compared with this species in the previous chapter. *S. rosetta* shared with its closest relative 20 codons, all except AUU (Ile) and GGC (Gly), while it shared with *S. infusionum* 19 codons, all except AUU (Ile), CCU and CCC (Pro). Although the optimal codons are highly conserved among these species, *S. rosetta* presented the highest percentage of GC3, as the 69% of *M. roanoka* and 70% of *S. infusionum* increases to 77% in this choanoflagellate.

The organism *S. arctica* presented 25 optimal codons, with 14 of them ended in GC (56%), and *T. trahens* showed 23 optimal codons, which included 21 ended in GC (91.3%). The three species of Table 69 exhibited a correlation in the optimal codons of their dataset, sharing at least one optimal codon per amino acid, with only three exceptions, Ile in *S. rosetta*, and Leu and Asp in *S. arctica*. These results were also consistent with the most used optimal codons in the choanoflagellates species of Chapter 3, presenting in each amino acid the most frequent optimal codon except Ile in

S. rosetta, and Leu and Cys in *S. arctica*. The fact that there is a clear preference for optimal codons across choanoflagellates, and that this is compatible with the ichthyosporean and apusomonadean species, sharing *S. rosetta* 18 optimal codons with *S. arctica* and 19 with *T. trahens*, is consistent with this trait being conserved from an ancestor early in the Opisthokonta evolution.

Table 69. Comparison of optimal codons determined for the unicellular species *S. rosetta*, *S. arctica* and *T. trahens*. Results were based on both methodologies, correspondence analysis in CodonW and expression level (Exp. Level). Codons in bold highlight the difference between methodologies.

Amino Acid	<i>S. rosetta</i>		<i>S. arctica</i>		<i>T. trahens</i>	
	CodonW	Exp. level	CodonW	Exp. level	CodonW	Exp. level
Phe	-	UUC	UUC	UUC	-	UUC
Leu	CUC, CUG	CUC	CUC, CUG	CUU	CUC	CUC
Ile	AUU, AUC	AUU	AUC	AUU , AUC	AUC	AUC
Val	GUC, GUG	GUC	GUC, GUG	GUU , GUC	GUC	GUC
Ser	UCC, UCG , AGC	UCU , UCC	UCC, UCG , AGC	UCU , UCC	UCC, UCG	UCC, UCG
Pro	CCC, CCG	CCU , CCC	CCC, CCG	CCU , CCC	CCC, CCG	CCC, CCG
Thr	ACC, ACG	ACC	ACC, ACG	ACU , ACC	ACC	ACC, ACG
Ala	GCC	GCC	GCC, GCG	GCU , GCC	GCC	GCC
Tyr	UAC	UAC	UAC	UAC	UAC	UAC
His	CAC	CAC	CAC	CAC	CAC	CAC
Gln	CAG	CAG	CAG	CAG	CAG	CAG
Asn	AAC	AAC	AAC	AAC	AAC	AAC
Lys	AAG	AAG	AAG	AAG	AAG	AAG
Asp	GAC	GAC	GAC	GAU	GAC	GAC
Glu	GAG	GAG	GAG	GAG	GAG	GAG
Cys	UGC	UGC	UGC	-	UGC	UGC
Arg	CGC	CGU , CGC	CGU, CGC, CGG	CGU, CGC	CGC	CGU , CGC
Gly	GGC	GGU , GGC	GGC	GGU , GGA	GGC	GGU , GGC

Table 70. Comparison of optimal codons determined for the metazoan species *A. queenslandica* and *M. leidy*. The formatting is presented in a similar style as Table 69.

<i>A. queenslandica</i>			<i>M. leidy</i>	
Amino Acid	CodonW	Expression level	CodonW	Expression level
Phe	UUU	UUC	UUU	UUU
Leu	UUA, CUA	UUG, CUC, CUG	UUA, UUG, CUU, CUA	UUA
Ile	AUU, AUA	AUC, AUA	AUU, AUA	AUU, AUC
Val	GUU, GUA	GUC, GUG	GUU, GUA	GUU, GUC
Ser	UCA, AGU	UCU, UCC, UCG, AGC	UCU, UCA, AGU	UCU
Pro	CCU, CCA	CCC, CCG	CCU, CCA	CCC
Thr	ACU, ACA	ACC, ACG	ACU, ACA	ACC
Ala	GCU, GCA	GCC, GCG	GCU, GCA	GCU, GCC
Tyr	UAU	UAC	UAU	-
His	CAU	CAC	CAU	CAU
Gln	CAA	CAG	CAA	-
Asn	AAU	AAC	AAU	AAU
Lys	AAA	AAG	AAA	-
Asp	GAU	GAC	GAU	-
Glu	GAA	GAG	GAA	GAA
Cys	UGU	UGC	UGU	UGC
Arg	AGA	CGC, CGG,	CGU, AGA	CGU, CGA
Gly	GGU, GGA	AGG, GGC, GGG	GGU, GGA	GGU, GGC, GGA

Unlike the unicellular species seen in Table 69, in the metazoan detailed in Table 70 CodonW generated a high number of false positive a negative results that made the prediction not reliable, although it revealed a considerable amount of forces like random mutations and genetic drift influencing the genome of these species.

The poriferan *A. queenslandica* showed 31 optimal codons, with 29 ended in GC, while the ctenophoran *M. leidy* showed 20 optimal codons, with only 7 ended in GC and no result for Tyr, Gln, Lys or Asp. Both species were very different between them, with only 8 codons in common, in contrast with similarities found previously among unicellular organisms.

Table 71. Comparison of optimal codons determined for the metazoan species *N. vectensis*, *T. kitauei* and *T. adhaerens*. The formatting is presented in a similar style as Table 69.

Amino Acid	<i>N. vectensis</i>		<i>T. kitauei</i>		<i>T. adhaerens</i>	
	CodonW	Exp. level	CodonW	Exp. level	CodonW	Exp. level
Phe	UUU	UUU	UUU	UUC	UUU	UUC
Leu	UUA, CUU , CUC , CUA	UUA, CUA	UUA , UUG, CUU, CUA	UUG, CUU, CUC	UUA , CUU, CUA	UUG , CUU, CUC
Ile	AUU	AUU, AUA	AUU , AUA	AUC	AUA	AUC
Val	GUU	GUU, GUA	GUU , GUA	GUC , GUG	GUU , GUA	GUC
Ser	UCU, AGU	UCU	UCU, UCA , AGU	UCU, UCC	UCU, UCA	UCU, UCC , AGC
Pro	CCU	CCC , CCA	CCU, CCA	CCU, CCC	CCU, CCA	CCU, CCC
Thr	ACU, ACA	ACU	ACU, ACA	ACU, ACC	ACU, ACA	ACU, ACC
Ala	GCU, GCA	GCU, GCA	GCU, GCA	GCU	GCA	GCU , GCC
Tyr	UAU	UAU	UAU	UAC	UAU	UAC
His	CAU	-	CAU	CAC	CAU	CAC
Gln	CAA	CAA	CAA	CAG	CAG	CAG
Asn	AAU	AAU	AAU	AAC	AAU	AAC
Lys	AAA	AAA	AAA	AAG	AAA	AAG
Asp	GAU	GAU	GAU	GAC	GAU	GAC
Glu	GAA	GAA	GAA	GAG	-	GAA
Cys	UGU	-	UGU	-	UGU	UGC
Arg	CGU , CGA , AGA	AGA	CGA , AGA	CGU , AGG	AGA , AGG	CGU , CGC
Gly	GGU , GGA	GGA	GGU, GGA	GGU	GGA , GGG	GGU

Comparing Table 70 with previous results, *A. queenslandica* showed closer results to unicellular organisms, both on high GC3 content (93.5%) and presenting one of the choanoflagellates most prevalent optimal codon in 16 of its amino acids. *M. leidy* in comparison with choanoflagellates, contained the most frequent optimal codons for 8 of its amino acids and reduced its GC3 content to 30%. As the poriferan species resulted more similar in optimal codons to unicellular organisms than ctenophoran ones, it suggest that either it is the closest view on the ancestral state, or that the earliest branching poriferan allowed them to retain the ancestral traits.

In agreement with the results presented previously for the metazoan species of Table 70, CodonW optimal codons predictions resulted completely unreliable in the metazoan of Table 71. Although *N. vectensis* showed best predictions that other species in the table, it still detected 10 false negatives. Also, as in the previous metazoan species, this analysis revealed the influence of genetic drift and random mutations shaping the genome of these organisms.

The cnidarian *N. vectensis* showed 21 optimal codons, with only 1 ended in GC (4.76%), while the other cnidarian, *T. kitauei*, showed 24 optimal codons with 17 ended in GC (70.8%). Neither of the species showed result for Cys, but only *N. vectensis* did not show result for His. The representative of placozoan *T. adhaerens* (shown also in Table 71) presented 26 optimal codons, with 18 GC3 (69.2%). The species *N. vectensis* shared with *T. kitauei* only 4 optimal codons, while it shared with *M. leidy* 10 optimal codons. *T. kitauei* instead shared 21 optimal codons with *T. adhaerens* and 17 with *A. queenslandica*.

Considering the information presented in this section, a comparison of the similarities found among choanoflagellates, with the poriferan *A. queenslandica*, the placozoan *T. adhaerens* and the cnidarian *T. kitauei* suggest that these species retained the ancestral state of the common ancestor. On the contrary, the species *N. vectensis* and *M. leidy* diverged from that ancestral shared traits. In this hypothesis, cnidarian species of the class Myxozoa could have been more conserved, while the class Anthozoa, with *N. vectensis*, diverged in evolution.

4.2.3 Codon usage statistics.

4.2.3.1 General codon statistic values.

Codon usage statistic values were obtained with the software CodonW using the coding genes of unicellular and metazoan species (as explained in 2.2.3). The parameters GC content, GC at third position (GC3), codon bias index (CBI), effective number of codons (Nc) and frequency of optimal codons (Fop) are gathered in the tables below for each organism (in Tables 72 to 76) showing average and standard deviation values. The abbreviations used are LE (lowest expressed genes), HE (highest expressed genes) and All (for the full data set of CDS).

Table 72. GC content mean and standard deviation for unicellular and metazoan species. The full set of CDS (All) is compared with its value in the 5% of lowest (LE) and highest expressed genes (HE).

GC content			
Species	LE	All	HE
<i>S. rosetta</i>	0.598 ± 0.032	0.598 ± 0.028	0.601 ± 0.032
<i>S. arctica</i>	0.510 ± 0.045	0.509 ± 0.039	0.509 ± 0.036
<i>T. trahens</i>	0.657 ± 0.030	0.648 ± 0.029	0.627 ± 0.028
<i>A. queenslandica</i>	0.383 ± 0.034	0.403 ± 0.049	0.409 ± 0.071
<i>M. leidy</i>	0.465 ± 0.034	0.457 ± 0.042	0.459 ± 0.065
<i>N. vectensis</i>	0.481 ± 0.032	0.482 ± 0.053	0.462 ± 0.041
<i>T. kitauei</i>	0.377 ± 0.052	0.392 ± 0.047	0.429 ± 0.035
<i>T. adhaerens</i>	0.368 ± 0.028	0.379 ± 0.029	0.411 ± 0.032

GC content results (presented in Table 72) gave no correlation with expression levels in unicellular species, but the datasets showed lower values in LE and higher values in HE than in All for the species *A. queenslandica*, *T. kitauei* and *T. adhaerens*. The purpose of this analysis was to infer if differences among groups were exclusive in GC3 or could be influenced by general GC levels, although this values also could be the result of GC3 biasing the entire GC content of the datasets. This analysis revealed a change in the tendency from unicellular organisms of this and the previous chapter, while the metazoan species no presenting correlation (*M. leidy* and *N. vectensis*) were the ones diverging from the rest of species in optimal codons.

Table 73. GC3 mean and standard deviation for unicellular and metazoan species. The formatting is similar in style as Table 72.

GC3 content			
Species	LE	All	HE
<i>S. rosetta</i>	0.699 ± 0.072	0.695 ± 0.086	0.778 ± 0.085
<i>S. arctica</i>	0.521 ± 0.089	0.538 ± 0.082	0.547 ± 0.079
<i>T. trahens</i>	0.751 ± 0.052	0.771 ± 0.069	0.817 ± 0.068
<i>A. queenslandica</i>	0.280 ± 0.063	0.321 ± 0.085	0.355 ± 0.095
<i>M. leidy</i>	0.483 ± 0.073	0.463 ± 0.085	0.469 ± 0.111
<i>N. vectensis</i>	0.491 ± 0.065	0.456 ± 0.095	0.432 ± 0.071
<i>T. kitauei</i>	0.344 ± 0.077	0.365 ± 0.069	0.402 ± 0.065
<i>T. adhaerens</i>	0.276 ± 0.046	0.287 ± 0.044	0.323 ± 0.051

GC3 analysis revealed a correlation of expression level and this parameter in unicellular organisms (as detailed in Table 73), as was demonstrated in the previous chapter for all the choanoflagellates species studied, while no correlation of expression and GC content was found. In metazoan species, those presenting a correlation GC and expression (*A. queenslandica*, *T. kiatauei* and *T. adhaerens*) maintained a correlation for the GC3 level, while those species not showing it previously (*M. leidy* and *N. vectensis*) neither presented correlation for GC3.

Comparing species, the GC3 level was higher in all the unicellular species than in all the pluricellular ones. As in choanoflagellates, the data in unicellular species are consistent with natural selection, and in metazoan this is also true for the species presenting high GC3 content in optimal codons, correlating GC3 and expression except for *M. leidy* and *N. vectensis*. These data are consistent with the higher amount of GC3 found in optimal codons of unicellular organisms, while the species not presenting any correlation also showed a very low GC3 level in optimal codons, and with the metazoan, which correlation in general GC showed high GC3 content in optimal codons.

Table 74. CBI mean and standard deviation for unicellular and metazoan species. The formatting is similar in style as Table 72.

Codon Bias Index			
Species	LE	All	HE
<i>S. rosetta</i>	0.383 ± 0.134	0.376 ± 0.165	0.587 ± 0.178
<i>S. arctica</i>	0.054 ± 0.163	0.086 ± 0.154	0.138 ± 0.161
<i>T. trahens</i>	0.368 ± 0.123	0.416 ± 0.155	0.538 ± 0.172
<i>A. queenslandica</i>	0.377 ± 0.121	0.306 ± 0.156	0.248 ± 0.180
<i>M. leidy</i>	0.017 ± 0.154	0.059 ± 0.179	0.046 ± 0.235
<i>N. vectensis</i>	0.004 ± 0.104	0.051 ± 0.176	0.105 ± 0.123
<i>T. kitauei</i>	0.304 ± 0.162	0.266 ± 0.142	0.190 ± 0.134
<i>T. adhaerens</i>	0.236 ± 0.079	0.220 ± 0.084	0.133 ± 0.106

CBI values (specified in Table 74) showed consistently higher values in HE than in others datasets for unicellular organisms, similarly to the results obtained in choanoflagellates in 3.2.3.1. In the same way, the data obtained were consistent with natural selection for these organisms. In a comparison across species of the HE genes, CBI was much higher in *S. rosetta* and *T. trahens*. In metazoan this correlation was found exclusively in the species *N. vectensis*. The differences in data for *N. vectensis* from other metazoan were consistent with the previous results, as presented a very different GC3 content (4.76%) from the other organisms, and diverged from the ancestral form and other cnidarian class (theory explained in 4.2.2).

Table 75. Nc mean and standard deviation for unicellular and metazoan species. The formatting is similar in style as Table 72.

Effective number of codons			
Species	LE	All	HE
<i>S. rosetta</i>	45.718 ± 5.617	45.169 ± 5.428	36.700 ± 6.097
<i>S. artica</i>	54.389 ± 6.713	54.416 ± 6.346	51.553 ± 7.539
<i>T. trahens</i>	40.567 ± 4.607	38.710 ± 5.082	34.963 ± 5.970
<i>A. queenslandica</i>	46.611 ± 4.396	47.722 ± 6.325	48.517 ± 7.407
<i>M. leidy</i>	55.010 ± 2.782	54.027 ± 4.900	50.351 ± 7.529
<i>N. vectensis</i>	56.722 ± 4.382	50.710 ± 10.27	52.958 ± 6.931
<i>T. kitauei</i>	52.972 ± 5.062	53.847 ± 5.309	54.418 ± 5.411
<i>T. adhaerens</i>	49.189 ± 4.427	50.019 ± 4.465	50.114 ± 5.196

For Nc values (shown in Table 75) the three unicellular species gave consistent results with GC3, CBI and the results seen in choanoflagellates. Both *S. rosetta* and *T. trahens* showed a complete inverse correlation between expression levels and Nc value (and also presented the lowest values across all the species), while *S. artica* only showed a lower value for HE genes. These results were also compatible with natural selection. In metazoans, the only organisms showing a similar correlation was *M. leidy*, and not in *N. vectensis*, which conflicts with CBI values obtained. To explain this is necessary to remember that CBI measures the codon bias towards the use of optimal codons using the expected usage as scaling factor, while Nc measures the difference between the codon usage and the absence of bias. Although both parameters use similar inputs, the formulas are different and the scaling factor trait in CBI could transform a variation between observed and expected into larger numbers. These results, nonetheless, are not in conflict with GC levels or the optimal codon analysis, where both species showed to diverge from the other metazoan studied, but should be taken carefully due to the large numbers obtained for the standard deviation in all the species.

Table 76. Fop mean and standard deviation for unicellular and metazoan species. The data were obtained with customised optimal codon files, which were based on the expression level analysis. The formatting is similar in style as Table 72.

Frequency of optimal codons			
Species	LE	All	HE
<i>S. rosetta</i>	0.506 ± 0.063	0.507 ± 0.085	0.668 ± 0.112
<i>S. arctica</i>	0.454 ± 0.058	0.464 ± 0.068	0.567 ± 0.115
<i>T. trahens</i>	0.600 ± 0.083	0.637 ± 0.099	0.729 ± 0.113
<i>A. queenslandica</i>	0.328 ± 0.063	0.368 ± 0.084	0.403 ± 0.091
<i>M. leidy</i>	0.395 ± 0.042	0.407 ± 0.060	0.461 ± 0.081
<i>N. vectensis</i>	0.394 ± 0.056	0.418 ± 0.091	0.468 ± 0.077
<i>T. kitauei</i>	0.349 ± 0.048	0.370 ± 0.055	0.418 ± 0.061
<i>T. adhaerens</i>	0.360 ± 0.040	0.379 ± 0.045	0.447 ± 0.074

The Fop analysis (detailed in Table 76) presented higher values in HE than any other dataset for all the unicellular and metazoan species, although comparatively the values were higher in all the unicellular organisms. These results are consistent with natural selection for all the species, and although it was consistent with all the statistics seen previously (GC, GC3, CBI and Nc) in unicellular species, and was similar to the results obtained for choanoflagellates in 3.2.3, these values are in conflict with previous statistics in metazoan. An explanation for this could be that previous statistics are based on a formula more focused on measure the differences between the observed and expected codons, without considering the actual optimal codons, in a large dataset with the influx of other forces shaping the genome that could mask the effect of optimal codons (like random mutation and genetic drift, which was determined changing almost completely the prediction of optimal codons with CodonW in 4.2.2). It can be determined that optimal codons are more frequent in HE genes in metazoan, that could be a product of natural selection, but this may not be a major driver shaping the genome in their evolution.

4.2.3.2 Effective number of codons and GC3 plots.

The effective number of codons (Nc) and GC3 values calculated in the coding region analysis with CodonW (as described in 2.2.3) were plotted together in single graphs. These plots generated depict the distribution of the genes presenting general codon bias and GC at third position, over a line that represent the expected position of genes under a neutral mutation model (as in 3.2.3.2). Figures with these data for all the unicellular and metazoan species studied are showed below in Figures 28 to 35.

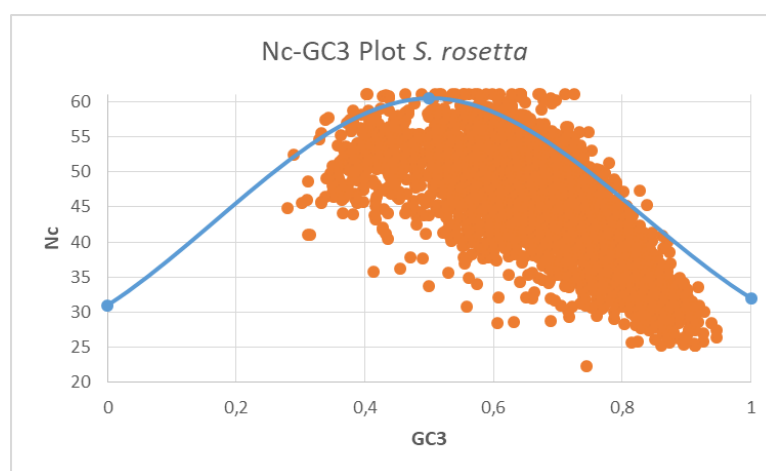


Figure 28. Nc and GC3 plot for *S. rosetta* genes. The graph represents these values for each gene of the CDS dataset as orange dots, and the neutral mutation line is showed in blue.

S. rosetta presented (in Figure 28) genes forming a single cloud of dots around the centre of the graph and skewed to the right side, mainly under the neutral mutation line (thus, showing most genes in an area predicted to be linked with bias by natural selection and/or genetic drift). This asymmetry evidences a strong correlation of genes increasing both in general bias and GC3 content. Results were similar to those of its sibling species *M. roanoka*, and also with the close related species *S. infusionum*, although presenting a stronger trend. These three species showed a strong correlation in the plots, like in the related species *C. perplexa*, thus considering the lowest bias in the *Hartaetosiga* genus as a divergence from an ancestral state within Craspedida Clade 1 (this lower codon bias was supported by the statistics of 3.2.3.1).

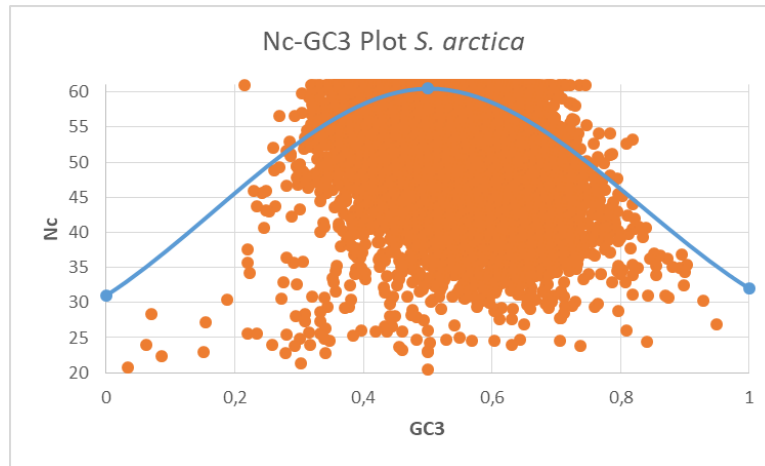


Figure 29. Nc and GC3 plot for *S. arctica* genes. The formatting is similar in style as Figure 28.

S. arctica showed (in Figure 29) genes placed mainly around the centre of the plot and skewed to the right side, mainly below the neutral mutation line. Although some scattered dots reflected high biased genes and were skewed, no clear correlation of general bias and GC3 can be determined in the graph. This is in agreement with previous results in 4.2.3.1, which showed statistics with evidence for natural selection (GC3, CBI, Nc and Fop), but with lowest codon bias parameters than in the other unicellular species.

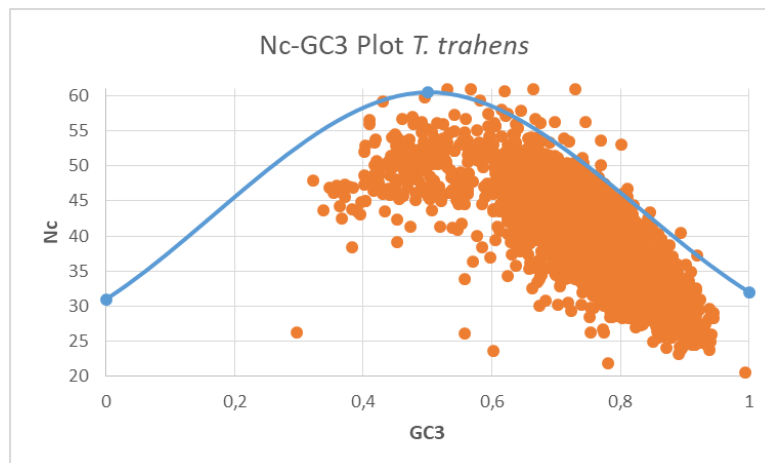


Figure 30. Nc and GC3 plot for *T. trahens* genes. The formatting is similar in style as Figure 28.

T. trahens (shown in Figure 30), in a similar way to *S. rosetta* (Figure 28), manifested a cloud of dots highly skewed to the right side of the graph, mainly below the neutral mutation line. This showed a clear strong correlation of general bias and GC3 for the genes of this species. This result was consistent with the data showed in GC3, CBI, Nc and Fop, the presented evidence for natural selection, and the similarities in bias with *S. rosetta*.

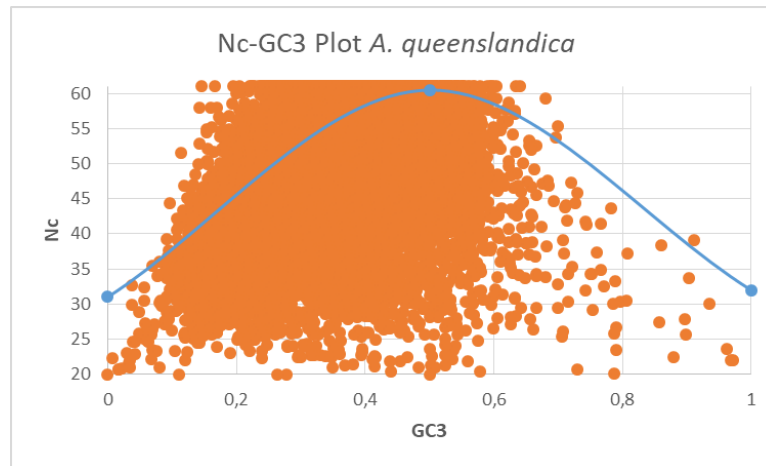


Figure 31. Nc and GC3 plot for *A. queenslandica* genes. The formatting is similar in style as Figure 28.

A. queenslandica exhibited genes (represented in Figure 31) grouped mainly forming a cloud from the centre to the left side of the plot, mainly under the neutral mutation line, but with some scattered genes towards the right side. No clear correlation was found between codon bias and GC3 in this plot. These results were in agreement with the low level of GC and GC3 content found in the different datasets, and with the presence of highly use of GC3 optimal codons in HE, but with a full set of genes masking this trend in Nc and CBI values.

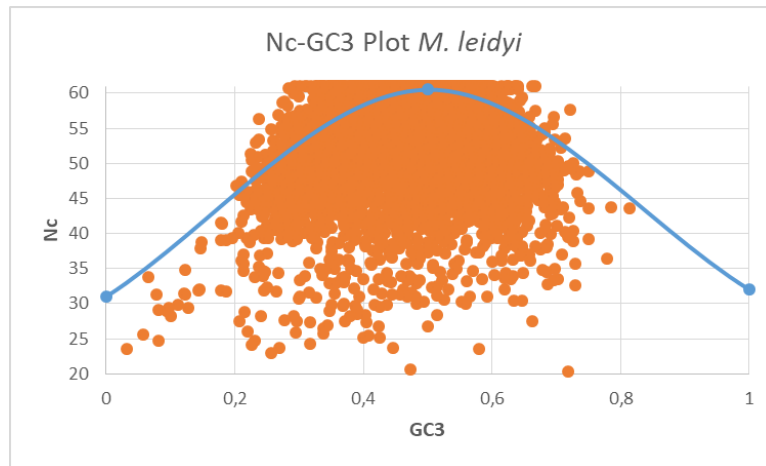


Figure 32. Nc and GC3 plot for *M. leidy* genes. The formatting is similar in style as Figure 28.

M. leidy presented genes (depicted in Figure 32) forming a cloud of dots around the centre of the plot, mostly under the neutral mutation line, and showing no clear tendency for GC3 content in biased genes. For this species natural selection was compatible with the results presented in optimal codons, GC3, Nc and Fop in HE, but in the plot, as with the statistics GC3 and CBI, this correlation was masked by characteristics of the complete set of genes. This was similar to the case of *A. queenslandica* (Figure 31), although with the additional support of the Nc values.

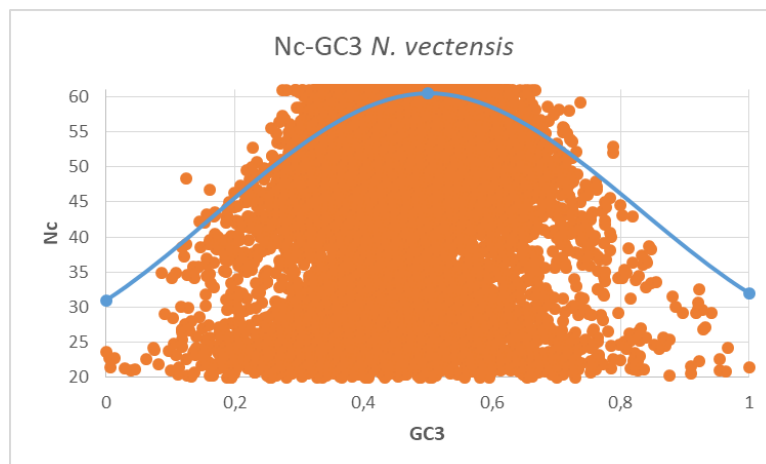


Figure 33. Nc and GC3 plot for *N. vectensis* genes. The formatting is similar in style as Figure 28.

N. vectensis showed a high number of biased genes under the neutral mutation line, but those were distributed symmetrically around the centre of the graph (as it can be seen in Figure 33). In this species two clouds of dots could be distinguished, corresponding to different sets of genes showing different level of bias. The plot did not revealed any reverse correlation of biased genes and GC3 content, which could be expected due to the very low level of GC3 in optimal codons of this species. Similar to the case of other metazoan studied like *M. leidy*, this species showed support for natural selection in GC3, Fop and CBI values (Nc in the case of *M. leidy*), but any possible correlation in this plot was masked by the characteristics of the whole set of genes.

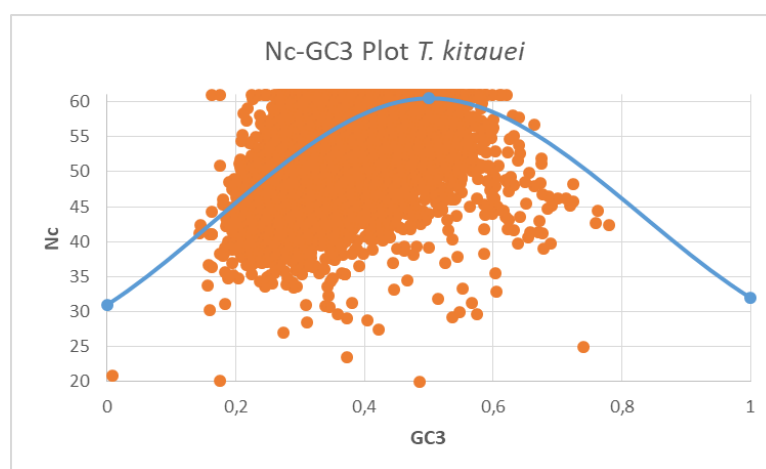


Figure 34. Nc and GC3 plot for *T. kitauei* genes. The formatting is similar in style as Figure 28.

T. kitauei manifested a set of genes (represented in Figure 34) forming a cloud skewed to the left side of the plot, under the neutral mutation line, but with scattered genes towards the right side. Although not very strong, a tendency for biased genes to contain a low GC3 could be suggested, and correlates with the low level of GC and GC3 found in this species. The parameters CBI and Nc did not evidenced any natural selection, although both Fop and a high GC3 content in optimal codons suggest that a small set of GC3 biased highly expressed genes is masked within the dataset. This situation is similar to that of the metazoan species presented previously.

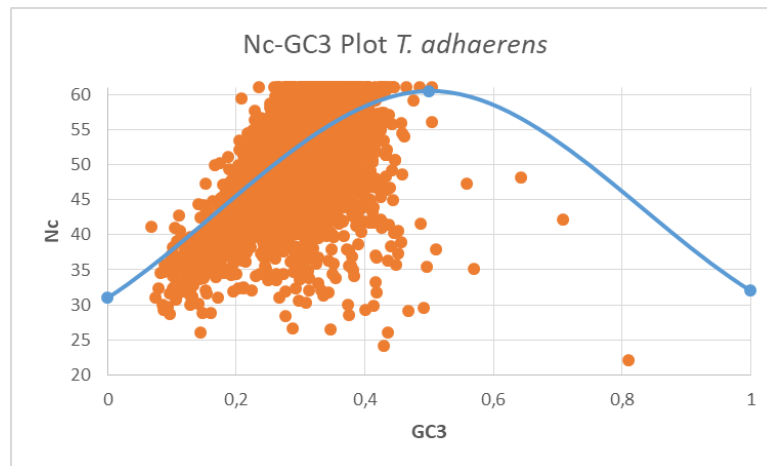


Figure 35. Nc and GC3 plot for *T. adhaerens* genes. The formatting is similar in style as Figure 28.

T. adhaerens showed genes (depicted in Table 35) in a cloud of dots in the left side of the plot, reflecting the dataset with the lowest GC content of the species studied, and with a group of genes that may be a product of low GC3 by random neutral mutations. The statistics revealed that even if the general GC3 is low in this species, the statistics GC, GC3 and Fop increases in HE, which also could be consistent with the high GC3 content in optimal codons for this species. Although Fop values were consistent with natural selection, the whole set of genes masked any correlation of the graph in agreement with previous data, as happened in all the metazoan species studied.

As a summary, the unicellular eukaryotic species showed data consistent with the statistics calculated and the evidence of natural selection found in 4.2.3.1, presenting a correlation GC3 and overall bias graphically, in a similar way that was demonstrated in most choanoflagellates. In metazoan species, although some statistics (mainly Fop) were consistent with natural selection, correlation in plots was not evidenced for any species, as most of the statistics calculated in 4.2.3.1 were unclear in demonstrating codon usage bias. The effect seen in metazoan could be a product of other forces like random mutation and genetic drift also driving the evolution of these genomes and masking their bias. In the set of species studied in this chapter, no double cloud of dots reflecting two different set of genes were found (except in *N. vectensis*), either because the sequencing of these species is further developed against contamination, or that the presence of horizontal transfer events that could happen through evolution acquired the traits of the host genome. In the case of *N. vectensis*, there must be considered that the double could also be a product of false annotations, as seen for *M. brevicolis* previously (Southworth et al., 2018).

4.2.3.3 Codon usage in protein domain and non-domain regions.

The 100 genes with highest and lowest relative expression in unicellular and metazoan species were separated into domains and non domains regions for the predicted proteins as described in 2.2.3. The sequences for each of the four categories were analysed with CodonW for codon usage statistics. Comparisons of the values obtained are listed below for every statistic, category and organism (in Tables 77 to 81). Abbreviations used here are NDR (predicted protein non-domain region), DR (predicted protein domain region), LE (lowest expressed genes) and HE (highest expressed genes).

The GC content was studied in DR and NDR (detailed in Table 77), as previously, with the purpose of correlate GC3 differences with optimal codon bias events or see if it could be attributed to a general GC content. In this analysis two main comparisons can be made. The comparison of DR in HE and LE genes revealed higher results in HE for the species *S. rosetta*, *S. arctica*, *A. queenslandica*, *M. leidy*, *T. kitauei* and *T. adhaerens*. The comparison of DR and NDR in HE genes showed highest results in DR for *S. rosetta*, *T. trahens*, *M. leidy*, *N. vectensis* and *T. kitauei*. These results were not completely consistent with the previous analysis in 4.2.3.1, as expression was only correlated previously with metazoan species (except in *T. kitauei*). Nonetheless, the increment of this parameters among the datasets was small and only must be considered for the GC3 interpretation.

The GC3 values (presented in Table 77) had an increment in DR over NDR of HE in all the unicellular species and in the metazoan *M. leidy*, *N. vectensis* and *T. kitauei*. As this increment was stronger than in the general GC, the GC3 could be interpreted as influencing the GC content. DR were also enriched in GC3 in HE genes over LE, in all the species except for *N. vectensis*. The first described results are compatible with natural selection acting on translational accuracy, while the second results addressed are consistent with translational efficiency. This conclusion is accurate for the unicellular species studied, but *M. leidy* and *N. vectensis* did not possess optimal codons enriched in GC3, thus this increment for metazoan only was relevant in *T. kitauei*. The lower GC3 value in *N. vectensis* would be expected in case of translational efficiency due to the high AU3 content in optimal codons for this species.

Table 77. GC content mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species. Results are based on the 100 highest (Highest Exp.) and lowest expressed genes (Lowest Exp.).

GC content				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>S. rosetta</i>	0.602 ± 0.040	0.603 ± 0.086	0.599 ± 0.044	0.607 ± 0.026
<i>S. arctica</i>	0.512 ± 0.053	0.518 ± 0.113	0.511 ± 0.044	0.517 ± 0.032
<i>T. trahens</i>	0.658 ± 0.038	0.612 ± 0.073	0.650 ± 0.030	0.622 ± 0.029
<i>A. queenslandica</i>	0.414 ± 0.027	0.458 ± 0.100	0.384 ± 0.034	0.436 ± 0.036
<i>M. leidy</i>	0.466 ± 0.039	0.483 ± 0.138	0.466 ± 0.039	0.485 ± 0.044
<i>N. vectensis</i>	0.487 ± 0.037	0.452 ± 0.064	0.494 ± 0.038	0.453 ± 0.054
<i>T. kitauei</i>	0.386 ± 0.081	0.421 ± 0.069	0.387 ± 0.056	0.438 ± 0.036
<i>T. adhaerens</i>	0.363 ± 0.052	0.431 ± 0.126	0.361 ± 0.023	0.414 ± 0.027

Table 78. GC3 mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species. The formatting is similar in style as Table 77.

GC3 content				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>S. rosetta</i>	0.707 ± 0.092	0.789 ± 0.200	0.766 ± 0.108	0.831 ± 0.056
<i>S. arctica</i>	0.517 ± 0.115	0.521 ± 0.253	0.543 ± 0.086	0.584 ± 0.069
<i>T. trahens</i>	0.739 ± 0.079	0.771 ± 0.147	0.785 ± 0.058	0.852 ± 0.046
<i>A. queenslandica</i>	0.339 ± 0.053	0.406 ± 0.193	0.272 ± 0.069	0.403 ± 0.085
<i>M. leidy</i>	0.474 ± 0.072	0.444 ± 0.240	0.491 ± 0.084	0.545 ± 0.095
<i>N. vectensis</i>	0.494 ± 0.073	0.442 ± 0.110	0.522 ± 0.070	0.450 ± 0.101
<i>T. kitauei</i>	0.346 ± 0.126	0.383 ± 0.165	0.356 ± 0.102	0.417 ± 0.084
<i>T. adhaerens</i>	0.273 ± 0.079	0.354 ± 0.198	0.271 ± 0.043	0.339 ± 0.053

Table 79. CBI mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species. The formatting is similar in style as Table 77.

Codon Bias Index				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>S. rosetta</i>	0.403 ± 0.166	0.616 ± 0.328	0.507 ± 0.199	0.702 ± 0.111
<i>S. arctica</i>	0.042 ± 0.210	0.137 ± 0.478	0.097 ± 0.156	0.268 ± 0.143
<i>T. trahens</i>	0.355 ± 0.168	0.479 ± 0.272	0.440 ± 0.129	0.608 ± 0.154
<i>A. queenslandica</i>	0.086 ± 0.093	0.131 ± 0.363	0.394 ± 0.131	0.151 ± 0.169
<i>M. leidy</i>	0.038 ± 0.154	0.099 ± 0.510	-0.002 ± 0.175	-0.117 ± 0.206
<i>N. vectensis</i>	-0.003 ± 0.121	0.086 ± 0.199	-0.039 ± 0.121	0.079 ± 0.167
<i>T. kitauei</i>	0.303 ± 0.265	0.227 ± 0.336	0.273 ± 0.222	0.154 ± 0.172
<i>T. adhaerens</i>	0.256 ± 0.137	0.073 ± 0.393	0.225 ± 0.076	0.086 ± 0.093

Table 80. Nc mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species. The formatting is similar in style as Table 77.

Effective Number of Codons				
	Non-domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>S. rosetta</i>	44.359 ± 6.382	32.225 ± 5.289	40.795 ± 7.756	31.087 ± 4.131
<i>S. arctica</i>	54.084 ± 7.477	44.262 ± 11.205	54.109 ± 6.900	42.970 ± 6.453
<i>T. trahens</i>	41.177 ± 6.264	37.468 ± 6.117	38.028 ± 6.089	31.756 ± 5.436
<i>A. queenslandica</i>	48.865 ± 6.038	44.72 ± 9.264	44.858 ± 6.143	49.527 ± 6.424
<i>M. leidy</i>	55.032 ± 4.553	45.488 ± 9.351	53.686 ± 5.040	45.908 ± 5.987
<i>N. vectensis</i>	57.022 ± 3.545	53.369 ± 6.389	56.047 ± 4.932	52.607 ± 7.815
<i>T. kitauei</i>	51.917 ± 7.110	50.577 ± 9.121	50.628 ± 6.920	50.944 ± 8.755
<i>T. adhaerens</i>	47.116 ± 9.530	44.978 ± 7.925	47.836 ± 4.781	48.865 ± 6.038

The CBI data (specified in Table 79) revealed that unicellular species presented higher values in HE than in LE and even higher values in DR than in NDR, showing results both compatible with translational accuracy and efficiency. Metazoan species presented higher values in DR than in NDR in *A. queenslandica* and *T. adhaerens*, but these species also showed higher values in one or both regions of low expressed genes, avoiding to establish a solid correlation. The data showed are consistent with previous results, as CBI in 4.2.3.1 also presented strong correlation in unicellular organisms and not in metazoan, except for the expression levels for *N. vectensis*, that also were maintained here (although showing higher values in NDR).

The Nc results (shown in Table 80) were lower in HE genes, and lower in DR over NDR consistently through the unicellular species. These results were consistent with the CBI values obtained and with the values obtained in 4.2.3.1, showing both evidence of natural selection upon translational accuracy and efficiency. For the metazoan organisms, this correlation was also seen in *N. vectensis* for both comparisons, and only showed lower values in HE genes for *M. leidy*. Therefore, these data were compatible with both kinds of selection for the first one and only translational efficiency in the second case. In *N. vectensis*, these data correlates with the evidence for translational efficiency found in the Fop and CBI parameters of the previous analysis, while *M. leidy* showed similar evidence in Fop and Nc in 4.2.3.1.

Fop values (detailed in Table 81) in unicellular species were higher in HE and DR, which was consistent with the previous data showed in GC3, CBI and Nc, with the parameters of 4.2.3.1, and with the correlation found in Nc-GC3 plots. Comparing the Fop levels obtained with those presented in the choanoflagellates, these three species are at the level of the top one-third of the most biased species studied. In this analysis, results were consistent with selection upon translational accuracy and efficiency, as also were detected in the choanoflagellates of Chapter 3. Also, including the mentioned choanoflagellates, the metazoan species showed the lowest values of all the species studied. In addition, among all metazoans the only organism showing higher results in DR over NDR was *T. kitauei*, while *M. leidy*, *N. vectensis* and *T. adhaerens* just showed higher results correlating with HE.

Table 81. Fop mean and standard deviation in protein domain and non-domain regions of unicellular and metazoan species. Results are based on customised Fop files using optimal codons determined by expression levels. The formatting is similar in style as Table 77.

Frequency of Optimal Codons				
	Non-Domain regions		Domain regions	
Species	Lowest Exp.	Highest Exp.	Lowest Exp.	Highest Exp.
<i>S. rosetta</i>	0.525 ± 0.076	0.735 ± 0.230	0.560 ± 0.083	0.775 ± 0.061
<i>S. arctica</i>	0.459 ± 0.068	0.699 ± 0.229	0.468 ± 0.053	0.723 ± 0.101
<i>T. trahens</i>	0.595 ± 0.109	0.680 ± 0.179	0.647 ± 0.084	0.778 ± 0.099
<i>A. queenslandica</i>	0.475 ± 0.070	0.439 ± 0.188	0.318 ± 0.070	0.438 ± 0.082
<i>M. leidy</i>	0.390 ± 0.039	0.532 ± 0.253	0.401 ± 0.049	0.498 ± 0.060
<i>N. vectensis</i>	0.386 ± 0.060	0.451 ± 0.125	0.369 ± 0.064	0.445 ± 0.095
<i>T. kitauei</i>	0.349 ± 0.106	0.422 ± 0.148	0.355 ± 0.055	0.442 ± 0.077
<i>T. adhaerens</i>	0.357 ± 0.077	0.511 ± 0.249	0.352 ± 0.044	0.475 ± 0.070

In summary, all the unicellular eukaryotes studied in this chapter, *S. rosetta*, *S. arctica* and *T. trahens*, showed higher levels of GC3, CBI, Fop and lower Nc first in DR, and later in HE genes. These results were consistent with optimal codons presenting high GC3 content, no correlation with GC, but with GC3, CBI, Nc and Fop correlating with expression levels, and with the overall bias correlating with GC3 in the plots of 4.2.3.2. All these data are compatible with both natural selection acting upon translational efficiency (higher bias in HE) and accuracy (higher bias in DR), a similar conclusion that was reached for the choanoflagellates studied in 3.2.3.3.

The metazoan species showed less homogenous results than the unicellular ones. In this section, across Tables 77 to 81 *A. queenslandica* only showed correlation in GC3 with HE, which was consistent with the GC3 correlation obtained in 4.2.3.1, high GC3 in optimal codons, a GC level with low bias in HE, and Fop correlating with expression levels. This information leads to conclude that there was no evidence for translational accuracy in this species, while the translational efficiency is only supported by GC3 and Fop levels.

For *M. leidy*, previous analysis revealed a correlation of Nc and Fop with expression, while in the present section GC3 was higher in DR and HE, while Nc and Fop were still correlated with expression. In this species selection for translational efficiency is supported by Nc and Fop level of both analysis, with translational accuracy only supported by GC3 levels.

In cnidarian species, *N. vectensis* only presented correlation of CBI and Fop with expression, while in the present section showed also this CBI and Fop correlation with expression, plus an inverse correlation of Nc with HE and DR. In this species, selection on translational efficiency is supported by both CBI and Fop analysis, and the second Nc analysis, while selection on translational accuracy is only backed up by Nc levels. The other cnidarian studied, *T. kitauei*, showed previously GC3 and Fop correlated with expression level, while in the present analysis revealed correlation of GC3 and Fop with DR and HE genes. Both evidence for translational accuracy and efficiency is backed up by GC3 and Fop in this species.

The relative of Cnidaria, the placozoan *T. adhaerens* showed previously GC3 and Fop correlated with expression, while in the present analysis showed the same parameters also correlated with HE. This species only showed evidence for selection upon translational efficiency supported by GC3 and Fop values.

4.2.4 Strength of selection.

With the aim of evaluating the strength of selection, CodonW data were used to calculate odds ratio and statistic $\hat{S}$ values for two-degenerated amino acids (as explained in 2.2.5). Results for the optimal codons determined by expression levels are presented below (in Tables 82 to 84), showing data per species and amino acid, and distributed by phylogenetic distance as in 4.2.2. Similarly to the choanoflagellates results in 3.2.4, only the statistic $\hat{S}$ values will be addressed, as both statistics are correlated and the formula of this value describes selection more accurately (this statement was already addressed in 2.2.5).

Table 82. Strength of selection values for *S. rosetta*, *S. arctica* and *T. trahens*. Odds ratio (OR) and Statistic $\hat{S}$ (St. $\hat{S}$) values are shown for the optimal codon of each of the two-degenerated amino acids. Only data based on the expression level analysis was used. The sum and arithmetic average values for the statistics are showed at the bottom of the columns.

Strength of selection in optimal codons by expression levels									
	<i>S. rosetta</i>			<i>S. arctica</i>			<i>T. trahens</i>		
AA	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	2.449	0.824	UUC	2.728	0.836	UUC	0.773	-0.249
Tyr	UAC	24.385	2.069	UAC	2.927	0.738	UAC	36.241	2.464
His	CAC	22.240	1.785	CAC	3.111	0.556	CAC	7.302	1.099
Gln	CAG	14.562	2.634	CAG	2.442	0.687	CAG	48.875	3.434
Asn	AAC	111.186	4.120	AAC	1.696	0.491	AAC	52.218	3.677
Lys	AAG	102.005	6.725	AAG	3.621	1.821	AAG	268.146	7.695
Asp	GAC	8.723	3.261	GAU	1.166	0.211	GAC	8.311	3.455
Glu	GAG	28.070	5.193	GAG	2.596	1.428	GAG	65.986	6.414
Cys	UGC	13.833	1.266	-	-	-	UGC	19.156	1.306
Sum		327.454	27.878		20.287	6.769		507.008	29.294
Aver		36.384	3.098		2.536	0.846		56.334	3.255

S. rosetta presented an average for statistic $\hat{S}$ of 3.098 (detailed in Table 82), showing as strongest selected codons AAG (Lys), GAG (Glu) and AAC (Asn), while the weakest ones were UUC (Phe), UGC (Cys) and CAC (His). This species showed a strength of selection much higher than its sibling *M. roanoka* (0.554), and that its close related *S. infusionum* (1.209) and *C. perplexa* (1.18). As the *Hartaetosiga* genus was determined low biased in comparison with the species studied, *S. rosetta* is the strongest biased species studied in Craspedida Clade 1. All the species of this clade showed Asn and Lys as strongest biased optimal codons, showing that selection is favouring similar codons despite acting with different efficiency on them. In comparison with species of other clades, *S. rosetta* only showed lower results than *M. fluctuans* and *S. punica* of the highly biased Craspedida Clade 2, situating this organism as one of the most biased ones studied.

S. arctica showed an average for statistic $\hat{S}$ of 0.846 (presented in Table 82). Its strongest selected codons were AAG (Lys) and GAG (Glu), while the weakest codons were GAC (Asp), AAC (Asn) and CAC (His). The strength of selection in this species could be similar to that found in *S. diplocostata*, which belongs to Stephanoecidae, one of the groups of organisms determined as less biased in Chapter 3. In the opposite situation, *T. trahens* presented the strongest selected codons of all the species of this chapter (3.255 on average for statistic $\hat{S}$), also shown in Table 82. The strongest selected codons in this species were AAG (Lys) and GAG (Glu), while the weakest ones were UUC (Phe), CAC (His) and UGC (Cys). The results showed for this species could be similar those found in *S. rosetta*, only surpassed by *M. fluctuans* and *S. punica*.

Table 83. Strength of selection values for *A. queenslandica* and *M. leidy*. The formatting is similar in style as Table 82.

Strength of selection in optimal codons by expression levels						
	<i>A. queenslandica</i>			<i>M. leidy</i>		
AA	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUC	0.296	-1.040	UUU	0.855	-0.161
Tyr	UAC	0.335	-0.839	-	-	-
His	CAC	0.402	-0.491	CAU	0.481	-0.376
Gln	CAG	0.575	-0.497	-	-	-
Asn	AAC	0.205	-1.650	AAU	0.304	-1.360
Lys	AAG	0.576	-0.912	-	-	-
Asp	GAC	0.197	-1.970	-	-	-
Glu	GAG	0.618	-0.765	GAA	0.975	-0.045
Cys	UGC	0.216	-0.690	UGC	0.399	-0.482
Sum		3.419	-8.852		3.014	-2.424
Aver		0.380	-0.984		0.603	-0.485

A. queenslandica (results shown in Table 83) exhibited the second lowest statistic $\hat{S}$ studied (-0.984). The strongest selected codons in this species were CAC (His), CAG (Gln) and GAG (Glu), while the weakest ones were GAC (Asp), AAC (Asn) and UUC (Phe). In opposition to the poriferan described, the ctenophoran *M. leidy* (also detailed in Table 83) was the second metazoan with strongest codon selection (-0.485 of average for statistic $\hat{S}$), which also was much lower than the lowest value for choanoflagellates (-0.073 in *H. balthica*). The codons with higher statistic $\hat{S}$ values were GAA (Glu) and UUU (Phe), while the lowest values were AAU (Asn) and UGC (Cys).

Table 84. Strength of selection values for *N. vectensis*, *T. kitauei* and *T. adhaerens*. The formatting is similar in style as Table 82.

Strength of selection in optimal codons by expression levels									
	<i>N. vectensis</i>			<i>T. kitauei</i>			<i>T. adhaerens</i>		
AA	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$	Codon	OR	St. $\hat{S}$
Phe	UUU	1.271	0.200	UUC	0.482	-0.731	UUC	0.247	-1.208
Tyr	UAU	0.713	-0.195	UAC	0.364	-0.774	UAC	0.255	-0.994
His	-	-	-	CAC	0.360	-0.589	CAC	0.111	-0.965
Gln	CAA	1.017	0.015	CAG	0.298	-1.064	CAG	0.143	-1.634
Asn	AAU	0.899	-0.093	AAC	0.447	-0.775	AAC	0.161	-2.041
Lys	AAA	0.990	-0.014	AAG	0.263	-1.868	AAG	0.282	-2.260
Asp	GAU	1.225	0.217	GAC	0.248	-1.562	GAC	0.097	-3.028
Glu	GAA	0.880	-0.172	GAG	0.229	-1.909	GAA	11.999	3.798
Cys	-	-	-	-	-	-	UGC	0.607	-0.199
Sum		6.995	-0.042		2.691	-9.272		13.903	-8.531
Aver		0.999	-0.006		0.336	-1.159		1.545	-0.948

N. vectensis showed the highest average for statistic $\hat{S}$ of the metazoan studied (-0.006), as specified in Table 84, although this value was lower than in all the choanoflagellates except for *H. balthica*. The strongest selected codons identified were UUU (Phe), GAU (Asp) and CAA (Gln), while its weakest were UAU (Tyr) and GAA (Glu). The other cnidarian, *T. kitauei* (also in Table 84), presented the lowest statistic $\hat{S}$ average values of all the species studied in this project (-1.159). Its strongest selected codon was CAC (His), and its weakest ones were GAG (Glu), AAG (Lys) and GAC (Asp).

Both cnidarian presented very different strength of selection and different codons with highest and lowest bias, which is consistent with the great differences found previously since the first section, showing divergence of *N. vectensis* in the identification of optimal codons in 4.2.2. The placozoan *T. adhaerens* (shown in Table 84) presented a strength of selection of -0.948 of statistic $\hat{S}$ average, close to the result in *A. queenslandica* (-0.984). The strongest selected codons in this species was GAA (Glu), and the weakest selected were GAC (Asp), AAG (Lys) and AAC (Asn). Although close related to Cnidaria, no similarities were found among these species. On the contrary, the strongest biased codon in *T. adhaerens* was one of the weakest biased in both cnidarian.

As a summary, the unicellular species showed a high level of strength of selection, especially in *S. rosetta* and *T. trahens*, which were situated among the species with highest results for statistic $\hat{S}$, while *S. artica* had more moderate results, similar in bias to *S. diplocostata*. These results were consistent with the values showed previously for the statistics GC3, CBI, Nc and Fop, with the GC3-Nc correlation, with the analysis of DR and NDR and the evidence for selection upon translational efficiency and accuracy. On the contrary, the metazoan species exhibited lowest values in strength of selection than those obtained for all the unicellular species (with the exception of *H. balthica*), which were consistent with the statistical values presented through the section 4.2.3, and the lowest evidence presented for natural selection.

4.3 tRNA AND ANTICODON ANALYSIS.

4.3.1 Indirect evidence for tRNA deamination through selective pressure.

Evidence for tRNA deamination was searched separating the domains and non domains regions for the predicted proteins of the 100 genes with highest and lowest relative expression. These categories were used to compare domains and non domains on the search for overrepresented codons ended in cytosine (as described in 2.3.1), selected as an advantage produced by the use of deaminated codons for the most abundant tRNA (as also was analysed in 3.3.1). The overrepresented codons for TAPSILVR amino acids (threonine, alanine, proline, serine, isoleucine, leucine, valine, and arginine) in the domains regions are listed below for predicted proteins of highest and lowest expressed genes in Tables 85 and 86.

Table 85. Overrepresented cytosine ended codons in protein domain regions of unicellular species. Codons for TAPSILVR amino acids in domains of highest expressed genes (HE) and lowest expressed genes (LE) are compared.

Overrepresented codons in protein domain regions of unicellular organisms						
	<i>S. rosetta</i>		<i>S. arctica</i>		<i>T. trahens</i>	
aa	HE	LE	HE	LE	HE	LE
Leu	CUC	CUC	-	-	CUC	-
Ile	-	-	AUC	-	AUC	AUC
Val	-	-	-	-	GUC	GUC
Ser	-	UCC	UCC	UCC	-	UCC
Pro	CCC	-	CCC	-	-	CCC
Thr	-	ACC	-	-	ACC	ACC
Ala	GCC	GCC	-	-	GCC	-
Arg	-	-	-	-	CGC	-

Table 86. Overrepresented cytosine ended codons in protein domain regions of metazoan species. The formatting is similar in style as Table 85.

Overrepresented codons in protein domain regions of metazoan organisms										
	<i>A. queenslandica</i>		<i>M. leidy</i>		<i>N. vectensis</i>		<i>T. kitauei</i>		<i>T. adhaerens</i>	
aa	HE	LE	HE	LE	HE	LE	HE	LE	HE	LE
Leu	-	-	-	CUC	-	CUC	-	-	-	-
Ile	AUC	-	AUC	AUC	AUC	AUC	AUC	AUC	AUC	-
Val	-	-	-	-	-	-	-	-	-	-
Ser	-	-	-	-	-	UCC	-	-	-	-
Pro	-	-	CCC	-	-	-	-	CCC	-	-
Thr	-	-	-	-	-	-	-	-	-	-
Ala	-	-	GCC	-	-	-	-	-	-	-
Arg	-	-	-	-	-	-	-	CGC	-	-

The analysis of C ending codons in TAPSILVR amino acids, reflecting a selective advantage to find a correlation with deamination process, presented 16 overrepresented codons in unicellular species (Table 85), 12 in HE and 10 in LE (as 6 results were shared). This proportion is much lower to that found in choanoflagellates in 3.3.1, which showed a presence of these codons 4 times higher in HE. The species showing lower number of overrepresented codons was *S. artica*, while the highest number was in *T. trahens*. A difference to the results found across the choanoflagellates study, this proportion correlates with the level of bias found for these species in section 4.2.3 and 4.2.4.

The overrepresented codons in metazoan (detailed in Table 86) showed a lower amount of results in comparison with the unicellular species. In this case 12 codons were detected across the 5 species, 7 in HE and 8 in LE (as 3 results were shared in datasets of the same species). On the contrary, in unicellular species of this and Chapter 3 overrepresented codons were not found more abundant in HE genes. A lower number of these codons in metazoan correlates with the lowest codon bias found in the sections 4.2.3 and 4.2.4, although this correlation was not clear in the comparison across choanoflagellates. Metazoan species also exhibited a preference for overrepresentation of the codon AUC for Ile, which was found in all the species analysed.

In general, the studied effect as an indirect product of deamination was detected in all the species studied, as it was detected in all the choanoflagellates in Chapter 3. Results were found more prevalent in unicellular species and more biased HE genes, but considering groups of species this tendency was not clear while comparing the lowest biased unicellular and most biased metazoan. This result is consistent with choanoflagellates in 3.3.1, where although there was a tendency linked to bias levels, it failed to match a phylogenetic organisation or a ranking of bias parameters like Fop. As in the previous chapter choanoflagellates presented a higher abundance of the overrepresented codon GCC for Ala, which was found in all the species except for *H. balthica* and *H. nana*, this tendency was kept in *S. rosetta* and the apusomonadean *T. trahens*, but not in the ichthyosporean *S. arctica*. Metazoan consistently changed this preference for AUC codons.

The overrepresented codons showed in this section were determined indirectly to provide an advantage for using deaminated tRNA with inosine. As it was shown also in 3.3.1, this evidence found in all the species studied did not correlate the phylogenetic position, but offered an insight about the impact of deamination, clearly different between unicellular and pluricellular organisms.

4.3.2 Direct evidence for tRNA deamination through transcriptome screening.

The direct evidence for deamination process was performed, as explained in 2.3.2, mapping the transcriptomes into tRNA genes to observe adenine in position 34, and compare it with the mapping onto a modified gene containing guanine in the wobble position. The replacement of A for G is seen in cDNA in case of deamination to inosine. Results obtained for presence or absence of transcripts (modified and unmodified) for this sequences are showed next in Tables 87 and 88.

Table 87. Detection of direct deamination in transcripts of tRNA genes in unicellular species. The detection of reads using an unmodified version of the gene (containing A at the wobble position) and a modified one (containing G at the wobble position) was determined in the transcriptome sequencing described in 2.1.1. The classification is made for the categories “P” for present, and “ND” in non-detected transcripts.

Detection of A or G at wobble position in unicellular species						
	<i>S. rosetta</i>		<i>S. arctica</i>		<i>T. trahens</i>	
aa	Unmod.	Modified	Unmod.	Modified	Unmod.	Modified
Leu	P	ND	P	ND	P	ND
Ile	ND	ND	P	P	P	ND
Val	ND	ND	P	P	P	ND
Ser	P	ND	P	P	P	ND
Pro	P	ND	P	P	ND	ND
Thr	ND	ND	P	P	ND	ND
Ala	P	ND	P	P	P	ND
Arg	ND	ND	P	P	P	ND

The analysis of modified transcript in unicellular organisms showed that although the three species contained the unmodified transcript (for half or more of the studied amino acids), only in *S. arctica* was present the modified version of the gene. Thus, Table 87 shows that only *S. arctica* evidenced wobble positions containing G and A for all the TAPSLVR amino acids except for Leu.

Table 88. Detection of direct deamination in transcripts of tRNA genes in metazoan species. The formatting is similar in style as Table 87.

Detection of A or G at wobble position in metazoan species										
	<i>A. queenslandica</i>		<i>M. leidy</i>		<i>N. vectensis</i>		<i>T. kitauei</i>		<i>T. adhaerens</i>	
aa	Unmod.	Modif.	Unmod.	Modif.	Unmod.	Modif.	Unmod.	Modif.	Unmod.	Modif.
Leu	ND	ND	ND	ND	P	P	ND	ND	ND	ND
Ile	ND	ND	P	P	P	P	ND	ND	ND	ND
Val	P	P	P	P	P	P	ND	ND	ND	ND
Ser	P	P	P	P	P	P	ND	ND	ND	ND
Pro	P	P	P	P	P	P	ND	ND	ND	ND
Thr	P	P	P	P	P	P	ND	ND	ND	ND
Ala	ND	ND	ND	ND	P	P	P	ND	ND	ND
Arg	P	P	P	P	P	P	ND	ND	ND	ND

The results for deamination in metazoan species (detailed in Table 88) revealed the presence of unmodified transcript in half or more of the amino acids studied for the species *A. queenslandica*, *M. leidy* and *N. vectensis*. These species showed also the modified version for all the unmodified detected. The unmodified versions of the genes not detected belonged in both *A. queenslandica* and *M. leidy* to Leu and Ala, although this first species also did not show presence for Ile. These three species evidenced wobble position showing directly consequences of the deamination process. Results in *T. kitauei* only showed the presence of one of the unmodified genes, and *T. adhaerens* did not show any presence of tRNA genes, but this must be taken cautiously as the absence of results cannot rule out the existence of modified or unmodified genes that were not represented in the transcriptome sequencing.

As a general evaluation of this analysis, direct evidence for deamination events were found in the species *S. arctica*, *A. queenslandica*, and in *N. vectensis*. Correlation of these results with the level of codon usage bias, phylogenetic position or indirect evidence for deamination in 4.3.1 were not found. Nonetheless, as mentioned, the absence of results for this analysis does not exclude the presence of deamination events in not present genes, as cannot be ruled out that those were not sequenced in the transcriptomes used for the search, especially in cDNA sequencing not enriched for non-coding RNA (that was used for all the species except for *N. vectensis*, where tRNA is more rare). The presence of overrepresented codons in all the species studied in 4.3.1 supports this explanation and render the *S. rosetta*, *T. trahens*, *T. kitauei* and *T. adhaerens* as not conclusive for direct evidence of deamination.

4.3.3 Differential anticodon abundance based on tRNA copy number.

Anticodon abundance for unicellular and metazoan species was based on the calculation of tRNA copy number and performed with tRNAscan-SE 2.0 local version as detailed in 2.3.3. The process was carried out over DNA toplevel files from EnsemblProtist and EnsemblMetazoa (obtained as in 2.1.1). The results were compared with the optimal codons obtained by expression levels (presented in 4.1.2) and are listed below in Tables 89, 90 and 91.

The DNA toplevel files analysis (as detailed next in Table 89) revealed 122 anticodons in the choanoflagellate *S. rosetta*, with 63.4% corresponding to optimal codons. The amino acids presenting higher values were Phe, Asn, Asp and His (100%), and the lower ones Ala, Thr, Val and Leu (0%). The ichthyosporean *S. arctica* showed 334 anticodons, with 55.4% corresponding to optimal codons. In this species the amino acids with highest scores were Phe, Asn, His and Tyr (100%), and with lowest Asp and Cys (0%). For the last unicellular species studied, the apusomonadean *T. trahens*, presented 89 anticodons, corresponding to optimal codons in 58.4%. The highest correlation values were found in Phe, Asn, His, Tyr and Cys (100%), and the lowest in Ala, Val, Leu and Ile (0%).

Table 89. Anticodons identified in DNA toplevel files for unicellular species. Results are shown comparing all anticodons identified per amino acid (Ac), amount of anticodons pairing with optimal codons identified by expression (OC), and the proportion between both as OC/Ac.

Anticodons in DNA toplevel files for unicellular organisms									
	<i>S. rosetta</i>			<i>S. arctica</i>			<i>T. trahens</i>		
	Ac	OC	OC/Ac	Ac	OC	OC/Ac	Ac	OC	OC/Ac
Ala	9	0	0	27	12	0.444	8	0	0
Gly	9	7	0.778	28	12	0.429	7	5	0.714
Pro	6	3	0.5	19	10	0.526	4	2	0.5
Thr	5	0	0	20	10	0.5	5	1	0.2
Val	8	0	0	23	8	0.348	5	0	0
Ser	7	3	0.429	19	7	0.368	9	4	0.444
Arg	8	5	0.625	25	12	0.48	9	4	0.444
Leu	10	0	0	15	3	0.2	8	0	0
Phe	4	4	1	11	11	1	2	2	1
Asn	4	4	1	12	12	1	1	1	1
Lys	7	6	0.857	23	15	0.652	5	4	0.8
Asp	6	6	1	15	0	0	3	3	1
Glu	6	4	0.667	17	12	0.706	4	3	0.75
His	3	3	1	8	8	1	2	2	1
Gln	4	3	0.75	14	7	0.5	3	2	0.667
Ile	5	4	0.8	17	14	0.824	4	0	0
Met/iMet	11	-	-	19	-	-	5	-	-
Tyr	5	5	1	8	8	1	2	2	1
Supres	0	-	-	0	-	-	0	-	-
Cys	2	2	1	7	0	0	1	1	1
Trp	2	-	-	6	-	-	1	-	-
SelCys	1	-	-	1	-	-	1	-	-
Sum	122	59	11.405	334	161	9.977	89	36	10.520
Average	5.545	3.278	0.634	15.182	9.471	0.554	4.045	2	0.584
Stdev	2.94	2.191	0.390	7.902	3.842	0.321	2.768	1.572	0.398

In addition to the results obtained for anticodons in Table 89, the effect of deamination events must be considered. Although in *S. rosetta* evidence for direct deamination was not found in section 4.3.2, overrepresented codons in 4.3.1 revealed data consistent with deamination for the amino acids Ala, Pro, Thr, Ser and Leu, all showing a low number of anticodons corresponding to optimal codons (from 0 to 50%). As these results are not consistent with the high bias presented for this species through section 4.2.3 (GC3, CBI, Nc, Fop, GC3-Nc correlation and strength of selection), these numbers could be explained for this species using deaminated tRNA to match optimal codons. This situation would be very consistent with the high number of optimal codons found in HE genes for *S. rosetta*.

In a similar way to *S. rosetta*, the species *S. arctica* showed overrepresented codons for Pro, Ser and Ile, while also direct evidence was found for Ala, Pro, Thr, Val, Ser, Arg and Ile. As in the case of *S. rosetta*, these codons presented low level of coincidence with optimal codons (0-52% except for Ile), and due to the evidence for natural selection in this species, including high Fop in HE genes and DR, this species also could be showing lower levels on this amino acids due to deamination events. In the last unicellular species studied, *T. trahens*, no direct evidence for deamination was found, but overrepresented codons were present for all the TAPSILVR amino acids. Considering the level of bias presented in this species previously (see section 4.2.3), similar to that found in *S. rosetta*, the strong evidence for natural selection, and the high use of GC3 optimal codons, all indicates that this species is also showing a reduced number of anticodons matching optimal codons in TAPSILVR amino acids as consequence of the deamination process.

The tRNA-scan-SE 2.0 analysis for metazoan species (shown next in Table 90) revealed in the poriferan *A. queenslandica* 158 anticodons, 65.1% corresponding to optimal codons. The amino acids with higher values were Asn, Asp, His, Ile, Tyr and Cys (100%), while the lowest values were presented in Ala (16.7%), Arg (26.7%) and Pro (28.6%). The ctenophoran *M. leidy* showed 719 anticodons, but with only 36.4% matching the optimal codons for this species. Although no amino acids were detected using 100% optimal codons, and also less amino acids using 0% were found, the highest value was for Ser (83.5%), while the lowest were in Pro, Asn and His (0%).

Table 90. Anticodons identified in DNA toplevel files for *A. queenslandica* and *M. leidy*. The formatting is similar in style as Table 89.

Anticodons in DNA toplevel files for metazoan organisms (PART 1)						
	<i>A. queenslandica</i>			<i>M. leidy</i>		
	Ac	OC	OC/Ac	Ac	OC	OC/Ac
Ala	6	1	0.167	18	11	0.611
Gly	14	7	0.5	123	28	0.288
Pro	7	2	0.286	9	0	0
Thr	10	3	0.3	20	3	0.15
Val	6	2	0.333	12	8	0.667
Ser	10	8	0.8	340	284	0.835
Arg	15	4	0.267	31	6	0.194
Leu	27	19	0.704	14	2	0.143
Phe	15	14	0.933	13	4	0.308
Asn	4	4	1	8	0	0
Lys	5	3	0.6	17	-	-
Asp	4	4	1	38	-	-
Glu	7	3	0.429	22	12	0.545
His	1	1	1	4	0	0
Gln	5	2	0.4	6	-	-
Ile	3	3	1	8	6	0.75
Met/iMet	7	-	-	13	-	-
Tyr	4	4	1	5	-	-
Supres	1	-	-	2	-	-
Cys	3	3	1	9	6	0.667
Trp	1	-	-	6	-	-
SelCys	3	-	-	1	-	-
Sum	158	87	11.718	719	370	5.097
Average	7.182	4.833	0.651	32.682	26.429	0.364
Stdev	6.123	4.681	0.317	73.101	74.487	0.303

Table 91. Anticodons identified in DNA toplevel files for *N. vectensis*, *T. kitauei* and *T. adhaerens*.

The formatting is similar in style as Table 89.

Anticodons in DNA toplevel files for metazoan organisms (PART 2)									
	<i>N. vectensis</i>			<i>T. kitauei</i>			<i>T. adhaerens</i>		
	Ac	OC	OC/Ac	Ac	OC	O/A	Ac	OC	OC/Ac
Ala	634	527	0.831	2	2	1	3	1	0.333
Gly	517	173	0.335	7	0	0	4	0	0
Pro	1148	761	0.663	4	2	0.5	3	1	0.333
Thr	448	177	0.395	4	2	0.5	3	1	0.333
Val	284	186	0.655	6	1	0.167	3	0	0
Ser	1083	588	0.543	5	2	0.4	4	2	0.5
Arg	1003	309	0.308	8	4	0.5	6	1	0.167
Leu	1692	600	0.355	6	3	0.5	5	2	0.4
Phe	447	113	0.253	3	3	1	1	1	1
Asn	400	8	0.02	5	5	1	1	1	1
Lys	505	155	0.307	5	3	0.6	2	1	0.5
Asp	208	38	0.183	4	4	1	1	1	1
Glu	981	342	0.343	9	5	0.556	3	2	0.667
His	1180	-	-	2	2	1	1	1	1
Gln	279	87	0.312	2	1	0.5	2	1	0.5
Ile	336	302	0.899	3	0	0	2	1	0.5
Met/iMet	2250	-	-	3	-	-	2	-	-
Tyr	1512	11	0.007	3	3	1	1	1	1
Supres	56	-	-	0	-	-	0	-	-
Cys	343	-	-	2	-	-	2	2	1
Trp	729	-	-	1	-	-	1	-	-
SelCys	1	-	-	2	-	-	1	-	-
Sum	16036	4377	6.413	86	42	10.222	51	20	10.233
Average	728.909	273.563	0.401	3.909	2.471	0.601	2.318	1.111	0.569
Stdev	570.127	232.480	0.257	2.287	1.505	0.350	1.492	0.583	0.355

For *A. queenslandica*, as seen in the unicellular species of this section, deamination events could also be influencing the anticodon results. In this species (as specified in Table 90), overrepresented codon was found for Ile, while direct deamination was evidenced in Pro, Thr, Val, Ser and Arg, all showing from 0.2 to 0.33% of coincidence with optimal codons (except Ser with 80%). In the species *M. leidy* (also in Table 90), deamination was found through overrepresented codons for Ala, Pro, Leu and Ile, while the direct deamination evidenced Pro, Thr, Val, Ser, Arg and Ile. In these amino acids the percentage of coincidence with optimal codons ranges from 0 to 83% for these amino acids. The difference of correlation between these two species for deamination could be explained with the fact that while *A. queenslandica* contain 88% of GC3 in optimal codons for TAPSILVR amino acids, *M. leidy* only contains 36%. Thus, both species can take advantage of deamination to inosine to different extents.

In the cnidarian (both presented in Table 91), 16036 anticodons were identified for *N. vectensis*, with a coincidence for optimal codons of 40.1%. In this species, no amino acids with 0% values were found, and the amino acids with highest anticodon-optimal codon score were Ile (29.9%) and Ala (83.1%), while the lowest scores were found in Tyr (0.7%) and Asn (2%). The other cnidarian species, *T. kitauei*, showed 86 anticodons with an optimal codon coincidence of 60.1%. The amino acids with highest score were Ala, Phe, Asn, Asp, His and Tyr (100%), while the lowest score corresponded to Gly, Ile (0%) and Val (16.7%). Finally, for the placozoan *T. adhaerens* (shown in Table 91), 51 anticodons were identified, showing a coincidence with optimal codons of 56.9%. The amino acids presenting the highest coincidence were Phe, Asn, Asp, His, Tyr and Cys (100%), and the lowest found were Gly, Val (0%) and Arg (16.7%).

The deamination in *N. vectensis* presented the strongest evidence of the species studied. Although only Ser, Leu and Ile were detected for the overrepresented codon analysis, all TAPSILVR amino acids presented direct evidence of deamination. These amino acids also showed the highest anticodon-optimal codon correlation of all the species studied, ranging from 30 to 89.9%. The higher results in this species than in any other, indicating less advantage on using deamination for inosine, is consistent with the low GC3 level in optimal codons for this species (only 6.2%). The data suggest that although deamination is present in *N. vectensis*, it is also less determinant for the efficiency of translation in this species.

For the species *T. kitauei* only overrepresented codons for Pro, Arg and Ile were found, while for *T. adhaeres* overrepresented codons were limited to the amino acid Ile. In both species, indirect deamination was found in amino acids ranging in correspondence anticodons-optimal codons from 0 to 50%. Thus, it is possible that these species showed lower correspondence due to deamination process, but as discussed in 4.3.2, these datasets may not be representative enough to determine the extent of deamination in these organisms.

In conclusion, anticodon abundance levels were shown in this chapter, detailing its correlation with the optimal codons, but also gathering all the results obtained in this chapter that could be influencing the numbers presented. This summary for tRNA and anticodon analysis established evidence for adenosine deamination both directly and indirectly, highlighting its role in the use of optimal codons.

4.3.4 Chimeric tRNA.

Evidence for chimeric tRNA was searched for all the unicellular and metazoan species described in 4.1 using DNA toplevel files. The process was made using tRNAscan-SE 2.0 on its online version as explained in 2.3.4. Next, the list of tRNA identified per species will be presented (Table 92), followed by lists for the isotype-anticodon disagreements found (Tables 93 to 99). This second set of tables contain exclusively the amino acids found with disagreement events and the abnormal anticodons that were found codifying them. Although chimeric tRNA is presented by species, *N. vectensis* was separated in two different lists, due to the large amount of data found for this species the separation could help to clarify if a correlation between chimeric tRNA and highly degenerated TAPSILVR amino acid exists.

The data in Table 92 show the amount of pseudogenes identified, with a clear distinction of tRNA identified as truncated. This table also evidenced that although some sequences were categorised as pseudogenes, the truncation event is a separate phenomenon that could present characteristics of pseudogenes at the same time, thus opening the possibility that pseudogenes could be misidentified, as chimeric tRNA has not been deeply studied before. At the end column of the table the proportion of this tRNA over the total is shown, presenting *M. leidy* the highest score (60.19 %).

Table 92. Types of tRNA detected in unicellular and metazoan species. The analysis was carried out using tRNAscan-SE 2.0 online version. These tables present the amount of all tRNA detected per species (All tRNA), the amount of pure pseudogenes (Pseudo), truncated pseudogenes (T-Pseudo), pure truncated genes (Trunc.), the total number of genes detected with isotype-anticodon disagreement (tRNA-DIS), and the proportion of tRNA-DIS relative to All tRNA (% DIS).

Types of tRNA detected in DNAtoplevel data						
	All tRNA	Pseudo	T-pseudo	Trunc.	tRNA-DIS	% DIS
<i>S. rosetta</i>	122	1	4	4	3	2.46 %
<i>S. arctica</i>	333	32	5	102	91	27.33 %
<i>T. trahens</i>	89	0	0	25	25	28.09 %
<i>A. queenslandica</i>	154	1	32	19	51	33.12 %
<i>M. leidy</i>	746	108	153	303	449	60.19 %
<i>N. vectensis</i>	16193	1109	1021	968	1704	10.52 %
<i>T. kitauei</i>	86	3	0	6	6	6.97 %
<i>T. adhaerens</i>	51	1	0	1	1	1.96 %

In the analysis searching for chimeric tRNA (detailed next in Table 93), the species *S. rosetta* showed a single amino acid presenting isotype-anticodon disagreements (Thr), codified by 3 abnormal anticodons. In the species *S. arctica* 91 disagreements were found, distributed among 11 amino acids, with special relevance of Thr. In the case of *T. trahens*, 25 tRNAs presented disagreements, found in 9 amino acids. These species did not show any preference for anticodons in any of the presented amino acids. The results obtained in this section for unicellular eukaryotes were negatively correlated with the codon bias, as *S. arctica* obtained lower values in analysis like strength of selection (0.846 of statistic $\hat{S}$ on average, while *S. rosetta* and *T. trahens* gave 3.098 and 3.255 respectively) or the results for GC3, CBI, Nc and Fop statistics in 4.2.3.1. An explanation for this phenomenon could be that when the pressure of selection is less efficient in a species, organisms could rely on other mechanisms granting some biological advantage during the translation process, like chimeric tRNA, or in the case that it suppose a negative effect that selection is not effective enough to make these elements disappear.

Table 93. Isotype-anticodon mismatches identified in the species *S. rosetta*, *S. arctica* and *T. trahens*. The table shows the relevant amino acids found with this type of disagreements, and a list of the anticodons that presented the mismatch.

Isotype-anticodon mismatches in unicellular eukaryotes			
Isotype	<i>S. rosetta</i>	<i>S. arctica</i>	<i>T. trahens</i>
Ala	-	ACG, ACC, TGT	CAC
Gly	-	CTG (x2)	CCA
Thr	GCA, TAA, TTT	ACG (x9), CAT, CCT (x2), GTG (x7), GTT, TCC, TCG (x5), TCT (x3)	TCG, CCT, TAT
Val	-	-	TCG, TTG
Arg	-	TAA, CCA	TTT, AGA, TGA, CGA (x2)
Lys	-	-	GAA (x2)
Asp	-	TTC (x5), CTC (x10), GAA (x2)	CTC (x3), TTC
His	-	GCA, CAT (x2)	TCT, CAT (x3)
Gln	-	GTT, GAC	AGT (x3)
Ile	-	CAT (x8), TCT (x2)	-
Met	-	GTT (x10), ACG, TCT, CCA (x5)	-
Cys	-	GAT, TTT (x2), TTG	-
Trp	-	TGT, TGG	-

A. queenslandica (shown next in Table 94) presented 51 isotype disagreements, distributed among 11 amino acids, but with clear results biased towards Thr. This species showed previously the second lowest average for strength of selection (-0.984), no direct evidence for deamination and only one overrepresented codon (for Ile). Also, it was presented evidence for translational efficiency based only on GC3 and Fop statistics. These data are consistent with the hypothesis established previously for *S. arctica*.

Table 94. Isotype-anticodon mismatches identified in the species *A. queenslandica*. The table shows the relevant amino acids found with this type of disagreements, and a list of the anticodons that presented the mismatch.

Isotype-anticodon mismatches in <i>A. queenslandica</i>	
Isotype	Anticodons mismatching
Thr	AAA, AGA (x2), CAA (x6), CAC, CAG (x8), CCG, CTT, GAA, TAC, TAG (x2), TCA, TTG
Val	CAA
Arg	TGC, GCA, TAG
Leu	TGA
Asp	CTC, AGA
His	CCC
Gln	TTC (x2), TCC (x3), GTA
Ile	ACG
Tyr	CAT (x2)
Cys	GTT (x2), GAA (x3), CGT
Trp	GTC, CGG

M. leidy (detailed next in Table 95) exhibited 449 isotype-anticodon disagreements, the second highest number in the species studied, distributed in 18 amino acids, but with a clear bias data towards Thr. This species showed low strength of selection, but the second highest within the metazoan studied (-0.485 on average for statistic $\hat{S}$), with direct evidence of deamination in 6 amino acids and 4 overrepresented codons. In this species translational efficiency evidence was supported by Nc and Fop statistics, while translational accuracy only by GC3 levels. These data are also consistent with the hypothesis established in *A. queenslandica* and *S. arctica*, where this could benefit less efficiently selected organisms or that selection is not efficient enough to avoid its apparition.

Table 95. Isotype-anticodon mismatches identified in the species *M. leidy*. The formatting is similar in style as Table 94.

Isotype-anticodon mismatches in <i>M. leidy</i>	
Isotype	Anticodons mismatching
Ala	ACA, AGA (x2), GGA (x2), TCC (x2), TCT (x2), TTT
Gly	ACA, CCA, CCT (x3), CTC (x3), GGA, TCT (x12), TTC
Pro	CCC (x3), GGT
Thr	AAA (x4), AAG, ACA, AGA (x266), AGC, CAT (x2), CGA, CTA, CTT, GAA (x3), GAC, GCG, GGA (x28), GTC (x2), TAA, TGA (x2), TTC
Val	CAT, CCC, CGT (x2), GGA (x4)
Arg	AGC, CGT (x2), GGA (x2), GGT, TGT
Leu	CTA
Asn	AGA (x4), AGC
Lys	AGA, ATC (x3), GTC (x6), TGG
Asp	CCC, CGA, CTT, GGC
Glu	CTG
His	CGA (x3), CGC, CGT, GCA, TGT
Gln	CGA, TCT
Ile	AGA (x4), CAT, CTC, GAC, GCT, GGT, GTC, TCC (x3), TTC, TTT
Met	ATC
Tyr	AAT, GAA, GGA
Cys	AAC, AGA (x2), ATC (x2), CCC, GCC, GTC (x13), TCC (x7), TTC (x2)
Trp	CAA

Table 96. Isotype-anticodon mismatches identified in the species *N. vectensis* for TAPSILVR amino acids. The formatting is similar in style as Table 94.

Isotype-anticodon mismatches in <i>N. vectensis</i> – TAPSILVR amino acids	
Isotype	Anticodons mismatching
Ala	ACC, AGA, ATC, CAC, CGT (x3), CTC (x2), TAA (x2), TAC, TTC, TTG, TTT
Pro	AAG (x2), AGT, ATG, CGC (x3), CTC (x8), CTG (x2), TGA, TGC (x2), TGT (x2), TTG (x2)
Thr	AAA (x2), ATC, ATT (x2), CAA (x20), CAT (x4), CCC, CTC (x35), CTG, GCT, TAA (x5), TTC (x5), TTG
Val	AAG (x4), AAT (x3), ATC, CAA (x65), CCC, CCT, CGA, CTC (x178), CTG (x9), CTT, GAA, GTC, TAA, TAG, TCT (x19), TTC, TTT
Ser	AAA (x50), ACA, CAA (x10), CAG, CCA (x11), CTC (x2), GAA (x29), GAC, TAA (x6), TAG, TCA (x2), TTA (x2)
Arg	AAC, AAG, ACT, AGA (x2), AGG (x22), AGT (x3), ATG (x2), CAT (x2), CCA (x4), CGT (x2), CTC (x17), GCT (x2), TAA (x3), TAG, TCC, TGG (x29), TGT (x3), TTT (x7)
Leu	AAA (x8), AAC (x2), AAT (x2), CAT (x32), CGA (x7), CTA, CTC, GAA (x3), TAT, TCT (x2), TGA (x3), TTA (x5)
Ile	AAA, GTC, TAC (x3), TCT (x3), TTT

N. vectensis was the species presenting a higher number of results with a large difference (Tables 96 and 97). It presented 1704 isotype-anticodon disagreements, distributed among 19 amino acids and iMet (initiator methionine). Due to the large dataset, no clear trends were found, although amino acids like His and Met presented a particular high number of disagreements. To clarify the interpretation, results were separated by TAPSILVR amino acids (Table 96). A comparison of Table 96 and 97 let us reveal no evidence for chimeric tRNA being more prevalent in highly degenerated amino acids. This species presented the highest strength of selection in metazoan species (-0.006 on average for statistic $\hat{S}$), but in comparison with unicellular species only higher than the same value in *H. balthica* (-0.073). In the deamination analysis, it showed direct deamination for all the 8 amino acids studied, and indirect evidence through overrepresentation in 3 codons. For this species, evidence for translational efficiency was based on CBI and Fop values, while translational accuracy on Nc levels. Although highest than in other metazoan studied, the low value for selection could be promoting the use of alternative strategies like seen in *A. queenslandica* and *M. leidyi*.

Table 97. Isotype-anticodon mismatches identified in the species *N. vectensis* for non-TAPSILVR amino acids. The formatting is similar in style as Table 94.

Isotype-anticodon mismatches in <i>N. vectensis</i> – Non-TAPSILVR amino acids	
Isotype	Anticodons mismatching
Gly	CAA (x8), CAT (x6), CGT (x9), CTC (x80), CTG (x2), CTT (x3), GAC (x3), GCT (x9), GTC, TAC (x2), TCA (x3), TCT (x16), TTC (x8), TTT
Phe	GAT (x3), GTA, TAA (x2)
Asn	ATC, TCT (x2), GTC, CAA, ACC, TCT, TTC (x7), CAT, TCT, CGT (x5), GCT (x2), GTG, ATA, GTA, TCT, GTG, CTC, TTT, TCT, CTT, GGT, TCT (x10)
Lys	AAT (x2), CAT, CTC, CTG, GTA (x3), GTG (x13), GTT (x2), TTC
Asp	CTC (x13)
Glu	CAG, CAT (x3), CGT (x5), CTT, GTC (x7), TAA, TGT, TTA (x3), TTT (x3)
His	AAA, CTC (x7), CTG (x119), GAG (x30), GCG (x6), GGG (x3), GTA (x3), GTT (x46), TCT (x3), TGG, TGT, TTG (x5)
Gln	CAG, CAT, CGT, CTA (x6), CTC, CTT (x2), GCA, GTA (x13), GTG, TGT (x2), TTA (x3), TTC (x2), TTT (x13)
iMet	AAT (x7), CAA (x6), TAT
Met	AAT (x14), AGA, AGC (x16), AGT (x4), ATA (x8), ATC (x32), ATT, CAA (x11), CAC (x8), CAG, CCT (x5), CGT, CTT (x9), GAA, GAC, GAT (x29), GTG (x2), TAC, TAG (x3), TAT (x39), TCA, TCG (x3), TGC, TGT (x2)
Tyr	AAA, AGA (x6), CAT, CTA (x2), GAA (x5), GCA (x9), GGG, GTG (x4), GTT (x2), TAG (x7), TAT (x7), TCA, TGG (x2), TGT (x2), TTA (x9)
Cys	CAA (x7), CAT, CGG, CGT (x5), CTA, CTC (x65), CTG (x3), GCG, GTA (x2), GTC, TTC (x3)
Trp	AAA (x42), ACA, ACG, CAA (x11), CCC (x4), CCT (x2), CGA (x2), CTA (x2), CTC (x18), CTG (x3), GCA (x2), GTA (x3), GTG (x2), TAA (x2), TCA (x11), TCC, TCT, TTA

Table 98. Isotype-anticodon mismatches identified in the species *T. kitauei*. The formatting is similar in style as Table 94.

Isotype-anticodon mismatches in <i>T. kitauei</i>	
Isotype	Anticodons mismatching
Thr	GAA, TCT
Glu	GTC
His	CCC
Cys	GTA (x2)

The species *T. kitauei* (detailed in Table 98) presented only 6 anticodon disagreements, distributed among 4 amino acids and with no particular trend. This species presented one of the lowest number of mismatches identified despite being the metazoan showing lowest strength of selection (-1.159 for statistic $\hat{S}$ on average), presenting three overrepresented codons and no direct evidence of deamination. As this species showed evidence both for translational accuracy and efficiency supported by GC3 and Fop levels, *T. kitauei* either could rely less on chimeric tRNA than the previous metazoan described, or these events have not been reflected in the sequencing carried out in this organism.

Table 99. Isotype-anticodon mismatches identified in the species *T. adhaerens*. The formatting is similar in style as Table 94.

Isotype-anticodon mismatches in <i>T. adhaerens</i>	
Isotype	Anticodons mismatching
Cys	CCG

T. adhaerens (shown in Table 99) presented only 1 isotype-anticodon disagreement, as in *S. rosetta*, although found in the amino acid Cys. This species presented a strength of selection slightly lower to that in *A. queenslandica* (-0.948 compared to -0.984 on average for statistic $\hat{S}$), although this species presented 51 disagreements. *T. adhaerens* only showed 1 overrepresented codon, no direct evidence of deamination, and evidence for selection upon translational efficiency supported by GC3 and Fop values.

The isotype-anticodon analysis was revealed for *T. adhaerens* not consistent with the results found in species under similar selection or related species, although as theorised for *T. kitauei*, this metazoan could be relying less on chimeric tRNA, or the sequencing carried out in this organism could not be suitable to evidence this events.

As a conclusion, all the species studied presented some level of tRNA isotype-anticodon disagreement. Unicellular species presented less results than metazoan organisms, where the use of chimeric tRNA could be a strategy to increase their translational advantage, compensating a lower efficiency in other areas like codon usage bias. Although trends in this analysis are difficult to infer and presented no correlation with highly degenerated amino acids, Thr resulted particularly relevant for most species, as *S. rosetta* only presented this amino acid as disagreement, 3 species contain it in highest proportion than for any other amino acid (*S. arctica*, *A. queenslandica* and *M. leidy*), and other 3 species present it in similar amount to other amino acids with high results (*T. trahens*, *N. vectensis* and *T. kitauei*). Thus, chimeric tRNA for this amino acid was relevant in all the species studied except for *T. adhaerens*. Results in *T. kitauei* and *T. adhaerens* must be taken carefully, as the sequencing carried out in this species is may not be reflecting the non-functional RNA correctly and the levels of chimeric tRNA could be much higher. Also, the numerous results in *N. vectensis* in comparison with the rest of the species must be highlighted, as this species could be using different evolutionary advantages that could explain its big divergence from related species. Overall, chimeric tRNA could be playing a role as one of the strategies increasing the translational advantage and influencing the evolution of organisms, possibly increasing its importance when natural selection over codon bias is less effective.

4.4 FINAL RESULTS OVERVIEW.

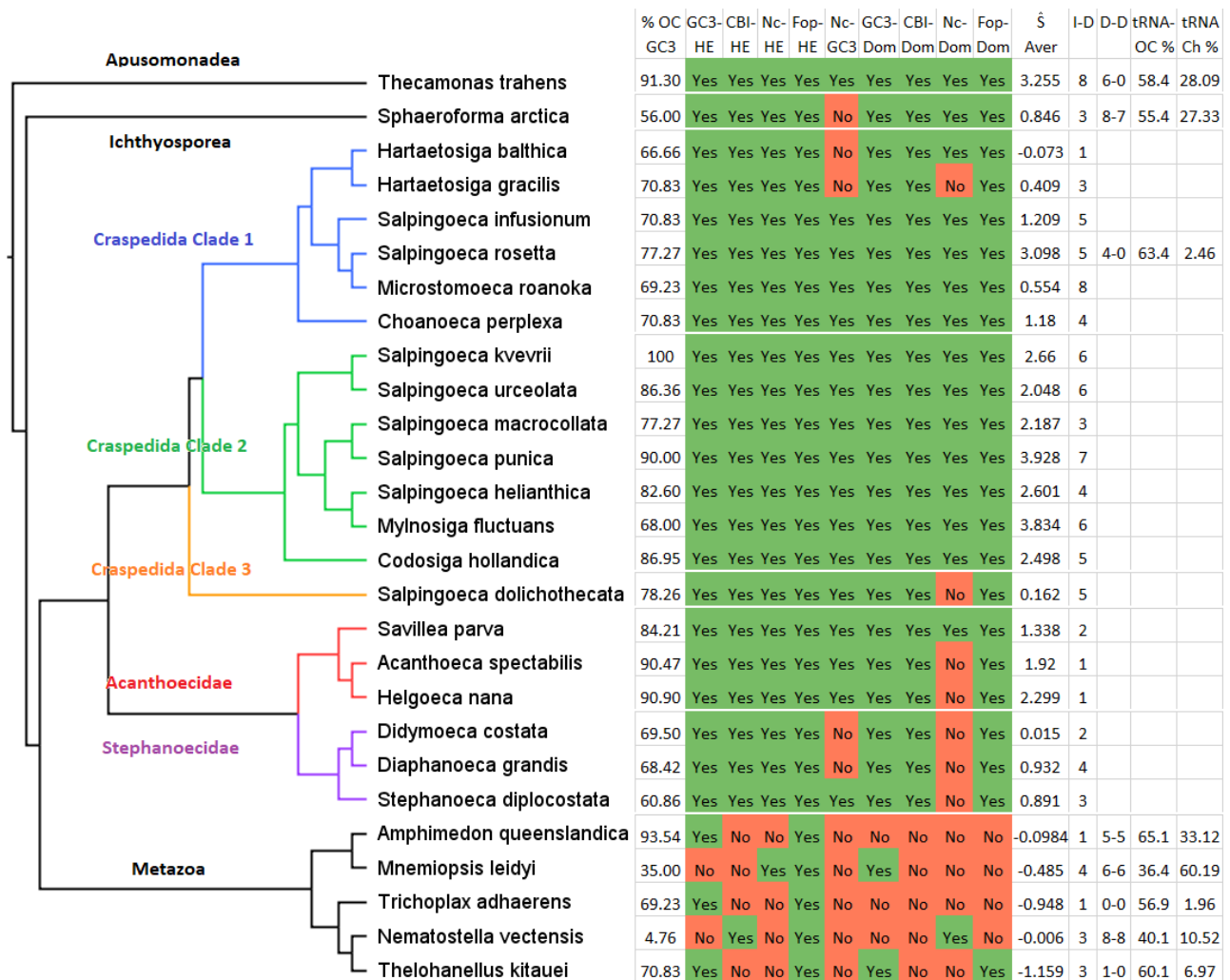


Figure 36. Final phylogeny with main results. The phylogeny was built in NewickTrees format and visualised using FigTree v.1.4.4, based on phylogenetic relationships previously established (Philippe et al., 2009; Derelle et al., 2015; and Carr et al., 2017). The choanoflagellates groups have been differentiated with different colours. Abbreviations stand for: percentage of optimal codons (% OC), codons ended in G or C (GC3), codon bias index (CBI), effective number of codons (Nc), frequency of optimal codons (Fop), status of the presented parameter highest in highly expressed genes (-HE), parameter highest in domains over non-domains of the predicted proteins (-Dom), clear correlation of Nc-GC3 in plots (Nc-GC3), average for statistic $\hat{S}$ ($\hat{S}$ Aver), number of deaminated amino acid found indirectly (D-I), number of deaminated amino acids found directly, with tRNA unmodified (left) and modified (right) (DD), percentage of tRNA found corresponding to optimal codons (tRNA-OC %), percentage of chimeric tRNA detected over the total tRNA (tRNA Chim %).

As a final word from the sections relative to results, Figure 36 shows the species studied in a phylogeny, together with some relevant element of all the sections addressed. Especially in choanoflagellates, we can appreciate the diversity found within Craspedida Clade 1, the high selection pressure in Craspedida Clade 2, its big differences with Craspedida Clade 3, the low selective pressure in Stephanoecidae, and how this differs from its closes Clade Acanthoecida, showing a selective pressure more close to Craspedida Clade 2. In the results we can also appreciate the differences between *S. arctica* and *T. trahens*, acting as outgroup of choanoflagellates within and without Opisthokonta respectively, and the contrast of unicellular and pluricellular species, showing also the two Cnidarian species (*N. vectensis* and *T. kitauei*) many differences between them.

The thorough description and summaries for the results along Chapters 3 and 4, plus their graphic representation in this section, led us to finally discuss the implications of this findings in the next chapter, addressing also its impact in future research.

CHAPTER 5. DISCUSSION AND CONCLUSIONS

This chapter aims to discuss the results presented in Chapters 3 and 4, highlighting its importance for the scientific community and extracting the main conclusions of this research. The first section will be focused on discussing the data sources from Chapters 3 and 4 and the treated data from choanoflagellates (5.1). This is followed by the results obtained from the coding region analysis in all unicellular and metazoan species (5.2) that contained the mapping of reads for relative expression, determination of optimal codons, codon usage, and strength of selection). The third section will address the results for the anticodon and tRNA analysis (5.3), relative to the evidence for tRNA deamination through anticodon quantification and transcriptome screening, differential anticodon abundance and chimeric tRNA. To finalise the chapter, a section discussing the impact (5.4) and the main conclusions of this research (5.5) will be presented.

5.1 DATA SOURCES FOR EUKARYOTE ORGANISMS AND TREATED DATA OF CHOANOFLLAGELLATES.

The choanoflagellates sequencing data from Table 01, which were the base for Chapter 3 analysis, belonged to a single sequencing project (Richter et al., 2018). These data files, after a filtering for contamination, were deemed reliable as the BUSCO analysis gave a representation of conserved eukaryotic genes over 88.1 for all the species. Although this analysis showed evidence for the completeness of the dataset and was tested against bacterial and cross contamination, evidence showed through Chapter 3 support the hypothesis of contaminated CDS files.

Initially, the difficulties of sequencing choanoflagellates without bacterial contamination were addressed in 1.5.4, where it was explained that bacteria are choanoflagellates prey. Bacteria are added to the medium as a food source necessary for choanoflagellates survival and multiplication, but during their separation the bacteria could be attached to the cell surfaces, or could be incorporated into the choanoflagellates through their feeding, remaining within vacuoles, which still can contain DNA and RNA, contaminating the early steps of sequencing. Although antibiotic treatment was applied in different steps, this could have not eliminated the bacteria in vacuoles.

Thus, transcriptome size calculated for choanoflagellates did not correlate with the number of CDS, presenting a 17.5% of variation between two isolate of species that were deemed identical in protein content, and were not consistent with the choanoflagellates genome size published for the species *S. rosetta* and *M. brevicollis* (King et al., 2008; Fairclough et al., 2013). It is improbable that the size of most transcriptomes sequenced were much bigger than the whole genome of the most reliable choanoflagellates sequenced until date (55.4 and 41.6 Mb respectively for the cited genomes, in comparison with the transcriptome of *S. diplocostata* Australia isolate, showing 217.04 Mb).

During the screening process performed in 2.1.3, a portion of the sequences were identified by databases like Ecofind as previously described exclusively in bacteria, which were completely discarded. In the optimal codon identification, CodonW seemed to fail to predict results correctly due to the dataset more than to the proper methodology in some species. Thus, the prediction accuracy increased post-screening considerably in species like *D. grandis*, *D. costata* and *H. gracilis*. Beyond methodology discussions, optimal codons post-screening showed a higher consistency for optimal codons ended in GC, as was determined for *S. rosetta*, *M. brevicollis* and the filasterean *C. owczarzaki* in previous publications (Southworth et al., 2018), and suggesting that this could be a conserved ancestral trait.

For the statistics calculated in 3.2.3.1 considering expression, and in 3.2.3.3 based on predicted protein domain and non-domain regions, species showing a strong codon bias showed more consistent and expected results post-screening. In the same way, in 3.2.3.2, GC3-Nc correlations fail in the original datasets for most species (except for *M. fluctuans*), while it was clear in sets post screening (except in the *Hartaetosiga* genus). In this same section, a double cloud of dots in species of the family Acanthoecidae (Figures 08, 15 and 25) and in *M. roanoka* (Figure 16) reflected the presence of groups of genes with different characteristics, which could be a subset of genes within the same species or product of contamination (all diverging sets of genes were eliminated post-screening). Finally, the strength of selection was more consistent with the analysis performed post-screening, correlating with the position of the species across the phylogenetic tree. As all the comparisons described here have been detailed meticulously in Chapter 3, and has been summarised, there is enough evidence and consistency through the screened data to support the hypothesis of bacterial contamination in Richter's filtered CDS.

The transcriptome screening performed in this project resulted in a fraction of the initial CDS, ranging from 2845 to 4079 genes (while in *S. rosetta* 11731 genes were analysed). Nonetheless, using only proven eukaryotic genes to avoid bacterial originated ones is essential to avoid interference and statistical biases. For instance, this is reflected in a reduction of the double cloud of dots in the GC3-Nc plots seen in 3.2.3.2, or in the correlation of post-screening results with phylogenetic positions. The strict filtering could also remove sequences of choanoflagellates origin due to the fact that no similar protein could have never been previously identified in any eukaryotic organisms. This balance between losing genuine sequences and avoiding contamination was necessary due to the direct identification of an array of genes described before exclusively for bacteria. Some of these bacterial genes also could have originated from horizontal transfer events, but as those hypothetical genes are indistinguishable from bacterial ones, finding a method to improve sequencing would be necessary to determine those genes. Although some proteins may not have been previously described, the protein family classification was required for the process of identifying protein origins, avoiding sequences that may have their origin in the choanoflagellates prey. Screening therefore, was vital in removing dubious transcripts from bacterial prey.

The stability presented in the number of filtered coding regions could be due to different reasons: it could be a product of removing bias from contamination, showing that all genomes sizes could be more similar in size to each other than the variability displayed; it also could be due to a correspondence of the genes identified with a limited number of proteins identified in other organisms, keeping during the filtering a set of similar conserved genes among the different choanoflagellates; or it could be a combination of both situations. Similar genes between choanoflagellates and metazoan could include novel genes in these organisms, as the number on both was close, around 2463 for the first one and 1944 for metazoan (Richter et al., 2018), thus some genes may be shared by horizontal gene transfer events, as choanoflagellates have also received genes from bacteria and algae (Tucker, 2013). Independently of the origin, the legitimacy of the sequences and transcripts must be tested, as in genuine horizontal transfer events from bacteria, a small amount of transcripts from the prey could overrepresent these sequences in an expression analysis of the host organism.

5.2 CODING REGIONS ANALYSIS.

Optimal codons for choanoflagellates have been determined only in *S. rosetta* and *M. brevicollis* (Southworth et al., 2018), both species belonging to Craspedida Clade 1. The optimal codons determined for *S. rosetta* in this work presented some variations from the codon identified previously. In previous research 23 optimal codons have been identified for *S. rosetta* (Southworth et al., 2018), while in this project only 22 were determined (section 4.2.2), sharing 19 codons with the previous analysis. The differences were the lack of CUG (Leu), UCG (Ser), CCG (Pro) and ACG (Thr), the addition of CCU (Pro), CGU (Arg) and GGU (Gly). Despite these differences, all optimal codons complementing the major tRNA genes determined by Southworth et al. (2018) were also identified in this project. The origin of these discrepancies is methodological, while in the present study genes are classified based on expression, Southworth analysis classified genes based on Nc levels.

The species *M. brevicollis* was analysed exclusively by correspondence analysis due to the absence of expression levels, relying on the high accuracy of this method presented in the same work with only 2 disagreements in *S. rosetta* between methodologies. Also, assuming that in a similar way most choanoflagellates post-screening and its sister species *C. perplexa*, presented a low number of false positives in this method (1 pre and 3 post-screening for *C. perplexa*, with no false negatives), these data also will be considered. *M. brevicollis* presented 18 optimal codons shared with its sister species (of 24 for this species), suggesting a divergence in optimal codon selection for these species. Nonetheless, while *S. rosetta* from Southworth et al. (2018) contained the most frequent optimal codons across choanoflagellates in 16 amino acids, *M. brevicollis* contained those codons for all the amino acids, and their GC3 content were 90.9% and 87.5% respectively, higher than the 77% of the *S. rosetta* of this study and the 70.84% of *C. perplexa*. GC3-Nc plots in both *S. rosetta* were very similar, showing a clear correlation between those two parameters, but in this project the approximately 200 genes that Southworth identified as false annotations were not found.

In other previous analysis of choanoflagellates, Nc values were determined in a broad range of species using four genes (Carr et al., pers. commun. See figure 06). This analysis offered a first broad view on choanoflagellates bias, although results were not completely coincident with the data of this project using thousands of genes. Thus, species like *S. infusionum* were revealed in Chapter 3 with a higher level of overall codon bias (from around 50 in Figure 06 to 40.4 in HE genes), while species like *S. urceolata* showed a lower Nc value (from less than 30 in Figure 06 to 41.3 in HE genes).

The codon usage has been studied also in other groups of organisms in Opisthokonta, with higher focus in metazoans and fungi. In fungi, species like *A. nidulans* showed optimal codons enriched in GC3 (Lloyd and Sharp, 1991), and in metazoan, GC3 was correlated with expression levels in species like *T. pisiformis* (Chen et al., 2013) (see 1.4.4 for a review with more examples). In this project, although most metazoans showed an increment of GC3 in optimal codons, this disappeared in the cnidarian *N. vectensis* (to 6.2%) and in the ctenophoran *M. leidy*. This divergence revealed that the poriferan are more closely to the ancestral state in the pluricellular transition than ctenophora in optimal codon use. Codon bias towards GC ending codons appear very conserved through Opisthokonta, but the related organism of this project, *T. trahens*, revealed that this conserved trait could be traced even before the divergence between Opisthokonta and Apusomonadea. Although the analysis presented here showed how codon usage have been selected in unicellular organisms prior to the transition to pluricellularity, these data also suggest that a similar selection could be present in related groups to Opisthokonta.

Evidence for natural selection upon translational efficiency and accuracy have been showed in all the unicellular eukaryotes studied, even in those presenting the weakest bias. This has been reflected through the parameters GC3, CBI, Nc and Fop, in highly expressed genes against the whole set of CDS and lowest expressed genes, for the first statement, and in the study of domains against non-domains for the second one. Results showing selection upon translational accuracy were present even in lowest expressed genes of all the unicellular studied except for *D. grandis*, *S. dolichothecata* and *S. parva*. Selection for translational accuracy in low codon biased genes was also found previously in choanoflagellates, filasterean (Southworth et al., 2018) and *E. coli* (Yannai et al., 2018). Codon bias data were in agreement with the determination of the strength of selection for the two-degenerated optimal codons, which distributed across the phylogenetic tree

that revealed trends for the different clades. The highest biased organisms studied belonged to Craspedida Clade 2, while the lowest biased were found in the family Stephanoecidae, the Craspedida Clade 3 showed an intermediate bias more close to that of the family Acanthoecidae, and Craspedida Clade 1 showed species with diversity of traits. Within Craspedida Clade 1 are included both species with high codon bias, like *S. rosetta*, and species with low codon bias, like the *Hartaetosiga* genus. An explanation for the differences in this selection could be based on differences in the effective population size across species, although there is not enough ecological or population data to clarify the main cause of this variation.

In metazoans similar results were obtained showing natural selection upon translational efficiency for all the species, but with comparatively lower values both for codon usage bias statistics and strength of selection. *A. queenslandica*, *T. kitauei* and *T. adhaerens* showed a similar effect to that seen in the fungus *A. nidulans* (Lloyd and Sharp, 1991), where all GC content mean is less than 50%, but GC3 was still increased in optimal codons. Results in *A. queenslandica*, *M. leidy*, *T. kitauei* and *N. vectensis* also can be compared to the effect seen in Nematoda and Cestoda. In Nematoda 32 species revealed evidence for natural selection, but also the general GC composition appeared driving codon bias (Marais and Duret, 2001; Mitreva et al., 2006), while in Cestoda similar results were showed in species of the genus *Taenia* (Chent et al., 2003; Yang et al., 2014), the same correlation found between GC and expression in the aforementioned species.

Evidence for selection upon translational accuracy was detected poorly and only in *M. leidy*, *N. vectensis* and *T. kitauei*. Nonetheless, this kind of selection in metazoan can be still valid despite the lower bias addressed, as it has been seen in phylum like the mentioned Nematoda (Marais and Duret, 2001; Mitreva et al., 2006). Thus, translational efficiency and accuracy appeared in this project as a major driver of codon usage bias in choanoflagellates, ichthyosporeans and an apusomonadean, but only translational efficiency as a major driver in metazoan. Metazoans are considered to have smaller population size than unicellular organisms, and also longer generation times (Thomas et al., 2010), two factors able to have reduced the effect of natural selection in these organisms.

5.3 ANTICODON AND tRNA ANALYSIS.

In the present study, analysis to detect deamination of adenosine to inosine in high-degenerated TAPSILVR amino acids was carried out. Due to the limitation of the data available for most choanoflagellates species, only indirect evidence of these events were revealed in Chapter 3. Nonetheless, previously direct evidence was found in *S. rosetta* (Southworth et al., 2018). The amino acid showing deamination indirectly, more consistently across unicellular organisms was Ala, detected in all the species except *H. balthica*, *H. nana*, and *S. arctica*, but GCC was detected as optimal codon in all of them, including the ones not presenting deamination. In metazoan also indirect evidence was found, with Ile as the most common result, as it was present in all the species. The codon AUC for Ile was the optimal codon for all the metazoans except for *N. vectensis*, which also showed this codon overrepresented. Direct deamination evidence was found in *S. arctica*, *A. queenslandica*, *M. leidy* and *N. vectensis*. Previous work suggested that deamination of tRNA appeared early in Holozoa (Porath et al., 2017, Rafels-Ybern et al., 2017; Southworth et al., 2017), and this is supported by the direct evidence found, but the indirect evidence in the organisms studied in this work suggest its apparition even before the divergence of Opisthokonta and Apusomonadida.

In relation with deamination, the tRNA copy number analysis was studied and revealed some species like *A. queenslandica* not showing a high percentage of anticodons matching the optimal codons in TAPSILVR amino acids, while it contains a high GC3 level in its optimal codons. On the contrary, the percentage in these amino acids were higher in species with lower GC3 like *N. vectensis*, suggesting that these species rely on tRNA deamination events to increase the optimal codon use. These data suggest that species with higher level of bias take more advantage of the ambiguity of recognition due to deamination to improve efficiency through optimal codons.

Finally, chimeric tRNA was detected in all the species of Chapter 4, which means *S. rosetta*, apusomonadean, ichthyosporean and metazoan. No particular correlation has been found between isotype-anticodon disagreements and data on deamination. Thus, the significance of these elements have been speculated in 4.3.4. If these elements are genuine, either could appear granting some advantage in the translation process, or could be a disadvantageous element that remain due to the different efficiency of selection in various species. If the elements could be

artefacts, this could help to improve to identify the source of the problem and improve the analysis on tRNA elements. In *E. coli*, chimeric tRNAs have been shown the cause for the higher expression in certain genes (Zhenlin et al., 2017), while in human it has been related with HIV inhibition, suggesting a relation of this mechanism with cell defence (Zongli et al., 2002). The knowledge on chimeric tRNA has resulted useful for the development of techniques for the introduction of random mutations in proteins (Geslain et al., 2010) and the incorporation of unnatural amino acids (Ding et al., 2020).

5.4 IMPACT OF THIS PROJECT IN FUTURE RESEARCH.

This project has evidenced the remaining difficulty on choanoflagellates sequencing. Despite a thorough physical and bioinformatic cleaning carried out in the 19 species studied in Chapter 3, there is solid evidence that the CDS obtained were contaminated. In Chapter 2 a method for filtering transcriptomes to obtain reliable sequences for eukaryotic organisms has been proposed, although not exempt from flaws. New strategies for choanoflagellates sequencing are needed to refine the current knowledge on their genomes and transcriptomes, and finding a methodology to elucidate the genuine horizontal transfer genes and transcripts, as the presence of a few dozens of abundant false genes can produce bias like a complete misidentification of optimal codons.

A technique that could solve the presented problems about contamination is the selective whole genome amplification (SWGA), that is based on amplifying genomes using sequences motifs characteristics of the target species, but not of the background organisms (Leichty and Brisson, 2014). This technique could address specifically the problem of bacteria remaining in choanoflagellates vacuoles, not considered in methods like single-cell sequencing (López-Escardó et al., 2017; 2019). This work could evidence the difficulty of sequencing new species, but also the fact that apparently solid protocols for this process must not remain unquestioned.

The analysis on codon usage across a broad range of choanoflagellates, an apusomonadean, an ichthyosporean and early branching metazoan species gave insight into the evolution of these organisms. Conserved characteristics identified like preferred optimal codon with high GC3 content in unicellular eukaryotes possibly could be traced back even before the divergence between Apusomonadida and Opisthokonta, which would require a new investigation of related groups to Eukaryotes. This next research step could be carried out with the software RASP, a tool to reconstruct ancestral states (Yu et al., 2015), that although it is primarily used in biogeography, functions like Tree View can be used to create pie charts with internal nodes, showing the probabilities of the ancestral traits.

The data obtained on codon usage bias, besides offering evolutionary information about the genome and population of the species studies, can benefit future software approaches and applications in biotechnology. CodonW predictions were reliable in *S. rosetta* and the apusomonadean, but had a poor performance in the ichthyosporean and metazoan studied (even in the earliest branching organisms), setting a precedent of the groups that could be analysed in the absence of relative expression.

The biotechnology applications of codon bias branch from the knowledge established about controlling the translation process and the importance of deamination, evidenced in all the species studied, and important for the efficiency of TAPSLVR amino acids translation in all the species with major tRNA presenting adenosine at the wobble position in these amino acids. Thus, a next step considering these results is the correction of optimal codon and deamination over sequences, in projects like heterologous expression involving these or related species. Any investigation producing transgenic organisms or focused on efficiency of protein production must consider the characteristics of the species used. The consequence of not considering optimal codons or deamination could be a poor performance in translation by a limitation produced by the primarily consume of the less abundant tRNAs, while the overuse of these codons could cause errors due to a protein misfolding produced by an excess of speed in key moments of translation (Andersson and Kurland, 1990; Drummond and Wilke, 2008; Sharp et al., 2010; Novoa et al., 2012; Zhao et al., 2017; Lyu and Liu, 2020). Thus, understanding codon usage bias is a key stage on any research involving protein production, including synthetic biology.

Processes affecting the translation of proteins have been identified in all the organisms studied. Deamination events of adenosine into inosine in tRNA wobble position have been evidenced indirectly in all the species studied and directly in *S. arctica*, *A. queenlandica*, *M. leidy* and *N. vectensis*. Nonetheless, indirect evidence could be benefited from a direct research, and considering the nature of the sequencing used, direct evidence results are likely to be incomplete. Transcriptome sequencing of cDNA not enriched for non-functional tRNA, like in the sequences available for *T. kitauei* and *T. adhaerens*, represents poorly this kind of elements (although for other species, NCBI also gathers data for non-coding RNA and small RNA). An evidence of the lack of completion of deamination identification in unicellular organisms can be seen in the deamination discovered previously in *S. rosetta* (Southworth et al., 2018), and could also be expected in metazoan due to the frequency of these events across the clade (Porath et al., 2017; Rafels-Ybern et al., 2017). Thus, the analysis based on a reliable sequencing enriched for non-functional RNA elements can clarify the further existence of adenosine deamination in these organisms.

Another event that may affect the translation and that have been evidenced in all the species studied is the chimeric tRNA. Although these results could be a product of pseudogenes detection, errors during sequencing or even artefacts of the software, some characteristics like the lack of pseudogenes traits in many of these sequences, its prevalence and its variation across organisms could suggest that these elements have another function. Hypothetically, chimeric tRNA either must grant some advantage during translation, or could be correlated with the efficiency of selective pressure and the ability of the organisms to overcome the apparition of tRNA with abnormal anticodons. In the case that these elements are artefact, those can be used to detect its source and improve the tRNA sequencing or identification, and correcting data that could bias tRNA analysis. These hypothesis must be tested in future research focused on identifying the origin and significance of isotype-anticodons disagreements, making use of directed sequencing like SWGA to avoid contamination, and also of an enrichment for non-functional RNA.

5.5 CONCLUSIONS.

The screening of choanoflagellates transcriptomes published *de novo* allowed a more reliable analysis of the codon usage characteristics of this group. The transcriptome analyses carried out in this project support the importance of natural selection on codon usage evolution in the Opisthokonta group. These data also help to elucidate the changes in codon evolution strategies that appeared as a consequence of the pluricellular transition, when the effect of natural selection over codon bias, although present, became less apparent. Codon usage bias presented a preference for certain optimal codons ending in GC, that were consistent across choanoflagellates and other unicellular eukaryotes, with this ancestral trait retained in metazoans like poriferans. Results revealed high GC3 content in optimal codons selected in unicellular eukaryotes, favoured by natural selection upon both translational accuracy and efficiency even in the less codon biased species. In metazoan this preference was lost in organisms like ctenophorans, and in cnidarians of the class Anthozoa. Nonetheless, although with less strength than in unicellular organisms, evidence for selection upon translational efficiency was relevant in all the metazoan studied, with a lower evidence for selection on translational accuracy for *M. leidy*, *N. vectensis* and *T. kitauei*, and no evidence in other metazoans.

There is evidence showing that the translation process can be influenced by other elements related but not limited to the codon usage. The deamination of adenosine into inosine in tRNA molecules is important in eukaryotes. Indirect evidence of deamination in the high degenerated TAPSILVR amino acids was found in all the species studied, while direct evidence was found in *S. arctica*, *A. queenslandica*, *M. leidy* and *N. vectensis*. Additionally, tRNAs presenting isotype-anticodon disagreements were found in apusomonadean, ichthyosporean, metazoan species and in *S. rosetta*. Deamination events of adenosine in tRNA molecules, and possibly chimeric tRNA appeared early in evolution, before the divergence of Apusomonadea and Opisthokonta. Both mechanisms, hypothetically in the second case, have granted flexibility to tRNA molecules to benefit the use of optimal codons through evolution until the apparition of the latest metazoan organisms.

Both in unicellular and pluricellular organisms, the population size appeared influencing the natural selection. In choanoflagellates the levels of codon bias are correlated with its phylogenetic position, finding the effect of a more efficient selection in Craspedida Clade 2, while this was lower for Stephanoecidae and the *Hartaetosiga* genus. In metazoans, the efficiency of selection has been reduced in comparison with all the unicellular species studied, likely as an effect of the reduced effective population size in these organisms. It is believed that the cell division rates in metazoan, due to an increase of the generation times, also reduced the pressure of selection upon codon usage (Thomas et al., 2010).

In general, the Opisthokonta group survival during its evolutionary history has relied on different strategies that granted advantages during the translation process, the most patent one and shaping the genome is the codon usage bias, but also deamination and possible chimeric tRNA are strategies seen from the early branching of Opisthokonta to metazoan organisms. The evidence presented in this project could benefit both the unveiling of further data on the evolution of Opisthokonta, and biotechnology with an improvement of the translation control over heterologous and synthetic proteins. Furthermore, the deeper knowledge of the nuances governing the genetic code could enlighten our understanding of the very origin of life.

BIBLIOGRAPHY

- Abel, P. (1964, January). Evidence for the Universality of the Genetic Code. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 29, pp. 185-187). Cold Spring Harbor Laboratory Press.
- Adl, S. M., Simpson, A. G., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., ... & Taylor, M. F. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology*, 52(5), 399-451.
- Adl, S. M., Simpson, A. G., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., ... & Heiss, A. (2012). The revised classification of eukaryotes. *Journal of eukaryotic microbiology*, 59(5), 429-514.
- Adler, F. (1951). Yates' correction and the statisticians. *Journal of the American Statistical Association*, 46(256), 490-501.
- Agris, P. F. (1996). The importance of being modified: roles of modified nucleosides and Mg²⁺ in RNA structure and function. *Progress in nucleic acid research and molecular biology*, 53, 79-129.
- Agris, Paul F. "Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications." *EMBO reports* 9.7 (2008): 629-635.
- Agris, P. F., Vendeix, F. A., & Graham, W. D. (2007). tRNA's wobble decoding of the genome: 40 years of modification. *Journal of molecular biology*, 366(1), 1-13.
- Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics*, 136(3), 927-935.
- Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics*, 139(2), 1067-1076.
- Alazami, A. M., Hijazi, H., Al-Dosari, M. S., Shaheen, R., Hashem, A., Aldahmesh, M. A., ... & Alkuraya, F. S. (2013). Mutation in ADAT3, encoding adenosine deaminase acting on transfer RNA, causes intellectual disability and strabismus. *Journal of medical genetics*, 50(7), 425-430.

- Allen, W. V., & Ponnamperna, C. (1967). A possible prebiotic synthesis of monocarboxylic acids. *Biosystems*, 1(1), 24-28.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403-410.
- Anderson, N. G. (1970). Evolutionary significance of virus infection. *Nature*, 227(5265), 1346-1347.
- Andersson, S. G., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2), 198-210.
- Arndt, N. T., & Nisbet, E. G. (2012). Processes on the young Earth and the habitats of early life. *Annual Review of Earth and Planetary Sciences*, 40, 521-549.
- Atkins, J. F., & Bjork, G. R. (2009). A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. *Microbiology and Molecular Biology Reviews*, 73(1), 178-210.
- Bada, J. L. (2004). How life began on Earth: a status report. *Earth and Planetary Science Letters*, 226(1-2), 1-15.
- Badet, T., Peyraud, R., Mbengue, M., Navaud, O., Derbyshire, M., Oliver, R. P., ... & Raffaele, S. (2017). Codon optimization underpins generalist parasitism in fungi. *Elife*, 6, e22472.
- Baldauf, S. L., & Palmer, J. D. (1993). Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proceedings of the National Academy of Sciences*, 90(24), 11558-11562.
- Baldauf, S., Romeralo, M., & Carr, M. (2013). The evolutionary origin of animals and fungi. In *Evolution from the Galapagos* (pp. 73-106). Springer, New York, NY.
- Barbhuiya, R. I., Uddin, A., & Chakraborty, S. (2020). Codon usage pattern and its influencing factors for mitochondrial CO genes among different classes of Arthropoda. *Mitochondrial DNA Part A*, 31(8), 313-326.
- Baross, J. A., & Hoffman, S. E. (1985). Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. *Origins of Life and Evolution of the Biosphere*, 15(4), 327-345.
- Barrell, B. G., Bankier, A. T., & Drouin, J. (1979). A different genetic code in human mitochondria. *Nature*, 282(5735), 189-194.

- Bartel, D. P., & Chen, C. Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics*, 5(5), 396-400.
- Bass, B. L., & Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell*, 55(6), 1089-1098.
- Becker, H. F., Corda, Y., Mathews, M. B., Fourrey, J. L., & Grosjean, H. (1999). Inosine and N1-methylinosine within a synthetic oligomer mimicking the anticodon loop of human tRNA^{Ala} are major epitopes for anti-PL-12 myositis autoantibodies. *Rna*, 5(7), 865-875.
- Beckmann, A., & Özbek, S. (2012). The nematocyst: a molecular map of the cnidarian stinging organelle. *International Journal of Developmental Biology*, 56(6-7-8), 577-582.
- Benner, S. A., Ellington, A. D., & Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences*, 86(18), 7054-7058.
- Bennetzen, J. L., & Hall, B. D. (1982). Codon selection in yeast. *Journal of Biological Chemistry*, 257(6), 3026-3031.
- Benzécri, J. P. (1992). *Correspondence analysis handbook*. CRC Press LLC.
- Bernhardt, H. S., & Tate, W. P. (2010). The transition from noncoded to coded protein synthesis: did coding mRNAs arise from stability-enhancing binding partners to tRNA?. *Biology Direct*, 5(1), 1-18.
- Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., ... & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6), 1111-1120.
- Bessho, Y., Hodgson, D. R., & Suga, H. (2002). A tRNA aminoacylation system for non-natural amino acids based on a programmable ribozyme. *Nature biotechnology*, 20(7), 723-728.
- Bianconi, G., Zhao, K., Chen, I. A., & Nowak, M. A. (2013). Selection for replicases in protocells. *PLoS computational biology*, 9(5).

- Björk, G. R., Jacobsson, K., Nilsson, K., Johansson, M. J., Byström, A. S., & Persson, O. P. (2001). A primordial tRNA modification required for the evolution of life?. *The EMBO journal*, 20(1-2), 231-239.
- Bland, J. M., & Altman, D. G. (2000). The odds ratio. *Bmj*, 320(7247), 1468.
- Boccaletto, P., Machnicka, M. A., Purta, E., Piątkowski, P., Bagiński, B., Wirecki, T. K., ... & Bujnicki, J. M. (2018). MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic acids research*, 46(D1), D303-D307.
- Boraas, M. E., Seale, D. B., & Boxhorn, J. E. (1998). Phagotrophy by a flagellate selects for colonial prey: a possible origin of multicellularity. *Evolutionary Ecology*, 12(2), 153-164.
- Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., & Madden, T. L. (2012). Domain enhanced lookup time accelerated BLAST. *Biology direct*, 7(1), 1-14.
- Borchiellini, C., Chombard, C., Lafay, B., & Boury-Esnault, N. (2000). Molecular systematics of sponges (Porifera). In *Marine Genetics* (pp. 15-27). Springer, Dordrecht.
- Borchiellini, C., Manuel, M., Alivon, E., Boury-Esnault, N., Vacelet, J., & Le Parco, Y. (2001). Sponge paraphyly and the origin of Metazoa. *Journal of Evolutionary Biology*, 14(1), 171-179.
- Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular ecology resources*, 16(1), 176-182.
- Brocks, J. J., Logan, G. A., Buick, R., & Summons, R. E. (1999). Archean molecular fossils and the early rise of eukaryotes. *science*, 285(5430), 1033-1036.
- Brown, M. W., Spiegel, F. W., & Silberman, J. D. (2009). Phylogeny of the “forgotten” cellular slime mold, *Fonticula alba*, reveals a key evolutionary branch within Opisthokonta. *Molecular Biology and Evolution*, 26(12), 2699-2709.
- Brown, M. W., Sharpe, S. C., Silberman, J. D., Heiss, A. A., Lang, B. F., Simpson, A. G., & Roger, A. J. (2013). Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proceedings of the Royal Society B: Biological Sciences*, 280(1769), 20131755.

- Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature*, 325(6106), 728-730.
- Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3), 897-907.
- Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S. I., Jakobsen, K. S., & Pawlowski, J. (2007). Phylogenomics reshuffles the eukaryotic supergroups. *PloS one*, 2(8).
- Cairns-Smith, A. G. (1982). Genetic takeover: and the mineral origins of life.
- Cairns-Smith, A. G., & Davies, C. J. (1977). *The design of novel replicating polymers* (pp. 391-403). New York: Pergamon Press.
- Camiolo, S., Melito, S., & Porceddu, A. (2015). New insights into the interplay between codon bias determinants in plants. *DNA Research*, 22(6), 461-470.
- Camiolo, S., Toome-Heller, M., Aime, M. C., Haridas, S., Grigoriev, I. V., Porceddu, A., & Mannazzu, I. (2019). An analysis of codon bias in six red yeast species. *Yeast*, 36(1), 53-64.
- Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., ... & Agris, P. F. (2010). The RNA modification database, RNAMDB: 2011 update. *Nucleic acids research*, 39(suppl_1), D195-D201.
- Carr, M., Leadbeater, B. S., Hassan, R., Nelson, M., & Baldauf, S. L. (2008a). Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proceedings of the National Academy of Sciences*, 105(43), 16641-16646.
- Carr, M., Nelson, M., Leadbeater, B. S., & Baldauf, S. L. (2008b). Three families of LTR retrotransposons are present in the genome of the choanoflagellate *Monosiga brevicollis*. *Protist*, 159(4), 579-590.
- Carr, M., Richter, D. J., Fozouni, P., Smith, T. J., Jeuck, A., Leadbeater, B. S., & Nitsche, F. (2017). A six-gene phylogeny provides new insights into choanoflagellate evolution. *Molecular Phylogenetics and Evolution*, 107, 166-178.
- Carr, M., Richter, D. J., Fozouni, P., Smith, T. J., Jeuck, A., Leadbeater, B. S., & Nitsche, F. (2017). A six-gene phylogeny provides new insights into choanoflagellate evolution. *Molecular Phylogenetics and Evolution*, 107, 166-178.

- Carrington, J. C., & Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science*, 301(5631), 336-338.
- Cavalier-Smith, T. (1987). The origin of fungi and pseudofungi. In *Evolutionary Biology of the FUNgi, British Mycological Society Symposium* (Vol. 12, pp. 339-353). Cambridge Univ. Press.
- Cavalier-Smith, T. (1997). Amoeboflagellates and mitochondrial cristae in eukaryote evolution: megasystematics of the new protozoan subkingdoms Eozoa and Neozoa. *Archiv für Protistenkunde*, 147(3-4), 237-258.
- Cavalier-Smith, T. (1998). Neomonada and the origin of animal and fungi. *Evolutionary relationships among Protozoa*, 56, 375-407.
- Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International journal of systematic and evolutionary microbiology*, 52(2), 297-354.
- Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoon root of the eukaryotic tree. *Biology letters*, 6(3), 342-345.
- Cavalier-Smith, T., & Chao, E. E. (1995). The opalozoan Apusomonas is related to the common ancestor of animals, fungi, and choanoflagellates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 261(1360), 1-6.
- Cavalier-Smith, T., & Chao, E. E. Y. (2003). Phylogeny of choanozoa, apusozoa, and other protozoa and early eukaryote megaevolution. *Journal of Molecular Evolution*, 56(5), 540-563.
- Cavalier-Smith, T., & Chao, E. E. (2010). Phylogeny and evolution of apusomonadida (protozoa: apusozoa): new genera and species. *Protist*, 161(4), 549-576.
- Chakraborty, S., Nag, D., Mazumder, T. H., & Uddin, A. (2017). Codon usage pattern and prediction of gene expression level in Bungarus species. *Gene*, 604, 48-60.
- Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research*, 44(D1), D184-D189.
- Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2019). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *BioRxiv*, 614032.

- Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *BioRxiv*, 614032.
- Chang, Z., Westaway, S., Li, S., Zaia, J. A., Rossi, J. J., & Scherer, L. J. (2002). Enhanced expression and HIV-1 inhibition of chimeric tRNA^{Lys3}-ribozymes under dual U6 snRNA and tRNA promoters. *Molecular Therapy*, 6(4), 481-489.
- Chen, I. A., & Szostak, J. W. (2004). Membrane growth can generate a transmembrane pH gradient in fatty acid vesicles. *Proceedings of the National Academy of Sciences*, 101(21), 7965-7970.
- Chen, I. A., Roberts, R. W., & Szostak, J. W. (2004). The emergence of competition between model protocells. *Science*, 305(5689), 1474-1476.
- Chen, I. A., Salehi-Ashtiani, K., & Szostak, J. W. (2005). RNA catalysis in model protocell vesicles. *Journal of the American Chemical Society*, 127(38), 13213-13219.
- Chen, L., Liu, T., Yang, D., Nong, X., Xie, Y., Fu, Y., ... & Peng, X. (2013). Analysis of codon usage patterns in *Taenia pisiformis* through annotated transcriptome data. *Biochemical and biophysical research communications*, 430(4), 1344-1348.
- Chun, K. (1879). *Die im Golf von Neapel erscheinenden Rippenquallen*.
- Conaco, C., Tsoulfas, P., Sakarya, O., Dolan, A., Werren, J., & Kosik, K. S. (2016). Detection of prokaryotic genes in the *Amphimedon queenslandica* genome. *PloS one*, 11(3), e0151092.
- Copeland, H. F. (1956). The classification of lower organisms. *The classification of lower organisms*.
- Corliss, J. B., Dymond, J., Gordon, L. I., Edmond, J. M., von Herzen, R. P., Ballard, R. D., ... & van Andel, T. H. (1979). Submarine thermal springs on the Galapagos Rift. *Science*, 203(4385), 1073-1083.
- Cox, E., & Bonner, J. (2001). The advantages of togetherness. *Science*, 292(5516), 448-449.
- Crick, F. H. (1958, January). On protein synthesis. In *Symp Soc Exp Biol* (Vol. 12, No. 138-63, p. 8).
- Crick, F. H. (1966). Codon--anticodon pairing: the wobble hypothesis.
- Crick, F. H. (1968). The origin of the genetic code. *Journal of molecular biology*, 38(3), 367-379.

- Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 192(4809), 1227-1232.
- Dayel, M. J., Alegado, R. A., Fairclough, S. R., Levin, T. C., Nichols, S. A., McDonald, K., & King, N. (2011). Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. *Developmental biology*, 357(1), 73-82.
- De Duve, C. (1995). *Vital dust: life as a cosmic imperative*. Basic Books.
- Deamer, D. W. (1985). Boundary structures are formed by organic components of the Murchison carbonaceous chondrite. *Nature*, 317(6040), 792-794.
- Deamer, D. W., & Pashley, R. M. (1989). Amphiphilic components of the Murchison carbonaceous chondrite: surface properties and membrane formation. *Origins of Life and Evolution of the Biosphere*, 19(1), 21-38.
- Deamer, D., Dworkin, J. P., Sandford, S. A., Bernstein, M. P., & Allamandola, L. J. (2002). The first cell membranes. *Astrobiology*, 2(4), 371-381.
- Derelle, R., Torruella, G., Klimeš, V., Brinkmann, H., Kim, E., Vlček, Č., ... & Eliáš, M. (2015). Bacterial proteins pinpoint a single eukaryotic root. *Proceedings of the National Academy of Sciences*, 112(7), E693-E699.
- Di Giulio, M. (1992). On the origin of the transfer RNA molecule. *Journal of theoretical biology*, 159(2), 199-214.
- Di Giulio, M. (1995). Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule?. *Journal of theoretical biology*, 177(1), 95-101.
- Di Giulio, M. (1999). The non-monophyletic origin of the tRNA molecule. *Journal of theoretical biology*, 197(3), 403-414.
- Di Giulio, M. (2004). The origin of the tRNA molecule: implications for the origin of protein synthesis. *Journal of theoretical biology*, 226(1), 89-93.
- Di Giulio, M. (2006). The non-monophyletic origin of the tRNA molecule and the origin of genes only after the evolutionary stage of the last universal common ancestor (LUCA). *Journal of theoretical biology*, 240(3), 343-352.
- Diederichsen, U. (1996). Pairing properties of alanyl peptide nucleic acids containing an amino acid backbone with alternating configuration. *Angewandte Chemie International Edition in English*, 35(4), 445-448.

- Ding, W., Zhao, H., Chen, Y., Zhang, B., Yang, Y., Zang, J., ... & Lin, S. (2020). Chimeric design of pyrrolysyl-tRNA synthetase/tRNA pairs and canonical synthetase/tRNA pairs for genetic code expansion. *Nature communications*, 11(1), 1-13.
- Dodd, M. S., Papineau, D., Grenne, T., Slack, J. F., Rittner, M., Pirajno, F., ... & Little, C. T. (2017). Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature*, 543(7643), 60-64.
- Doherty, A., & McInerney, J. O. (2013). Translational selection frequently overcomes genetic drift in shaping synonymous codon usage patterns in vertebrates. *Molecular biology and evolution*, 30(10), 2263-2267.
- dos Reis, M., & Wernisch, L. (2009). Estimating translational selection in eukaryotic genomes. *Molecular Biology and Evolution*, 26(2), 451-461.
- Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2), 341-352.
- Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., ... & Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188), 745-749.
- Duret, L. (2000). tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics*, 16(7), 287-289.
- Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6), 640-649.
- Dworkin, J. P., Deamer, D. W., Sandford, S. A., & Allamandola, L. J. (2001). Self-assembling amphiphilic molecules: Synthesis in simulated interstellar/precometary ices. *Proceedings of the National Academy of Sciences*, 98(3), 815-819.
- Edgcomb, V. P., Simpson, A. G., Zettler, L. A., Nerad, T. A., Patterson, D. J., Holder, M. E., & Sogin, M. L. (2002). Pelobionts are degenerate protists: insights from molecules and morphology. *Molecular biology and evolution*, 19(6), 978-982.
- Egholm, M., Buchardt, O., Christensen, L., Behrens, C., Freier, S. M., Driver, D. A., ... & Nielsen, P. E. (1993). PNA hybridizes to complementary oligonucleotides obeying the Watson–Crick hydrogen-bonding rules. *Nature*, 365(6446), 566-568.

- Ehrenberg, M., & Kurland, C. G. (1984). Costs of accuracy determined by a maximal growth rate constraint. *Quarterly reviews of biophysics*, 17(1), 45-82.
- Eitel, M., & Schierwater, B. (2010). The phylogeography of the Placozoa suggests a taxon-rich phylum in tropical and subtropical waters. *Molecular Ecology*, 19(11), 2315-2327.
- Eitel, M., Francis, W. R., Varoqueaux, F., Daraspe, J., Osigus, H. J., Krebs, S., ... & Wörheide, G. (2018). Comparative genomics and the nature of placozoan species. *PLoS biology*, 16(7), e2005359.
- Elias, Y., & Huang, R. H. (2005). Biochemical and structural studies of A-to-I editing by tRNA: A34 deaminases at the wobble position of transfer RNA. *Biochemistry*, 44(36), 12057-12065.
- Emery, L. R., & Sharp, P. M. (2011). Impact of translational selection on codon usage bias in the archaeon *Methanococcus maripaludis*. *Biology letters*, 7(1), 131-135.
- Eschscholtz, F. (1825). Bericht über die zoologische Ausbeute der Reise von Kronstadt bis St.-Peter und Paul. *Oken's Isis*, 6, 734-747.
- Exposito, J. Y., Cluzel, C., Garrone, R., & Lethias, C. (2002). Evolution of collagens. *The Anatomical Record: An Official Publication of the American Association of Anatomists*, 268(3), 302-316.
- Eyre-Walker, A. C. (1991). An analysis of codon usage in mammals: selection or mutation bias?. *Journal of molecular evolution*, 33(5), 442-449.
- Eyre-Walker, A., & Bulmer, M. (1995). Synonymous substitution rates in enterobacteria. *Genetics*, 140(4), 1407-1412.
- Fairclough, S. R., Dayel, M. J., & King, N. (2010). Multicellular development in a choanoflagellate. *Current Biology*, 20(20), R875-R876.
- Fairclough, S. R., Chen, Z., Kramer, E., Zeng, Q., Young, S., Robertson, H. M., ... & King, N. (2013). Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome biology*, 14(2), 1-15.
- Fan, Z., Song, J., Guan, T., Lv, X., & Wei, J. (2018). Efficient expression of glutathione peroxidase with chimeric tRNA in amber-less *Escherichia coli*. *ACS synthetic biology*, 7(1), 249-257.

- Fenchel, T. (1982). Ecology of heterotrophic microflagellates. IV. Quantitative occurrence and importance as bacterial consumers. *Mar. Ecol. Prog. Ser.*, 9(3), 5.
- Ferris, J. P. (2002). Montmorillonite catalysis of 30–50 mer oligonucleotides: laboratory demonstration of potential steps in the origin of the RNA world. *Origins of Life and Evolution of the Biosphere*, 32(4), 311-332.
- Ferris, J. P., Hill, A. R., Liu, R., & Orgel, L. E. (1996). Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*, 381(6577), 59-61.
- Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., ... & Pisani, D. (2017). Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology*, 27(24), 3864-3870.
- Figuet, E., Ballenghien, M., Romiguier, J., & Galtier, N. (2015). Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome biology and evolution*, 7(1), 240-250.
- Fitzpatrick, D. A., Logue, M. E., Stajich, J. E., & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC evolutionary biology*, 6(1), 1-15.
- Flores, E., & Herrero, A. (2010). Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nature Reviews Microbiology*, 8(1), 39-50.
- Forterre, P. (2002). The origin of DNA genomes and DNA replication proteins. *Current opinion in microbiology*, 5(5), 525-532.
- Forterre, P. (2006). Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proceedings of the National Academy of Sciences*, 103(10), 3669-3674.
- Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., ... & Zablen, L. B. (1980). The phylogeny of prokaryotes. *Science*, 209(4455), 457-463.
- Frenkel-Pinter, M., Haynes, J. W., Mohyeldin, A. M., Martin, C., Sargon, A. B., Petrov, A. S., ... & Leman, L. J. (2020). Mutually stabilizing interactions between proto-peptides and RNA. *Nature communications*, 11(1), 1-14.

- Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., ... & Liu, Y. (2016). Codon usage affects the structure and function of the *Drosophila* circadian clock protein PERIOD. *Genes & development*, 30(15), 1761-1775.
- Galtier, N., Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., ... & Duret, L. (2018). Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. *Molecular biology and evolution*, 35(5), 1092-1103.
- Gamow, G. (1954). Possible relation between deoxyribonucleic acid and protein structures. *Nature*, 173(4398), 318-318.
- Gams, H. (1947). Die Protochlorinae als autotrophe Vorfahren von Pilzen und Tieren. *Mikroskopie (Wien)*, 2, 383.
- Gazave, E., Lapébie, P., Ereskovsky, A. V., Vacelet, J., Renard, E., Cárdenas, P., & Borchellini, C. (2011). No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. In *Ancient Animals, New Challenges* (pp. 3-10). Springer, Dordrecht.
- Gerber, A. P., & Keller, W. (1999). An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, 286(5442), 1146-1149.
- Geslain, R., Cubells, L., Bori-Sanz, T., Alvarez-Medina, R., Rossell, D., Martí, E., & de Pouplana, L. R. (2010). Chimeric tRNAs as tools to induce proteome damage and identify components of stress responses. *Nucleic acids research*, 38(5), e30-e30.
- Gilbert, W. (1986). Origin of life: The RNA world. *nature*, 319(6055), 618-618.
- Gilbert, W. (1987). The exon theory of genes. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 52, pp. 901-905). Cold Spring Harbor Laboratory Press.
- Gile, G. H., Faktorová, D., Castlejohn, C. A., Burger, G., Lang, B. F., Farmer, M. A., ... & Keeling, P. J. (2009). Distribution and phylogeny of EFL and EF-1 α in Euglenozoa suggest ancestral co-occurrence followed by differential loss. *PLoS One*, 4(4).
- Glansdorff, N., Xu, Y., & Labedan, B. (2008). The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology direct*, 3(1), 29.
- Goel, N. S., Rao, G. S., Yčas, M., Bremermann, H. J., & King, L. (1972). A method for calculating codon frequencies in DNA. *Journal of theoretical biology*, 35(3), 399-457.

- Goldberg, A. L., & Wittes, R. E. (1966). Genetic code: aspects of organization. *Science*, 153(3734), 420-424.
- Goldenfeld, N., & Woese, C. (2007). Biology's next revolution. *Nature*, 445(7126), 369-369.
- Goñi, N., Iriarte, A., Comas, V., Soñora, M., Moreno, P., Moratorio, G., ... & Cristina, J. (2012). Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. *Virology Journal*, 9(1), 1-8.
- Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic acids research*, 8(1), 197-197.
- Gray, M. W., & Doolittle, W. F. (1982). Has the endosymbiont hypothesis been proven?. *Microbiological Reviews*, 46(1), 1.
- Gray, M. W., Burger, G., & Lang, B. F. (1999). Mitochondrial evolution. *Science*, 283(5407), 1476-1481.
- Grobben, K. (1908). Die systematische einteilung des tierreichs. *Verh. Zool. Bot. Ges. Wien*, 58(49), 1-5.
- Groebke, K., Hunziker, J., Fraser, W., Peng, L., Diederichsen, U., Zimmermann, K., ... & Eschenmoser, A. (1998). Why Pentose-and Not Hexose-Nucleic Acids? Purine-Purine Pairing in homo-DNA: Guanine, Isoguanine, 2, 6-Diaminopurine, and Xanthine. *ChemInform*, 29(23), no-no.
- Grosberg, R. K., & Strathmann, R. R. (2007). The evolution of multicellularity: a minor major transition?. *Annu. Rev. Ecol. Evol. Syst.*, 38, 621-654.
- Grosjean, H., Söll, D. G., & Crothers, D. M. (1976). Studies of the complex between transfer RNAs with complementary anticodons: I. Origins of enhanced affinity between complementary triplets. *Journal of molecular biology*, 103(3), 499-519.
- Guan, D. L., Ma, L. B., Khan, M. S., Zhang, X. X., Xu, S. Q., & Xie, J. Y. (2018). Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. *BMC genomics*, 19(1), 1-14.
- Guerrier-Takada, C., & Altman, S. (1984). Catalytic activity of an RNA molecule prepared by transcription in vitro. *Science*, 223(4633), 285-286.

- Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3), 849-857.
- Guidi, L., Eitel, M., Cesarini, E., Schierwater, B., & Balsamo, M. (2011). Ultrastructural analyses support different morphological lineages in the phylum Placozoa Grell, 1971. *Journal of Morphology*, 272(3), 371-378.
- Gustafsson, C., Minshull, J., Govindarajan, S., Ness, J., Villalobos, A., & Welch, M. (2012). Engineering genes for predictable protein expression. *Protein expression and purification*, 83(1), 37-46.
- Haeckel, E. (1874). Memoirs: the Gastraea-Theory, the phylogenetic classification of the animal kingdom and the homology of the germ-lamellæ. *Journal of Cell Science*, 2(54), 142-165.
- Hagervall, T. G., & Björk, G. R. (1984). Undermodification in the first position of the anticodon of supG-tRNA reduces translational efficiency. *Molecular and General Genetics MGG*, 196(2), 194-200.
- Hambuch, T. M., & Parsch, J. (2005). Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. *Genetics*, 170(4), 1691-1700.
- Hanczyc, M. M., Fujikawa, S. M., & Szostak, J. W. (2003). Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science*, 302(5645), 618-622.
- Hartl, D. L., Moriyama, E. N., & Sawyer, S. A. (1994). Selection intensity for codon bias. *Genetics*, 138(1), 227-234.
- Haviland, M. G. (1990). Yates's correction for continuity and the analysis of 2× 2 contingency tables. *Statistics in medicine*, 9(4), 363-367.
- Hehenberger, E., Tikhonenkov, D. V., Kolisko, M., Del Campo, J., Esaulov, A. S., Mylnikov, A. P., & Keeling, P. J. (2017). Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Current Biology*, 27(13), 2043-2050.
- Van Hemert, F. J., & Berkhout, B. (1995). The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. *Journal of molecular evolution*, 41(2), 132-140.

- Henry, I., & Sharp, P. M. (2007). Predicting gene expression level from codon usage bias. *Molecular biology and evolution*, 24(1), 10-12.
- Herr, R. A., Ajello, L., Taylor, J. W., Arseculeratne, S. N., & Mendoza, L. (1999). Phylogenetic analysis of *Rhinosporidium seeberi*'s 18S small-subunit ribosomal DNA groups this pathogen among members of the protocistan Mesomycetozoa clade. *Journal of Clinical Microbiology*, 37(9), 2750-2754.
- Hibberd, D. (1975). Observations on the ultrastructure of the choanoflagellate *Codosiga botrytis* (Ehr.) Saville-Kent with special reference to the flagellar apparatus. *Journal of cell science*, 17(1), 191-219.
- Hill, A. R., Böhler, C., & Orgel, L. E. (1998). Polymerization on the rocks: negatively-charged α -amino acids. *Origins of Life and Evolution of the Biosphere*, 28(3), 235-243.
- Hirabayashi, J., & Kasai, K. I. (1993). The family of metazoan metal-independent β -galactoside-binding lectins: structure, function and molecular evolution. *Glycobiology*, 3(4), 297-304.
- Horn, D. (2008). Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids. *BMC genomics*, 9(1), 1-11.
- Huang, X., Xu, J., Chen, L., Wang, Y., Gu, X., Peng, X., & Yang, G. (2017). Analysis of transcriptome data reveals multifactor constraint on codon usage in *Taenia multiceps*. *BMC genomics*, 18(1), 1-13.
- Hui, M., Cheng, J., & Sha, Z. (2018). Adaptation to the deep-sea hydrothermal vents and cold seeps: insights from the transcriptomes of *Alvinocaris longirostris* in both environments. *Deep Sea Research Part I: Oceanographic Research Papers*, 135, 23-33.
- Hunter, T., & Sefton, B. M. (1980). Transforming gene product of Rous sarcoma virus phosphorylates tyrosine. *Proceedings of the National Academy of Sciences*, 77(3), 1311-1315.
- Ichihashi, N., Usui, K., Kazuta, Y., Sunami, T., Matsuura, T., & Yomo, T. (2013). Darwinian evolution in a translation-coupled RNA replication system within a cell-like compartment. *Nature communications*, 4(1), 1-7.

- 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. *Journal of molecular biology*, 151(3), 389-409.
- Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *Journal of molecular biology*, 158(4), 573-597.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), 13-34.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), 13-34.
- James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., ... & Lumbsch, H. T. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113), 818-822.
- James-Clark, H. (1866). On the nature of sponges. In *Proceedings of the Boston Society of Natural History* (Vol. 11, pp. 16-17).
- James-Clark, H. (1868). XXVII.—On the Spongiæ ciliatæ as Infusoria flagellata; or observations on the structure, animality, and relationship of Leucosolenia botryoides, Bowerbank. *Annals and magazine of natural history*, 1(3), 188-215.
- James-Clark, H. (1871). Note on the infusoria flagellata and the spongiæ ciliatæ. *Journal of Natural History*, 7(39), 247-248.
- Javaux, E. J., Knoll, A. H., & Walter, M. (2003). Recognizing and interpreting the fossils of early eukaryotes. *Origins of Life and Evolution of the Biosphere*, 33(1), 75-94.
- Javaux, E. J. (2007). The early eukaryotic fossil record. In *Eukaryotic Membranes and Cytoskeleton* (pp. 1-19). Springer, New York, NY.
- Jiang, W., Lü, B., He, J., Wang, J., Wu, X., Wu, G., ... & Tang, X. (2014). Codon usage bias in the straw mushroom *Volvariella volvacea*. *Sheng wu gong cheng xue bao= Chinese journal of biotechnology*, 30(9), 1424-1435.

- Jukes, T. H., & Osawa, S. (1990). The genetic code in mitochondria and chloroplasts. *Experientia*, 46(11), 1117-1126.
- Jurkowski, T. P., & Jeltsch, A. (2011). On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. *PloS one*, 6(11), e28104.
- Justice, S. S., Hunstad, D. A., Cegelski, L., & Hultgren, S. J. (2008). Morphological plasticity as a bacterial survival strategy. *Nature Reviews Microbiology*, 6(2), 162-168.
- Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, 238(1), 143-155.
- Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., ... & Ikemura, T. (2001a). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene*, 276(1-2), 89-99.
- Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., & Ikemura, T. (2001b). Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of molecular evolution*, 53(4-5), 290-298.
- Karcher, D., & Bock, R. (2009). Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *Rna*, 15(7), 1251-1257.
- Karpov, S. A., & Coupe, S. J. (1998). A revision of choanoflagellate genera *Kentrosiga* Schiller, 1953 and *Desmarella* Kent, 1880. *Acta Protozoologica*, 37, 23-28.
- Karpov, S. A., & Leadbeater, B. S. (1997). Cell and nuclear division in a freshwater choanoflagellate, *Monosiga ovata* Kent. *European Journal of Protistology*, 33(3), 323-334.
- Karpov, S. A., & Leadbeater, B. S. (1998). Cytoskeleton structure and composition in choanoflagellates. *Journal of Eukaryotic Microbiology*, 45(3), 361-367.
- Katz, L. A. (2002). Lateral gene transfers and the evolution of eukaryotes: theories and data. *International Journal of Systematic and Evolutionary Microbiology*, 52(5), 1893-1900.

- Katze, J. R., & Beck, W. T. (1980). Administration of queuine to mice relieves modified nucleoside queuosine deficiency in Ehrlich ascites tumor tRNA. *Biochemical and biophysical research communications*, 96(1), 313-319.
- Kawabe, A., & Miyashita, N. T. (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes & genetic systems*, 78(5), 343-352.
- Kayal, E., Bentlage, B., Pankey, M. S., Ohdera, A. H., Medina, M., Plachetzki, D. C., ... & Ryan, J. F. (2018). Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. *BMC evolutionary biology*, 18(1), 1-18.
- Keeling, P. J., & Inagaki, Y. (2004). A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1 α . *Proceedings of the National Academy of Sciences*, 101(43), 15380-15385.
- Kent, W. S. (1880). *A Manual of the Infusoria: Including a Description of All Known Flagellate, Ciliate, and Tentaculiferous Protozoa, British and foreign, and an Account of the Organization and the Affinities of the Sponges* (Vol. 1). David Bogue.
- Kessler, M. D., & Dean, M. D. (2014). Effective population size does not predict codon usage bias in mammals. *Ecology and evolution*, 4(20), 3887-3900.
- Kim, E., Simpson, A. G., & Graham, L. E. (2006). Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. *Molecular biology and evolution*, 23(12), 2455-2466.
- Kimura, M. (1981). Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. *Proceedings of the National Academy of Sciences*, 78(9), 5773-5777.
- Kimura, M. (1983). Rare variant alleles in the light of the neutral theory. *Molecular Biology and Evolution*, 1(1), 84-93.
- King, N. (2004). The unicellular ancestry of animal development. *Developmental cell*, 7(3), 313-325.
- King, N., & Carroll, S. B. (2001). A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proceedings of the National Academy of Sciences*, 98(26), 15032-15037.

- King, N., Hittinger, C. T., & Carroll, S. B. (2003). Evolution of key cell signaling and adhesion protein families predates animal origins. *Science*, 301(5631), 361-363.
- King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., ... & Rokhsar, D. (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, 451(7180), 783-788.
- Knoll, A. H., Javaux, E. J., Hewitt, D., & Cohen, P. (2006). Eukaryotic organisms in Proterozoic oceans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1470), 1023-1038.
- Kobayashi, M., Wada, H., & Satoh, N. (1996). Early evolution of the metazoa and phylogenetic status of diploblasts as inferred from amino acid sequence of elongation factor-1 α . *Molecular Phylogenetics and Evolution*, 5(2), 414-422.
- Kohn, A. B., & Moroz, L. L. (2018). Polymorphism within the mitochondrial genome of the ctenophore, *Pleurobrachia bachei* and its ongoing rapid evolution. bioRxiv, 366880.
- Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2), 127-136.
- Koonin, E. V., & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB life*, 61(2), 99-111.
- Koppitz, M., Nielsen, P. E., & Orgel, L. E. (1998). Formation of Oligonucleotide–PNA–Chimeras by Template-Directed Ligation. *Journal of the American Chemical Society*, 120(19), 4563-4569.
- Koufopanou, V., & Bell, G. (1993). Soma and germ: an experimental approach using *Volvox*. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 254(1340), 107-113.
- Kozlov, I. A., Orgel, L. E., & Nielsen, P. E. (2000). Remote enantioselection transmitted by an achiral peptide nucleic acid backbone. *Angewandte Chemie*, 112(23), 4462-4465.
- Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., & Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *cell*, 31(1), 147-157.

- Kruse, M., Leys, S. P., Müller, I. M., & Müller, W. E. (1998). Phylogenetic position of the Hexactinellida within the phylum Porifera based on the amino acid sequence of the protein kinase C from *Rhabdocalyptus dawsoni*. *Journal of Molecular Evolution*, 46(6), 721-728.
- Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *science*, 324(5924), 255-258.
- Lafay, B., Sharp, P. M., Lloyd, A. T., McLean, M. J., Devine, K. M., & Wolfe, K. H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic acids research*, 27(7), 1642-1649.
- Lafay, B., Atherton, J. C., & Sharp, P. M. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, 146(4), 851-860.
- Lagerkvist, U. L. F. (1978). "Two out of three": an alternative method for codon reading. *Proceedings of the National Academy of Sciences*, 75(4), 1759-1762.
- Lake, J. A. (1990). Origin of the Metazoa. *Proceedings of the National Academy of Sciences*, 87(2), 763-766.
- Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W., & Burger, G. (2002). The closest unicellular relatives of animals. *Current biology*, 12(20), 1773-1778.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), 1-10.
- Larsen, J., & Patterson, D. J. (1990). Some flagellates (Protista) from tropical marine sediments. *Journal of Natural History*, 24(4), 801-937.
- Latorre, A., & Manzano-Marín, A. (2017). Dissecting genome reduction and trait loss in insect endosymbionts. *Annals of the New York Academy of Sciences*, 1389(1), 52-75.
- Laumer, C. E., Gruber-Vodicka, H., Hadfield, M. G., Pearse, V. B., Riesgo, A., Marioni, J. C., & Giribet, G. (2018). Placozoa and Cnidaria are sister taxa. *bioRxiv*, 200972.

- Law, N. C., Donaubauer, E. M., Zeleznik, A. J., & Hunzicker-Dunn, M. (2017). How protein kinase A activates canonical tyrosine kinase signaling pathways to promote granulosa cell differentiation. *Endocrinology*, 158(7), 2043-2051.
- Laxman, S., Sutter, B. M., Wu, X., Kumar, S., Guo, X., Trudgian, D. C., ... & Tu, B. P. (2013). Sulfur amino acids regulate translational capacity and metabolic homeostasis through modulation of tRNA thiolation. *Cell*, 154(2), 416-429.
- Leadbeater, B. S. C. (1981). Ultrastructure and deposition of silica in loricate choanoflagellates. In *Silicon and siliceous structures in biological systems* (pp. 295-322). Springer, New York, NY.
- Leadbeater, B. C. (1983). Distribution and chemistry of microfilaments in choanoflagellates, with special reference to the collar and other tentacle systems. *Protistologica (Paris. 1965)*, 19(2), 157-166.
- Leadbeater, B. S. (2008). Choanoflagellate evolution: the morphological perspective. *Protistology*, 5(4).
- Leadbeater, B. S. (2015). *The choanoflagellates*. Cambridge University Press.
- Leadbeater, B. S., & Karpov, S. A. (2000). Cyst formation in a freshwater strain of the choanoflagellate *Desmarella moniliformis* Kent. *Journal of Eukaryotic Microbiology*, 47(5), 433-439.
- Leadbeater, B. S., Hassan, R., Nelson, M., Carr, M., & Baldauf, S. L. (2008). A new genus, *Helgoeca* gen. nov., for a nudiform choanoflagellate. *European journal of protistology*, 44(3), 227-237.
- Leichty, A. R., & Brisson, D. (2014). Selective whole genome amplification for resequencing target microbial species from complex natural samples. *Genetics*, 198(2), 473-481.
- Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2010). The sequence read archive. *Nucleic acids research*, 39(suppl_1), D19-D21.
- Leipe, D. D., Aravind, L., & Koonin, E. V. (1999). Did DNA replication evolve twice independently?. *Nucleic acids research*, 27(17), 3389-3401.
- Lenhoff, H. M., & Lenhoff, S. G. (1988). Trembley's polyps. *Scientific American*, 258(4), 108-113.

- Leslie E, O. (2004). Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology*, 39(2), 99-123.
- Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147-151.
- Li, W. H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of molecular evolution*, 24(4), 337-345.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14), 1754-1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.
- Liu, Y., Steenkamp, E. T., Brinkmann, H., Forget, L., Philippe, H., & Lang, B. F. (2009). Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. *BMC Evolutionary Biology*, 9(1), 272.
- Liu, R. J., Zhou, M., Fang, Z. P., Wang, M., Zhou, X. L., & Wang, E. D. (2013). The tRNA recognition mechanism of the minimalist SPOUT methyltransferase, TrmL. *Nucleic acids research*, 41(16), 7828-7842.
- Liu, Y., Zhou, Z., Pan, J., Baker, B. J., Gu, J. D., & Li, M. (2018). Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *The ISME journal*, 12(4), 1021-1031.
- Lloyd, A. T., & Sharp, P. M. (1991). Codon usage in *Aspergillus nidulans*. *Molecular and General Genetics MGG*, 230(1), 288-294.
- Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular biology and evolution*, 13(5), 660-665.
- Loison, A., Dubant, S., Adam, P., & Albrecht, P. (2010). Elucidation of an iterative process of carbon–carbon bond formation of prebiotic significance. *Astrobiology*, 10(10), 973-988.
- Long, M. (2001). Evolution of novel genes. *Current opinion in genetics & development*, 11(6), 673-680.

- López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M. E., & Ruiz-Trillo, I. (2017). Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate *Monosiga brevicollis*. *Scientific reports*, 7(1), 1-14..
- López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M. E., & Ruiz-Trillo, I. (2019). Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philosophical Transactions of the Royal Society B*, 374(1786), 20190088.
- Lundin, L. G. (1999). Gene duplications in early metazoan evolution. In *Seminars in cell & developmental biology* (Vol. 10, No. 5, pp. 523-530). Academic Press.
- Lyu, X., & Liu, Y. (2020). Nonoptimal Codon Usage Is Critical for Protein Structure and Function of the Master General Amino Acid Control Regulator CPC-1. *Mbio*, 11(5).
- Machado, H. E., Lawrie, D. S., & Petrov, D. A. (2020). Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. *Genetics*, 214(2), 511-528.
- Magistrelli, G., Poitevin, Y., Schlosser, F., Pontini, G., Malinge, P., Josserand, S., ... & Fischer, N. (2017, February). Optimizing assembly and production of native bispecific antibodies by codon de-optimization. In *MAbs* (Vol. 9, No. 2, pp. 231-239). Taylor & Francis.
- Mansy, S. S., Schrum, J. P., Krishnamurthy, M., Tobé, S., Treco, D. A., & Szostak, J. W. (2008). Template-directed synthesis of a genetic polymer in a model protocell. *Nature*, 454(7200), 122-125.
- Marais, G., & Duret, L. (2001). Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. *Journal of molecular evolution*, 52(3), 275-280.
- Marck, C., & Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *Rna*, 8(10), 1189-1232.
- Margulis, L. (1967). On the origin of mitosing cells. *Journal of theoretical biology*, 14(3), 255-274.

- Martin, W., & Kowallik, K. (1999). Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *European Journal of Phycology*, 34(3), 287-295.
- Martin, W., & Russell, M. J. (2003). On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1429), 59-85.
- Martin, W., & Russell, M. J. (2007). On the origin of biochemistry at an alkaline hydrothermal vent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1486), 1887-1926.
- Martin, A., Bertranpetit, J., Oliver, J. L., & Medina, J. R. (1989). Variation in G+ C-content and codon choice: differences among synonymous codon groups in vertebrate genes. *Nucleic acids research*, 17(15), 6181-6189.
- Maside, X., Lee, A. W., & Charlesworth, B. (2004). Selection on codon usage in *Drosophila americana*. *Current Biology*, 14(2), 150-154.
- Matsuura, T., Yamaguchi, M., Ko-Mitamura, E. P., Shima, Y., Urabe, I., & Yomo, T. (2002). Importance of compartment formation for a self-encoding system. *Proceedings of the National Academy of Sciences*, 99(11), 7514-7517.
- Mayer, C., Schreiber, U., Dávila, M. J., Schmitz, O. J., Bronja, A., Meyer, M., ... & Meckelmann, S. W. (2018). Molecular evolution in a peptide-vesicle system. *Life*, 8(2), 16.
- McCollom, T. M., Ritter, G., & Simoneit, B. R. (1999). Lipid synthesis under hydrothermal conditions by Fischer-Tropsch-type reactions. *Origins of Life and Evolution of the Biosphere*, 29(2), 153-166.
- Mcinerney, J. O. (1997). Codon usage patterns in *Trichomonas vaginalis*. *European Journal of Protistology*, 33(3), 266-273.
- Medina, M., Collins, A. G., Silberman, J. D., & Sogin, M. L. (2001). Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proceedings of the National Academy of Sciences*, 98(17), 9707-9712.
- Medina, M., Collins, A. G., Taylor, J. W., Valentine, J. W., Lipps, J. H., Amaral-Zettler, L., & Sogin, M. L. (2003). Phylogeny of Opisthokonta and the evolution of

multicellularity and complexity in Fungi and Metazoa. *International Journal of Astrobiology*, 2(3), 203-211.

- Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, 8(8), 1551-1566.
- Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1), D183-D189.
- Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, 117(3046), 528-529.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315-327.
- Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., ... & Marshall, D. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in bioinformatics*, 14(2), 193-202.
- Mitreva, M., Wendl, M. C., Martin, J., Wylie, T., Yin, Y., Larson, A., ... & McCarter, J. P. (2006). Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biology*, 7(8), 1-19.
- Moore, D. (1987). *Evolutionary biology of the fungi*. A. D. Rayner, & C. M. Brasier (Eds.). Cambridge: Cambridge University Press.
- Moriyama, E. N., & Powell, J. R. (1997). Codon usage bias and tRNA abundance in *Drosophila*. *Journal of molecular evolution*, 45(5), 514-523.
- Moriyama, E. N., & Powell, J. R. (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic acids research*, 26(13), 3188-3193.
- Moriyama, E. N., & Hartl, D. L. (1993). Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics*, 134(3), 847-858.
- Morowitz, H. J., Kostelnik, J. D., Yang, J., & Cody, G. D. (2000). The origin of intermediary metabolism. *Proceedings of the National Academy of Sciences*, 97(14), 7704-7708.

- Moroz, L. L., Kocot, K. M., Citarella, M. R., Dosung, S., Norekian, T. P., Povolotskaya, I. S., ... & Kohn, A. B. (2014). The ctenophore genome and the evolutionary origins of neural systems. *Nature*, 510(7503), 109-114.
- Müller, W. E. G. (1998). Molecular phylogeny of Eumetazoa: genes in sponges (Porifera) give evidence for monophyly of animals. In *Molecular Evolution: Evidence for Monophyly of Metazoa* (pp. 89-132). Springer, Berlin, Heidelberg.
- Müller, W. E. (2001). How was metazoan threshold crossed? The hypothetical Urmetazoa. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, 129(2-3), 433-460.
- Müller, W. E., Wiens, M., Adell, T., Gamulin, V., Schröder, H. C., & Müller, I. M. (2004). Bauplan of urmetazoa: basis for genetic complexity of metazoa. *International Review of Cytology*, 235, 53-92.
- Murakami, H., Kourouklis, D., & Suga, H. (2003). Using a solid-phase ribozyme aminoacylation system to reprogram the genetic code. *Chemistry & biology*, 10(11), 1077-1084.
- Murray, E. E., Lotzer, J., & Eberle, M. (1989). Codon usage in plant genes. *Nucleic Acids Research*, 17(2), 477-498.
- Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic acids research*, 28(1), 292-292.
- Naotake, O. (1985). Markedly unbiased codon usage in *Bacillus subtilis*. *Gene*, 40(1), 145-150.
- NÄSVALL, S. J., Chen, P., & BJÖRK, G. R. (2004). The modified wobble nucleoside uridine-5-oxyacetic acid in tRNA^{Pro}mo5UGG promotes reading of all four proline codons in vivo. *Rna*, 10(10), 1662-1673.
- Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933-2935.
- Nedialkova, D. D., & Leidel, S. A. (2015). Optimization of codon translation rates via tRNA modifications maintains proteome integrity. *Cell*, 161(7), 1606-1618.

- Ngo, H., Tschudi, C., Gull, K., & Ullu, E. (1998). Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei*. *Proceedings of the National Academy of Sciences*, 95(25), 14687-14692.
- Nitsche, F., Carr, M., Arndt, H., & Leadbeater, B. S. (2011). Higher level taxonomy and molecular phylogenetics of the Choanoflagellata. *Journal of Eukaryotic Microbiology*, 58(5), 452-462.
- Nitsche, F., Carr, M., Arndt, H., & Leadbeater, B. S. (2011). Higher level taxonomy and molecular phylogenetics of the Choanoflagellata. *Journal of Eukaryotic Microbiology*, 58(5), 452-462.
- Noller, H. F. (2004). The driving force for molecular evolution of translation. *Rna*, 10(12), 1833-1837.
- Noma, A., & Suzuki, T. (2006). Ribonucleome analysis identified enzyme genes responsible for wybutosine synthesis. In *Nucleic acids symposium series* (Vol. 50, No. 1, pp. 65-66). Oxford University Press.
- Norris, R. E. (1965). Neustonic marine craspedomonadales (choanoflagellates) from Washington and California. *The Journal of Protozoology*, 12(4), 589-602.
- Novoa, E. M., & de Pouplana, L. R. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends in Genetics*, 28(11), 574-581.
- Oba, T., Andachi, Y., Muto, A., & Osawa, S. (1991). CGG: an unassigned or nonsense codon in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences*, 88(3), 921-925.
- Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, 246(5428), 96-98.
- Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences*, 99(25), 16134-16137.
- Ono, K., Suga, H., Iwabe, N., Kuma, K. I., & Miyata, T. (1999). Multiple protein tyrosine phosphatases in sponges and explosive gene duplication in the early evolution of animals before the parazoan–eumetazoan split. *Journal of molecular evolution*, 48(6), 654-662.
- Oparin, A. I. (1924). *Proischogdenie zhizni. Moscovsky Robotchii, Moscow*.

- Orgel, L. E. (1968). Evolution of the genetic apparatus. *Journal of molecular biology*, 38(3), 381-393.
- Orgel, L. E. (1998). Polymerization on the rocks: theoretical introduction. *Origins of Life and Evolution of the Biosphere*, 28(3), 227-234.
- Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiological reviews*, 56(1), 229-264.
- Osigus, H. J., Rolfes, S., Herzog, R., Kamm, K., & Schierwater, B. (2019). Polyplacotoma mediterranea is a new ramified placozoan species. *Current Biology*, 29(5), R148-R149.
- Pandis, N. (2016). The chi-square test. *American journal of orthodontics and dentofacial orthopedics*, 150(5), 898-899.
- Paps, J., Medina-Chacón, L. A., Marshall, W., Suga, H., & Ruiz-Trillo, I. (2013). Molecular phylogeny of unikonts: new insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. *Protist*, 164(1), 2-12.
- Parfrey, L. W., & Lahr, D. J. (2013). Multicellularity arose several times in the evolution of eukaryotes (response to DOI 10.1002/bies. 201100187). *Bioessays*, 35(4), 339-347.
- Parfrey, L. W., Lahr, D. J., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences*, 108(33), 13624-13629.
- Patterson, D. J. (1988). The evolution of protozoa. *Memorias do Instituto Oswaldo Cruz*, 83, 580-600.
- Pearse, V. B., & Voigt, O. (2007). Field biology of placozoans (Trichoplax): distribution, diversity, biotic interactions. *Integrative and Comparative Biology*, 47(5), 677-692.
- Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157-175.
- Peden, J. (1997). CodonW. *Trinity College*.

- Peden, J. (2005). CodonW version 1.4. 2. *University of Nottingham: Nottingham, UK*.
- Peterson, K. J., Cotton, J. A., Gehling, J. G., & Pisani, D. (2008). The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1496), 1435-1443.
- Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchellini, C., Boury-Esnault, N., ... & Da Silva, C. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, 19(8), 706-712.
- Pitsch, S., Wendeborn, S., Krishnamurthy, R., Holzner, A., Minton, M., Bolli, M., ... & Jaun, B. (2003). Pentopyranosyl Oligonucleotide Systems. 9th Communication: The β -d-Ribopyranosyl-(4'→ 2')-oligonucleotide System ('Pyranosyl-RNA'): Synthesis and Resumé of Base-Pairing Properties. *Helvetica chimica acta*, 86(12), 4270-4363.
- Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1), 32-42.
- Ponce-Toledo, R. I., Deschamps, P., López-García, P., Zivanovic, Y., Benzerara, K., & Moreira, D. (2017). An early-branching freshwater cyanobacterium at the origin of plastids. *Current Biology*, 27(3), 386-391.
- Ponstingl, H., & Ning, Z. (2010). SMALT-a new mapper for DNA sequencing reads. *F1000 Posters*, 1(L313).
- Poole, A. M., & Logan, D. T. (2005). Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome?. *Molecular biology and evolution*, 22(6), 1444-1455.
- Porath, H. T., Knisbacher, B. A., Eisenberg, E., & Levanon, E. Y. (2017). Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. *Genome biology*, 18(1), 1-12.
- Porceddu, A., Zenoni, S., & Camiolo, S. (2013). The signatures of selection for translational accuracy in plant genes. *Genome biology and evolution*, 5(6), 1117-1126.
- Porter, S. M. (2004). The fossil record of early eukaryotic diversification. *The Paleontological Society Papers*, 10, 35-50.

- Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H., & Dennis, P. P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 76(4), 1697-1701.
- Powell, J. R., & Moriyama, E. N. (1997). Evolution of codon usage bias in *Drosophila*. *Proceedings of the National Academy of Sciences*, 94(15), 7784-7790.
- Proskurowski, G., Lilley, M. D., Seewald, J. S., Früh-Green, G. L., Olson, E. J., Lupton, J. E., ... & Kelley, D. S. (2008). Abiogenic hydrocarbon production at Lost City hydrothermal field. *Science*, 319(5863), 604-607.
- Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., ... & Rokhsar, D. S. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *science*, 317(5834), 86-94.
- Rafels-Ybern, À., Torres, A. G., Grau-Bove, X., Ruiz-Trillo, I., & Ribas de Pouplana, L. (2018). Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla. *RNA biology*, 15(4-5), 500-507.
- Ragan, M. A., Goggin, C. L., Cawthorn, R. J., Cerenius, L., Jamieson, A. V., Plourde, S. M., ... & Gutell, R. R. (1996). A novel clade of protistan parasites near the animal-fungal divergence. *Proceedings of the National Academy of Sciences*, 93(21), 11907-11912.
- Rajamani, S., Vlassov, A., Benner, S., Coombs, A., Olasagasti, F., & Deamer, D. (2008). Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Origins of Life and Evolution of Biospheres*, 38(1), 57-74.
- Ramos-Morales, E., Bayam, E., Del-Pozo-Rodríguez, J., Salinas-Giegé, T., Marek, M., Tilly, P., ... & Romier, C. (2021). The structure of the mouse ADAT2/ADAT3 complex reveals the molecular basis for mammalian tRNA wobble adenosine-to-inosine deamination. *Nucleic Acids Research*, 49(11), 6529-6548.
- Rassat, J., & Ruthmann, A. (1979). *Trichoplax adhaerens* FE Schulze (Placozoa) in the scanning electron microscope. *Zoomorphologie*, 93(1), 59-72.

- Reader, J. S., Metzgar, D., Schimmel, P., & de Crécy-Lagard, V. (2004). Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *Journal of Biological Chemistry*, 279(8), 6280-6285.
- Redmond, A. K., & McLysaght, A. (2021). Evidence for sponges as sister to all other animals from partitioned phylogenomics with mixture models and recoding. *Nature communications*, 12(1), 1-14.
- Reisser, T., Langgut, W., & Kersten, H. (1994). The nutrient factor queuine protects HeLa cells from hypoxic stress and improves metabolic adaptation to oxygen availability. *European journal of biochemistry*, 221(3), 979-986.
- Rezgui, V. A. N., Tyagi, K., Ranjan, N., Konevega, A. L., Mittelstaet, J., Rodnina, M. V., ... & Pedrioli, P. G. (2013). tRNA tKUUU, tQUUG, and tEUUC wobble position modifications fine-tune protein translation by promoting ribosome A-site binding. *Proceedings of the National Academy of Sciences*, 110(30), 12289-12294.
- Richards, T. A., & Cavalier-Smith, T. (2005). Myosin domain evolution and the primary divergence of eukaryotes. *Nature*, 436(7054), 1113-1118.
- Richter, D. J. (2013). *The gene content of diverse choanoflagellates illuminates animal origins* (Doctoral dissertation, UC Berkeley).
- Richter, D. J., Fozouni, P., Eisen, M. B., & King, N. (2018). Gene family innovation, conservation and loss on the animal stem lineage. *Elife*, 7, e34226.
- Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A. J., Gray, M. W., Philippe, H., & Lang, B. F. (2007). Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Current Biology*, 17(16), 1420-1425.
- Rodriguez-Hernandez, A., Spears, J. L., Gaston, K. W., Limbach, P. A., Gamper, H., Hou, Y. M., ... & Perona, J. J. (2013). Structural and mechanistic basis for enhanced translational efficiency by 2-thiouridine at the tRNA anticodon wobble position. *Journal of molecular biology*, 425(20), 3888-3906.
- Rogalski, M., Karcher, D., & Bock, R. (2008). Superwobbling facilitates translation with reduced tRNA sets. *Nature structural & molecular biology*, 15(2), 192-198.
- Roger, A. J., & Simpson, A. G. (2009). Evolution: revisiting the root of the eukaryote tree. *Current biology*, 19(4), R165-R167.

- Rogozin, I. B., Basu, M. K., Csürös, M., & Koonin, E. V. (2009). Analysis of rare genomic changes does not support the unikont–bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome biology and evolution*, 1, 99-113.
- Romero, H., Zavala, A., & Musto, H. (2000). Compositional pressure and translational selection determine codon usage in the extremely GC-poor unicellular eukaryote *Entamoeba histolytica*. *Gene*, 242(1-2), 307-311.
- Rowe, G. W., & Trainor, L. E. H. (1983). A thermodynamic theory of codon bias in viral genes. *Journal of theoretical biology*, 101(2), 171-203.
- Roy, S. W. (2003). Recent evidence for the exon theory of genes. *Genetica*, 118(2-3), 251-266.
- Ruiz-Trillo, I., Y., Davis, L. A., Sperstad, S., Landfald, B., & Roger, A. J. (2004). *Capsaspora owczarzaki* is an independent opisthokont lineage. *Current Biology*, 14(22), R946-R947.
- Ruiz-Trillo, I., Lane, C. E., Archibald, J. M., & Roger, A. J. (2006). Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *Journal of Eukaryotic Microbiology*, 53(5), 379-384.
- Ruiz-Trillo, I., Roger, A. J., Burger, G., Gray, M. W., & Lang, B. F. (2008). A phylogenomic investigation into the origin of metazoa. *Molecular biology and evolution*, 25(4), 664-672.
- Ryan, J. F., Pang, K., Schnitzler, C. E., Nguyen, A. D., Moreland, R. T., Simmons, D. K., ... & Baxevanis, A. D. (2013). The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*, 342(6164).
- Saha, R., Pohorille, A., & Chen, I. A. (2014). Molecular crowding and early evolution. *Origins of Life and Evolution of Biospheres*, 44(4), 319-324.
- Saint-Léger, A., Bello, C., Dans, P. D., Torres, A. G., Novoa, E. M., Camacho, N., ... & De Pouplana, L. R. (2016). Saturation of recognition elements blocks evolution of new tRNA identities. *Science advances*, 2(4), e1501860.

- Salim, H. M., Ring, K. L., & Cavalcanti, A. R. (2008). Patterns of codon usage in two ciliates that reassign the genetic code: *Tetrahymena thermophila* and *Paramecium tetraurelia*. *Protist*, 159(2), 283-298.
- Sato, N. (2020). Complex origins of chloroplast membranes with photosynthetic machineries: multiple transfers of genes from divergent organisms at different times or a single endosymbiotic event?. *Journal of plant research*, 133(1), 15-33.
- Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., ... & Ye, J. (2010). Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl_1), D38-D51.
- Schöning, K. U., Scholz, P., Guntha, S., Wu, X., Krishnamurthy, R., & Eschenmoser, A. (2000). Chemical etiology of nucleic acid structure: the α -threofuranosyl-(3'→2') oligonucleotide system. *Science*, 290(5495), 1347-1351.
- Schopf, J. W. (1993). Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. *Science*, 260(5108), 640-646.
- Schopf, J. W. (2006). Fossil evidence of Archaean life. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1470), 869-885.
- Schopf, J. W., Kudryavtsev, A. B., Czaja, A. D., & Tripathi, A. B. (2007). Evidence of Archean life: stromatolites and microfossils. *Precambrian Research*, 158(3-4), 141-155.
- Schwartz, A. W., & Orgel, L. E. (1985). Template-directed synthesis of novel, nucleic acid-like structures. *Science*, 228(4699), 585-587.
- Shalchian-Tabrizi, K., Minge, M. A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K. S., & Cavalier-Smith, T. (2008). Multigene phylogeny of choanozoa and the origin of animals. *PloS one*, 3(5).
- Sharp, P. M., & Devine, K. M. (1989). Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do [prefer [optimal codons. *Nucleic Acids Research*, 17(13), 5029-5040.
- Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of molecular evolution*, 24(1-2), 28-38.

- Sharp, P. M., Tuohy, T. M., & Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic acids research*, 14(13), 5125-5143.
- Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., & Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic acids research*, 16(17), 8207-8211.
- Sharp, P. M., & Lloyd, A. T. (1993). Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucleic Acids Research*, 21(2), 179-183.
- 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 research*, 33(4), 1141-1153.
- Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544), 1203-1212.
- Shoshan, E., Mobley, A. K., Braeuer, R. R., Kamiya, T., Huang, L., Vasquez, M. E., ... & Bar-Eli, M. (2015). Reduced adenosine-to-inosine miR-455-5p editing promotes melanoma growth and metastasis. *Nature cell biology*, 17(3), 311-321.
- Shrivastava, S., Poddar, R., Shukla, P., & Mukhopadhyay, K. (2009). Study of codon bias perspective of fungal xylanase gene by multivariate analysis. *Bioinformation*, 3(10), 425.
- Sievers, A., Beringer, M., Rodnina, M. V., & Wolfenden, R. (2004). The ribosome as an entropy trap. *Proceedings of the National Academy of Sciences*, 101(21), 7897-7901.
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma*. 31, 3210–3212.

- Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... & Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, 27(7), 958-967.
- Sloan, D. B., Nakabachi, A., Richards, S., Qu, J., Murali, S. C., Gibbs, R. A., & Moran, N. A. (2014). Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular biology and evolution*, 31(4), 857-871.
- Smith, N. G., & Eyre-Walker, A. (2001). Synonymous codon bias is not caused by mutation bias in G+ C-rich genes in humans. *Molecular Biology and Evolution*, 18(6), 982-986.
- Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195-197.
- Snoke, M. S., Berendonk, T. U., Barth, D., & Lynch, M. (2006). Large global effective population sizes in Paramecium. *Molecular biology and evolution*, 23(12), 2474-2479.
- Southworth, J., Armitage, P., Fallon, B., Dawson, H., Bryk, J., & Carr, M. (2018). Patterns of ancestral animal codon usage Bias revealed through Holozoan Protists. *Molecular biology and evolution*, 35(10), 2499-2511.
- Spiess, F. N., Macdonald, K. C., Atwater, T., Ballard, R., Carranza, A., Cordoba, D., ... & Hawkins, J. (1980). East Pacific Rise: hot springs and geophysical experiments. *Science*, 207(4438), 1421-1433.
- Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., ... & Rokhsar, D. S. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*, 466(7307), 720-726.
- Staden, R., & McLachlan, A. D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic acids research*, 10(1), 141-156.
- Stahl, G., McCarty, G. P., & Farabaugh, P. J. (2002). Ribosome structure: revisiting the connection between translational accuracy and unconventional decoding. *Trends in biochemical sciences*, 27(4), 178-183.

- Stechmann, A., & Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. *Science*, 297(5578), 89-91.
- Steenkamp, E. T., Wright, J., & Baldauf, S. L. (2006). The protistan origins of animals and fungi. *Molecular biology and evolution*, 23(1), 93-106.
- Stenico, M., Lloyd, A. T., & Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic acids research*, 22(13), 2437-2446.
- Suga, H., Koyanagi, M., Hoshiyama, D., Ono, K., Iwabe, N., Kuma, K. I., & Miyata, T. (1999). Extensive gene duplication in the early evolution of animals before the parazoan–eumetazoan split demonstrated by G proteins and protein tyrosine kinases from sponge and hydra. *Journal of molecular evolution*, 48(6), 646-653.
- Suga, H., Katoh, K., & Miyata, T. (2001). Sponge homologs of vertebrate protein tyrosine kinases and frequent domain shufflings in the early evolution of animals before the parazoan–eumetazoan split. *Gene*, 280(1-2), 195-201.
- Sun, L., Cui, Z., Gottlieb, R. L., & Zhang, B. (2002). A selected ribozyme catalyzing diverse dipeptide synthesis. *Chemistry & biology*, 9(5), 619-628.
- Suzuki, T., Ueda, H., Okada, S., & Sakurai, M. (2015). Transcriptome-wide identification of adenosine-to-inosine editing using the ICE-seq method. *Nature protocols*, 10(5), 715-732.
- Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. *Journal of theoretical Biology*, 112(2), 333-343.
- Szathmáry, E., & Smith, J. M. (1995). The major evolutionary transitions. *Nature*, 374(6519), 227-232.
- Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3), 227.
- Tamura, K., & Schimmel, P. (2003). Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *Proceedings of the National Academy of Sciences*, 100(15), 8666-8669.
- Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338), 631-637.

- Taylor, D. J., Ballinger, M. J., Bowman, S. M., & Bruenn, J. A. (2013). Virus-host co-evolution under a modified nuclear genetic code. *PeerJ*, 1, e50.
- Thiaville, P. C., El Yacoubi, B., Köhrer, C., Thiaville, J. J., Deutsch, C., Iwata-Reuyl, D., ... & de Crécy-Lagard, V. (2015). Essentiality of threonylcarbamoyladenosine (t6 A), a universal t RNA modification, in bacteria. *Molecular microbiology*, 98(6), 1199-1221.
- Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research*, 13(9), 2129-2141.
- Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., ... & Doremieux, O. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*, 31(1), 334-341.
- Thomas, J. A., Welch, J. J., Lanfear, R., & Bromham, L. (2010). A generation time effect on the rate of molecular evolution in invertebrates. *Molecular biology and evolution*, 27(5), 1173-1180
- Tikhonenkov, D. V., Mikhailov, K. V., Hehenberger, E., Karpov, S. A., Prokina, K. I., Esaulov, A. S., ... & Keeling, P. J. (2020). New lineage of microbial predators adds complexity to reconstructing the evolutionary origin of animals. *Current Biology*, 30(22), 4500-4509.
- Tomassetti, P., Voigt, O., Collins, A. G., Porrello, S., Pearse, V. B., & Schierwater, B. (2005). Placozoans (*Trichoplax adhaerens* Schulze 1883) in the Mediterranean sea. *Meiofauna Marina*, 14, 5-7.
- Torruella, G., Derelle, R., Paps, J., Lang, B. F., Roger, A. J., Shalchian-Tabrizi, K., & Ruiz-Trillo, I. (2012). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Molecular Biology and Evolution*, 29(2), 531-544.
- Torruella, G., De Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., Del Campo, J., ... & Ruiz-Trillo, I. (2015). Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Current Biology*, 25(18), 2404-2410.

- Tucker, R. P. (2013). Horizontal gene transfer in choanoflagellates. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 320(1), 1-9.
- Tuorto, F., Herbst, F., Alerasool, N., Bender, S., Popp, O., Federico, G., ... & Lyko, F. (2015). The tRNA methyltransferase Dnmt2 is required for accurate polypeptide synthesis during haematopoiesis. *The EMBO journal*, 34(18), 2350-2362.
- Vetsigian, K., Woese, C., & Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences*, 103(28), 10696-10701.
- Vicario, S., Moriyama, E. N., & Powell, J. R. (2007). Codon usage in twelve species of *Drosophila*. *BMC evolutionary biology*, 7(1), 1-17.
- Vicario, S., Mason, C. E., White, K. P., & Powell, J. R. (2008). Developmental stage and level of codon usage bias in *Drosophila*. *Molecular biology and evolution*, 25(11), 2269-2277.
- Vischer, W. (1945). *Über einen pilzähnlichen, autotrophen Mikroorganismus, Chlorochytridion, einige neue Protococcales und die systematische Bedeutung der Chloroplasten*. Buchdruckerei Emil Birkhäuser.
- Voigt, O., Collins, A. G., Pearse, V. B., Pearse, J. S., Ender, A., Hadrys, H., & Schierwater, B. (2004). Placozoa—no longer a phylum of one. *Current Biology*, 14(22), R944-R945.
- Walker, G., Dacks, J. B., & Martin Embley, T. (2006). Ultrastructural description of *Breviata anathema*, n. gen., n. sp., the organism previously studied as “*Mastigamoeba invertens*”. *Journal of Eukaryotic Microbiology*, 53(2), 65-78.
- Warnecke, T., & Hurst, L. D. (2007). Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Molecular biology and evolution*, 24(12), 2755-2762.
- Warrick, H. M., & Spudich, J. A. (1988). Codon preference in *Dictyostelium discoideum*. *Nucleic acids research*, 16(14), 6617-6635.
- Weiss, M. C., Sousa, F. L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., & Martin, W. F. (2016). The physiology and habitat of the last universal common ancestor. *Nature microbiology*, 1(9), 1-8.

- Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Annals of the New York Academy of Sciences*, 1360(1), 16.
- Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences*, 112(18), 5773-5778.
- Whittle, C. A., Sun, Y., & Johannesson, H. (2012). Genome-wide selection on codon usage at the population level in the fungal model organism *Neurospora crassa*. *Molecular biology and evolution*, 29(8), 1975-1986.
- Wiedemann, B., & Boles, E. (2008). Codon-optimized bacterial genes improve L-arabinose fermentation in recombinant *Saccharomyces cerevisiae*. *Applied and environmental microbiology*, 74(7), 2043-2050.
- Woese, C. R. (1965). On the evolution of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 54(6), 1546.
- Woese, C. R., & Fox, G. E. (1977). The concept of cellular evolution. *Journal of molecular evolution*, 10(1), 1-6.
- Woese, C. R., Hinegardner, R. T., & Engelberg, J. (1964). Universality in the genetic code. *Science*, 144(3621), 1030-1031.
- Woese, C. R., Dugre, D. H., Dugre, S. A., Kondo, M., & Saxinger, W. C. (1966, January). On the fundamental nature and evolution of the genetic code. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 31, pp. 723-736). Cold Spring Harbor Laboratory Press.
- Wolf, Y. I., & Koonin, E. V. (2007). On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization. *Biology Direct*, 2(1), 1-25.
- Wolf, J., Gerber, A. P., & Keller, W. (2002). tadA, an essential tRNA-specific adenosine deaminase from *Escherichia coli*. *The EMBO journal*, 21(14), 3841-3851.
- Wong, J. T. F. (1975). A co-evolution theory of the genetic code. *Proceedings of the National Academy of Sciences of the United States of America*, 72(5), 1909.
- Wong, J. T. F. (1991). Origin of genetically encoded protein synthesis: a model based on selection for RNA peptidation. *Origins of Life and Evolution of the Biosphere*, 21(3), 165-176.

- Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene*, 87(1), 23-29.
- Yang, X., Luo, X., & Cai, X. (2014). Analysis of codon usage pattern in *Taenia saginata* based on a transcriptome dataset. *Parasites & vectors*, 7(1), 1-11.
- Yang, X., Ma, X., Luo, X., Ling, H., Zhang, X., & Cai, X. (2015). Codon usage bias and determining forces in *Taenia solium* genome. *The Korean journal of parasitology*, 53(6), 689.
- Yannai, A., Katz, S., & Hershberg, R. (2018). The codon usage of lowly expressed genes is subject to natural selection. *Genome biology and evolution*, 10(5), 1237-1246.
- Yarus, M. (1998). Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *Journal of molecular evolution*, 47(1), 109-117.
- Yarus, M., Caporaso, J. G., & Knight, R. (2005). Origins of the genetic code: the escaped triplet theory. *Annu. Rev. Biochem.*, 74, 179-198.
- Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2), 217-235.
- Yokoyama, S., & Nishimura, S. (1995). Modified nucleosides and codon recognition. In *tRNA: Structure, Biosynthesis and Function* (Soëll, D. & RajBhandary, UL, eds).
- Yu, Y., Harris, A. J., Blair, C., & He, X. (2015). RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. *Molecular phylogenetics and evolution*, 87, 46-49.
- Yuen, G. U., & Kvenvolden, K. A. (1973). Monocarboxylic acids in Murray and Murchison carbonaceous meteorites. *Nature*, 246(5431), 301-303.
- Yuen, G. U., Lawless, J. G., & Edelson, E. H. (1981). Quantification of monocarboxylic acids from a spark discharge synthesis. *Journal of molecular evolution*, 17(1), 43-47.
- Zaher, H. S., & Unrau, P. J. (2007). Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. *Rna*, 13(7), 1017-1026.
- Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., ... & Ettema, T. J. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541(7637), 353-358.

- Zaug, A. J., & Cech, T. R. (1986). The intervening sequence RNA of *Tetrahymena* is an enzyme. *Science*, 231(4737), 470-475.
- Zettler, L. A. A., Nerad, T. A., O'Kelly, C. J., & Sogin, M. L. (2001). The nucleariids amoebae: more protists at the animal-fungal boundary. *Journal of Eukaryotic Microbiology*, 48(3), 293-297.
- Zhang, L., Yong, G. U. O., Ling, L., Yue-Ping, W. A. N. G., Zhi-Min, D. O. N. G., Shou-Hong, S. U. N., & Li-Juan, Q. I. U. (2011). Analysis of nuclear gene codon bias on soybean genome and transcriptome. *Acta Agronomica Sinica*, 37(6), 965-974.
- Zhang, X., Walker, R. C., Phizicky, E. M., & Mathews, D. H. (2014). Influence of sequence and covalent modifications on yeast tRNA dynamics. *Journal of chemical theory and computation*, 10(8), 3473-3483.
- Zhao, F., Yu, C. H., & Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic acids research*, 45(14), 8484-8492.
- Zhao, H., & Chen, T. (2013). Tet family of 5-methylcytosine dioxygenases in mammalian development. *Journal of human genetics*, 58(7), 421-427.
- Zhao, F., Yu, C. H., & Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. *Nucleic acids research*, 45(14), 8484-8492.
- Zhou, J., Liu, W. J., Peng, S. W., Sun, X. Y., & Frazer, I. (1999). Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability. *Journal of virology*, 73(6), 4972-4982.
- Zrzavý, J., Mihulka, S., Kepka, P., Bezděk, A., & Tietz, D. (1998). Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics*, 14(3), 249-285.