Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations | Graduate School

2015

MYSTERIES OF THE TRYPANOSOMATID MAXICIRCLES: CHARACTERIZATION OF THE MAXICIRCLE GENOMES AND THE EVOLUTION OF RNA EDITING IN THE ORDER KINETOPLASTIDA

Preethi Ranganathan Iyengar
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

Part of the Bioinformatics Commons, Biology Commons, Computational Biology Commons, Microbiology Commons, and the Molecular Biology Commons

© The Author

Downloaded from
https://scholarscompass.vcu.edu/etd/4010

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

©Preethi Ranganathan Iyengar 2015

All rights reserved

MYSTERIES OF THE TRYPANOSOMATID MAXICIRCLES: CHARACTERIZATION OF THE MAXICIRCLE GENOMES AND THE EVOLUTION OF RNA EDITING IN THE ORDER KINETOPLASTIDA

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

By

Preethi Ranganathan Iyengar
M. Sc. (Biotechnology), University of Mumbai, India, 2009
B. Sc. (Biotechnology), University of Mumbai, India, 2007

Director: Dr. Gregory A. Buck
Professor,
Department of Microbiology and Immunology and
Center for the Study of Biological Complexity

Virginia Commonwealth University
Richmond, Virginia
November, 2015

ACKNOWLEDGEMENT

I am indebted to Dr. Gregory Buck BIG time! I have no words to express my appreciation for his encouragement and support all through my graduate years here at VCU. For some reason, it appeared as though he always believed in me more than I ever have. I admired his laid-back attitude and non-pushy ways; especially because this project was a huge leap for me, given my background; and he waited patiently for me to conquer that steep learning curve, while still providing ample space, time and motivation! Despite his hectic schedule and travel, he has always made an effort to help me, be there and offer some much-needed advice and suggestions from time-to-time. Be it offering to help me rehearse a talk over the phone on a weekend night, or working with me tirelessly through our endless corrections and re-corrections of our manuscripts, he has been a fabulous advisor! He has been very considerate and inspiring.

I want to thank all my committee members. I do not have enough good words for Dr. Kimberly Jefferson and Dr. Francine Cabral. I have known them both the longest. Kim had been a co-mentor, and literally worked on the bench next to me and taught me, during my first rotation. I got to know Francine through one of the Microbiology classes in my first year. Both these ladies deserve a special mention for quickly befriending me and always being around for any kind of help and support. I have always felt very lucky to be held in such high regards by both of them. They are both an epitome of peacefulness and resourcefulness; and have always been the first people to talk for me and have my back. I got to know Dr. Ping Xu and Dr. Maria Rivera only after they got aboard my committee; but I am ever grateful to them for their advice through some of my hard times, their suggestions and their constructive criticism of my work. I owe all my knowledge of Phylogeny to Maria, and she has been a fantastic teacher and course director. I have to admit that I knew nothing about Phylogeny before my PhD, and now, a big part of my work involves phylogeny! Overall, I have always known that my committee has been extremely supportive of me and patient with me through my lag phase. They have seen me through thick and thin, and have always made time to meet with me, talk and advise.

I have to thank my Program Director, Dr. Gail Christie, for her time and help since even before I joined VCU. In a way, she is responsible for me even getting my visa! I have ALWAYS received a reply to all my emails in less than 10 mins! She has seen me grow, seen me break down, heard me out patiently through my sobs and frustrations, motivated me, and has always been a shadow over my head! I don't think I would have even finished my Ph.D., to say the least, without her support!

In terms of my work and the kDNA project, Dr. Myrna Serrano has been the closest to me. I cannot thank her enough for all our hours of discussions, brain-storming, and trouble-shooting. From failures to some exciting results and fantastic discoveries, we have seen them all, together! She has also been my "go-to" advisor in Dr. Buck's absence. Dr. Andrey Matveyev is easily one of the most resourceful persons I've ever met! He was among the few people I was initially introduced to, and had the privilege of working closely with. I have come to believe that

he apparently has this "Midas' touch" in his fingers (no, literally!), and all of his ideas, methods and molecular techniques ALWAYS work, no matter what! He is easily one of the funniest, honest and most genuine persons around (despite that sarcastic and wry sense of humor!); and working with him was easily my nicest phase in the lab. I also hold Dr. Vishal Koparde in very high regards. Knowing where I come from, he has been immensely helpful and extra-ordinarily patient, understanding and nice with me! From running small bioinformatic analyses to helping me make some cheat-sheets of shell commands, he has literally saved my day… And several times that is! What deserves a mention here is that he has never denied help to me, and has tried his best to rescue me from my situations in whatever way he can. I also thank Dr. Joao Alves, our lab's ex-post-doctoral research associate and bioinformaticist, for his continued help and support even years after his departure. Jay (as we call him) has saved so much of my time and taught me so much, despite the fact that he's in a different country and time zone now. He still sat patiently through many of our video camera-shared screen sessions, each of several hours, to help me decipher some of my results and add great inputs.

I have also worked closely with Ruth Carvalho and YingPing Wang from the Sequencing Core (NARF-VCU). They are both incredibly nice and helpful people, in their own special ways. From a million orders of primers and oligos, to innumerable sequencings and re-sequencings, I think a big chunk of my project can easily be dedicated to both of them!

I do not think I can ever thank our lab-tech, Ania, enough. She has been a friend, a sister, a critique, a family and even my bridesmaid! From teaching me many lab shortcuts to some complicated techniques, she has been incredibly supportive and unbelievably reliable. Her ability to adapt, learn, work hard and excel is truly inspiring. She kept me grounded, and was always there if I needed to just talk and vent it out to someone. We both agree that dogs are the best things on this planet, and if all else failed, we'd just go shopping and dinner! Best time ever! Katie, my fellow grad student and a great friend; I and she have more than one thing in common. Taking quick coffee-breaks, going out for lunches, gushing over our significant others, crying to each other over frustrations… we have done it all… in the true spirit of a graduate student! Life has been incredibly hard (yes Katie, I hear you!) and we have both been through several ups and downs, common difficult situations, we have talked, cried, hugged and moved on. We have our share of secrets that we shall keep! Finally, special thanks to Ana Lara for teaching me parasite culturing and maintenance; and also patiently doing the ordering and management of all my reagents and stock supplies.

We keep the best for the last! I do not think any of this would have been possible if it was not for my parents, who are undeniably my biggest pillars of strength, my moral support and my unending reservoirs of all good things in life! I know for a given that I am what I am only because of their prayers and blessings. They've always let me pursue my interests and it was their encouragement and belief in me that saw me through everything in my life. They insist on talking to me to just hear my voice (and to ensure that I'm not starving!) EVERY SINGLE DAY! They were also the ones to introduce me to my now husband, Vishal. I have no doubts that they are my guardian angels. My maternal Granny deserves her own special little place here. She has raised me since I was born and has seen me grow. Around the exam seasons back in

India, she would drop everything and travel to be around me, pamper me, feed me and make me endless teas and coffees as I mulled over my books all my life! She was the most happy and proud when her favorite granddaughter received an admit from the US for my Ph.D!

I have to thank my loving husband for all his patience, calmness, practicality and unconditional love. But for his sacrifices, I do not know if our long-distance relationship (of almost 5 years now!) could have even survived. Having gone through a difficult Ph.D. himself, he is all eyes and ears to my tiniest troubles, and always comes in with valuable suggestions. When our virtual realities do not appease me, he quietly puts everything on hold and takes a flight or drives all the way from Atlanta just to wipe a tear, and for a simple comforting bear hug, and for his "this too shall pass" reassurance. I can say that my parents know me well and have chosen well! Not just my husband, but my mom-in-law, papa-in-law and grandparents-in-law have also been a terrific addition to my life! They are always available to hear me out patiently, advise and keep tabs on my health and wellbeing. My new family has known me for only about 4 years now, but they are already my other biggest source of strength. My families are undoubtedly my biggest blessings and they are the ones who keep me going! I can't wait to finish my Ph.D. and move to Atlanta to start a new life with my husband (after almost 3 years of marriage)!! It's about time!

TABLE OF CONTENTS

S. No.		Page
	List of Tables…….	ix
	List of Figures……	xiv
	List of Alignments…….	xvii
	List of Abbreviations…….	xx
	Abstract……..	xxii
1.	**Background, Objectives and Purpose of the study……...**	**1**
2.	**Literature Review……**	**4**
	2.1 Order Kinetoplastida: Diversity and Evolution……	4
	2.2 The curious case of trypanosomatids……	8
	2.3 Evolution of parasitism in Order Kinetoplastida……...	9
	2.4 Important trypanosomatid parasites and their life cycles……..	11
	2.4.1 Human African Trypanosomiasis……...……	12
	2.4.2 Chagas Disease……..	12
	2.4.3 Leishmaniases……	13
	2.5 The trypanosomatid mitochondrion and kinetoplast…….	14
	2.5.1 Maxicircles…….	16
	2.5.2 Minicircles……..	18
	2.5.3 Why do the kinetoplastids have kDNA networks?.....	19
	2.6 Trypanosomatid mitochondrial RNA editing……	20
	2.7 Mechanism of U insertion-deletion RNA editing……..	21
	2.8 Evolution of RNA editing in the trypanosomatids…….	24
	2.9 Functional implications of RNA editing……	25
	2.10 RNA editing and life-cycle stages of *T. brucei*……	27
	2.11 Trypanosomatid mitochondrial bioenergetics……..	29

2.12 Variety in the trypanosomatid energy metabolism…….. 31
2.13 Mitochondrion as a chemotherapeutic target……... 34
2.13.1 Electron transport chain……. 34
2.13.2 Trypanosomatid alternative oxidase (TAO)....…….. 35
2.13.3 kDNA replication…….. 35
2.13.4 Mitochondrial tRNA import……... 36
2.13.5 Mitochondrial fatty acid and sterol synthesis…… 36
2.13.6 Mitochondrial Membrane Potential (MMP)…….. 37
2.13.7 Mitochondrial Reactive Oxygen Species (ROS)…..... 37
2.13.8 Induction of Apoptosis……..... 38

3. The maxicircle genomes of endosymbiont-bearing and related insect trypanosomatids…… 39
Abstract……... 39
Introduction……. 40
Abbreviations……...……... 44
Materials and Methods…… 45
Results……. 50
Maxicircle genome characteristics…….. 50
RNA editing patterns show distinct similarities and differences 58
Genes not known to undergo editing……... 58
Genes that undergo partial (5’ or internal) editing….. 62
Genes that undergo extensive (pan) editing……. 71
Phylogenetic relationships and the evolution of RNA editing. 75
Discussion……..... 78
Conclusion……..... 84
Supplementary Tables for Chapter 3…… 85
Supplementary figures for Chapter 3……... 87
Supporting Information (SI) for Chapter 3……... 90

4. The maxicircle genomes of several trypanosomatids of the *Trypanosoma* group (*T. cruzi* related and unrelated parasites)…… 185

Abstract…… 185

Introduction……. 187

Abbreviations.....…... 192

Materials and Methods…… 193

Results…….. 198

Maxicircle genome characteristics……... 198

The RNA editing patterns are remarkably similar among all the members of the *Trypanosoma* genus in this report…….. 211

Genes that are not known to undergo editing……... 211

Genes that undergo partial (5' or internal) editing…… 216

Genes that undergo extensive (pan) editing…….. 223

Phylogenetic relationships and evolution of RNA editing…... 229

Discussion……. 233

Conclusion…… 242

Supplementary Tables for Chapter 4…… 243

Supplementary Figures for Chapter 4…….. 245

Supporting Information (SI) for Chapter 4…….. 249

5. References…… 369

6. Vita…… 392

LIST OF TABLES

Table 1.1	List of organisms in this report……..	2
Table 3.1	Maxicircle genome statistics……..	51
Table 3.2A	Coding gene lengths with co-ordinates and start - stop codons for *L.tar*, *E. sch*, *L. cos, C. aca, H. mus* and *P. ser*……	53
Table 3.2B	Coding gene lengths with co-ordinates and start - stop codons for *A. dea, A. des*, *S. cul, S. gal* and *S. onc*……..	55
Table 3.3	Percent identity matrix of the syntenic 20-gene region of maxicircle genomes…….	57
Table 3.4A	Summary of the inferred editing patterns in genes (*L. tar, E. sch, L. cos, C. aca, H. mus, P. ser*)……..	67
Table 3.4B	Summary of the inferred editing patterns in genes (*A. dea, A. des, S. cul, S. gal, S. onc*)……..	69
S. Table 3.1	Primers used for gap-closures and circularization of the maxicircle contigs……...	85
S. Table 3.2	Percent identity matrix for 18S and GAPDH (nuclear) genes for the trypanosomatids……	86
3.SI-S1b	12S nucleotide percent identity matrix…….	93
3.SI-S2b	9S nucleotide percent identity matrix……...	96
3.SI-S3c	COI nucleotide percent identity matrix……	103
3.SI-S3d	COI protein percent identity matrix…….....	103
3.SI-S4c	MURF1 nucleotide percent identity matrix…….	109
3.SI-S4d	MURF1 protein percent identity matrix……...	109
3.SI-S5c	ND1 nucleotide percent identity matrix……	114
3.SI-S5d	ND1 protein percent identity matrix…….	114
3.SI-S6c	ND4 nucleotide percent identity matrix……	120
3.SI-S6d	ND4 protein percent identity matrix…….	120
3.SI-S7c	ND5 nucleotide percent identity matrix……	128

3.SI-S7d	ND5 protein percent identity matrix…….	128
3.SI-S8c	A6 edited transcript percent identity matrix…….	132
3.SI-S8d	A6 protein percent identity matrix……	132
3.SI-S9c	COII edited transcript percent identity matrix…….	136
3.SI-S9d	COII protein percent identity matrix…….	136
3.SI-S10c	COIII edited transcript percent identity matrix……	141
3.SI-S10d	COIII protein percent identity matrix……	141
3.SI-S11c	CYb edited transcript percent identity matrix……	146
3.SI-S11d	CYb protein percent identity matrix…….	146
3.SI-S12c	MURF2 edited transcript percent identity matrix……	151
3.SI-S12d	MURF2 protein percent identity matrix……	151
3.SI-S13c	ND7 edited transcript percent identity matrix……..	157
3.SI-S13d	ND7 protein percent identity matrix……	157
3.SI-S14c	MURF5 nucleotide percent identity matrix…….	161
3.SI-S14d	MURF5 protein percent identity matrix……	161
3.SI-S15c	ND8 edited transcript percent identity matrix……..	165
3.SI-S15d	ND8 protein percent identity matrix……	165
3.SI-S16c	ND9 edited transcript percent identity matrix…….	170
3.SI-S16d	ND9 protein percent identity matrix……	170
3.SI-S17c	RPS12 edited transcript percent identity matrix…….	173
3.SI-S17d	RSP12 protein percent identity matrix……	173
3.SI-S18c	ND3 edited transcript percent identity matrix (conserved motif)……...	177
3.SI-S18d	ND3 protein percent identity matrix (conserved motif)……..	177
3.SI-S19c	G3 edited transcript percent identity matrix……	180
3.SI-S19d	G3 protein percent identity matrix…….	180

3.SI-S20c	G4 edited transcript percent identity matrix……	184
3.SI-S20d	G4 protein percent identity matrix……	184
Table 4.1A	Maxicircle genome statistics (Non *T. cruzi Trypanosoma* members)..	199
Table 4.1B	Maxicircle genome statistics (*T. cruzi* strains)…….	199
Table 4.2A	Coding gene lengths with co-ordinates and start - stop codons for non-*T. cruzi Trypanosoma* members……	203
Table 4.2B	Coding gene lengths with co-ordinates and start - stop codons for *T. cruzi* DTUs VI, III and IV……	205
Table 4.2C	Coding gene lengths with co-ordinates and start - stop codons for *T. cruzi* DTUs I, II and *TcMar*…….	207
Table 4.3	Percent identity matrix of the syntenic 20-gene region of maxicircle genomes……	209
Table 4.4A	Inferred RNA editing patterns in genes of non-*T. cruzi Trypanosoma* members……	219
Table 4.4B	Inferred RNA editing patterns in genes of *T. cruzi* DTUs VI, III and IV	220
Table 4.4C	Inferred RNA editing patterns in genes of *T. cruzi* DTUs I, II and *TcMar*…….	221
S. Table 4.1	Primers used for gap-closures and circularization of the maxicircle contigs……	243
S. Table 4.2	Percent identity matrix for GAPDH gene for the trypanosomatids…..	243
S. Table 4.3	Percent identity matrix for DHFR gene for the trypanosomatids……..	244
S. Table 4.4	Percent identity matrix for GSH1 gene for the trypanosomatids……..	244
4.SI-S1b	12S nucleotide percent identity matrix……..	254
4.SI-S2b	9S nucleotide percent identity matrix……	257
4.SI-S3c	MURF1 nucleotide percent identity matrix……..	265
4.SI-S3d	MURF1 protein percent identity matrix……	265
4.SI-S4c	ND1 nucleotide percent identity matrix……	272
4.SI-S4d	ND1 protein percent identity matrix…….	272
4.SI-S5c	COI nucleotide percent identity matrix…….	281
4.SI-S5d	COI protein percent identity matrix……..	281

4.SI-S6c	ND4 nucleotide percent identity matrix…….	289
4.SI-S6d	ND4 protein percent identity matrix……..	289
4.SI-S7c	ND5 nucleotide percent identity matrix…….	298
4.SI-S7d	ND5 protein percent identity matrix……..	298
4.SI-S8c	CYb edited transcript percent identity matrix……	306
4.SI-S8d	CYb protein percent identity matrix……..	306
4.SI-S9c	COII edited transcript percent identity matrix……	311
4.SI-S9d	COII protein percent identity matrix……..	311
4.SI-S10c	MURF2 edited transcript percent identity matrix……..	319
4.SI-S10d	MURF2 protein percent identity matrix…….	319
4.SI-S11c	MURF5 nucleotide percent identity matrix…….....	322
4.SI-S11d	MURF5 protein percent identity matrix…….	322
4.SI-S12e	ND8 edited transcript percent identity matrix……	329
4.SI-S12f	ND8 protein percent identity matrix……..	329
4.SI-S13c	ND9 edited transcript percent identity matrix……	334
4.SI-S13d	ND9 protein percent identity matrix……..	334
4.SI-S14c	ND7 edited transcript percent identity matrix……	342
4.SI-S14d	ND7 protein percent identity matrix……..	342
4.SI-S15c	COIII edited transcript percent identity matrix…….	349
4.SI-S15d	COIII protein percent identity matrix……	349
4.SI-S16c	A6 edited transcript percent identity matrix……..	354
4.SI-S16d	A6 protein percent identity matrix……	354
4.SI-S17c	G3 edited transcript percent identity matrix…….	357
4.SI-S17d	G3 protein percent identity matrix……	357
4.SI-S18c	G4 edited transcript percent identity matrix…….	361
4.SI-S18d	G4 protein percent identity matrix……	361
4.SI-S19c	ND3 edited transcript percent identity matrix (for the conserved stretch)……	365

4.SI-S19d	ND3 protein percent identity matrix (for the conserved motif)……..	365
4.SI-S20c	RPS12 edited transcript percent identity matrix…….	368
4.SI-S20d	RPS12 protein percent identity matrix……	368

S. Table = Supplementary Table

SI = Supporting Information

LIST OF FIGURES

Figure 2.1 Infrakingdom Excavata and Phylum Euglenozoa…….. 5

Figure 2.2 Order Kinetoplastida…….. 6

Figure 2.3 Kinetoplast and proteins associated with kDNA replication……. 15

Figure 2.4 Single and multi-round kRNA editing……... 23

Figure 3.1 Syntenic 20-gene region of the trypanosomatid maxicircle genomes… 61

Figure 3.2 Maximum likelihood phylogenies depicting the relationships between the various trypanosomatids……..……. 77

S. Figure 3.1 Binary approach to quantitate editing events……. 87

S. Figure 3.2A *L. tar* 12S folding…… 88

S. Figure 3.2B *E. sch* 12S folding…… 88

S. Figure 3.2C *L. cos* 12S folding…… 88

S. Figure 3.2D *C. aca* 12S folding……... 88

S. Figure 3.2E *H. mus* 12S folding…….. 88

S. Figure 3.2F *P. ser* 12S folding…… 88

S. Figure 3.2G *A. dea* 12S folding…… 88

S. Figure 3.2H *A. des* 12S folding…… 88

S. Figure 3.2I *S. cul* 12S folding…… 88

S. Figure 3.2J *S. gal* 12S folding…… 88

S. Figure 3.2K *S. onc* 12S folding…… 88

S. Figure 3.3A *L. tar* 9S folding…….. 89

S. Figure 3.3B *E. sch* 9S folding…….. 89

S. Figure 3.3C *L. cos* 9S folding…….. 89

S. Figure 3.3D *C. aca* 9S folding…….. 89

S. Figure 3.3E *H. mus* 9S folding……. 89

S. Figure 3.3F *P. ser* 9S folding…… 89

S. Figure 3.3G *A. dea* 9S folding…….. 89

S. Figure 3.3H *A. des* 9S folding…….. 89

S. Figure 3.3I *S. cul* 9S folding…….. 89

S. Figure 3.3J *S. gal* 9S folding…….. 89

S. Figure 3.3K *S. onc* 9S folding…….. 89

Figure 4.1 The syntenic 20-gene region of the trypanosomatid maxicircle genomes 201

Figure 4.2 Maximum likelihood phylogenies depicting the relationships between the trypanosomatids……... 232

Figure 4.3 The revisited model for *Trypanosoma cruzi* evolution……..... 236

S. Figure 4.1A *L. tar* 12S folding……. 245

S. Figure 4.1B *Tbru* 12S folding…….. 245

S. Figure 4.1C *Tser* 12S folding…… 245

S. Figure 4.1D *Tlew* 12S folding…… 245

S. Figure 4.1E *Tran* 12S folding…… 245

S. Figure 4.1F *Tcon* 12S folding…… 245

S. Figure 4.1G *Tdio* 12S folding…… 245

S. Figure 4.1H *Tern* 12S folding…… 245

S. Figure 4.1I *TcCLB* 12S folding…… 246

S. Figure 4.1J *TcCL* 12S folding…….. 246

S. Figure 4.1K *TcTula* 12S folding…… 246

S. Figure 4.1L *TcM62* 12S folding…… 246

S. Figure 4.1M *TcCAN* 12S folding…….. 246

S. Figure 4.1N *TcG* 12S folding…… 246

S. Figure 4.1O *TcSyl* 12S folding……. 246

S. Figure 4.1P *TcJR* 12S folding…….. 246

S. Figure 4.1Q *TcEsm* 12S folding…… 246

S. Figure 4.1R *TcMar* 12S folding…… 246

S. Figure 4.2A *L. tar* 9S folding…… 247

S. Figure 4.2B *Tbru* 9S folding…… 247
S. Figure 4.2C *Tser* 9S folding…… 247
S. Figure 4.2D *Tlew* 9S folding…… 247
S. Figure 4.2E *Tran* 9S folding…… 247
S. Figure 4.2F *Tcon* 9S folding…… 247
S. Figure 4.2G *Tdio* 9S folding…… 247
S. Figure 4.2H *Tern* 9S folding…… 247
S. Figure 4.2I *TcCLB* 9S folding…… 248
S. Figure 4.2J *TcCL* 9S folding…….. 248
S. Figure 4.2K *TcTula* 9S folding…… 248
S. Figure 4.2L *TcM62* 9S folding…… 248
S. Figure 4.2M *TcCAN* 9S folding…….. 248
S. Figure 4.2N *TcG* 9S folding…… 248
S. Figure 4.2O *TcSyl* 9S folding…….. 248
S. Figure 4.2P *TcJR* 9S folding…….. 248
S. Figure 4.2Q *TcEsm* 9S folding…… 248
S. Figure 4.2R *TcMar* 9S folding…… 248

S. Figure = Supplementary Figure

LIST OF ALIGNMENTS

3.SI-S1a	GeneDoc 12S gene alignment……	90
3.SI-S2a	GeneDoc 9S gene alignment……..	94
3.SI-S3a	GeneDoc COI nucleotide alignment……..	97
3.SI-S3b	GeneDoc COI protein alignment……	101
3.SI-S4a	GeneDoc MURF1 / ND2 nucleotide alignment……	104
3.SI-S4b	GeneDoc MURF1 / ND2 protein alignment…….	108
3.SI-S5a	GeneDoc ND1 nucleotide alignment…….	110
3.SI-S5b	GeneDoc ND1 protein alignment……..	113
3.SI-S6a	GeneDoc ND4 nucleotide alignment…….	115
3.SI-S6b	GeneDoc ND4 protein alignment……..	119
3.SI-S7a	GeneDoc ND5 nucleotide alignment…….	121
3.SI-S7b	GeneDoc ND5 protein alignment……..	126
3.SI-S8a	GeneDoc ATPase6 / MURF4 nucleotide alignment……..	129
3.SI-S8b	GeneDoc ATPase6 / MURF4 protein alignment……	131
3.SI-S9a	GeneDoc COII nucleotide alignment…….	133
3.SI-S9b	GeneDoc COII protein alignment……..	135
3.SI-S10a	GeneDoc COIII nucleotide alignment……	137
3.SI-S10b	GeneDoc COIII protein alignment…….	140
3.SI-S11a	GeneDoc CYb nucleotide alignment…….	142
3.SI-S11b	GeneDoc CYb protein alignment……..	145
3.SI-S12a	GeneDoc MURF2 nucleotide alignment……	147
3.SI-S12b	GeneDoc MURF2 protein alignment…….	150
3.SI-S13a	GeneDoc ND7 nucleotide alignment…….	152
3.SI-S13b	GeneDoc ND7 protein alignment……..	156
3.SI-S14a	GeneDoc MURF5 nucleotide alignment……	158

3.SI-S14b	GeneDoc MURF5 protein alignment…….	160
3.SI-S15a	GeneDoc ND8 / G1 nucleotide alignment…….	162
3.SI-S15b	GeneDoc ND8 / G1 protein alignment……..	164
3.SI-S16a	GeneDoc ND9 / G2 nucleotide alignment…….	166
3.SI-S16b	GeneDoc ND9 / G2 protein alignment……..	169
3.SI-S17a	GeneDoc RPS12 nucleotide alignment…….	171
3.SI-S17b	GeneDoc RPS12 protein alignment……..	172
3.SI-S18a	GeneDoc ND3 / G5 nucleotide alignment……	174
3.SI-S18b	GeneDoc ND3 / G5 protein alignment of the conserved motif…….	176
3.SI-S19a	GeneDoc G3 nucleotide alignment……..	178
3.SI-S19b	GeneDoc G3 protein alignment……	179
3.SI-S20a	GeneDoc G4 nucleotide alignment……..	181
3.SI-S20b	GeneDoc G4 protein alignment……	183
4.SI-S1a	GeneDoc 12S gene alignment…….	249
4.SI-S2a	GeneDoc 9S gene alignment……	255
4.SI-S3a	GeneDoc MURF1 / ND2 nucleotide alignment……..	258
4.SI-S3b	GeneDoc MURF1 / ND2 protein alignment……	263
4.SI-S4a	GeneDoc ND1 nucleotide alignment……	266
4.SI-S4b	GeneDoc ND1 protein alignment…….	270
4.SI-S5a	GeneDoc COI nucleotide alignment……	273
4.SI-S5b	GeneDoc COI protein alignment…….	279
4.SI-S6a	GeneDoc ND4 nucleotide alignment……	282
4.SI-S6b	GeneDoc ND4 protein alignment…….	257
4.SI-S7a	GeneDoc ND5 nucleotide alignment……	290
4.SI-S7b	GeneDoc ND5 protein alignment……	296
4.SI-S8a	GeneDoc CYb nucleotide alignment……	299
4.SI-S8b	GeneDoc CYb protein alignment…….	304

4.SI-S9a	GeneDoc COII nucleotide alignment……	307
4.SI-S9b	GeneDoc COII protein alignment…….	310
4.SI-S10a	GeneDoc MURF2 nucleotide alignment……..	312
4.SI-S10b	GeneDoc MURF2 protein alignment……	317
4.SI-S11a	GeneDoc MURF5 gene alignment……	320
4.SI-S11b	GeneDoc MURF5 protein alignment……	321
4.SI-S12a	GeneDoc ND8 / G1 nucleotide alignment…….	323
4.SI-S12b	GeneDoc ND8 / G1 protein alignment……..	325
4.SI-S12c	GeneDoc ND8 / G1 nucleotide alignment…….	326
4.SI-S12d	GeneDoc ND8 / G1 protein alignment……..	328
4.SI-S13a	GeneDoc ND9 / G2 nucleotide alignment…….	330
4.SI-S13b	GeneDoc ND9 / G2 protein alignment……..	333
4.SI-S14a	GeneDoc ND7 nucleotide alignment……	335
4.SI-S14b	GeneDoc ND7 protein alignment…….	340
4.SI-S15a	GeneDoc COIII nucleotide alignment……..	343
4.SI-S15b	GeneDoc COIII protein alignment……	347
4.SI-S16a	GeneDoc ATPase / MURF4 nucleotide alignment……	350
4.SI-S16b	GeneDoc ATPase / MURF4 protein alignment…….	353
4.SI-S17a	GeneDoc G3 nucleotide alignment……	355
4.SI-S17b	GeneDoc G3 protein alignment…….	356
4.SI-S18a	GeneDoc G4 nucleotide alignment……..	358
4.SI-S18b	GeneDoc G4 protein alignment……	360
4.SI-S19a	GeneDoc ND3 / G5 nucleotide alignment……	362
4.SI-S19b	GeneDoc ND3 / G5 protein alignment……	364
4.SI-S20a	GeneDoc RPS12 nucleotide alignment……..	366
4.SI-S20b	GeneDoc RPS12 protein alignment……	367

SI = Supporting Information

LIST OF ABBREVIATIONS

A6	ATPase 6
CO	Cytochrome Oxidase
CYb	(Apo) Cytochrome b
G	G-rich region
DHFR	Dihydrofolate reductase
GAPDH	Glyceraldehyde 3-Phosphate Dehydrogenase
GSH1	Gamma-Glutamylcysteine Synthetase
MURF	Mitochondrial Unidentified Reading Frame
ND	NADH Dehydrogenase
RPS12	Ribosomal Protein Subunit 12
SSU rDNA	Small subunit ribosomal RNA gene
AA(s)	amino-acid(s)
Avg	average
bp(s)	base-pair(s)
gDNA	genomic DNA
indels	insertion-deletion mutations
kb	kilo bases
min	minutes
MSA	Multiple Sequence Alignment
No.	Number
nt(s)	nucleotide(s)
PCR	polymerase chain reaction
s	seconds

T	Thymidylate residue
T_m	Temperature of melting
U	Uridylate residue

ABSTRACT

MYSTERIES OF THE TRYPANOSOMATID MAXICIRCLES: CHARACTERIZATION OF THE MAXICIRCLE GENOMES AND THE EVOLUTION OF RNA EDITING IN THE ORDER KINETOPLASTIDA

By Preethi Ranganathan Iyengar, M. Sc.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2015.

Major Director: Dr. Gregory A. Buck,
Professor,
Department of Microbiology and Immunology, and
Center for the Study of Biological Complexity.

The trypanosomatid protists belonging to Order Kinetoplastida are some of the most successful parasites ever known to mankind. Their extreme physiological diversity and adaptability to different environmental conditions and host systems make them some of the most widespread parasites, causing deadly diseases in humans and other vertebrates.

This project focuses on their unique mitochondrion, called the kinetoplast, and more specifically involves the characterization of a part of their mitochondrial DNA (also called kinetoplast DNA or kDNA), the maxicircles, which are functional homologs of eukaryotic mitochondrial DNA in the kinetoplastid protists. We have sequenced and characterized the maxicircle genomes of 20 new trypanosomatids and compared them with 8 previously published maxicircle genomes of other trypanosomatids. Transcripts of ~13 of the 20 total genes in these maxicircles undergo post-transcriptional modifications involving the insertion and deletion of U residues at precise sites, to yield the final, fully-edited, translatable mRNA. We have deciphered the diverse patterns

and extents of RNA editing of each edited gene in the maxicircle of each organism, and inferred the sequences of the putative fully edited mitochondrial transcripts and proteins. Using a binary value - based strategy (1/0), we quantified the RNA editing in all these trypanosomatids and estimated the evolution of RNA editing in the group. Additionally, we conducted phylogenetic analyses using a subset of unedited maxicircle genes to predict the relationships between the various trypanosomatids in this project, and compared them to the previously published nuclear gene-based phylogenies.

For convenience of analysis, the 28 total trypanosomatids in this work were divided into two groups: the first group consisting of the endosymbiont-bearing and related insect trypanosomatids, which constitute the first half of the project, and the second group consisting of trypanosomatids of the *Trypanosoma* genus, including *T. cruzi*-related and unrelated parasites, constituting the latter half of the project.

In summary, most of the trypanosomatid maxicircles showed a syntenic panel of 20 protein-coding genes (excluding any guide RNA genes), beginning with the mitochondrial ribosomal genes and ending with the gene encoding NADH dehydrogenase-5. Although some genes were partially or completely absent in the maxcircles of some species, the remaining genes were completely syntenic. The total number of genes edited and their editing patterns varied considerably among the first group of insect trypanosomatids, but were remarkably similar in the second group of the *Trypanosoma* genus. On a broad scale, the mitochondrial phylogeny reflects the nuclear phylogeny for these trypanosomatids, except within the *T. cruzi* population. Similarly, RNA editing appears to have evolved in parallel with the nuclear genes, although subtle differences are again noticeable within the *T. cruzi* family.

CHAPTER 1
Background, Objectives and Purpose of the study

The infrakingdom Excavata comprises organisms that are believed to be some of the earliest diverging eukaryotes. One of its phyla, Phylum Euglenozoa, constitutes an extensive group of flagellated protists that are characterized by extremely diverse morphologies, and show unique genetic and ecological features. Members of Phylum Euglenozoa can be free-living phototrophs, osmotrophs or phagotrophs. They can also be symbionts, commensals or obligate parasites of most of the other taxa of animals, plants, insects and other invertebrates. Please refer to **Chapter 2** for more details on Phylum Euglenozoa.

Our lab had been working on assembling a part of the eukaryotic Tree of Life (Supported by grants from the National Science Foundation's 'Assembling the Tree of Life' program through grant DEB-0830056 to G. Buck). As a part of this project, we have sequenced: a) the whole genomes of several kinetoplastid clades, and b) cDNA libraries of several euglenids and diplonemids. All the data from our high-throughput sequencing is available in our internal databases. As a sub-project, we decided to look into the maxicircles (mitochondrial DNA) of the various kinetoplastid clades. Please refer to **Chapter 2** for a literature review on the mitochondrial genomes of kinetoplastids. Since the kinetoplast DNA is a network that accounts for up to 30% of the total trypanosomatid cellular DNA, protocols were developed to eliminate a major portion of the kDNA network before sequencing the whole genomes. Despite this step, kDNA formed a significant subset of the total sequencing data (often showing over 30X coverage), and partial maxicircle contigs could be assembled from the sequencing reads. These resulting maxicircle contigs were used as the starting point for this project.

Table 1.1 below is a list of the organisms described in this work, as well as some published maxcircle sequences. To study this group of organisms, they were categorized as follows:

a) Endosymbiont-bearing and related insect trypanosomatids [**Chapter 3**], and

b) Trypanosomatids of the *Trypanosoma* genus (*T. cruzi* related and unrelated parasites) [**Chapter 4**].

Table 1.1. List of organisms in this project

	Organisms	Strain	Lineage		GenBank	
1	*Leishmania tarentolae*			Published	M10126.1	Dixenous
2	*Endotrypanum schaudinni*					Dixenous
3	*Leptomonas costaricensis*					Monoxenous
4	*Crithidia acanthocephali*					
5	*Herpetomonas muscarum*					
6	*Phytomonas serpens*			Published	AF079967.2	Dixenous
7	*Angomonas deanei*					Endosymbiont bearing group
8	*Angomonas desouzai*					
9	*Strigomonas culicis*					
10	*Strigomonas galati*					
11	*Strigomonas oncopelti*					
12	*Trypanosoma cruzi*	Sylvio X10	I	Published	FJ203996.1	Dixenous
13	*Trypanosoma cruzi*	G	I			
14	*Trypanosoma cruzi*	JR	I			
15	*Trypanosoma cruzi*	Esmeraldo	II	Published	DQ343646.1	
16	*Trypanosoma cruzi*	M6241	III			
17	*Trypanosoma cruzi*	CAN III	IV			
18	*Trypanosoma cruzi*	CL Brener	VI	Published	DQ343645.1	
19	*Trypanosoma cruzi*	CL	VI			
20	*Trypanosoma cruzi*	Tulahuen	VI			
21	*Trypanosoma cruzi marinkellei*	B7		Published	KC427240.1	
22	*Trypanosoma dionisii*					
23	*Trypanosoma erneyi*	sp. 1946				
24	*Trypanosoma rangeli*	SC-58		Published	KJ803830.1	
25	*Trypanosoma conorrhini*					
26	*Trypanosoma lewisi*					
27	*Trypanosoma serpentis*					
28	*Trypanosoma brucei*			Published	M94286.1	

Endosymbiont - bearing and insect trypanosomatids

Trypanosoma cruzi and related parasites

Non - *T. cruzi* parasites

The specific aims of this project include:

1. Characterization of the maxicircle genomes of the above organisms with respect to their size, synteny, gene-content and gene sequences.
2. Analyses of the extent and diversity of the RNA editing patterns in each gene in each of the organisms listed above; and examination of the evolution of RNA editing among these kinetoplastids.
3. Phylogenetic analyses using the mitochondrial (maxicircle) genes to examine the relationships among these organisms and to compare their mitochondrial phylogeny relative to the phylogeny of the nuclear-genes.

Kinetoplastid protozoa parasitize their hosts differently and seem to have life-cycles that are dependent on the composition of their kDNA. The study of the kinetoplast and kDNA of these parasites promises to identify potential chemotherapeutic targets. Phylogenetic studies establish closely related parasites, permitting the use of similar drugs to treat infections with different parasites. Our work furthers the understanding of the role, importance and the dynamics of RNA editing in the kinetoplast of these organisms and their evolutionary relationships to shed more light on their life-cycles, differentiation and infectivity. Please refer to **Chapter 2** for more details.

CHAPTER 2
Literature Review

2.1 Order Kinetoplastida: Diversity and Evolution

Members of the Order Kinetoplastida (Kingdom Excavata, Superphylum Discicristata, Phylum Euglenozoa) comprise an extensive group, that are often considered as some of the most successful parasites. They contain a range of free-living species and obligate parasites of invertebrates, vertebrates and plants (Vickerman, 1994). Initially, kinetoplastids were considered to be very close to the euglenids (also Phylum Euglenozoa), as estimated by several early small ribosomal subunit (SSU) rRNA analyses. As molecular phylogenies based on multiple genes or proteins were applied, it was discovered that members of Phylum Euglenozoa were closer to the heterolobosеid amoeboflagellates (Heterolobosea, Phylum Percolozoa), and the free-living, bacterivorous jakobid flagellates (Jakobida, Phylum Loukozoa) (Baldauf et al., 2000) (Simpson et al., 2006). Eventually, the kinetoplastids were found to be closest to an overlooked euglenozoan group, the diplonemids (Simpson and Roger, 2004), which comprise a poorly defined group of free-living protozoans that occasionally parasitize invertebrates (von der Heyden et al., 2004) (Marande et al., 2005). Recently, a fourth group of symbionts (Symbiontida) was added to the Phylum Euglenozoa, along with euglenids, diplonemids and kinetoplastids (Adl et al., 2012), and currently, the symbiontids appear to be the closest relative of the kinetoplastids.

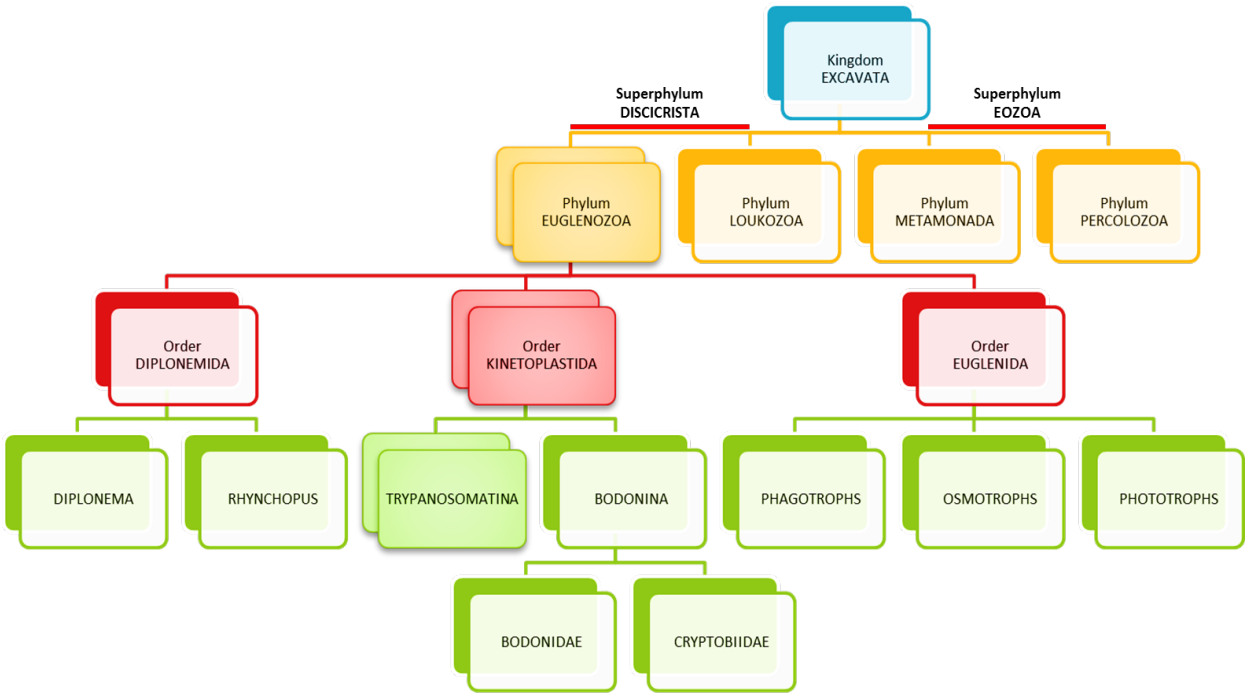


Fig 2.1: Infrakingdom Excavata and Phylum Euglenozoa

Phylogenies based on heat shock proteins (HSPs) and SSU rRNA indicate that kinetoplastids can be divided into two subclasses: Prokinetoplastina and Metakinetoplastina (Moreira et al., 2004). The Metakinetoplastina could be further subdivided into four groups: three clades of bodonids that include Neobodonida, Parabodonida and Eubodonida; and one clade of trypanosomatids, that apparently diverged 'late' within the kinetoplastida (Simpson et al., 2002) (Simpson et al., 2004a).

Despite all the molecular phylogenies, two controversies continue to exist: 1) the monophyly of trypanosomatids, which is currently widely accepted (Lukes et al., 1997) (Simpson et al., 2004a) (Hughes and Piontkivska, 2003a) (Simpson et al., 2006) (Deschamps et al., 2011); and 2) the closest bodonid relative of the trypanosomatids. Initially, SSU rRNA phylogenies indicated that the trypanosomatids diverged early from the other Metakinetoplastina (von der Heyden et al.,

2004) (Moreira et al., 2004). However, the HSP70/90-based trees reported that the trypanosomatids are nested within the Bodonids, and are, in fact, a sister clade to Eubodonida (*Bodo saltans*) (Blom et al., 1998) (Simpson et al., 2004a).

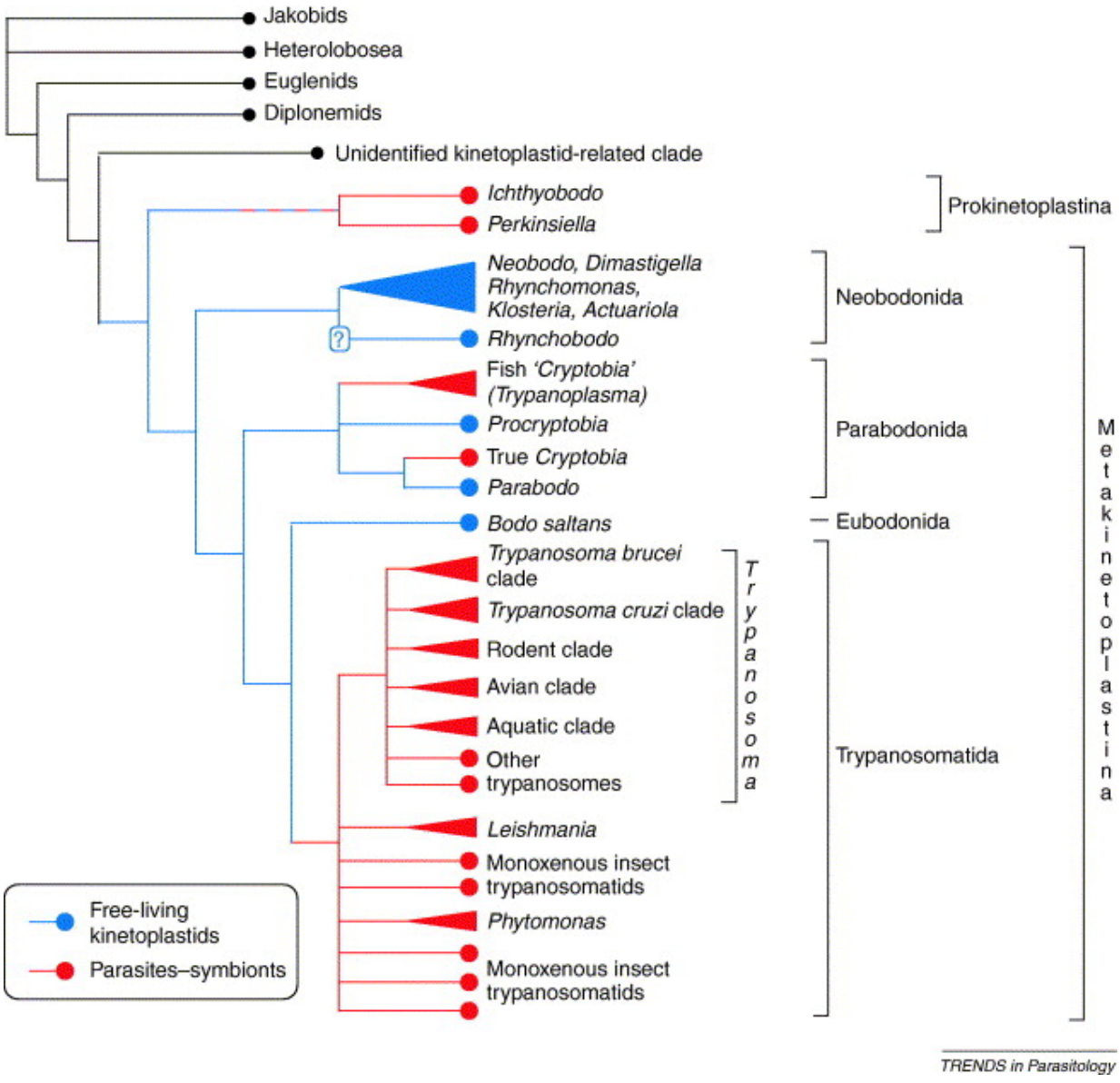


Figure 2.2: Order Kinetoplastida: The evolutionary relationships among kinetoplastids as estimated by taxon-rich SSU rRNA gene trees and protein phylogenies. Figure reprinted from Trends in Parasitology, Volume 22, Issue 4, April 2006, Alastair G.B. Simpson, Jamie R. Stevens, Julius Lukeš, The evolution and diversity of kinetoplastid flagellates, Pages 168-174, Copyright © 2006, with permission from Elsevier (License No. 3746081080918).

Within the trypanosomatids, the *Trypanosoma* comprise one of the most important genera, and its members have been sub-grouped into several clades using different molecular markers (Stevens et al., 2001) (Hamilton et al., 2004). Some of the most well-known (but not very robustly supported) sub-groupings are as follows: the African Salivarian trypanosomes (including *T. brucei* and its relatives) that are often transmitted by tsetse flies; the New World group of *T. cruzi* and its relatives that are transmitted by triatomine bugs; an 'aquatic clade' transmitted by leeches, an 'avian clade' transmitted by hippoboscid and black flies; and a 'rodent clade' presumed to be transmitted by fleas. Other than the *Trypanosoma*, the trypanosomatids include other genera that include symbiont-harboring trypanosomatids [SHTs], which contain a single bacterial endosymbiont (termed the trypanosomatid proteobacterial endosymbiont or TPE) in their cytoplasm, that replicates synchronously with the host cell and is closely related to the β–proteobacterium *Taylorella* (de Souza and Motta, 1999) (Motta et al., 2010) (Alves et al., 2013a); monoxenous insect infecting parasites; and dixenous parasites that alternate between insects and vertebrates or plants. Please refer to the 'Introduction' sections in **Chapters 3 and 4** for a background on the organisms included in this study. Curiously, the monoxenous genera are not monophyletic; and the dixenous parasites are found interspersed with the monoxenous genera in molecular phylogenies, implying that parasitism and newer host acquisitions may have evolved independently at several times by a convergent evolutionary mechanism (Simpson et al., 2006). The known SHTs, however, do form a monophyletic group, indicating that the event leading to endosymbiosis may have occured just once, and was retained due the advantage it confers to its protist host (de Souza and Motta, 1999).

There had been an ongoing search for a link between the trypanosomatids and the bodonids. Very recently, a novel, putatively monoxenous parasite has been isolated from mosquitoes and

characterized. This new species, *Paratrypanosoma confusum*, is not positioned with the insect trypanosomatids, as was thought earlier, but rather, it is robustly positioned at the base of the trypanosomatids and falls aptly between the bodonids and the trypanosomatids, thereby providing that missing link between the ancestral bodonids and the derived group of parasitic trypanosomatids (Flegontov et al., 2013) (Stevens, 2014).

2.2 The curious case of trypanosomatids

Ribosomal RNA lineages have indicated that Kinetoplastids are one of the most ancient eukaryotes, that may have diverged even prior to fungi, animals or plants (Fernandes et al., 1993) (Beverley, 1996). These organisms have evolved some interesting features that markedly differ from other higher eukaryotes. Many important biological phenomena that were first discovered in trypanosomatids include but are not limited to: a) a unique organelle called 'kinetoplast' that houses their unique mitochondrial DNA network (see below); b) a complex process of mitochondrial RNA editing (detailed below) (Benne et al., 1986); c) trans-splicing of all their transcripts (Sutton and Boothroyd, 1986); d) polycistronic transcription in a eukaryote (Johnson et al., 1987); e) a striking paucity of introns (Myler et al., 1999) (Simpson et al., 2006); f) several types of unique nucleotide modification mechanisms; g) bent regions of DNA (Marini et al., 1983); h) expansion and contraction of telomeric DNA repeats (Bernards et al., 1983); i) compartmentalization of glycolysis in unique organelles called 'glycosomes' (modified peroxisome-like organelles) (Opperdoes and Borst, 1977); j) glycosylphosphatidylinositol anchors of membrane proteins (Holder and Cross, 1981) (Ferguson and Cross, 1984) (Ferguson et al., 1985); k) evasion of the host immune responses using a "programmed" antigenic variation of surface glycoproteins; l) an ability to escape from phagocytic vacuoles (Bridgen et al., 1976)

(Pays et al., 2004); m) novel metabolic pathways (Bacchi et al., 1980); and n) protein encoding genes without an obvious promoter region (Myler et al., 1999).

Overall, it has been speculated that because of their evolutionary placement between prokaryotes and eukaryotes, the trypanosomatids have some prokaryotic features including polycistronic transcription and intron poverty; some eukaryotic features including pre-mRNA splicing; and some features found in neither prokaryotes nor eukaryotes, including a lack of promoters for protein-encoding genes. It is quite possible that more of such unusual features will be discovered; and what is known may well be just the tip of an iceberg!

2.3 Evolution of parasitism in Order Kinetoplastida

While the closest relatives of kinetoplastids are almost always phototrophic or phagotrophic, many of the kinetoplastids are commensals or obligate parasites. It is believed that parasitism has evolved independently, several times in the kinetoplastids by a convergent evolutionary mechanism. At least four independent adoptions of commensalism and parasitism have been upheld: once in the Prokinetoplastids (*Ichthyobodo-Perkinsiella* clade); twice independently in the Parabodonids (once in the fish-infecting *Cryptobia* species, i.e. *Trypanoplasma*; and once in the 'true' *Cryptobia* which share a commensal relationship with snails); and once in the trypanosomatids (Simpson et al., 2006).

Since most of the euglenids, diplonemids, symbiontids and even most of the bodonids are predominantly free-living, it is likely that the ancestral state of the kinetoplastids was also free-living. Eventually however, as parasitism arose, it is unclear as to whether the trypanosomatids originally parasitized the insects or vertebrate. Since the genus *Trypanosoma* (a present day digenetic genus) is the most basal among the trypanosomatids, it was first presumed that the

ancestral trypanosomatids parasitized the vertebrates, and later gradually invaded the insects via their blood-sucking class (Minchin, 1908). However, recent molecular data refuted this hypothesis and the 'insect-first' model of parasitism is now generally accepted, since the dixenous *Leishmania* and *Phytomonas* are nested within the insect-infecting, monoxenous trypanosomatids. The insect-first model proposes that digenetic parasites descended from parasites of blood-sucking insects that somehow survived an accidental transmission to vertebrate hosts during the infected insects' feeding on the blood of the vertebrates (Lukeš et al., 2014) (Simpson et al., 2006). Although the transmission of insect-trypanosomatids to vertebrates may have occurred several times, it is believed that the establishment of infection in the vertebrate host was successful only rarely.

The existence of a separate aquatic clade within the *Trypanosoma* genus strengthens the hypothesis that the trypanosomes infecting the aquatic vertebrates are derived from the trypanosomes infecting the aquatic leeches that feed on the blood of aqatic vertebrates (Stevens et al., 2001) (Hamilton et al., 2004). Similarly, the more recent description of the putatively monoxenous *P. confusum*, forming a link between the free-living Bodonids and the parasitic trypanosomatids (Flegontov et al., 2013) (Stevens, 2014), also asserts the theory that the ancestral kinetoplastids may have invaded the insects and other invertebrate hosts first, before entering the vertebrates thorough the blood-feeding mechanism of their primary insect hosts. Finally, several current phylogenies show that the derived, dixenous lifestyle evolved from the monoxenous lifestyle independently in each of the *Leishmania*, *Trypanosoma* and *Phytomonas* genera (Maslov et al., 2013) (Flegontov et al., 2013) (Stevens, 2014). There are slos theories that several monoxenous (insect) trypanosomatids switched to dixeny, opportunistically, in response to decrease in immunity of the vertebrate hosts (Chicharro and Alvar, 2003) (Dedet and Pratlong,

2000) (Jiménez et al., 1996). Several insect trypanosomatids have been found within one insect; and conversely, the same trypanosomatid species has been found in a wide range of insect hosts across different geographical locations; thereby disproving the 'one host, one parasite' theory (Podlipaev et al., 2004). In summary though, molecular analyses demonstrate that many of the peculiar characteristics of the kinetoplastids were independently acquired after evaluating their advantage, several times during their evolution. However, since most of these features were already present in the free-living and early diverging bodonids, it does not look like they were developed as an adaptation to a parasitic lifestyle (Lukeš et al., 2014).

2.4 Important trypanosomatid diseases, parasites and their life-cycles

Most of the kinetoplastid protozoans have similar cellular structures, genomic organizations and undergo morphological changes during their life-cycles. However, they cause distinct human diseases and are spread by different insect vectors. The most common human diseases are caused by the two species *Trypanosoma* and *Leishmania*. The three major diseases caused include: (i) Human African Trypanosomiasis (also known as African sleeping sickness), caused by *T. brucei gambiense* and *T. brucei rhodesiense*; (ii) American Chagas disease caused by *T. cruzi*; and (iii) various types of leishmaniasis, caused by infection with different types of *Leishmania* species. About half a billion people in the tropical and sub-tropical regions in the world are at risk of contracting these diseases, with an estimated 20 million infected individuals at any time, leading to about 100,000 deaths per year, around the world.

2.4.1 Human African Trypanosomiasis

Human African Trypanosomiasis is transmitted by the bite of infected Tsetse flies, and rarely by blood transfusion and congenitally. The early stage of HAT is the hemolymphatic stage, which is characterized by fever, headache, adenopathy, joint pain and pruritus. The parasite rapidly propagates in the host, although it can be countered by the host's immune system. However, the parasite's antigenic variation mechanism enables it to escape host immunity, thereby causing relapses and waves of parasitemia. The disease thus progresses to the late CNS stage resulting in severe neurological symptoms that cause mental, sensory and sleep anomalies. HAT can be lethal if untreated. The Tsetse fly takes a bloodmeal of parasite-containing blood from an infected human or animal and ingests the bloodstream trypomastigotes, which convert in to procyclic trypomastigotes in the insect's midgut and multiply. Eventually, the parasites transform into epimastigotes and reach the insect's salivary gland to multiply there and transform into infective metacyclic trypomastigotes. When the fly takes its next blood meal, the metacyclic trypomastigotes are injected into the next human or the animal, thereby continuing its transmission. Inside the human, the metacyclic trypomastigotes convert into bloodstream trypomastigotes and are carried to the other sites. These trypomastigotes multiply rapidly and circulate in the blood, lymph and spinal fluid of the human host, which can later be picked up by another Tsetse fly, thereby completing its life cycle. See (Stuart et al., 2008) and references therein for details.

2.4.2 Chagas disease

Chagas disease is spread in humans by contaminated feces of infected Reduviid bugs at either the bite site, in mucus membranes or in food and drink. It can also be transmitted congenitally,

by blood transfusion and rarely by organ transplant. The disease results from persistent infection with *T. cruzi*, and the resultant myriad interactions between the pathogen and the host immune system. A good immune system is able to clear out an acute infection; although in the indeterminate or chronic phase, a complete elimination of the pathogen is almost impossible (Tarleton, 2003). About 30% of individuals with *T. cruzi* infection develop clinical Chagas disease, which causes a cardiac disease or gut enlargement, several years after the initial infection (Andrade and Andrews, 2005) (Burleigh, 2005) (Combs et al., 2005). Infected blood-sucking triatomine bugs deposit infective trypomastigotes in their feces when they feed on the human blood, at the bite site. These trypomastigotes invade the surrounding cells and tissues, and differentiate into intracellular amastigotes. The amastigotes multiply rapidly in the tissue and subsequently differentiate into trypomastigotes, which get released into the blood, travel all over the body and infect the cells of several organs like the heart, gut, adipose tissue, smooth muscles, CNS etc., before becoming amastigotes and multiplying in these tissues again. While in the human blood circulation, the trypomastigotes can be picked up by another triatomine bug during its bloodmeal. In the insect midgut, the trypomastigotes undergo physiological and morphological transformation to become infective trypomatigotes in the insect's hindgut, before being released in feces during the insect's next feeding, thereby completing and continuing the cycle (Stuart et al., 2008).

2.4.3 Leishmaniases

Leishmaniasis is transmitted in humans by the bite of infected Phebotomine sandflies. More than 21 species of *Leishmania* cause several diseases ranging from mild, self-healing cutaneous Leishmaniasis (CL), to extremely debilitating, and if left untreated, lethal visceral leishmaniasis (VL, also called as 'kala-azar) (Grevelink and Lerner, 1996). Less frequently, other forms of the

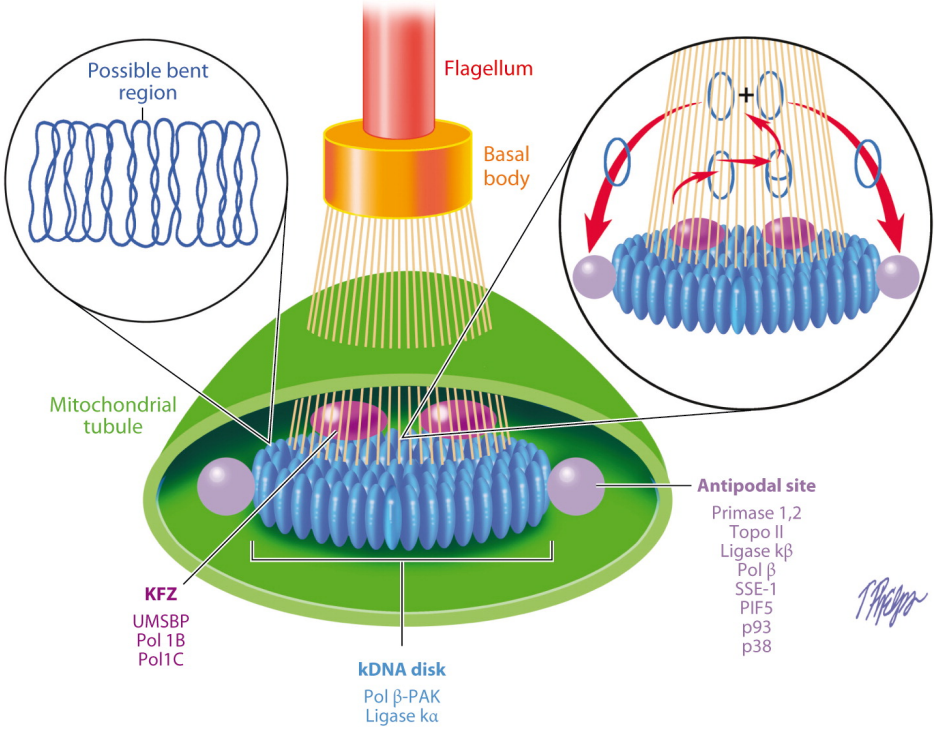
disease can occur, like mucosal leishmaniasis (ML), causing disfigurement (Marsden, 1990); or a diffuse CL, causing nodular lesions (Ramesh and Mukherjee, 1995); and a dermal leishmaniasis, which follows a VL (Zijlstra et al., 2003). CL may be caused by any of the different species of *Leishmania*, while VL is specifically caused only by either *L. infantum* or *L. donovani*. Infected sandflies inject the infective promastigotes when taking a blood meal. These promastigotes selectively invade the macrophages by receptor-mediated endocytosis, transform into amastigotes there and replicate in the macrophages. When another sandfly ingests blood and infected macrophages from an infected human, it gets infected. The amastigotes thus ingested transform into promastigotes, which further differentiate into infective metacyclic promastigotes in the insect gut. See review by Stuart et al., 2008 (Stuart et al., 2008).

2.5 The trypanosomatid mitochondrion and kinetoplast

Members of the Order Kinetoplastida show the presence of a single, ramified mitochondrion, which runs through the length of the cell. Although the fine structure of the mitochondrion varies among different genera and species, they are generally characterized by the presence of a dense matrix; and sparse, irregularly distributed, thin cristae. The mitochondrion itself appears to branch out under the subpellicular microtubules, but is dilated at the region where the mitochondrial DNA is present (de Souza et al., 2009).

The mitochondrial DNA of the trypanosomatids is densely packed in an unusual disc-shaped structure known as the ‘kinetoplast’ (and therefore being called as ‘kinetoplast DNA’ or kDNA), which is present near the basal body of the flagellum. The kinetoplast disc lies perpendicular to the axis of the flagellum within the kinetoplast matrix. However, electron microscopy has

revealed that the DNA strands themselves are aligned parallel to the network's axis (de Souza et al., 2009). The kDNA is one of the most complex mitochondrial DNA in nature, and has an extraordinarily enormous network structure. This dense network consists of thousands of circular DNA molecules that are topologically interlinked to form an apparent two dimensional 'fish-net type catenanes' (Shlomai, 2004) (Liu et al., 2005), with the exception of Bodonids, that do not show catenation (Lukes et al., 2002). Two types of DNA rings are present in the network: the larger maxicircles and the smaller minicircles.



Jensen RE, Englund PT. 2012.
Annu. Rev. Microbiol. 66:473–91

Figure 2.3: Kinetoplast and proteins associated with kDNA replication: A cross-section of the end of the mitochondrial tubule (green) near the basal body of the flagellum shows the compact kDNA disc, perpendicular to axis of the flagellum. The minicircles are stacked parallel to the axis of the disc and the location of several critical kDNA repliucation proteins are shown. The filament system linking the kDNA to the flagellar basal body is also shown. Figure republished from Annual Review of Microbiology, Volume 66, Oct 2002, Robert E. Jensen, Paul T. Englund, Network News: The Replication of Kinetoplast DNA, Pages 473-491, Copyright © 2012, with permission from Annual Reviews.

2.5.1 Maxicircles

The maxicircles are the functional homologs of the eukaryotic mitochondrial DNA, and depending upon the species, they range between about 20 kb to nearly 70 kb in size. They are generally homogeneous in size and sequence, and only a few dozen copies are present in each network. They encode the ribosomal RNA genes and several subunits of the mitochondrial respiratory complexes (like subunits of NADH dehydrogenase, ATP synthase, cytochrome oxidases and cytochrome *b*). Curiously, no tRNAs have been found to be encoded by the maxicircles, and all tRNAs required for translation are believed to be imported into the mitochondria from the cytosol (Alfonzo and Söll, 2009). The genes encoded in the maxicircle are generally highly syntenic and cluster in a contiguous stretch called the "consensus" or the conserved region; and the remaining maxicircle consists of a non-coding "divergent" region (DR) of variable lengths. There are 18 protein-encoding genes, 2 ribosomal RNA genes, and very few gRNA genes clustered in the consensus region. The genes could broadly be classified in to groups based on the respiratory complex they are subunits of. ND1, ND2 (MURF1), ND3 (G5), ND4, ND5, ND7, ND8 and ND9 are putative subunits of the Complex I (details of complexes in a subsequent section); COI, COII and COIII are subunits of Complex IV; CYb is a part of Complex III; A6 (MURF4) is a subunit of Complex V; 12S, 9S and RPS12 form the mitochondrial ribosomes; and the last four genes' functions have not been established (MURF2, MURF5, G3 and G4). Of these 20 maxicircle genes, the genes ND8, ND9, COIII, CYb, MURF4 (A6), G3, COII, MURF2, ND4, RPS12 and ND5 are transcribed from the same strand as the ribosomal genes (12S and 9S), in the same direction and polarity; while the genes ND9, MURF5, MURF1, ND1, COI, G4 and G5 (ND3) are transcribed from the opposite strand, in the opposite direction and polarity (Cruz et al., 1984) (Simpson et al., 1987) (Maslov et al., 1999).

The maxicircle divergent regions are highly variable among different species, no apparent conserved motifs exist in that region; and their functions are also relatively unknown. But peculiar types of repetitive sequences have been observed in the divergent regions (DRs) of all the trypanosomatids. DRs of *Leishmania-Leptomonas* group show two types of repeats: Type I, which are long GC-rich palindromic repeats with A_5C-elements at their center, separated by AT-rich regions of variable lengths, containing potential bent regions of DNA; and Type II, which are shorter than the Type I repeats, contain at least one A_5C-element towards their 5' or 3' ends; but no bent DNA regions (Flegontov et al., 2006). The DR of *Trypanosoma cruzi* also shows palindromic sequences and two A_5C-elements, associated with potential bent DNA sequences (Westenberger et al., 2006). The DR of *T. brucei* appears to show similar GC-rich repeats with some less GC-rich spacers (Myler et al., 1993); while that of *Strigomonas oncopelti* shows only GC-rich repeats in tandem clusters, but no AT-rich sequences (Gorbat et al., 1990). These repetitive structures are hypothesized to function in transcription initiation. Although a promoter functionality has not been demonstrated, their association with palindromic sequences and non-random presence in the DR led to the hypothesis that the A_5C-elements may be binding sites for transcription factors, even if they are not the promoters themselves (Flegontov et al., 2006). Also, bent DNA regions are commonly transcription initiation sites, and the palindromes that form hairpin loop, could act as binding sites for more proteins (Flegontov et al., 2006). The DR was also reported to contain the apparent promoter region for 12S rRNA gene in *Leptomonas seymouri* (Vasil'eva et al., 2004). Since the conserved sequence blocks (CSBs) in minicircles are replication origins (see below for details), it is believed that the CSBs in the maxicircles may act as the replication origin for the maxicircles, by the same analogy (Shapiro and Englund, 1995).

2.5.2 Minicircles

Over 90% of the mass of the kDNA network is made up by the minicircles, which are highly heterogeneous in their sequence in a network, and range in sizes between 0.5 to 10 kb, depending upon the species. There are a few thousand minicircles in a network and within a single network, all minicircles are of the same size. The most important function of the minicircles is to encode gRNAs that are absolutely important for editing of the maxicircle transcripts (See RNA editing section below). The gRNAs themselves are between 55-70 nts in length, lack any 5' processing and carry a short, non-encoded oligo U tail at their 3' ends. Alongside regions encoding gRNAs, the minicircles also show a conserved region that is about 100-180 nts in length that is identical among all the minicircles in a single network. These conserved regions are called "conserved sequence blocks" or CSBs. The number of CSBs in the minicircles are also species specific; in that, the minicircles of *T. brucei* and *Leishmania tarentolae* have a single CSB while those of *Crithidia fasciculata* show two; and those in *T. cruzi* show four CSBs. When multiple CSBs are present, they are generally positioned 90^{o} or 180^{o} apart. Additionally, the minicirles from most trypanosomatids show a region containing a bent-helix made up of regularly spaced adenine residues. While the CSBs are thought of as minicircle replication origins, the exact function of the bent helical region is unknown and is assumed to just help packaging the minicircles more compactly within a network (Shapiro and Englund, 1995) (Lukes et al., 2005). Unlike circular DNA molecules in nature, minicircles are never supercoiled. kDNA analysis throughout Kinetoplastida has revealed that regardless of catenation, minicircles always exist as relaxed molecules (Lukes et al., 2002).

2.5.3 Why do the kinetoplastids have kDNA networks?

The kDNA in *Bodo caudatus*, a free-living, ancient trypanosomatid is not catenated (Hajduk et al., 1986). However, the kDNAs of more recently evolved trypanosomatids are generally seen as catenated networks. There are several hypotheses as to why this could have evolved. One is that catenation of kDNA into a network promotes interaction between the transcripts of the maxicircles and minicircles, thereby helping to establish a more efficient RNA editing process (Shapiro and Englund, 1995). Alternatively, a network structure allows a rapid change in the minicircle repertoire. In *T. brucei*, minicircles are known to get exchanged during mating, and progeny parasites derive minicircles from both parents (Gibson and Garside, 1990) (Gibson et al., 1997). These exchanges are thought to be important since a mutation in the genes requiring editing in a maxicircle would become lethal to the cell, unless a corresponding mutation occurs in the minicircle encoding gRNAs to edit these genes. However, the probability of two such parallel mutations occurring in the same network is quite low. Thus, a constant influx of new minicircles during mating could help the cells by providing a minicircle that could potentially provide a gRNA for the correct editing of the mutated maxicircle gene. Therefore, a rapid change in the gRNA repertoire can be brought about by minicircle exchange between two networks; and such minicircle exchanges are easier in a network form, than recombining sequences in linear chromosomes (Chen et al., 1995) (Shapiro and Englund, 1995). Thirdly, since the trypanosomatid mitochondrial genetic system is heteroplasmic, not linking the discrete DNA circles together may lead to genetic instability under a less stringent replication and partitioning system, which could eventually result in homoplasmic mitochondrial genome in progeny. Unlinked maxicircles and minicircles could get lost due to random mitotic segregation during replication, which in turn may be fatal to the progeny due to the potential loss of specific

minicircle encoded gRNAs to edit their maxicircle genes. Covalently linking all the DNA molecules and maintaining a stringent and efficient replication mechanism helps prevent such segregation mediated loss of genetic material (Borst, 1991). Finally, networks are also thought to help in the efficient organization of the maxicircles and minicicles; thereby ensuring an orderly mechanism for replication of the kDNA (Shapiro and Englund, 1995).

2.6 Trypanosomatid mitochondrial RNA editing

Upon the realization that the kinetoplast houses the mitochondrial DNA of the trypanosomatds, efforts were concentrated on identifying the genes encoded in the kDNA, more specifically in the maxicircles. However, many genes found in the higher eukaryotes could not be identified in the maxicircles. Even the genes that were identified contained several frameshifts, lacked start and stop codons, and sometimes lacked an entire length of the open reading frame (Benne et al., 1986). Also, the genes present in the maxicircles of one trypanosomatid was absent in the corresponding locus in another trypanosomatid (Simpson et al., 1987). It was only after the discovery that the COII transcripts of some trypanosomatids contained an additional four U residues, not encoded in the maxicircle gene (Benne et al., 1986), that the entire realm of "RNA editing" came into being. Soon after this report, several reports showed that about 12 of the total 20 genes in the trypanosomatid maxicircle need to be post transcriptionally edited by insertion or deletion of U residues at specific locations of the transcript, to create translatable open reading frames (Estévez and Simpson, 1999). Therefore, this "U-insertion / deletion RNA editing" mechanism is able to (a) correct frameshifts; (b) create start codons and several subsequent codons in the 5' end of transcripts; (c) create an entire open reading frame, including the creation of start codons and stop codons, by insertion of several hundreds of U residues, and deletion of

tens of Us. Thus, the genes that require editing at varying levels in the 5' end of the transcript are called "5' edited genes", and genes that require U insertions and deletions across the entire lengths of their transcripts (start to stop, and even the untranslated regions), are called "pan-edited genes". Of course, there are also a few transcripts that do not require editing at all, the so-called "unedited genes" (Feagin et al., 1988a).

2.7 Mechanism of U insertion-deletion RNA editing

The process of RNA editing involves the addition and less frequently, the deletion of one or more U residues at precise sites within the coding regions of the transcript. The genetic function of the minicircles was relatively unknown until the discovery of a class of small RNA molecules, called the guide RNAs (gRNAs) (Blum et al., 1990). These gRNAs contain the information required for the correct insertions and deletion of U residues at precise sites. The 5' most sequence of a gRNA (~10 nts), termed the "anchor sequence", is complementary to the mRNA to be edited, just 3' to the site of editing, thereby ensuring a proper interaction between the two RNA molecules for the process of editing. The central portion of the gRNA contains the specific editing information, in such a way that the sequence of the mRNA is complementary to the sequence of the gRNA, except in the regions that involve addition or deletion of U residues in the mRNA. The base-pairing between the gRNA and the mRNA is wobble, allowing the pairing of G-U bases (Sturm and Simpson, 1990a), alongside the regular Watson-Crick pairing. The 3' end of the gRNA contains a stretch of U residues of variable length (added post-transcriptionally), which are thought to interact with the mRNA sequence just upstream to the region being edited (called the "editing block"), thereby preventing any secondary structures in the mRNA and stabilizing the interaction between the two RNA molecules at the site of editing

even further (Blum et al., 1990) (Blum and Simpson, 1990). Editing normally proceeds in the 3' to 5' direction for the mRNA, and as editing proceeds, the complementarity between the gRNA and the mRNA increases. The overall 3' to 5' polarity of editing (for an mRNA) is due to the formation of anchor sequences in the upstream (5') portions of the mRNA due to editing of the downstream (3') blocks.

Currently, the "enzyme-cascade" model (Blum et al., 1990) (Kable et al., 1996) (Alfonzo et al., 1997), is the most widely accepted model for the mechanism of the U insertion-deletion editing. As per this model, after the initial specific interaction between the gRNA and mRNA through their anchor region (just downstream of the editing block), a specific endonuclease recognizes and cleaves the mRNA, just 3' of the first mismatched base. After the cleavage, a terminal uridylyl transferase (TUTase) enzyme adds a large number of U residues to the 3' end of the 5' cleavage fragment of the mRNA. The added Us base pair with the A and G residues in the gRNA (thereby leading to an increased complementarity between the two interacting RNAs and increasing the length of the duplex), and any excessively added, unpaired U residues are chewed away by a 3' to 5' exonuclease. Finally, an RNA ligase joins the modified cleavage fragments of the mRNA. The mechanism of U deletion is similar, except for the direct trimming of the unpaired U residues at the 3' end of the mRNA by the exonuclease after the endonuclease cleavage, followed by ligation of the edited fragments. After once such cycle of an editing "event", the editing machinery now moves to the next upstream site to repeat the next cycle of endonuclease cleavage--U addition and exonuclease trimming (or U deletion by the exonuclease)--ligation event (Estévez and Simpson, 1999). A single gRNA can only direct the editing of a small stretch of the mRNA (~50 nts), called the "editing block". Thus, several gRNAs are needed to edit the complete stretch of a pan-edited or a 5' edited gene, the stretch

being called the “edited domain”. When editing mediated by one gRNA is completed, an RNA helicase (Missel and Göringer, 1994) promotes the unwinding between the gRNA and the edited mRNA, to allow the next gRNA to bind to the upstream sequences in the editing domain, or to allow a ribosome to bind and initiate translation, if the edited block was the 5’-most in the mRNA.

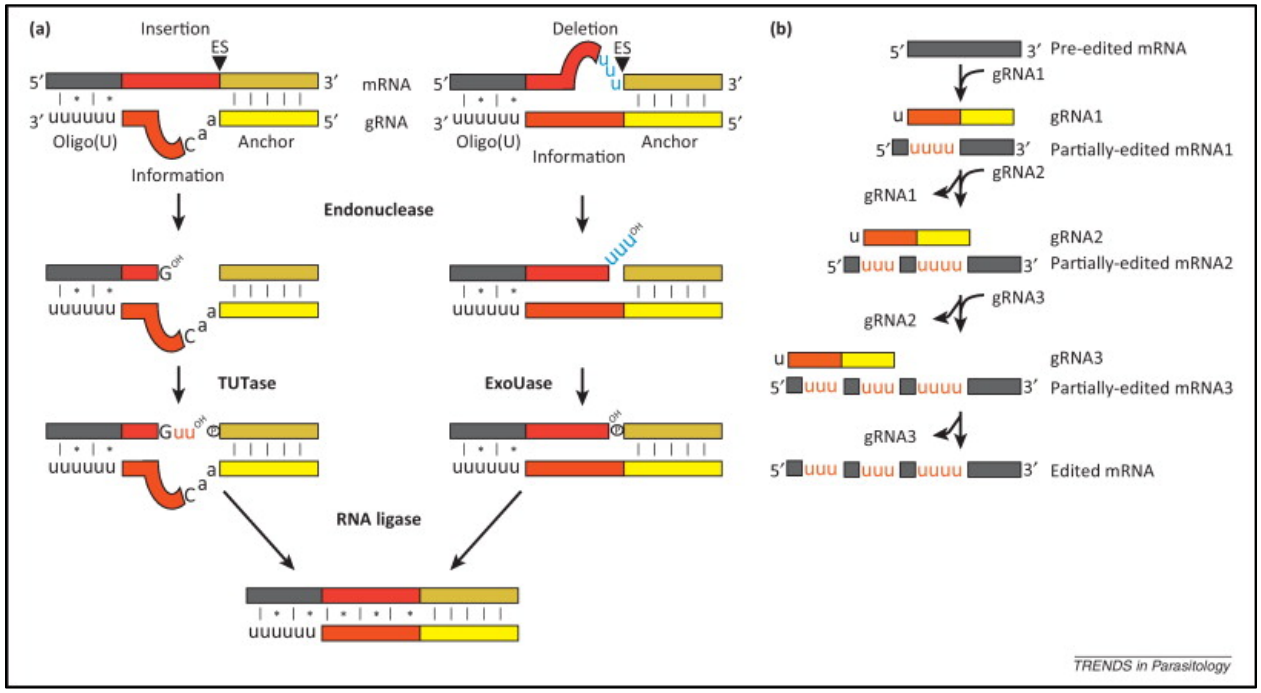


Figure 2.4: Single and multi-round kRNA editing: a) Mechanism of a single round of U-insertion-deletion RNA editing. b) Multiple rounds of kRNA editing in a pan-edited gene involving a cascade of gRNAs. Figure reprinted from Trends in Parasitology, Volume 29, Issue 2, Feb 2013, Hassan Hashimi, Sara L. Zimmer, Michelle L. Ammerman, Laurie K. Read, Julius Lukeš, Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex, Pages 91-99, Copyright © 2012, with permission from Elsevier (License No. 3746701338702).

2.8 Evolution of RNA editing in the trypanosomatids

Since its discovery, it has been proposed that RNA editing in the kinetoplastids is a relic of an ancient "RNA world" (Benne, 1990). More recently, this hypothesis has been questioned. U-insertion/deletion RNA editing is seen in the parasitic bodonid, *Trypanoplasma borreli* (Lukes et al., 1994) (Maslov and Simpson, 1994). It is also seen in the non-parasitic, free-living *Bodo saltans* (Blom et al., 1998). Therefore, it is clear that parasitism has not driven the mechanism of RNA editing in any measure, as was thought earlier.

Upon comparing and analyzing the editing patterns and extents among different members of the kinetoplastids, it was observed that the more recently diverged members appear to show lesser editing (Maslov et al., 1994a). Consequently, a 'retroposition model' for the evolution of RNA editing was proposed (Landweber, 1992) (Simpson and Maslov, 1994a) (Simpson et al., 2000). According to this model, there has been a gradual reduction in the edited domains (and editing levels) in the more recent species, which use their reverse transcriptase-like activities to convert partially and fully edited transcripts to cDNAs, which can then be incorporated into the maxicircles via homologous recombination, to replace the ancestral pan-edited or extensively edited genes. This "editing on-the-way-out" model was rationalized with arguments regarding the sloppiness of the editing machinery and the general energy load it imposed on the cell (Landweber and Gilbert, 1993). Also, there could be a loss of specific minicircles encoding particular gRNAs needed for editing, especially during the post-mitotic, random segregation (Simpson and Maslov, 1994b); as observed in a laboratory strain of *L. tarentolae* after its prolonged culture in the lab (Thiemann et al., 1994). Therefore, a reduction in the editing domain by homologous recombination is advantageous to the cell, since there is no longer a pressure to maintain the big minicircle repertoire. By that analogy, all the maxicircle genes would have been

ancestrally pan-edited, but the current unedited genes have been putatively formed by replacing the pan-edited ancestor with a fully-edited transcript complement; and similarly, 5' edited genes have resulted from the replacement of the pan-edited ancestor with partially edited transcript complements (since editing anyway follows a 3'-5' direction with respect to the mRNA) (Lukes et al., 2005).

However, the fact that genes like COII are unedited in the basal kinetoplastids like *T. borreli* (Lukes et al., 1994), but partially edited in the newer trypanosomatids and bodonids contradict the former theory, and rather proposes a "editing-on-the-way-in" model. Nonetheless, the COII editing domain is small compared to the other sites, and there is reason to believe that this editing may have evolved more recently to correct a genomic frameshift mutation, by utilizing the already existing editing machinery and a cis-acting COII-gRNA sequence (Blom et al., 1998). In conclusion, it is certain that the mechanism of RNA editing itself is at least as ancient as the trypanosomatids themselves and that it has evolved before parasitism (Estévez and Simpson, 1999).

2.9 Functional implications of RNA editing

Despite extensive studies on the several different aspects of RNA editing, the origin and the maintenance of such a complex RNA editing process remains a mystery. It has been proposed that retaining the RNA editing mechanism provides several advantages to the host: 1) RNA editing provides an extra level of regulation of for the expression of mitochondrial genes. Additionally, retention of the 5' editing domain in a few genes may allow translational regulation, by creating a methionine initiation codon and even a ribosomal-binding site (Simpson et al., 2000). 2) RNA editing promotes an accelerated rate of protein evolution (84% faster, on an

average), by creating more genetic variation, i.e., in different species, editing produces a slightly different mRNA sequence from the same unedited DNA sequence. (Landweber and Gilbert, 1993). 3) Many reports have indicated the creation of completely new proteins having totally novel and unrelated functions, from the same unedited gene, via the combinatorial employment of a whole cascade of novel gRNAs, or via editing of only some parts of the pre-edited transcript; the so-called "alternative RNA editing" mechanism (Read et al., 1994) (Ochsenreiter and Hajduk, 2006) (Ochsenreiter et al., 2008a) (Ochsenreiter et al., 2008b). 4) Editing also allows divergence in both the maxicircle and minicircle sequences, while still ensuring protein sequence conservation, at a level of DNA sequence divergence which would certainly be intolerable in the absence of editing (Read et al., 1993). 5) RNA editing is capable of fixing the mutations that accumulate in the mitochondrial genome when the organism lives in an anaerobic environment (no oxidative phosphorylation) (Cavalier-Smith, 1997). 6) Editing may help regulate the synthesis and maintenance of the mitochondrial respiratory system over the course of the life cycle of the organisms (see the following section).

On the other end of the spectrum, there is also a "neutralist hypothesis", that states that RNA editing does not necessarily provide any advantage to the organism. It is postulated that the current biological role of RNA editing is not what the machinery may have originated ancestrally for. One hypothesis suggests that the pre-existing enzymatic machinery capable of U-insertions and deletions may have performed different functions. Somehow, it evolved into the current RNA editing mechanism that could correct mutations involving T residues in the AT-rich kDNA genomes, and eventually, has become indispensable for the survival of the trypanosomatids (Covello and Gray, 1993). Currently, the well-established system allows the cell to accumulate more mutations, without the risk of lethality.

2.10 RNA editing and life-cycle stages of *T. brucei*

T. brucei, one of the well-studied trypanosomatids, expresses an obvious difference in the mitochondrial size and activity throughout its multiple life cycle stages. The overall mitochondrial activity and volume are higher when the parasite is in the insect midgut (epimastigote form), where the glucose levels are low. Both the activity and volume decrease as the trypanosome progresses to the infective metacyclic forms in the insect salivary gland; to finally being the lowest in the bloodstream slender form, where glucose levels are the highest. The glycosome dynamics are quite the opposite; and they are seen in highest numbers in the bloodstream forms (with lowest mitochondrial activity and size), and in least numbers in the insect stages of the parasite. This counterbalancing between the mitochondrion and glycosomes is universally observed in most of the cultured trypanosomatids (de Souza et al., 2009).

When *T. brucei* is in the bloodstream of its vertebrate host, the concentrations of glucose are the highest. Here, the parasite has no functional respiratory chain; most of the electron transport chain (ETC) complexes are absent, the tricarboxylic acid (TCA) system is largely repressed, and energy is obtained exclusively by aerobic glycolysis, or by anaerobic glycolysis, where the oxygen levels are low. Only the F0-F1 ATPase is active, and thus there is productive pan-editing of the A6 transcript. However, instead of the F0-F1 complex generating ATP, it utilizes ATP to maintain the electrochemical proton gradient across the mitochondrial membrane. Gradually, as the bloodstream forms differentiate into short, stumpy forms, along with the F0-F1 complex, Complex I is also additionally created. Here, the transcripts of ND3, ND7, ND8 and ND9 are pan-edited (subunits of Complex I), alongside an increase in the expression of the non-edited Complex I subunits-encoding genes like ND4, ND5, ND1 and possibly ND2 (MURF1). At this stage, the F0-F1 complex again generates ATP using the proton gradient generated by the

Complex I. When taken up by a tsetse fly during its blood-meal, only these stumpy forms survive and rapidly differentiate into the procyclic insect forms. Here, glucose levels are the lowest and amino acid levels are higher, and a complete panel of all the ETC complexes are present and fully functional, thereby indicating the occurrence of oxidative phosphorylation. A respiratory chain starting with Complex II is present and operational. There is also productive editing of transcripts of the subunits of Complex III (CYb) and Complex IV (COII and COIII), alongside an increase in the expression of the COI gene, which is unedited. A detailed review of energy metabolism in different life cycle stages of the parasite is available (Hannaert et al., 2003). In summary, CYb, COII and COIII are edited and necessary in the insect stage; all the ND-subunits are expressed in the bloodstream stage (the stumpy forms); and A6 and RPS12 are edited across all the developmental stages (Ochs et al., 1996) (Aphasizheva et al., 2013).

Although *T. brucei* demonstrates an elaborately distinct and differential RNA editing at different life stages, such striking differences have not been documented in other digenetic trypanosomatids like *T. cruzi* (Kim et al., 1994) or *L. tarentolae* (Carrillo et al., 2001). In another digenetic parasite, *Phytomonas serpens*, which infects plants, it has been demonstrated that Complexes III and IV are completely absent, and all the maxicircle genes encoding their subunits are missing. Thus, respiration, ATP production and Kreb's cycle are all absent, and all the ATP production is completely dependent on glycolysis (Maslov et al., 1999) (Nawathean and Maslov, 2000).

2.11 Trypanosomatid mitochondrial bioenergetics

The mitochondria are the powerhouses of eukaryotic cells that produce energy through oxidative phosphorylation. Without a mitochondrion, most eukaryotes may have had to depend on anaerobic glycolysis for all their energy, like many prokaryotes. In brief, eukaryotic mitochondrial respiration occurs via an electron transport chain (ETC), comprising four major transmembrane enzyme complexes in the mitochondrial inner membrane: NADH-ubiquinone oxidoreductase (Complex I), succinate-ubiquinone oxidoreductase (Complex II), ubiquinol-cytochrome c oxidoreductase (Complex III, also called cytochrome bc1) and cytochrome oxidase (Complex IV, also called cytochrome a3). Ubiquinone (coenzyme Q) and cytochrome c are the two major electron carriers between these complexes (Chance and Williams, 1956). Three of the four Complexes (I, III and IV) act as proton pumps and generate an electrochemical gradient across the mitochondrial membrane, which is used by the F0-F1 ATP synthase (Complex V) to generate ATP, thereby coupling the processes of respiration and oxidative phosphorylation (Schägger, 2001).

The presence and functioning of the Complex I in trypanosomatids is highly controversial and debated (Sloof et al., 1994) (Tielens and Van Hellemond, 1998) (Stuart et al., 2005) . In the higher eukaryotes, Complex I is the first protein complex in the mitochondrial electron transport chain, and it catalyzes the transfer of electrons from NADH to Coenzyme Q. It is also believed to be an important site for the production of mitochondrial superoxide via reverse electron transfer (Lambert and Brand, 2007). Its subunits are also involved in fatty-acid biosynthesis and regulation of apoptosis (Janssen et al., 2006). However, in the trypanosomatids, Complex I has not been detected, possibly due to the complicated life cycle stages the parasites go through, and their mitochondrion being repressed or active at different stages (Bringaud et al., 2006).

Although NADH oxidation has been reported in trypanosomatid mitochondrial extracts (Fang et al., 2001) (Fang and Beattie, 2003), whether the complex itself is involved in energy transduction is still inconclusive (Tielens and van Hellemond, 2009). Prolonged growth of trypanosomatids as lab cultures have been pinpointed to lead to loss of a functional Complex I, as in the case of the UC strain of *L. tarentolae* (Thiemann et al., 1994) or *T. cruzi* Esmeraldo (Westenberger et al., 2006). Additionally, wild type trypanosomatids that sustain without functional cytochromes, or even kDNA altogether, have also been reported, for e.g. the bloodstream form of *T. brucei* and *P. serpens,* and *T. equiperdum* and *T. evansi*, respectively (Maslov et al., 1999) (Nawathean and Maslov, 2000) (Lai et al., 2008). The bloodstream forms of *T. brucei* lacks a fully-developed mitochondrion, citric acid cycle, cytochromes, as also a cyanide sensitive electron transport (Bringaud et al., 2006), while the procyclic insect form has a fully active and functioning mitochondrion and electron transport chain (van Hellemond et al., 2005). Likewise, *P. serpens* also alternates between an insect stage and sugar-rich plant environments (Sanchez-Moreno et al., 1992). Curiously, *P. serpens* lacks Complexes III and IV, and all the maxicircle genes encoding their various subunits are also absent, as mentioned before. However, its mitochondrion oxidizes NADH, much like *T. brucei*. Both these parasites apparently use an aerobic type of glycolysis to obtain energy, by oxidizing the NADH synthesized in their glycosomes (in the bloodstream or plant stage), using a glycerol-3-phosphate dehydrogenase / dihydroxyacetone-phosphate shuttle, which utilizes both, a glycosomal NAD-dependent glycerol-3-phosphate dehydrogenase, and a mitochondrial FAD-dependent glycerol-3-phosphate dehydrogenase (Opperdoes et al., 1981) (Bringaud et al., 2006). In the mitochondrial membrane, the electrons are transferred to a pool of ubiquinone 9, which are oxidized by the trypanosome alternative oxidase (TAO), a nucleus-encoded cyanide insensitive oxidase (Clarkson et al., 1989)

(Chaudhuri et al., 1998). Overall, in *T. brucei* bloodstream form, synthesis of Complexes I-IV is suppressed alongside an upregulation of mitochondrial glycerol-3-phospahte dehydrogenase and TAO. In its procyclic form, Complexes I-IV are active and functional. In *P. serpens*, there is an upregulation of Complexes I, II, the mitochondrial glycerol-3-phospahte dehydrogenase and TAO, while Complexes III and IV are absent (Opperdoes and Michels, 2008). Additionally, an alternative dehydrogenase, NDH2, was also discovered in *T. brucei*, which could catalyze the same biochemical reaction as Complex I (Fang and Beattie, 2002). Therefore, it is being speculated that Complex I does not play a significant role in energy metabolism of procyclic *T. brucei*, similar to the observation in *T. cruzi* (Carranza et al., 2009). Enzymes like NDH2 and glycerol-3-phosphate dehydrogenase may be the major players and electron sources in the trypanosomatid respiratory chain. And if the procyclic stage parasites are grown in a glucose-rich environment, the mitochondrion uses its respiratory chain in the way it would in its bloodstream stage, i.e., ATP would primarily be synthesized via substrate level phosphorylation instead of oxidative phosphorylation and Complex V; and the respiratory system is reduced to just an electron sink for glycerol-3-phosphate dehydrogenase, which connects glycolysis and the mitochondrial respiration (Hill and Cross, 1973). The Complexes III and IV, however, become essential due to the overall reduced activity of Complex I, and the need to maintain a mitochondrial membrane potential by proton translocation (Verner et al., 2011).

2.12 Variety in the trypanosomatid energy metabolism

For the longest time, it was believed that the differences in the trypanosomatid energy metabolism are attributable to the differences in the available nutrient supply in their environments. However, this hypothesis was disproven, and additionally, there is no obvious

relationship between the trypanosomatid energy metabolism, their phylogenetic relationships or their mode of transmission and the number of hosts that can be infected (Tielens and van Hellemond, 2009). This same review divides the energy metabolism of trypanosomatids into **four** metabolic categories with an increasing complexity as under.

Parasites in the **first category** utilize glucose exclusively via glycolysis and produce pyruvate that is used to generate ATP. They do not consume amino acids and do not use a cytochrome-containing respiratory chain or oxidative phosphorylation for ATP production. The bloodstream stage of these trypanosomatids rely exclusively on the plant-like alternative oxidase (TAO) for the oxidation of NADH produced during pyruvate formation by glycolysis. This category of parasites include the long-slender bloodstream forms of the *T. brucei* family, including *T. b. brucei*, *T. b. gambiense*, *T. b. rhodesiense* and *T. evansi* (Ryley, 1956) (Flynn and Bowman, 1973) (Sutherland et al., 1995) (Schnaufer et al., 2002) (Lai et al., 2008).

Trypanosomatids of the **second category** also rely solely on glucose for their energy source. However, they do not produce pyruvate, but instead produce acetate as the end product. They also do not consume amino acids, or utilize a cytochrome containing respiratory chain and oxidative phosphorylation, and rely solely on the TAO to oxidize NADH, similar to the members of the first category. This category includes the bloodstream stage of *T. congolense* (Bienen et al., 1991) (Obungu et al., 1999) and plant stage of *Phytomonas* (Sanchez-Moreno et al., 1992) (Chaumont et al., 1994) (Van Hellemond et al., 1998).

The **third category** currently is limited to the bloodstream stage of *T. theileri*. They also do not use amino acids, but utilize only glucose for ATP production via glycolysis. The end products are acetate and succinate. However, in contrast to the first and second categories, these

trypanosomatids do not use their TAO, but instead utilize the classical cytochrome mediated respiratory chain and oxidative phosphorylation (van Hellemond et al., 2007).

The **fourth and final category** contains the trypanosomatids with the most complex metabolism. These trypanosomatids do not rely solely on glucose, and can additionally also utilize amino acids. Their metabolic end products also include acetate and succinate; and they use cytochrome mediated respiratory chain and oxidative phosphorylation for ATP production. This category includes the bloodstream form of *T. lewisi* (Ryley, 1951) (Ryley, 1956), procyclic form of *T. brucei* family (van Weelden et al., 2003) (van Weelden et al., 2005), choanomastigote forms of *Crithidia* species (Cazzulo et al., 1985) (Gilroy et al., 1988), and all the stages in the insect and mammalian hosts of *T. cruzi* (Cazzulo, 1992) (Cazzulo, 1994) and *Leishmania* species (Opperdoes and Coombs, 2007) (Tielens and Van Hellemond, 1998) (Hart and Coombs, 1982) (Rosenzweig et al., 2008). Although *T. cruzi* and *Leishmania* species solely depend on the cytochrome mediated respiratory chain, the procyclic forms of *T. brucei* family and the choanomastigote forms of *Crithidia* species can also utilize TAO.

In conclusion, it is hypothesized (Tielens and van Hellemond, 2009) that the reason for the extant variety in energy metabolism in trypanosomatids is that the parasites could discard many of their metabolic capabilities upon finding newer hosts including vertebrates or plants, that could provide those metabolic intermediates to them. As they started adapting to their new host environments, the extents to which they disposed their metabolic capabilities varied among the different trypanosomatid species. It is possible that eventually, all the trypanosomatids will follow the simplified, but distinct pathways and strategies utilized presently by *T. brucei*.

2.13 Mitochondrion as a chemotherapeutic target

Although clinical management of trypanosomatid diseases relies heavily on conventional chemotherapeutics, the drugs available are unable to meet the demands of the endemic population due to their high costs, toxicity, side effects, drug-resistance problems and so on. Therefore, there is an urgent need to develop novel vaccines as quickly as possible to contain the diseases. The cell biology of these protozoans are of a great interest since the parasites show unique organelles and cellular biochemical pathways, which promises to identify new drug targets and newer drugs (de Souza, 2008). In those lines, the trypanosomatid mitochondrion is an excellent candidate for studies to identify newer drug targets. It is important to note that in contrast to mammalian cells containing hundreds or thousands of mitochondria, the trypanosomatid cells contain only a single mitochondrion. Thus, proper functioning of this single organelle is of supreme importance for the parasite's survival (Mehta and Shaha, 2004). There are several structural and functional peculiarities exclusive to the trypanosomatid mitochondrion, which can be channeled into becoming effective chemotherapeutic targets. Following are some of the mitochondrial targets:

2.13.1 Electron Transport chain

An alternative rotenone insensitive Complex I, NADH:quinone oxidoreductase, has been described in trypanosomatids like *Trypanosoma* and *Leishmania* (Chen et al., 2001) (Abrahamsen et al., 2004). Since this alternative Complex I has no human counterpart, it could be a fascinating drug target. Additionally, parasites also normally have only a single NADH dehydrogenase gene, and thus, the gene could make an important drug target (Vaidya, 2004). Biochemical studies on these parasites have revealed that they have a functional Complex II and Complex V through all their life cycle stages; and these complexes and their subunits could

make an excellent candidate. Also, the parasite has a limited electron transport between Complexes I to III and thus, succinate might be their primary electron donor for energy transduction. Thus, succinate is very important for the parasites and *Trypanosoma* and *Leishmania* even have a novel enzyme, fumarate reductase, which recycles succinate from fumarate (Turrens, 1987) (Santhamma and Bhaduri, 1995). This enzyme's absence from the mammalian system makes it another important target. The amino acid sequence of the trypanosomatid CYb show differences in their ubiquinone interaction sites, when compared with the mammalian CYb, and these can act as an attractive target (Srivastava et al., 1997).

2.13.2 Trypanosomatid alternative oxidase

As mentioned before, the trypanosomatid alternative oxidase is found in the bloodstream form of *T. brucei* and to a smaller extent, also in its procyclic form. The TAO appears to be a non-proton pumping oxidase, which does not participate in generating a proton motive force across the mitochondrial membrane (Mather et al., 2007). Additionally, the corresponding enzyme appears to have been lost from the mammalian genome during evolution (Chaudhuri et al., 2006) (Clarkson et al., 1989). Therefore, the trypanosomatid TAO is an excellent drug target too.

2.13.3 kDNA replication

Again, the presence of a single mitochondrion implies that the kDNA needs to be faithfully replicated and segregated into the two new daughter cells. The kDNA replication is an elaborate process, where the DNA content of the kinetoplast network doubles in size before the network partitions into the progeny cells. The entire process is also synchronous to the nuclear replication. The DNA topoisomerase II has been implicated to be one of the most important enzymes in the kDNA replication, where it creates a double strand breakage and strand passing

to decatenate the minicircles in the network, to ensure a faithful segregation of the kDNA (Shapiro and Showalter, 1994). Several distinguishing features that differentiate the trypansomatid topoisomerase II from its prokaryotic and eukaryotic counterparts have been noticed (Das et al., 2006). Likewise, differences were also seen in the trypanosomatid topoisomerase IB, when compared to their mammalian homologs (Reguera et al., 2006). Thus the DNA topoisomerases are an excellent drug target. Two classes of topoisomerase drugs are being pursued: a) topoisomerase inhibitors that compete with ATP to bind to the catalytic site of the topoisomerase, and b) topoisomerase poisons that stabilize the topoisomerase-DNA complex and induce DNA breakage (Motta, 2008).

2.13.4 Mitochondrial tRNA import

Trypanosomatids have a unique scenario, where they have lost their complete set of mitochondrial tRNA genes, unlike most other organisms that still retain some of their tRNA genes. Thus, trypanosomatids solely rely of importing cytosolic tRNA into their mitochondrion via specialized systems for mitochondrial protein translation (Hancock and Hajduk, 1990) (Schneider and Maréchal-Drouard, 2000) (Bhattacharyya and Adhya, 2004) (Alfonzo and Söll, 2009). The RNA import complex (RIC) is the only multiprotein complex from the *Leishmania* mitochondria, that has been demonstrated to import tRNAs using phospholipid vesicles in an ATP dependent manner (Adhya, 2008).

2.13.5 Mitochondrial fatty acid and sterol synthesis

Trypanosomatid fatty acid synthesis (FAS) is remarkably different than that of the mammalian FAS and it is critical to their survival (Lee et al., 2006) (Morita et al., 2000). Recently, it has been indicated that the trypanosomatids are able to make fatty acids by a mitochondrial FAS

pathway (Morita et al., 2000) (Stephens et al., 2007), and that this pathway is essential for efficient respiration, although their relationship is unclear (Guler et al., 2008). *Leishmania* and *Trypanosoma* species produce ergosterol-related sterols, and are extremely susceptible to these sterol biosynthesis inhibitors. Inhibition of several enzymes in this pathway causes a depletion in their normal sterol levels and an accumulation of abnormally high levels of sterol precursors, which can be cytostatic or cytotoxic (Roberts et al., 2003).

2.13.6 Mitochondrial membrane potential (MMP)

Maintenance of the MMP is crucial for cellular survival and MMP variations can affect a wide range of mitochondrial functions like ETC inhibition, Complex V blockage, stimulation of uncoupling proteins or permeabilization and destabilization of the mitochondrial inner membrane (Gottlieb, 2003). Similar to the observations in mammalian mitochondria, inhibitors of the trypanosomatid Complexes II and III cause a dissipation of the MMP. Interestingly, although the functionality of the trypanosomatid Complex I has been questioned, the mammalian Complex I inhibitor, Rotenone, appears to also hyperpolarize the mitochondrial membrane of *Leishmania* (Mehta and Shaha, 2004). Therefore, destabilizing the MMP also appears to be an important target.

2.13.7 Mitochondrial Reactive Oxygen species (ROS)

ETC dysfunction results in an excessive release of ROS from the mitochondria, probably due to the deviation of electrons from the mitochondrial complexes (Gille and Nohl, 2001). Complex II inhibition has primarily been suspected to lead to ROS production in trypanosomatids. However, in higher eukaryotes, mitochondrial ROS are primarily generated in complexes I and III (Chen et

al., 2003) (Fidalgo and Gille, 2011), thereby making complex II another target to hit trypanosomatid survival by a different pathway.

2.13.8 Induction of Apoptosis

Besides the metabolic and respiratory roles of the mitochondria, their involvement in regulating apoptosis is also widely studied (Gottlieb, 2003). Although programmed cellular death are known to exist in *Trypanosoma* and *Leishmania*, their exact mechanism is again unclear. A few steps have been suggested and demonstrated to lead to apoptosis: i) a drop in the MMP and ATP levels; ii) increase in the production of ROS and hydrogen peroxide, iii) increase in the intracellular Ca^{2+} levels, and iv) the above steps leading to apoptosis initiation (Sen et al., 2007) (Arnoult et al., 2002).

Other than these major targets in the mitochondrion, other proteins associated with the kinetoplast could also be invaluable targets. These include, but are not limited to, DNA polymerase P, minicircle origin-binding protein, kDNA condensing proteins, kinetoplast associated proteins (KAPs), nicking enzymes, mitochondrial heat shock proteins, kinetoplast binding proteins, mitochondrial translation proteins, the RNA editing editosome complex, RNA ligase KREL1 or even inhibitors of insertional editing etc. (Engman et al., 1989) (González et al., 1990) (Torri and Englund, 1992) (Tzfati et al., 1992) (Xu and Ray, 1993) (Li et al., 1995) (Zavala-Castro et al., 2000) (Schnaufer et al., 2002) (Sen and Majumder, 2008) (Osato et al., 2009) (Durrant et al., 2010) (Silva Paes et al., 2011) (Niemann et al., 2011).

CHAPTER 3

The maxicircle genomes of endosymbiont-bearing and related insect trypanosomatids

Abstract:

We have sequenced and characterized the maxicircle genomes of two species of *Angomonas*, three species of *Strigomonas*, *Crithidia acanthocephali*, *Leptomonas costaricensis, Endotrypanum schaudinni* and *Herpetomonas muscarum* with respect to their sizes, gene-contents, synteny, similarities of their predicted proteins, RNA-editing patterns, and phylogenetic relationships. The maxicircles of these organisms exhibit a syntenic panel of 20 genes (excluding guide RNA genes) beginning with the ribosomal-RNA genes and ending with the gene encoding NADH dehydrogenase-5. The number of maxicircle genes that are edited, and their editing patterns are similar among closely related organisms, but vary widely among the more distantly related organisms. While *H. muscarum* has the greatest number of pan-edited genes, the symbiont-harboring trypanosomatids have the least number of pan-edited and highest number of unedited genes. The phylogeny of these sequences closely reflects the expected relationships based on nuclear genes. An analysis of the inferred RNA editing patterns among these organisms using a binary-value based strategy suggests that RNA editing has evolved in parallel with the evolution of their nuclear genes.

Introduction:

The protistan family Trypanosomatidae (Order Kinetoplastida) is extensively studied owing to its basal placement on the evolutionary tree and because it represents an extremely successful and important clade of parasites. Many animals including insects, invertebrates, virtually all members of the Phylum Chordata, and many plants are parasitized by one or more taxa of Kinetoplastida. The insect-infecting trypanosomatids probably represent the largest segment of the family. Invertebrates, insects and related arthropods are hosts to monogenetic parasites including the *Crithidia, Herpetomonas* and *Leptomonas*; and also act as the vector mediators of digenetic parasites; e.g., the *Leishmania, Endotrypanum* and *Trypanosoma* that infect vertebrates or the *Phytomonas* that infect plants (Vickerman, 1994). A smaller group of insect trypanosomatids; e.g., members of the genus *Angomonas* and *Strigomonas* are known as the symbiont-harboring trypanosomatids (SHTs) (Teixeira et al., 2011). Members of this group harbor a single betaproteobacterial symbiont in their cytoplasm (*Candidatus* Kinetoplastibacterium or trypanosomatid proteobacterial endosymbiont [TPEs]) that provide genes and functions that supplement the host metabolism (Alves et al., 2013a) (Alves et al., 2013b). These TPEs divide synchronously with the host cell and are transmitted vertically (Motta et al., 2010).

Members of the Order Kinetoplastida are characterized by the presence of a unique organelle called the “kinetoplast”, a single, large mitochondrion situated near the base of the flagellum. The kinetoplast contains the kinetoplast-DNA (kDNA) which is a complex network of thousands of small, interlocking circular DNA molecules of two types: a few dozen of maxicircles and upwards of 5000 minicircles, catenated in the network (Shapiro and Englund, 1995) (Lukes et al., 2002) (Shlomai, 2004) (Liu et al., 2005). The maxicircles that range from ~16 to over 40 kb

are the functional homologs of the eukaryotic mitochondrial DNA with genes that encode subunits of respiratory chain complexes and mitochondrial ribosomal RNAs. However, they do not encode any tRNAs and all tRNAs needed for translation within the kinetoplast are imported from the cytoplasm (Shapiro and Englund, 1995) (Maslov et al., 1984) (Alfonzo and Söll, 2009).

One of the most peculiar features of the kinetoplast is the complex mechanism of RNA editing that is required for maturation of many of the maxicircle transcripts. Primary transcripts of maxicircle genes often lack start and /or stop codons, contain frameshifts, lack open reading frames, or exhibit apparent insertions or deletions that disrupt the proper protein sequences. These transcripts are completed by insertion and/or deletion of one or more uridylate (U) residues at one or more sites across the length of the pre-edited transcript to create a mature, translatable mRNA (Benne et al., 1986) (Estévez and Simpson, 1999) (Simpson et al., 2003). The sites of these specific insertions and deletions are defined by small, 3' oligouridylated RNA molecules called guide RNAs (gRNAs) that are often encoded in the minicircle sequences and, less frequently, in the maxicircle genomes themselves (Blum et al., 1990) (Sturm and Simpson, 1990b).

The enzyme-cascade model for U-insertion/deletion editing is widely accepted (Blum et al., 1990) (Osato et al., 2009) (Simpson et al., 2004b) (Stuart et al., 2005) (Aphasizhev and Aphasizheva, 2008). A protein complex, the "editosome", mediates the interaction between the pre-edited transcript and the gRNA; and many of the proteins in the complex have been characterized (Aphasizhev et al., 2003) (Li et al., 2009). Editing begins by complementary interaction of the 5' end of a specific gRNA with a target ~15 nt sequence at the 3' end of the pre-edited transcript to form the "anchor-duplex" and U residues are added or deleted to the pre-edited RNA to generate a complementary sequence. Editing proceeds stepwise in the 3' to 5'

direction, often requiring multiple gRNAs until the mature mRNA is generated (Blum et al., 1990) (Maslov and Simpson, 1992). Each gRNA is responsible for the editing of only a short stretch of 30-50 nts termed as the "editing block" (Gao et al., 2001) (Corell et al., 1993). Some mRNAs, termed "pan-edited" transcripts, are edited across their full lengths.

Most trypanosomatid maxicircles contain 18 protein-encoding genes and two ribosomal RNA genes (Simpson et al., 1998). The location and the polarity of these genes are conserved among the trypanosomatids studied to date. Moreover, these syntenic genes are clustered tightly within a ~16 kb "conserved" or "consensus" region. Outside of this syntenic region in the maxicircles, is a non-coding "divergent region" (DR) of variable length (Muhich et al., 1985) (de Vries et al., 1988) (Sloof et al., 1992). Though the function of the divergent region remains unknown, there have been reports that this region contains sequences similar to the minicircle replication origin (Gorbat et al., 1990) (Myler et al., 1993) and promoters for the 12S rRNA gene (Vasil'eva et al., 2004). The genes ND8, ND7, COIII, CYb, MURF4 (ATPase 6), G3, COII, MURF2, ND4, RPS12 and ND5, and the 12S and 9S ribosomal genes are transcribed from the same strand (with the same polarity); whereas the genes ND9, MURF5, MURF1, ND1, COI, G4 and G5 are transcribed from the opposite strand (Cruz et al., 1984) (Ochs et al., 1996) (Maslov et al., 1999) (Yatawara et al., 2008). The transcripts of five of these protein-coding genes (MURF1, ND1, COI, ND4 and ND5) are not known to undergo editing, while transcripts from the remaining genes undergo 5', internal or extensive (pan) editing. The pattern and extent of editing for a given gene varies among different species (Alfonzo et al., 1997).

The whole length and partial maxicircle sequences of organisms including *Leishmania tarentolae* (GenBank: M10126.1) (Blum et al., 1990) (Maslov and Simpson, 1992) (Cruz et al., 1984) (de la Cruz et al., 1985a) (de la Cruz et al., 1985b) (Muhich et al., 1985) (Simpson et al.,

1985) (Simpson et al., 1987) (Feagin et al., 1988b) (Shaw et al., 1988) (Shaw et al., 1989) (Bhat et al., 1990) (Sturm and Simpson, 1990a) (Blum and Simpson, 1990) (Bhat et al., 1991) (Maslov et al., 1992) (Sturm et al., 1992); *Trypanosoma brucei brucei* (GenBank: M94286.1) (Benne et al., 1986) (Shaw et al., 1988) (Bhat et al., 1990) (Maslov et al., 1992) (Eperon et al., 1983) (Benne et al., 1983) (Hensgens et al., 1984) (Feagin et al., 1985) (Feagin et al., 1987) (Feagin et al., 1988a) (de Vries et al., 1988) (Koslowsky et al., 1990) (Sloof et al., 1992) (Souza et al., 1992); and *Phytomonas serpens* (GenBank: AF079967.2) (Maslov et al., 1999) (Maslov et al., 1998) are available in public databases. However, only a few other maxicircle sequences are available, as are the editing patterns for individual genes and transcripts for other important trypanosomatids.

Herein, we report the sequences and characterize the maxicircle genomes of *Angomonas deanei, A. desouzai, Strigomonas culicis, S. galati, S. oncopelti, Leptomonas costaricensis, Endotrypanum schaudinni, Herpetomonas muscarum* and *Crithidia acanthocephali*; and compare them with the published maxicircles of *L. tarentolae* and *P. serpens*. We have also inferred the apparent RNA editing patterns of each gene and performed phylogenetic analyses to ascertain if the phylogenetic relationships corroborate the relationships based on the nuclear-encoded genes. Finally, we have quantified the inferred RNA editing patterns of these genes using a binary value-based approach to probe the evolution of the RNA editing process.

Abbreviations for organisms herein

A. dea = Angomonas deanei; A .des = Angomonas desouzai; C. aca = Crithidia acanthocephali; E. sch = Endotrypanum schaudinni; H. mus = Herpetomonas muscarum; Lam = Leishmania mexicana amazonensis; L. cos = Leptomonas costaricensis; L. tar = Leishmania tarentolae; P. ser = Phytomonas serpens; *S. cul = Strigomonas culicis; S. gal = Strigomonas galati; S. onc = Strigomonas oncopelti*; *T. bru = Trypanosoma brucei.*

Gene Abbreviations

A6 = ATPase 6; CO = Cytochrome Oxidase; CYb = (Apo) Cytochrome b; G = G-rich regions; GAPDH = Glyceraldehyde 3-Phosphate Dehydrogenase; MURF = Mitochondrial Unidentified Reading Frame; ND = NADH Dehydrogenase; RPS12 = Ribosomal Protein Subunit 12; SSU rDNA = Small subunit ribosomal RNA gene.

Other abbreviations

AA(s) = amino-acid(s); Avg = average; bp(s) = base-pair(s); gDNA = genomic DNA; kb = kilo bases; min = minutes; MSA = Multiple Sequence Alignment; No. = Number; nt(s) = nucleotide(s); PCR = polymerase chain reaction; s = seconds; T = Thymidylate residue; T_m = Temperature of melting; U = uridylate residue

Materials and Methods:

Trypanosomatid strains, culture and DNA extraction

All the trypanosomatids cultures in this report [*C. aca* (TCC037E), *A. dea* (TCC036E), *A. des* (TCC079E), *S. cul* (TCC012E), *S. gal* (TCC219), *S. onc* (TCC290E), *L. cos* (TCC169E), *H. mus* (TCC001E) and *E. sch* (TCC224)] were obtained from the Trypanosomatid Culture Collection of the University of Sao Paulo (Brazil), TCC-USP. *C. aca*, *L. cos* and *H. mus* were grown in LIT medium (Camargo, 1964) supplemented with 5% fetal bovine serum (FBS). *S. onc* was grown in ATCC 483 Guttman's IIB medium. *E. sch* was grown in Grace's insect medium (GIBCO, Life Technologies, USA) supplemented with tryptose, hemin and 10% FBS. Cultures were incubated at 25^{o}C and culture volumes were expanded every two days. Cultures were maintained in logarithmic phase and harvested by centrifugation at 3000 rpm for 8 mins at 4^{o}C. Following centrifugation, the harvested pellets were washed three times with phosphate buffered saline (pH 7.4) and total genomic DNA (including kDNA) was extracted from the washed pellets (for *C. aca*, *L. cos*, *H. mus* and *E. sch*) using ZR Genomic DNA – Tissue MiniPrep kit (Zymo Research, USA), as described by the manufacturer.

Shotgun sequencing of trypanosomatid genomes

Total DNA from all the organisms described above was subjected to shotgun sequencing using the Roche 454 FLX pyrosequencing platform, essentially as described by the manufacturer, and the reads were assembled using Newbler assembler (Roche) software (version 2.3) as previously described (Alves et al., 2013a) (Alves et al., 2013b) (Alves et al., 2011). Contigs bearing putative maxicircle sequences were pre-selected by alignment to published maxicircle sequences of *L. tar*

using BLASTN version 2.2.25 (Altschul et al., 1990), with a maximum E-value threshold of $1x10^{-6}$ (*-evalue 1e-6*).

Maxicircle contig closure

Gaps in maxicircle assemblies were closed by sequencing PCR amplicons generated using primers designed to target the ends of predicted adjacent contigs and amplify the gap sequences. Amplifications were performed using Platinum *Taq* DNA Polymerase (Life technologies, USA). For each reaction, 5 µl of 10X PCR Buffer (-Mg), 1 µl each of 5mM forward and reverse primers, 1 µl of 10 mM dNTP mix, 2 µl of 50 mM $MgCl_2$, 1 µl of 50 ng total genomic DNA, and 0. 2 ul (2 units per reaction) of Platinum *Taq* DNA Polymerase were combined with 38.8 µl double distilled water to a 50 µl total reaction volume in GeneMate 0.2 ml ultra-thin walled PCR tubes with flat caps (BioExpress, USA). PCR conditions were as follows: initial denaturation at 94°C for 2 min followed by 30 cycles of 30 s denaturation at 94°C, 30 s annealing at 48-58°C (See **Supplementary Table 3.1** for primers and their T_m) and extension at 72°C for 1 min per kb of the expected amplicon size. For sequences that failed to amplify, a magnesium-gradient PCR (range 0.5 mM – 5 mM Mg with 0.5 mM intervals) was performed with Platinum *Taq* DNA Polymerase with the same conditions as above. The size of each amplicon was visualized by electrophoresis in 1% agarose gels stained with SYBR Safe DNA Gel stain (Life technologies, USA). Reactions yielding a single sized product were purified using QIAquick PCR purification columns (QIAGEN, Germany). For reactions yielding multiple sized products, each band was individually excised and the amplicons were gel-extracted using QIAquick Gel Extraction kit (QIAGEN, Germany). The purified amplicons were sequenced in an Applied Biosystems 3730 capillary sequencer with BigDye terminator chemistry (Applied Biosystems, Life Technologies, USA), using the PCR primers as sequencing primers. Sequence reads were analyzed and

assembled with the maxicircle contigs using Sequencher version 5.0 sequence analysis software (Gene Codes Corporation, USA).

DNA sequence analyses

The maxicircle sequences were aligned using MUSCLE (Edgar, 2004a) (Edgar, 2004b) in MEGA 6.0 (Tamura et al., 2013), and subsequently curated manually on the MSA editing and visualization software, GeneDoc (Nicholas et al., 1997). Synteny was assessed using NCBI's BLAST (Altschul et al., 1990) (Altschul et al., 1997). Gene sequences were translated *in silico* using the ExPASy Translate tool with "mold, protozoan and coelenterate mitochondrial" genetic code (Gasteiger et al., 2003) (Artimo et al., 2012). Gene and protein percent identity matrices were created using Clustal 2.1 of ClustalW2 Phylogeny (Larkin et al., 2007) using the "Distance matrix" tree format and "Neighbor-joining" clustering method without gap-exclusions or distance correction.

RNA editing analyses

For the edited genes, putative start and stop codons, and open reading frames were inferred by aligning and comparing their pre-edited sequences with the previously published, edited transcripts from *L. tar, P. ser, Lam*, and *T. bru*. Using these alignments, it was possible to infer the open reading frames by manually inserting 'U' nucleotides or deleting encoded 'T' nucleotides parsimoniously to create open reading frames that encoded proteins with maximal identity to the published transcript and protein sequences. The resulting predicted coding sequences of the transcripts and protein alignments for all the genes and their identities are shown in the **Supporting Information (SI) for Chapter 3** section following this chapter.

RNA folding analyses

The 12S and 9S rDNA sequences were folded, and their secondary structures predicted, using the 'RNA folding form' application of the Mfold web server (Zuker and Jacobson, 1998) (Zuker, 2003), with default parameters. The folding constraints were calculated and specified in accordance with the secondary structures of *L. tar* 12S and 9S. If more than one secondary structure was possible despite the constraints, the structure with the lowest free energy (dG) was chosen for representation here, as it is expected to be the most stable. These secondary structures are consistent with the known secondary structure of the mitochondrial RNAs of *L. tarentolae* 12S (except an uncharacterized portion) and 9S (de la Cruz et al., 1985b) (de la Cruz et al., 1985a).

Phylogenetic analyses

Individual gene and protein sequences were aligned using MUSCLE (MEGA) and followed with manual curation, as described above, and the alignments were concatenated. Three types of alignment-datasets were generated for analysis using the maximum likelihood (ML) method, using the software raxmlGUI (Silvestro and Michalak, 2012) (Stamatakis, 2006): **1) the nuclear dataset** consisting of a concatenated nucleotide alignment of SSU rDNA and GAPDH genes; **2) the maxicircle dataset** consisting of a concatenated protein alignment of the following non-edited maxicircle genes: COI, MURF1, ND1, ND4 and ND5 (note that COI is absent and ND5 is incomplete in *P. ser*); and **3) the RNA editing quantification dataset** comprising a binary (1-0) concatenated matrix to quantify the editing events at each locus in each of the inferred mature coding sequences of the transcripts of the following edited maxicircle genes: G3, G4, ND3 (conserved motif), ND8, ND9 and RPS12 from the pan-edited genes; CYb, MURF2, A6 from the 5' edited genes; and the internally edited COII gene. ND7 and COIII were excluded from this

analysis because of their highly variable editing patterns and the consequent bias they introduced towards *H. mus*, thereby resulting in a long branch for *H. mus*. In brief, in this binary value based strategy, the presence of an editing event [insertion or deletion of U(s)] at a specific locus of an inferred mature coding sequence was scored as "1"; and the corresponding absence of editing at that same locus for a different organism was scored as "0" (Please refer to the **Supplementary Figure 3.1** for a pictorial explanation of the concept).

For the **nuclear dataset**, ambiguous sites in the curated concatenated nucleotide alignment of SSU rDNA and GAPDH genes were removed using Gblocks (Talavera and Castresana, 2007). The only change to the default parameters of Gblocks was the reduction in stringency with the use of "With Half" option in the "Allowed Gap Positions" parameter. Following this step, the alignment was imported into raxmlGUI, and "GTRGAMMA" option was used to apply the general time reversible model of nucleotide substitution (Tavaré, 1986) and among site rate heterogeneity (Yang, 1996). For the ML analysis, "ML + rapid bootstrap" was chosen with 1000 bootstrap replicates. The **maxicircle dataset** was imported into raxmlGUI and the software estimated MtZoa (Rota-Stabelli et al., 2009) as the best substitution model for this dataset through its "AUTO" option. MtZoa was used with the "PROTGAMMA" option to account for among site rate heterogeneity with empirical frequencies. Again, "ML + rapid bootstrap" option was used with 1000 bootstrap replicates. For the **RNA editing quantification (binary) dataset**, the concatenated matrix of editing events was imported into GeneDoc for a visual curation before being imported into raxmlGUI. ML analysis was carried out using the "ML + rapid bootstrap" option with 1000 bootstrap replicates and "BINGAMMA" option for among site rate heterogeneity. Trees were visualized and formatted on MEGA 6.0. All alignment-datasets are available upon request.

Results:

Maxicircle genome characteristics

The maxicircle genomes from *C. aca*, *A. dea*, *A. des*, *S. cul*, *S. gal*, *S. onc*, *L. cos*, *H. mus* and *E. sch* were assembled as outlined in Materials and Methods, analyzed and compared to published maxicircle sequences of *L. tar* and *P. ser*. The estimated maxicircle sizes range from ~16 kb for *H. mus* to ~20 kb for *E. sch* and generally assort according to the predicted phylogenetic relationships; e.g., the maxicircle of *E. sch* is closest in size to its closest relative, *L. tar*, and most of the genome characteristics follow a similar pattern (**Table 3.1**). The maxicircles of only *A. des* and *S. cul* are full length and circularized (provided the primers did not span a repetitive sequence in the DR).

The GC content of the maxicircles ranged between 21 to 29% for all the organisms studied. The 20 known maxicircle genes (including the 12S and 9S genes) were identified and appear to be perfectly syntenic in all the maxicircles studied. The maxicircle of *P. ser* (GenBank: AF079967.2) has a truncated ND5 gene (incomplete) and also reportedly misses four (CYb, COI, COII and COIII) of the twenty protein-coding genes (Maslov et al., 1999) (Nawathean and Maslov, 2000), which were also reported to be absent in the *Phytomonas* strains EM1 and HART1 (Porcel et al., 2014). Our preliminary analyses with *Phytomonas* JMA also confirms the absence of these four maxicircle genes. However, these genes are present in the maxicircle of *H. mus*, which is quite closely related to *Phytomonas*. A partial maxicircle sequence spanning the divergent region of *S. onc* (GenBank: X56015.1) overlaps with our *S. onc* sequence (COI to ND5 on one end and the 12S on the other) to form a virtual circular genome.

Table 3.1. Maxicircle genome statistics[+]

Organism	*L. tar*	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser^*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
Genome size [est. bps]	20992	20116	18176	18760	15680	10478	17454	16245	17449	17304	16911
Number of contigs	na	3	1	3	2	na	1	1	2	1	1
GC content (%) +	21	21	25	24	26	25	28	29	25	25	25
Coding sequence [bps] # >	15661	15613	15546	15487	14669	9931	15512	15587	15652	15832	15576
Coding sequence (%) # ~ +	75	78	86	83	94	95	89	96	90	92	92
Avg intergenic distance [bps] !+	30	33	35	31	32	21	18	11	29	8	19
5' edited genes*	6	6	6	6	4	4	6	6	6	6	6
Internally edited genes	1	1	1	1	1	0^ (absent)	1	1	1	1	1
Pan-edited genes	6	6	6	6	8	6	5	5	5	5	5
Non-edited genes	7	7	7	7	7	6	8	8	8	8	8
Total number of genes #	20	20	20	20	20	16	20	20	20	20	20

+ All values rounded to nearest integer
^ 4 (of 20) genes are absent in *P. ser*; internally edited COII gene is one of the genes absent; ND5 is incomplete. See text for details.
 # includes rRNA sequences, but not guide RNA genes
> sum of the lengths of pre-edited genes encoded by the maxicircle; excludes the intergenic regions
~ (coding sequence / genome size)*100; includes divergent region in the total genome size
! excludes the divergent region
*** Includes MURF5 gene that is putatively 5' edited to create a start codon**

Based on these maxicircle genome sizes, the lengths of the pre-edited coding sequences (excluding intergenic regions) summed to ~15 kb, occupying ~75-96% of the maxicircles. As shown for the other maxicircle genomes, the conserved gene-coding regions begin with the ribosomal RNA genes (12S and 9S) and end with the ND5 gene. The polarities of all of the maxicircle genes of all of the species examined were as previously reported in *L. tar* and *P. ser*

maxicircles. The co-ordinates of the pre-edited coding sequences, the pre-edited gene lengths and the putative start and stop codons of all the genes in each of the maxircircles studied are shown in **Tables 3.2A and 3.2B**. As described in the Materials and Methods section, start codons, stop codons and the editing patterns for the edited genes were inferred based on the alignment of the pre-edited maxicircle sequences with known fully-edited transcripts and protein sequences. Although the translation start codon for most of the maxicircle genes is AUG, non-canonical start codons including AUU, AUA, GUG and UUG also apparently function as translation start codons. The translation stop codons are universally UAA or UAG. The mitochondrial UGA codon was reported to encode tryptophan (Trp / W) for other trypanosomatids, and this seems to be true for all of these organisms as well (Osawa et al., 1992) (Jukes and Osawa, 1993).

Table 3.2A. Coding gene lengths with co-ordinates and start - stop codons for *L.tar*, *E. sch*, *L. cos*, *C. aca*, *H. mus* and *P. ser* (Genes ordered as per their location on the maxicircle)

GENES		*L. tar*	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*
12S*	**CC**	438-1610	749-1912	818-1984	59-1217	286-1439	211-1368
	CL	1173	1164	1167	1159	1154	1158
9S*	**CC**	1639-2249	1932-2543	2000-2611	1233-1840	1451-2055	1388-1996
	CL	611	612	612	608	605	609
ND8 (G1)*	**CC**	2348-2613	2653-2910	2694-2954	1941-2189	2166-2410	2034-2256
	CL	266	258	261	249	245	223
	Start-Stop	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG
ND9 (G2)*#	**CC**	2975-2685	3246-2971	3317-3054	2491-2249	2762-2479	2575-2312
	CL	291	276	264	243	284	264
	Start-Stop	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA
MURF5*#	**CC**	3268-2966	3458-3177	3614-3348	2773-2492	3023-2778	2825-2448
	CL	303	282	267	282	246	378
	Start-Stop	?-UAA	?-UAG	?-UAA	?-UAA	?-UAA	AUG-UAG
ND7 (MURF3)	**CC**	3318-4458	3584-4727	3648-4788	2809-3948	3067-3782	2900-4041
	CL	1141	1144	1141	1140	716	1142
	Start-Stop	AUU-UAA	AUU-UAA	AUU-UAA	AUU-UAA	AUU-UAG	AUU-UAA
COIII	**CC**	4511-5362	4760-5617	4832-5675	4001-4845	3843-4205	Gene
	CL	852	858	844	845	363	Absent
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	UUG-UAA	
CYb	**CC**	5404-6481	5699-6777	5717-6796	4891-5969	4268-5347	Gene
	CL	1078	1079	1080	1079	1080	Absent
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	
ATPase 6 (MURF4)	**CC**	6528-7123	6816-7420	6833-7432	6005-6638	5384-6013	4102-4788
	CL	596	605	600	634	630	687
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF1 (ND2)#	**CC**	8450-7125	8754-7429	8761-7436	7979-6654	7342-6017	6116-4791
	CL	1326	1326	1326	1326	1326	1326
	Start-Stop	UUG-UAA	GUG-UAA	UUG-UAA	AUG-UAA	UUG-UAA	AUG-UAA
G3*	**CC**	8446-8551	8756-8860	8757-8857	7975-8075	7343-7438	6120-6205
	CL	106	105	101	101	96	86
	Start-Stop	AUA-UAA	AUA-UAA	AUA-UAG	AUA-UAA	AUA-UAG	AUA-UAA
ND1#	**CC**	9489-8548	9798-8857	9802-8861	9013-8072	8389-7445	7156-6212
	CL	942	942	942	942	945	945

	Start-Stop	AUG-UAA	UUG-UAA	AUG-UAA	UUG-UAA	UUG-UAA	AUG-UAA
COII	**CC**	9496-10124	9804-10432	9842-10470	9014-9642	8392-9020	Gene
	CL	629	629	629	629	629	Absent
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAA	
MURF2	**CC**	10152-11190	10456-11502	10493-11538	9665-10710	9043-10088	7205-8250
	CL	1039	1047	1046	1046	1046	1046
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
COI#	**CC**	12830-11181	13142-11493	13178-11529	12350-10701	11728-10079	Gene
	CL	1650	1650	1650	1650	1650	Absent
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
G4*#	**CC**	13058-12871	13362-13206	13397-13252	12568-12410	11964-11783	8434-8279
	CL	188	157	146	159	182	156
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAG	AUG-UAA
ND4	**CC**	13152-14465	13479-14792	13525-14837	12690-14003	12098-13411	8547-9861
	CL	1314	1314	1314	1314	1314	1315
	Start-Stop	AUG-UAG	GUG-UAA	AUG-UAA	AUG-UAG	UUG-UAA	UUG-UAG
ND3 (G5)*#	**CC**	14695-14493	15022-14808	15078-14863	14184-14030	13664-13436	10088-9885
	CL	203	215	216	155	229	204
	Start-Stop	?-UAA	?-UAA	?-UAA	?-UAA	?-UAA	?-UAA
RPS12 (G6)	**CC**	14675-14854	15004-15180	15059-15215	14188-14340	13649-13804	10069-10230
	CL	180	177	157	153	156	162
	Start-Stop	AUG-UAA	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAA
ND5	**CC**	14924-16696	15244-17016	15298-17070	14401-16173	13819-15591	10259-10478^
	CL	1773	1773	1773	1773	1773	220^
	Start-Stop	AUG-UAA	AUG-UAA	UUG-UAG	AUG-UAA	AUG-UAA	AUG/?

Table 3.2B. Coding gene lengths with co-ordinates and start - stop codons for *A. dea*, *A. des*, *S. cul*, *S. gal* and *S. onc* (Genes ordered as per their location on the maxicircle)

GENES		*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*	NCS
12S*	**CC**	299-1439	113-1252	66-1207	151-1296	365-1505	
	CL	1141	1140	1142	1146	1141	
9S*	**CC**	1443-2044	1255-1852	1221-1822	1311-1909	1515-2113	
	CL	602	598	602	599	599	
ND8 (G1)*	**CC**	2127-2419	1938-2241	1914-2228	2001-2320	2204-2522	
	CL	293	304	315	320	319	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
ND9 (G2)*#	**CC**	2666-2444	2466-2235	2524-2272	2584-2348	2787-2547	
	CL	223	232	253	237	241	0
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAA	AUG-UAG	AUG-UAG	
MURF5*#	**CC**	2944-2714	2737-2471	2775-2485	2858-2355	3039-2797	
	CL	231	267	291	504	243	?
	Start-Stop	?-UAA	?-UAA	?-UAG	?-UAG	?-UAA	
ND7 (MURF3)	**CC**	2968-4109	2764-3903	2835-3974	2888-4028	3090-4229	
	CL	1142	1140	1140	1141	1140	10/11
	Start-Stop	AUG-UAA	AUU-UAG	AUU-UAA	AUU-UAA	AUU-UAA	
COIII	**CC**	4110-4976	3908-4774	3979-4845	4033-4899	4234-5100	
	CL	867	867	867	867	867	1/10
	Start-Stop	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
CYb	**CC**	4985-6064	4780-5859	4857-5936	4912-5991	5112-6191	
	CL	1080	1080	1080	1080	1080	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAA	
ATPase 6 (MURF4)	**CC**	6105-6787	5883-6567	6140-6818	6035-6713	6217-6895	
	CL	683	685	679	679	679	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
MURF1 (ND2)#	**CC**	8109-6784	7889-6564	8154-6829	8039-6714	8221-6896	
	CL	1326	1326	1326	1326	1326	4/11
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
G3*	**CC**	8098-8189	7878-7966	8150-8239	8035-8123	8217-8303	
	CL	92	89	90	89	87	100
	Start-Stop	AUA-UAG	AUA-UAG	AUA-UAA	AUA-UAA	AUA-UAG	
ND1#	**CC**	9190-8246	8912-7968	9186-8242	9068-8124	9248-8304	
	CL	945	945	945	945	945	4/11
	Start-Stop	AUG-UAG	UUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
COII	**CC**	9192-9820	8913-9541	9196-9824	9079-9707	9259-9887	
	CL	629	629	629	629	629	0
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAA	AUG-UAG	AUG-UAA	
MURF2	**CC**	9835-10880	9556-10601	9842-10887	9725-10770	9905-10950	

	CL	1046	1046	1046	1046	1046	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	
COI#	**CC**	12520-10871	12241-10592	12527-10878	12410-10761	12590-10941	
	CL	1650	1650	1650	1650	1650	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
G4*#	**CC**	12694-12542	12425-12277	12759-12591	12631-12468	12793-12641	
	CL	153	149	169	155	153	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	
ND4	**CC**	12796-14109	12519-13832	12863-14176	12728-14041	12889-14202	
	CL	1314	1314	1314	1314	1314	3/11
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
ND3 (G5)*#	**CC**	14282-14093	14015-13821	14404-14200	14251-14057	14412-14217	
	CL	190	195	205	195	196	?
	Start-Stop	?-UAA	?-UAA	?-UAG	?-UAG	?-UAA	
RPS12 (G6)	**CC**	14264-14395	13999-14130	14387-14516	14234-14361	14394-14533	
	CL	132	132	130	128	140	0
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	
ND5	**CC**	14404-16176	14136-15908	14530-16302	14370-16142	14544-16316	
	CL	1773	1773	1773	1773	1773	1/11
	Start-Stop	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	

CC = Co-ding Co-ordinates for the pre-edited gene in maxicircle; from start to stop codon.
CL = Unedited (or pre-edited) Coding Length of the gene; from start to stop codon.
*** Genes whose coding co-ordinates are approximate.**
 # Genes transcribed from the strand opposite to the ribosomal RNA genes.
? Codon uncharacterized / uncertain.
^ = truncated gene due to incomplete genome.
NCS = Frequency that a gene starts with a non-canonical start codon (not AUG).

The nucleotide sequence identities among the aligned syntenic 20-gene region (12S to ND5, with the intergenic regions) of the maxicircles generally followed expected patterns; i.e., organisms whose nuclear genes are more closely related, exhibited greater identities (**Table 3.3**). *L. tar* and *E. sch* showed the greatest identity between different genera (84%). Species of the same genera also exhibited high identities (76-86%). *P. ser* showed the least identity (68-72%) to all the other genomes. For comparison, we computed the identities of two nuclear genes, the SSU rDNA and the protein encoding GAPDH gene, for all the organisms. The SSU rDNA gene was between 88-99% identical and the GAPDH gene was between 79-94% identical. The nuclear genes also

showed similar patterns of identities as seen for the maxicircles, except that at the nuclear level, *H. mus* was the least identical to others (**Supplementary Table 3.2**). Although the nuclear ribosomal gene manifested higher identities between the organisms, the nuclear protein encoding GAPDH gene showed identities more comparable to the maxicircle syntenic region. However, note that both the nuclear genes used for comparison are discrete genes (no intergenic regions) and also do not undergo editing. Identities for each individual maxicircle gene is described below.

Table 3.3. Percent identity matrix of the syntenic 20-gene region of maxicircle genomes

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser**	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	84	**100**									
L. cos	79	80	**100**								
C. aca	80	81	78	**100**							
H. mus	76	78	75	77	**100**						
P. ser*	71	72	69	70	71	**100**					
A. dea	74	74	73	74	73	68	**100**				
A. des	74	74	73	74	73	68	76	**100**			
S. cul	76	77	75	77	77	71	76	76	**100**		
S. gal	77	77	75	77	77	72	76	77	84	**100**	
S. onc	77	77	75	77	77	71	76	76	84	86	**100**

The matrix above shows the percent identities of the syntenic 20-gene region of maxicircles (starting at 12S and ending at ND5; including the intergenic regions) between all the organisms studied. Identities within the members of the leishmaninae family are shaded green, while those within the SHTs are shaded yellow and identities of *P. ser* maxicircle to others are shown in gray.
*** indicates that the identities of *P. ser* maxicircle was obtained after elimination of the 4 absent genes (CYb, COI, COII and COIII) in all the other maxicircles.**

RNA editing patterns show distinct similarities and differences

We discuss the maxicircle genes in three groups: 1) genes that are not edited; 2) genes that are partially (5' or internally) edited; and 3) genes that are pan-edited.

Genes not known to undergo editing:

The 12S and 9S rRNA genes, COI, MURF1 (ND2), ND1, ND4 and ND5 are not edited in *L. tar* or *P. ser* and are apparently unedited in all of the organisms reported in this study. The characteristics of these genes are described briefly below:

12S and 9S mitochondrial rRNA genes: The 12S rRNA gene ranges in lengths between 1140-1170 nts and the 9S rRNA gene ranges between 598-612 nts in length. GeneDoc nucleotide alignments of 12S and 9S genes from all the organisms have been shown in the **Supporting Information for Chapter 3** (following this chapter), **3.SI-S1a** and **3.SI-S2a**. Their approximate co-ordinates have been assigned based on sequence homology in the beginning and the end of these genes with that of the published *L. tar* 12S and 9S genes. The published *P. ser* 12S gene starts at position 217 (GenBank: AF079967.2) but our alignment suggests that its 12S may start six bases upstream, at the 211th position. The 12S gene ranges between 76-93% in nucleotide identity (SI-S1b); and the 9S gene is 74-94% identical among the organisms studied (SI-S2b). The 12S and 9S sequences of all the organisms fold into the canonical rRNA secondary structures as previously reported for *L. tar* (de la Cruz et al., 1985a) (de la Cruz et al., 1985b) (**Supplementary Figures 3.2A through 3.2K** and **3.3A through 3.3K**, for 12S and 9S, respectively).

COI: The COI gene is transcribed on the strand opposite to the ribosomal genes. Its coding sequence is 1650 nts in length, apparently unedited and well conserved in all the organisms

(**3.SI-S3a** and **3.SI-S3b**). This gene was reported to be absent in *P. ser* (Nawathean and Maslov, 2000) and is consequently not found in the published maxicircle sequence (GenBank: AF079967.2). The start codon is AUG and the stop codon is UAA in all the organisms. The nucleotide sequences are 76-88% identical (**3.SI-S3c**) and the proteins are 85-100% identical (**3.SI-S3d**).

MURF1 / ND2: MURF1, also transcribed on the strand opposite to the ribosomal genes, was reported to encode ND2 (Kannan and Burger, 2008) and its coding sequence is 1326 nts in length. It is unedited in the published *L. tar*, *Lam* (Yatawara et al., 2008) and *P. ser* maxicircles (GenBank: AF079967.2). Our alignments confirm that the gene is unedited in all of the organisms we studied (see **3.SI-S4a** and **3.SI-S4b**). Based on homology, our alignments confirm that the (previously uncharacterized) putative start codon of *L. tar* is UUG, corroborating earlier observations (Maslov, 2010). Also, most trypanosomatids use AUG as the start codon for this gene, but *L. tar, L. cos* and *H. mus* apparently use UUG, and *E. sch* apparently uses GUG as start codon. The nucleotide sequences range between 66-83% in identity while the proteins are 64-98% identical (**3.SI-S4c** and **3.SI-S4d**).

ND1: ND1 is also transcribed on the strand opposite to the ribosomal genes. The ND1 gene transcript from *L. tar* (GenBank: M10126.1) is unedited and 942 bps from start to stop codon. However, the ND1 coding length in *P. ser* (GenBank: AF079967.2) is 945 bps despite being unedited and contains an additional AA residue at the third position. Based on our ND1 nucleotide alignment (**3.SI-S5a**), we observed that the coding sequence is 942 bps in *L. tar*, *C. aca*, *L. cos* and *E. sch* and 945 bps in all others. This difference is again attributable to the absence of the aforementioned third AA residue in the ND1 protein of these four organisms, which is present in the others (**3.SI-S5b**). Editing is not suspected since the remainder of the

protein is conserved and the protein sequences are between 61-96% identical. The nucleotide sequences are 65-83% identical (**3.SI-S5c** and **3.SI-S5d**). The start codon for ND1 is either AUG or UUG and the stop codon is predominantly UAA with UAG in just *A. dea*.

ND4: ND4 is one of the only two unedited genes that are transcribed from the same strand as the ribosomal genes in these organisms and its coding sequence is 1314 nts in length. Our ND4 gene and protein alignments (**3.SI-S6a** and **3.SI-S6b**) show that this gene remains unedited in all the organisms we report. *P. ser*'s ND4 gene (1315 nts in length) places an additional T residue at position 70 of the gene (represented as an asterisk in S6a) which is absent in the others. This indel leads to a truncation at the translated protein's (GenBank: AAF26929.1) 5' end. Eliminating this indel prevents this 5' truncation and creates a protein that is more identical to the other ND4 proteins (**3.SI-S6b**). Since this gene is unedited in all the other strains we have studied, and the additional T residue is also absent in the ND4 of the *Phytomonas* JMA strain, we believe that it is more likely that the additional T in the *P. ser* sequence stems from a sequencing error (we are unable to verify this because we did not sequence this organism and do not have its DNA). The start codon for this gene is AUG in most strains, but UUG is apparently used by *H. mus* and *P. ser* and GUG is used by *E. sch*. The stop codon is most frequently UAA, but occasionally UAG. The nucleotide sequences for ND4 range between 30-86% in identities (**3.SI-S6c**) and the resulting proteins (with 5' non-truncated *P. ser*) are between 68-97% identical (**3.SI-S6d**).

ND5: This gene is transcribed from the same strand as the ribosomal genes. The ND5 coding sequence is 1773 nts in length in these organisms, although as mentioned above, it is incomplete and measures only 220 nts in *P. ser* maxicircle. Our ND5 gene alignment (**3.SI-S7a**) shows that the gene remains unedited in all the organisms. The published partial protein for *P. ser* ND5

(GenBank: AAF26928.1) places an additional Phe (F) residue after the start codon (**3.SI-S7b**). Since Phe is encoded by "UUU" codon, it is possible that this codon is either edited out of *P. ser* transcript, or edited into the ND5 transcripts of other organisms. However, its absence in the maxicircle of all the other organisms, including that of the published *L. tar* suggests that the extra Phe residue may not be essential for the protein and therefore, we do not suspect the gene to be edited. The start codon is AUG in all organisms, except *L. cos*, where it is UUG, and the stop codon is UAA or UAG. Nucleotide alignments are between 59-87% identical (**3.SI-S7c**) and the proteins are between 40-97% identical (**3.SI-S7d**).

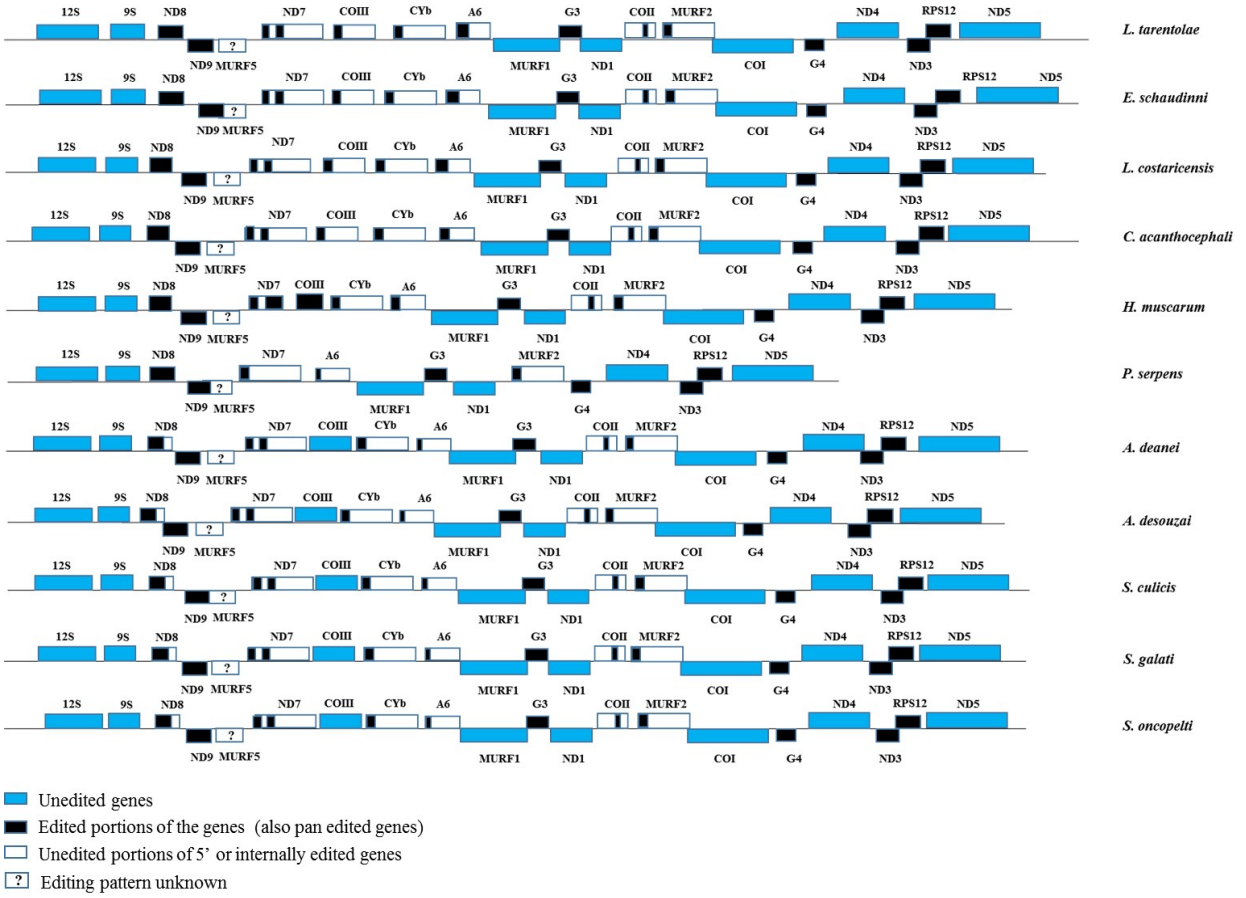


Fig 3.1: Syntenic 20-gene region of the trypanosomatid maxicircle genomes: The maxicircle genomes of all the organisms reported herein are shown along with the inferred editing patterns of the genes. Genes above the line are transcribed in the same direction as the ribosomal RNA genes and the ones underneath are transcribed in the opposite direction. The figure is not drawn to scale.

Genes that undergo partial (5' or internal) editing:

The genes A6, COII, COIII, CYb, MURF2 and ND7 have been reported to undergo editing in a fraction of their sequences. Although editing pattern(s) of the MURF5 gene have not been established, we believe that the MURF5 transcript is edited and have included it in this group (see below).

ATPase 6 / MURF4: The A6 transcript is transcribed from the same strand as the ribosomal genes and was reported to be 5' edited in *L. tar* and *T. bru* (Bhat et al., 1990) (Bhat et al., 1991). Alignment of the published A6 transcripts of *L. tar* (GenBank: M62821.1), *P. ser* (Maslov et al., 1999), *H. mus* and *S. cul* (Maslov et al., 1994a) with the pre-edited A6 sequences of the organisms characterized in this report show that the 5' ends of these transcripts are edited at varying levels to create mature mRNAs with coding sequences ranging from 699-711 nts (**3.SI-S8a**). A6 transcripts of *L. cos* and *E. sch* show editing patterns similar to that of the *L. tar* transcript, whereas the SHTs require significantly less editing. The *P. ser* A6 transcript requires the least editing. *H. mus* and *C. aca* are edited at levels intermediate between the *Leishmania* group and the SHTs. Our results corroborate the review that referred to the gradient A6 5' editing in the *Leishmania-Crithidia* clade (Simpson et al., 2000) and the editing extents correspond to the earlier report (Maslov et al., 1994a). The translation start and stop codons for A6 transcripts are universally AUG and UAA respectively. Their protein alignment is shown in **3.SI-S8b**. Gaps in the alignments identify codons that are absent in the transcripts of the other organisms. Post editing, the coding lengths of the transcripts are between 66-91% identical (**3.SI-S8c**) and the encoded proteins are 55-97% identical (**3.SI-S8d**).

COII: The COII transcript is also transcribed from the same strand as the ribosomal genes and was reported to be internally edited by addition of 4 Us to correct a frameshift in *L. tar* (Shaw et

al., 1989). Alignment of the pre-edited COII sequences of the organisms we report to the edited *L. tar* transcript [(GenBank: L07544.1); (Cruz et al., 1984)] shows that there are frameshifts in all of them that are corrected by the internal addition of four Us at the same locus as in *L. tar* transcript (**3.SI-S9a**). The final coding lengths of all the transcripts are 633 nts and their protein alignment is shown in **3.SI-S9b**. The start codon is universally AUG and the stop codons are UAG or UAA. The inferred coding lengths of the mature transcripts are between 77-90% identical (**3.SI-S9c**) while the translated proteins are 87-99% identical (**3.SI-S9d**).

COIII: COIII is also transcribed from the same strand as the ribosomal genes. Previous reports have shown the COIII transcript to be unedited in *S. cul* (Maslov et al., 1994a) but pan-edited in *H. mus* (Landweber and Gilbert, 1993) and the gene to be absent in *P. ser* (Maslov et al., 1999). Aligning the pre-edited COIII sequences of our organisms with the published fully-edited transcripts of *L. tar* (GenBank: L07539.1) (Cruz et al., 1984) (Shaw et al., 1988), *S. cul* and *H. mus* show that the COIII transcript remains unedited in the SHTs but is edited on its 5' end in *L. cos*, *E. sch* and *C. aca* (**3.SI-S10a**). The final coding length of the fully-edited COIII transcript is 864 or 867 nts. The start codon is universally AUG (UUG for *H. mus*, [published]) and the stop codon is UAA (UAG in *A. dea*). Their protein alignment is shown in **3.SI-S10b**. The mature coding lengths of COIII transcripts are between 71-90% identical (**3.SI-S10c**) and the proteins are between 69-97% identical (**3.SI-S10d**).

CYb: CYb is also transcribed from the same strand as the ribosomal genes. The published 5' end of the *L. tar* CYb transcript (GenBank: L07542.1); (Cruz et al., 1984) (Shaw et al., 1988), was used as a reference to align the pre-edited sequences of the organisms in this report. Our alignment indicates that all the transcripts undergo identical editing at their 5' ends to produce a translatable coding sequence of 1116 nucleotides (**3.SI-S11a**). Note that CYb is absent in *P. ser*

(Maslov et al., 1999). *In silico* translation of the inferred mature transcripts produced the proteins (**3.SI-S11b**) matching with previously published protein sequences (Horváth et al., 2000). The start codon is AUG and stop codon is predominantly UAA and less frequently UAG. The coding lengths of CYb fully edited transcripts range between 80-89% in identity (**3.SI-S11c**) and its protein is probably the most conserved protein with identities ranging between 91-99% (**3.SI-S11d**).

MURF2: MURF2 transcript is transcribed from the same strand as the ribosomal genes and is reportedly 5' edited in *L. tar* (Cruz et al., 1984) (Shaw et al., 1988) (GenBank: L07545.1) and *P. ser* (Nawathean and Maslov, 2000). Alignment of the pre-edited sequences of our organisms with these two published sequences, show that, like the CYb transcripts above, the MURF2 transcripts of all the organisms undergo an identical pattern of 5' editing (**3.SI-S12a**). The MURF2 protein alignment (**3.SI-S12b**) indicates that the AA residues at coding positions 110 and 111 that are present in all other organisms studied, are absent in the published *L. tar* transcript, thereby rendering the length of MURF2 protein to be 357 AAs in all the organisms rather than the 355 AAs in *L. tar*. Also, the *L. tar* MURF2 transcript shows an additional "C" at coding position 343, which we compensated for by placing a 'U' in our alignment to correct the resultant frameshift, and counted that U in **Table 3.4A**. AUG is the universal start codon for all the MURF2 transcripts and UAA is the stop codon (only *S. gal* uses UAG). Coding sequences of mature transcripts are between 64-88% identical and the translated proteins are 59-92% identical among the organisms studied (**3.SI-S12c** and **3.SI-S12d**).

ND7: ND7 gene is also transcribed from the same strand as the ribosomal genes. We aligned the pre-edited ND7 sequences of the organisms reported here with the published fully-edited transcripts of *L. tar* (GenBank: L07537.1) (Cruz et al., 1984) (Shaw et al., 1988) (Shaw et al.,

1989) and *P. ser* (Maslov et al., 1999). The alignment indicates that the ND7 transcripts from all the organisms, except *P. ser* and *H. mus*, are edited in two separate blocks at their 5' end. This result includes *S. cul*, for which only one of these two blocks was previously identified (Maslov et al., 1994a). ND7 of *P. ser* is edited in only one of the two blocks at the 5' end; and the ND7 of *H. mus* is not only edited in the first block, but also through its entire length, beginning at the second block, making it partially pan-edited (**3.SI-S13a** and **Fig 3.1**). The coding lengths of the completely edited, mature transcripts of all the organisms are 1164 nts and the alignment of the translated proteins are shown in **3.SI-S13b**. AUU is apparently the most common, non-canonical start codon for this gene [as reported earlier for *L. tar* and *P. ser* (Maslov et al., 1999)]. However, *A. dea* alone has an AUG start codon. The stop codon is predominantly UAA and less frequently UAG. The coding sequences from fully-edited transcripts are between 75-87% identical and the translated proteins are highly conserved, ranging between 80-100% in their identities (**3.SI-S13c** and **3.SI-S13d**).

MURF5: The MURF5 gene is transcribed on the strand opposite to the ribosomal genes. Editing of the MURF5 gene is unclear from the previously published maxicircles of *L. tar* and *P. ser*. It was postulated that the MURF5 start codon in *L. tar* is either the non-canonical AUA (Cruz et al., 1984) or UCU in an extended amino acid chain (GenBank: M10126.1); UUG in *L. mexicana amazonenesis (Lam)* (Maslov, 2010) and AUG in *P. ser* (GenBank: AF079967.2), although MURF5 transcripts were not detected in Northern blots from *P. ser* (Maslov et al., 1999). The MURF5 nucleotide and protein alignments of the organisms in this report (**3.SI-S14a** and **3.SI-S14b**) show a small internal conserved motif that encodes the conserved block previously observed in *Lam, L. tar, P. ser, T. bru* and *L. donovani* (Maslov, 2010). However, the regions flanking this internal conserved motif show minimal homologies. The 3' ends of all the putative

MURF5 transcripts have an in-frame UAA or UAG stop codons, at variable downstream lengths. The MURF5 transcript of *P. ser* (GenBank: AF079967.2) also has a previously unreported stop codon. It was also postulated earlier (Maslov, 2010), that polyadenylation of the unedited mRNA can create an in-frame stop codon for all these transcripts. However, no consensus in-frame start codon could be identified in the MURF5 genes we characterized. Therefore, we believe that the MURF5 transcript is extensively 5' and partially 3' edited to create start and stop codons and its co-ordinates in **Tables 3.2A and 3.2B** are putative. The overall identities of the pre-edited transcripts are between 44-72% (**3.SI-S14c**) and the in-frame proteins are between 21-60% identical (**3.SI-S14d**).

Table 3.4A. Summary of the inferred editing patterns in genes (*L. tar*, *E. sch*, *L. cos*, *C. aca*, *H. mus*, *P. ser*)

GENES		*L. tar*	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*
ND8 (G1)	U+[U-]	210[38]	213[33]	207[30]	211[22]	220[27]	229[14]
	FCTL	438	438	438	438	438	438
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND9 (G2)	U+[U-]	324[42]	324[27]	320[11]	338[8]	326[34]	337[25]
	FCTL	573	573	573	573	576	576
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
MURF5	PEP	5’	5’	5’	5’	5’	5’
ND7 (MURF3)	U+[U-]	23[0]	20[0]	23[0]	24[0]	476[28]	22[0]
	FCTL	1164	1164	1164	1164	1164	1164
	PEP	5'	5'	5'	5'	partially pan-edited	5'
COIII	U+[U-]	27[15]	24[18]	28[8]	26[7]	517[13]	Gene absent
	FCTL	864	864	864	864	867	
	PEP	5'	5'	5'	5'	pan-edited	NA
CYb	U+[U-]	38[0]	37[0]	36[0]	37[0]	36[0]	Gene absent
	FCTL	1116	1116	1116	1116	1116	
	PEP	5'	5'	5'	5'	5'	NA
ATPase 6 (MURF4)	U+[U-]	106[3]	102[8]	105[6]	65[0]	81[0]	24[0]
	FCTL	699	699	699	699	711	711
	PEP	5'	5'	5'	5'	5'	5'
G3	U+[U-]	134[6]	135[6]	137[4]	137[4]	147[9]	152[1]
	FCTL	234	234	234	234	234	237
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COII	U+[U-]	4[0]	4[0]	4[0]	4[0]	4[0]	Gene absent
	FCTL	633	633	633	633	633	
	PEP	internal	internal	internal	internal	internal	NA
MURF2	U+[U-]	29[0]	27[0]	28[0]	28[0]	28[0]	28[0]
	FCTL	1068	1074	1074	1074	1074	1074
	PEP	5'	5'	5'	5'	5'	5'
G4	U+[U-]	327[5]	290[3]	298[0]	290[5]	273[11]	293[5]
	FCTL	510	444	444	444	444	444
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND3	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
RPS12 (G6)	U+[U-]	110[32]	102[30]	105[10]	120[12]	105[9]	121[13]
	FCTL	258	249	252	261	252	270

	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
TOTAL	**U+**	**1332**	**1278**	**1291**	**1280**	**2213**	**1206**
	U-	**141**	**125**	**69**	**58**	**131**	**58**

Table 3.4B. Summary of the inferred editing patterns in genes (*A. dea*, *A. des*, *S. cul*, *S. gal*, *S. onc*)

GENES		*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
ND8 (G1)	U+[U-]	146[1]	137[3]	132[9]	128[10]	133[14]
	FCTL	438	438	438	438	438
	PEP	5'	5'	5'	5'	5'
ND9 (G2)	U+[U-]	359[6]	348[4]	339[16]	347[8]	339[4]
	FCTL	576	576	576	576	576
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
MURF5	PEP	5’	5’	5’	5’	5’
ND7 (MURF3)	U+[U-]	22[0]	24[0]	24[0]	23[0]	24[0]
	FCTL	1164	1164	1164	1164	1164
	PEP	5'	5'	5'	5'	5'
COIII	U+[U-]	0[0]	0[0]	0[0]	0[0]	0[0]
	FCTL	867	867	867	867	867
	PEP	unedited	unedited	unedited	unedited	unedited
CYb	U+[U-]	36[0]	36[0]	36[0]	36[0]	36[0]
	FCTL	1116	1116	1116	1116	1116
	PEP	5'	5'	5'	5'	5'
ATPase 6 (MURF4)	U+[U-]	21[2]	17[0]	23[0]	23[0]	23[0]
	FCTL	702	702	702	702	702
	PEP	5'	5'	5'	5'	5'
G3	U+[U-]	143[1]	145[0]	146[2]	145[0]	147[0]
	FCTL	234	234	234	234	234
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COII	U+[U-]	4[0]	4[0]	4[0]	4[0]	4[0]
	FCTL	633	633	633	633	633
	PEP	internal	internal	internal	internal	internal
MURF2	U+[U-]	28[0]	28[0]	28[0]	28[0]	28[0]
	FCTL	1074	1074	1074	1074	1074
	PEP	5'	5'	5'	5'	5'
G4	U+[U-]	295[4]	297[2]	277[2]	292[3]	294[3]
	FCTL	444	444	444	444	444
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND3	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
RPS12 (G6)	U+[U-]	120[3]	120[3]	123[1]	125[1]	118[6]
	FCTL	249	249	252	252	252

	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
TOTAL	**U+**	**1174**	**1156**	**1132**	**1151**	**1123**
	U-	**15**	**12**	**30**	**22**	**27**

U+ = Number of Us inserted in the coding region (i.e., start to stop codon).
U- = Number of Us deleted in the coding region (i.e., start to stop codon).
FCTL = Final Coding Transcript Length (start codon to stop codon).
PEP = Predicted Editing Pattern.

Genes that undergo extensive (pan) editing:

The transcripts of the genes ND8, ND9, RPS12, ND3 (G5), G3 and G4, have been previously reported to undergo editing across their entire lengths (pan-editing). Below, we describe the editing patterns predicted for those genes, in the organisms characterized herein:

ND8 / G1: This gene is transcribed from the same strand as the ribosomal genes. Early studies of the G1 transcript of *T. bru* found it to be pan-edited and to encode ND8 (Souza et al., 1992). Subsequently, the gene was extensively studied in many other trypanosomatids. Aligning the pre-edited ND8 sequences of the organisms reported herein with the published, fully-edited transcripts of *T. bru* (GenBank: M63820.1), *P. ser* (Nawathean and Maslov, 2000), *Lam* (Maslov, 2010), *L. tar* (GenBank: Z37535.1) (Thiemann et al., 1994), and *A. dea* and *S. onc* (Gerasimov et al., 2012a) indicate that the ND8 transcript needs extensive editing in *C. aca*, *L. cos*, *H. mus* and *E. sch*, corroborating the patterns of editing in *L. tar*, *Lam*, *P. ser* and *T. bru*. However, it was also reported that ND8 was not pan-edited in *S. onc* (Aravin et al., 1998). Our alignments re-confirm that the transcript is only 5' edited (more than half of its length) among the members of the SHTs (**3.SI-S15a** and **Fig 3.1**). The inferred protein alignment is shown in **3.SI-S15b**. The final coding lengths of the fully edited transcripts are 438 nts. Start codon is universally AUG and stop codons are UAA or UAG. The inferred mature coding sequences are between 76-95% identical (**3.SI-S15c**) and the proteins are between 66-97% identical (**3.SI-S15d**).

ND9 / G2: This gene is transcribed on the strand opposite to the ribosomal genes. Early studies of the G2 transcript in *T. bru* established it to be pan-edited (GenBank: L05586.1) and to encode ND9 (Souza et al., 1993). Following this, the ND9 transcripts of *L. tar* (GenBank: Z37536.1) and *Lam* were published (Maslov, 2010) (Thiemann et al., 1994). We used the ND9 transcripts from

these three organisms as references to align the pre-edited sequences of the trypanosomatids in this study. The transcript is extensively edited in all the organisms (**3.SI-S16a**) and the inferred proteins are shown in **3.SI-S16b**. The start codon is universally AUG and the stop codon was either UAA or UAG. The inferred mature coding sequences are between 83-94% identical and the translated proteins are 74-92% identical (**3.SI-S16c** and **3.SI-S16d**).

RPS12 / G6: The RPS12 gene is also transcribed from the same strand as the ribosomal genes. The G6 transcript of *L. tar* was reported to be pan-edited and to encode RPS12 (Maslov et al., 1992). Alignment of the pre-edited RPS12 sequences of the organisms reported herein with the published, fully edited sequences of *L. tar* (GenBank: L07546.1) (Cruz et al., 1984), *P. ser* (GenBank: AF034626.1) (Maslov et al., 1998), *Lam* (Maslov, 2010) and the putative fully edited RPS12 from *S. onc* (Maslov et al., 1992) indicate that all the RPS12 transcripts are pan-edited (**3.SI-S17a**), producing the putative proteins aligned in **3.SI-S17b**. The protein lengths range between 82-89 AAs, due to additional AA-insertions in some transcripts. The start codon is universally AUG and the stop codon is predominantly UAG and less frequently UAA. The mature coding sequences are about 58-96% identical (**3.SI-S17c**) and the translated proteins are 39-91% identical (**3.SI-S17d**).

ND3 / G5: This transcript is transcribed on the strand opposite to the ribosomal genes. Editing of the G5 transcript was initially studied in *T. bru* and the transcript was established to be pan-edited and to encode ND3 (Read et al., 1994). Subsequently, G5 editing was described in *L. tar* (GenBank: Z37539.1) (Thiemann et al., 1994) (Maslov et al., 1994b). Upon aligning the pre-edited ND3 sequences of the organisms in this report with the published edited ND3 sequence from *Lam* (Maslov, 2010), *T. bru* and *L. tar*, it is seen that the editing pattern varies widely even within the *Leishmania* group. However, the presence of an internal conserved motif helped

establish the presence of the gene in all the maxicircles (**3.SI-S18a**). The editing of *H. mus* and *P. ser* ND3 transcripts resembles that of *T. bru* while the ND3 editing in others resembles the ND3 transcript of *L. tar*. An accurate start position could not be assigned due to the lack of RNA sequence at the 5' end of the transcript, but a stop codon could be inferred putatively, based on the sequences of the published references. An extra U-residue (shown as "^" in our nucleotide alignment against a turquoise background) that shifted the frame in *L. tar* mRNA was described previously (Maslov, 2010). Deleting that residue in *L. tar* yields a transcript that is more identical to the *Lam* mRNA (**3.SI-S18b**), which we believe is the most likely product. Overall, the ND3 transcript is pan-edited and the most common stop codon is UAA. Identities were ascertained only for the conserved motif and this motif in the inferred, mature transcripts is between 61-92% identical (**3.SI-S18c**) and the translated partial proteins (conserved motif) are 37-85% identical (**3.SI-S18d**).

G3: Editing of the G3 transcript, which is transcribed from the same strand as the ribosomal genes, is not well characterized or established. The G3 transcript of *L. tar* was previously reported to be partially pan-edited and misedited in its 5' end, possibly due to loss of the 5' terminal gRNAs during its culture. (Thiemann et al., 1994). Therefore, we aligned the pre-edited G3 sequences of the organisms in this report with the published RNAs of *Lam* and *P. ser* (Maslov, 2010) and observed that the G3 transcript is pan-edited in all the organisms, with the putative fully-edited, mature coding sequences measuring either 234 or 237 nts in length (**3.SI-S19a**). The inferred G3 protein alignment is shown in **3.SI-S19b**. AUA is the universal non-canonical start codon and either UAA or UAG acts as the stop codon. The mature coding sequences are between 74-96% identical (**3.SI-S19c**) and the translated proteins are 49-95% identical (**3.SI-S19d**).

G4: Like for G3, not much editing information is available for G4, which is also transcribed on the strand opposite to the ribosomal genes. The published, fully-edited G4 transcript of *L. tar* (Thiemann et al., 1994) and *Lam* and *T. bru* (Maslov, 2010) were used as references for aligning the pre-edited G4 sequences of organisms herein. The putative mature transcript is shown in **3.SI-S20a** and the inferred protein alignment is shown in **3.SI-S20b**. The overall lengths of the coding sequences are 444 nts (except *L. tar* and *T. bru* that were published to be 510 and 441, respectively). AUG is the universal start codon, and stop codon is either UAA or UAG. The *L. tar's* mature G4 transcript appears out of frame for about a third of its total length (starting at the 5' end) when compared to the published transcripts of *Lam* or even *T. bru*. Consequently, based on nucleotide homology, the predicted editing patterns of the G4 transcripts of all the organisms in this report are more similar to the patterns of *Lam* and *T. bru* G4 transcripts, rather than to *L. tar* and this phenomenon is clearly reflected in their relative identities. The mature G4 coding sequences are between 69-94% identical (**3.SI-S20c**) while the inferred proteins are 39-90% identical (**3.SI-S20d**); the lowest identity values in both cases being the G4 of *L. tar* to that of the others.

Phylogenetic relationships and the evolution of RNA editing

Maximum Likelihood analyses were conducted as described in the Materials and Methods section, using a catenated subset of the unedited translated maxicircle genes; and nuclear genes for the SSU rRNA and GAPDH from these organisms. Not surprisingly, the pattern we observed from the nuclear genes (**Fig 3.2A**) resembles that of the previous analyses of nuclear genes from several reports (Teixeira et al., 2011) (Votýpka et al., 2012) (Yurchenko et al., 2009) (Hamilton et al., 2004) (Hughes and Piontkivska, 2003a) (Hughes and Piontkivska, 2003b). Similarly, the analysis with the catenated subset of the unedited maxicircle genes (**Fig 3.2B**) generated a nearly identical profile.

In parallel, we used a binary matrix generated to define the editing patterns in each of these organisms and to quantify the relationships among these organism with regards to the evolution of RNA editing. Thus, we anticipate that a ML analysis on this binary dataset would group the organisms based on their editing patterns and in effect, predict the evolution of RNA editing among these trypanosomatids. Interestingly, the phylogenetic tree generated using the binary editing dataset (**Fig 3.2C**) also closely reflects the relationships based on the maxicircle subset and the nuclear genes. A minor difference is observed in the topologies of *P. ser* and *H. mus* when compared to the maxicircle and the nuclear genes-based trees. According to the RNA editing binary tree, the RNA editing pattern in *H. mus* is more similar to the *Leishmania* group and intermediate between *P. ser* and the *Leishmania* group (excluding the editing in ND7 and COIII genes). But overall, these observations suggest that RNA editing has not only evolved to be similar in closely related organisms, but also has, in fact, evolved in parallel with the nuclear genes.

The one most important characteristic that stood out distinctly among all the three trees was the unambiguous grouping of the various trypanosomatids in this report: the SHTs group together (blue brace brackets), as do the slow-evolving *Leishmania* group (consisting of *L. tar, E. sch, L. cos* and *C. aca*; orange brace brackets); and the *Herpetomonas-Phytomonas* pair (black brace brackets), again reflecting their evolutionary relationships.

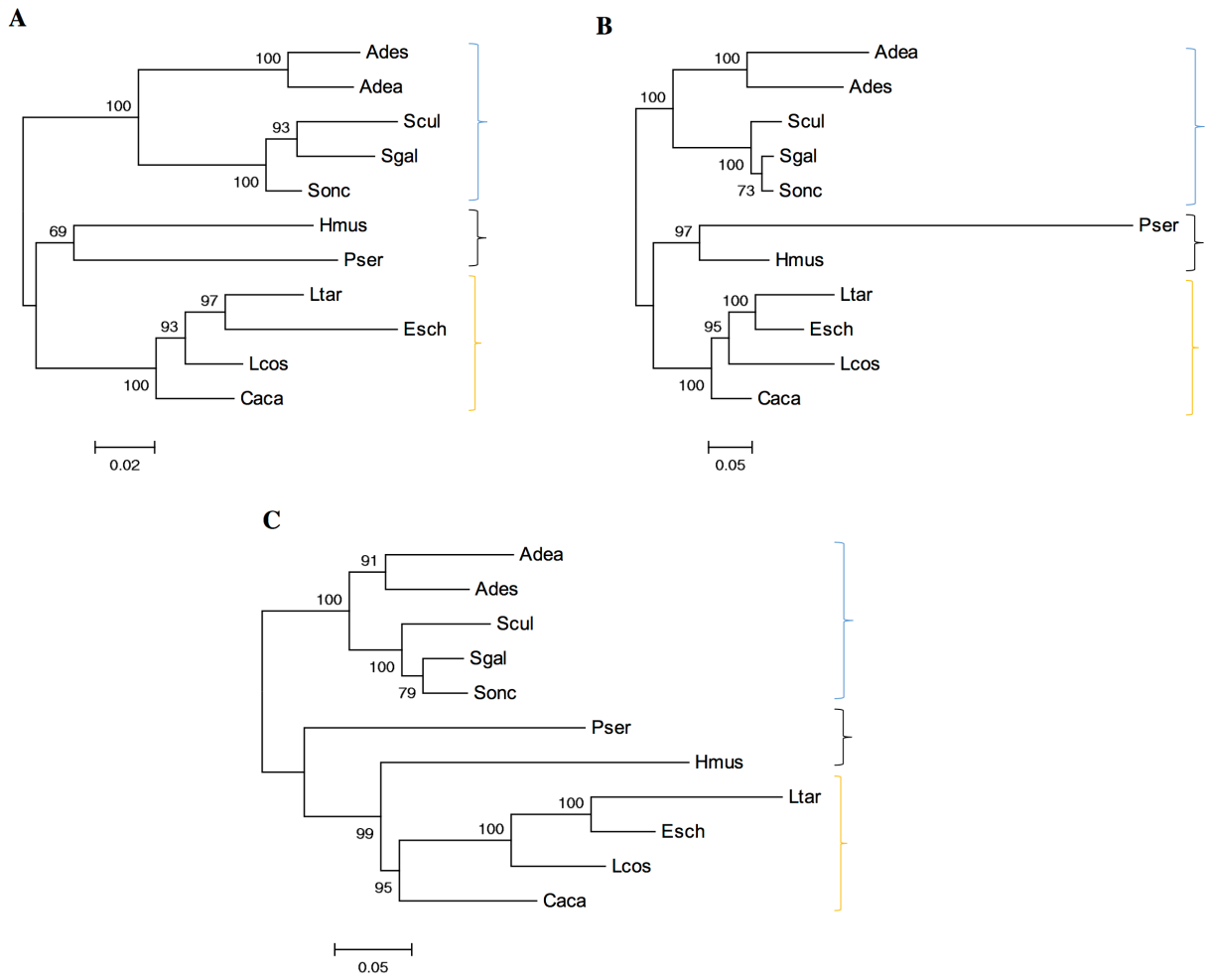


Fig 3.2. Maximum likelihood phylogenies depicting the relationships between the various trypanosomatids in this report. (A) Phylogenetic tree based on the nuclear dataset (3055 characters). **(B)** Phylogenetic tree based on the maxicircle dataset (2336 characters). **(C)** Phylogenetic tree based on the RNA editing quantification (binary) dataset (894 characters). Bootstrap values from 1000 pseudoreplicates are shown on the nodes. The scale bar represents the number of substitutions per site.

[Adea = *A. deanei*; Ades = *A. desouzai*; Caca = *C. acanthocephali*; Esch = *E. schaudinni*; Hmus = *H. muscarum*; Lcos = *L. costaricensis*; Ltar = *L. tarentolae*; Pser = *P. serpens*, Scul = *S. culicis*; Sgal = *S. galati*; Sonc = *S. oncopelti*]

Discussion

We report the syntenic maxicircle sequences of various insect-infecting and endosymbiont-bearing trypanosomatids and compare these with the previously reported maxicircle genomes of *Leishmania tarentolae* and *Phytomonas serpens*.

The maxicircle genomes showed the presence of the 18 protein coding genes and 2 ribosomal RNA genes with their transcriptional polarities as those reported earlier for *L. tar*, *P. ser*, *T. bru* etc. In general, the protozoan mitochondrial genomes, and particularly the trypanosomatid mitochondrial genomes have been reported to have some of the lowest GC contents (Shioiri and Takahata, 2001). As one can expect, the GC content was higher in organisms with relatively smaller maxicircle sizes and smaller intergenic lengths and correspondingly higher transcriptome to genome ratio (**Table 3.1**). Reportedly, the AT skew in the trypanosomatid mitochondrial genome is primarily due to the non-coding and divergent regions of the maxicircles rather than the coding regions themselves (Westenberger et al., 2006).

Further, the 20-gene syntenic regions of the maxicircles themselves shared higher identities (%) among the species of the same genera like *A. dea* to *A. des* and *S. cul* to *S. gal* and *S. onc*. Similarly, members of the *Leishmania* group (*L.tar, E. sch, L. cos* and *C. aca*) showed higher identities among each other (**Table 3.3).** It is noteworthy that maxicircles of members of the *Leishmania* group (belonging to different genera) share higher identities among each other than some members of the SHTs, e.g., *A. dea* and *A. des*. Post editing, the identities of both the inferred mature coding sequences as well as the *in silico* translated proteins of all the edited genes were comparable to that of their unedited counterparts. The inferred transcripts and

proteins of COI, CYb and ND9 (among the unedited, partially edited and pan-edited groups, respectively) shared the highest identities among each other.

Among the partially edited genes with clearly discernable editing patterns, the COII transcript is internally edited and the CYb and MURF2 transcripts are edited identically in all the organisms herein. However, the transcripts of A6, ND7 and COIII showed some striking variability in their editing patterns. While A6 showed a decreasing gradient of editing at the 5' end (**Fig 3.1**), the ND7 and COIII transcripts were pan edited in *H. mus* with a gradual reduction in their edited domains to being 5' or even unedited (COIII in SHTs) in others. Similarly, among the pan-edited genes, the ND8 transcript showed decreased editing in the SHTs while being pan edited in others. Overall, the SHTs exhibited the least editing of all the organisms reported herein, both in terms of the number of U insertions and deletions, as also by virtue of having the highest numbers of unedited genes. On the other end of the editing spectrum was *H. mus*, whose transcripts are the most edited, with the highest numbers of U insertions and deletions and it also has the highest number of pan-edited genes (**Tables 3.4A and 3.4B** and **Table 3.1**). Therefore, our results corroborate that *H. mus*, like *Trypanosoma*, represents the more primitive state of editing as compared to the *Leishmania* group or the SHTs that represent a more recent state of decreased editing (Maslov et al., 1994a) (Jan Arts and Benne, 1996) (Landweber and Gilbert, 1994). In general, closely related organisms showed similar RNA editing patterns, likely owing to higher sequence identities.

U insertion-deletion RNA editing is thought to have evolved in an ancestor to the kinetoplastids-bodonids, (Simpson et al., 2000) (Simpson and Maslov, 1999). Extensive (pan) RNA editing is believed to be a primitive process seen in the earlier branches of kinetoplastids including the bodonids and related cryptobiids and the trypanosomatids, like *Trypanosoma* and *Herpetomonas*;

whereas the loss of extensive editing in the more recently diverged members is attributed to the overall loss of RNA editing due to easing out the genetic load it imposed (Maslov et al., 1994a) (Landweber and Gilbert, 1993) (Jan Arts and Benne, 1996) (Landweber and Gilbert, 1994). It is hypothesized that many pan-edited genes were gradually replaced by their partially edited counterparts (Maslov and Simpson, 1992) (Simpson et al., 2000) (Landweber, 1992), when partially edited transcripts were reverse transcribed using their reverse transcriptase-like activity (Gabriel and Boeke, 1991) (González et al., 1997), and the resulting cDNA homologously recombined into the maxicircle genome. Such gene conversions would eliminate the pressure to maintain the repertoire of minicircles as also any potential adverse effects in case of minicircle mis-segregation during cell division or transkinetoplastidy-like scenarios. Alongside extensive editing, the complex genomic organization of the minicircles, their free, non-catenated state and their redundancy in encoding gRNAs as well as a stationary kinetoplast are all hypothesized as being ancestral states (Simpson et al., 2000) (Simpson and Maslov, 1994a).

It was also presumed that digenetic lifestyle in these parasites evolved from the older monogenetic lifestyle (Landweber and Gilbert, 1993) (Lake et al., 1988), but this theory was disregarded later when there were indications towards the evolution of digenetic life cycles independently at several times in the course of evolution of the trypanosomatids (Vickerman, 1994) (Maslov et al., 1994a) (Landweber and Gilbert, 1994) (Simpson and Maslov, 1999). Subsequently, the discovery of RNA editing in the free living *Bodo saltans* asserted that parasitism has not driven the process of RNA editing in any measure (Blom et al., 1998). Our results in this report also do not reveal any immediate correlation between the number of hosts an organism can infect and the levels of RNA editing among those trypanosomatids.

However, some studies (Kolesnikov et al., 2003) that analyzed reduction in the edited domains indicated a possibility that monogenetic parasites are more prone to reducing their edited domains compared to the digenetic parasites. This seemed probable due to two reasons: 1) in some digenetic parasites, some mitochondrial genes were expressed specifically in the insect stage or bloodstream stage while some other genes were constitutively expressed (Carrillo et al., 2001); and 2) editing is thought to regulate or switch the energy metabolism in digenetic trypanosomatids during their life cycles in different hosts; whereas monogenetic parasites do not need this type of regulation, due to their confinement to the insect host, thereby constantly needing RNA editing to maintain their oxidative phosphorylation. Thus, reducing the editing burden by reduction of the editing domain is logically helpful to the monogenetic parasites. Our results corroborate this theory to an extent, since a reduction in the editing is indeed observed in most monogenetic trypanosomatids analyzed in this report including the SHTs, *C. aca* and *L. cos*. However, since the actual degrees of reduction of the edited domains for partially edited genes (A6, COIII and ND7) are variable among the organisms, we believe that these reduction events have occurred independently within these different clades, but in parallel, following a convergent evolutionary mechanism, or homoplasy, via the retroposition event.

This brings us to the next variable phenomenon of energy metabolism in the trypanosomatids. There are bloodstream trypanosomatids that utilize carbohydrates exclusively via glycolysis and produce pyruvate or acetate as their metabolic end products using their plant-like alternative oxidases. There are other bloodstream forms that utilize carbohydrates via glycolysis but produce acetate and succinate using cytochrome mediated electron transport chain and oxidative phosphorylation. Finally, some bloodstream as well as insect stage trypanosomatids utilize both carbohydrates through glycolysis and amino acid catabolism to produce acetate and succinate as

the metabolic end products by either using the cytochrome or the alternative oxidases. For a long time, it was assumed that the trypanosomatids use different methods of energy metabolism depending upon the nutrient supply available to them. However, the available substrates didn't appear to control the metabolic pathway used. There also appeared to be no correlation between the metabolic pathways these organisms follow, their modes of transmission to their various hosts or their phylogenetic relationships. Clearly, some trypanosomatids show variable mitochondrial and RNA editing activities in the bloodstream stages as opposed to their insect stages (like *T. brucei*), while others express all the mitochondrial genes constitutively, thereby needing constant RNA editing (like *L. tarentolae* or *T. cruzi*). Therefore, the number of mitochondrial genes undergoing active editing at the different stages of the trypanosomatid life cycle, coupled with the metabolic pathways they follow appear to be unique to each organism. It is purported that *T. brucei* represents an "evolved" trypanosomatid, with a simplified but well-defined, and compartmentalized mitochondrial gene expression and function in its distinct host stages, while the others with a constitutive or an apparently less-defined mitochondrial gene expression profile at different stages of their life-cycle are still "in the process" of evolving to have a clear-cut and distinctly regulated gene-expression pattern [see (Tielens and van Hellemond, 2009) and references therein for a review on the energy metabolism].

Finally, despite the retroposition theory, the fact that RNA editing still exists implies that there is a functional significance attached to retaining and maintaining this complex mechanism. For instance, retaining the edited domain in a 5' edited gene, was postulated to allow regulation at the level of translation by creation of start codons and ribosomal binding sites via the process of editing (Simpson et al., 2000). Similarly, some other factors have been propounded to be the basis for the trypanosomes to retain their editing (Gerasimov et al., 2012b), for instance,

recently, there was the discovery of small non-coding (nc) 3'–oligouridylated gRNA-like molecules (Madej et al., 2007) encoded by the maxicircles, whose exact functions are unknown but whose positions on the maxicircles coincided with that of the edited domains of partially edited genes like CYb, ND7, A6 etc. This led to the hypothesis that these ncRNAs somehow prevented retroposition and gene-conversion to a completely unedited form. Also, many pan-edited genes were reported to encode alternate proteins by a mechanism of alternative editing, that employs a different cascade of gRNAs (Maslov, 2010) (Read et al., 1994) (Ochsenreiter et al., 2008a) (Ochsenreiter et al., 2008b) (Ochsenreiter and Hajduk, 2006). If such alternatively edited genes undergo gene-conversion to a lesser edited or unedited form, the alternate proteins with other potentially important functions in the cell cannot be synthesized.

Overall, more detailed analyses of the maxicircle genomes of Bodonids and Cryptobiids are necessary, coupled with details of their RNA editing. Only this shall firmly answer the question of whether editing is on its "way in" or "way out" or if it is simply a mechanism that originated in the ancestors of the kinetoplastids and continues to evolve convergently among different clades of the trypanosomatids. Similarly, extensive biochemical analyses on the metabolism in the different life cycle stages of several clades of the trypanosomatids are required to develop a clear picture of the roles of RNA editing in their life cycles and metabolism.

Conclusion

The data represented in this paper shows the genome organization and arrangement of all the protein-coding and ribosomal genes in the maxicircles of some endosymbiont-bearing and related insect parasitizing trypanosomatids. We have also characterized the editing patterns of all the genes of all the organisms studied and estimated the identities between their inferred fully edited coding sequences and proteins. Phylogenetic analyses reflect the expected relationships between the groups of organisms studied. Phylogenetic reconstruction on binary matrix of editing events revealed that the patterns and extents of the RNA editing of the individual genes are conserved between closely related species and that RNA editing has evolved in parallel to the nuclear genes.

Supplementary Tables for Chapter 3

Supplementary Table 3.1. Primers used for gap-closures and circularization of the maxicircle contigs (where circularization was possible)

Name	Primer sequence (5'-3')	T_m (^oC)
Scul-gap F	AAAACGCAAACACCTACG	55.3
Scul-gap R	GAGTCTTGAGTCATAGTAG	55.8
Scul-circ F	TCCCCATTTCTTACTAC	52.4
Scul-circ R	GAACAACTGTTTCATTAATAG	52.8
Caca-gap1 F	GCGTTTATTAAATGCGTTTG	54.2
Caca-gap1 R	GTAAAGGAGAGTAGGAC	54.8
Caca-gap2 F	TCGTTGTAAAGCAGATTTG	53.7
Caca-gap2 R	GCAACGGCTGGCATC	58.9
Ades-circ F	AATCACATAAACGATAAGTG	52.2
Ades-circ R	GGTACATATAGAACAACTGT	54.2
Hmus-gap F	GTGTTACTCAGTTATTATCTG	54.8
Hmus-gap R	GTGCATTCATCCACTTTTG	55.8
Esch-gap1 F	GCAGTGCCTGATAAGTTTACTGGA	62.9
Esch-gap1 R	GGTGCTATTCCGATTGTTTTCGG	62.8
Esch-gap2 F	ATGCATTAGATATGTTTTGTGTATTATTAC	57.8
Esch-gap2 R	TGGTGCTATTCCGATTGTTTTCG	61
Esch-circ F*	GAGACAAAGTGATCGAATGTAC	58.9
Esch-circ R*	CTTAGACCGACGCATTTAATAAAC	59.4

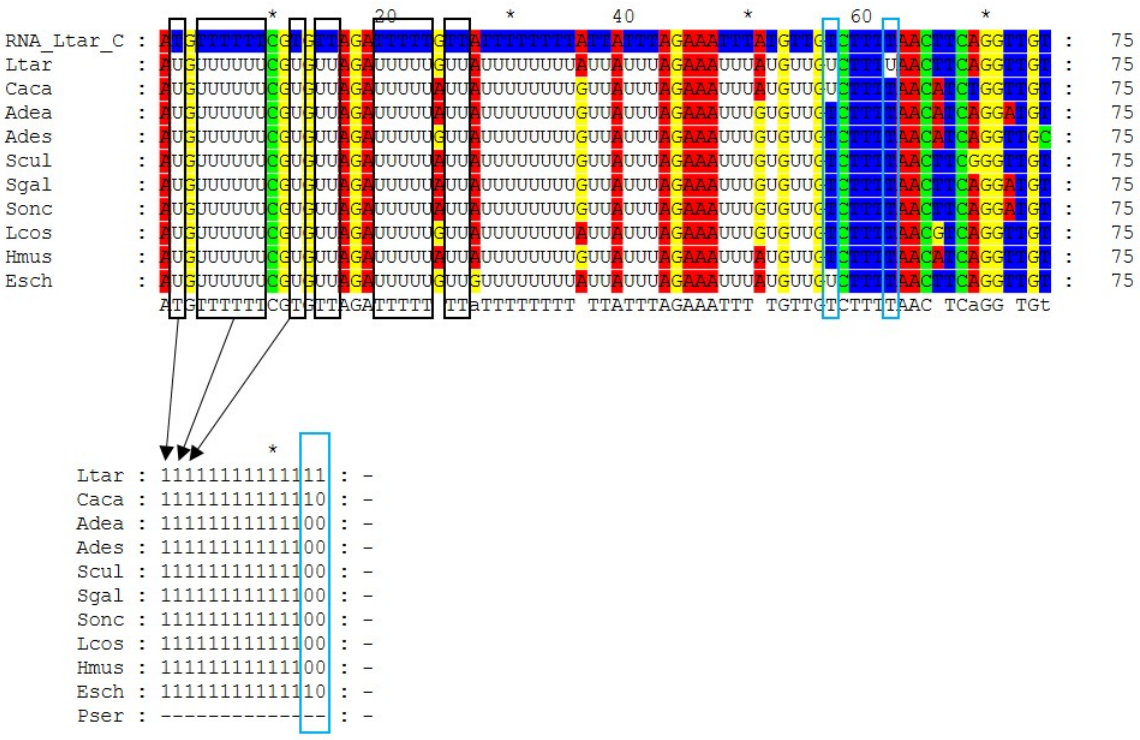
*** Primer set did not circularize the maxicircle, but helped extend the sequence at its 5' end.**

Supplementary Table 3.2. Percent identity matrix for 18S and GAPDH (nuclear) genes for the trypanosomatids

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H.mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100 (100)**										
E. sch	99 (87)	**100 (100)**									
L. cos	97 (90)	98 (87)	**100 (100)**								
C. aca	97 (88)	97 (85)	97 (91)	**100 (100)**							
H.mus	92 (82)	92 (80)	91 (84)	92 (84)	**100 (100)**						
P. ser	92 (80)	93 (79)	92 (83)	92 (82)	91 (82)	**100 (100)**					
A. dea	92 (81)	91 (81)	92 (82)	92 (83)	90 (83)	91 (82)	**100 (100)**				
A. des	92 (81)	91 (81)	92 (83)	92 (84)	90 (84)	91 (81)	98 (93)	**100 (100)**			
S. cul	89 (84)	90 (82)	89 (85)	89 (86)	88 (84)	89 (83)	92 (86)	91 (87)	**100 (100)**		
S. gal	90 (83)	90 (83)	90 (86)	90 (86)	88 (85)	89 (83)	92 (88)	92 (89)	96 (93)	**100 (100)**	
S. onc	89 (85)	91 (83)	90 (87)	89 (87)	88 (86)	89 (83)	92 (88)	92 (89)	96 (93)	96 (94)	**100 (100)**

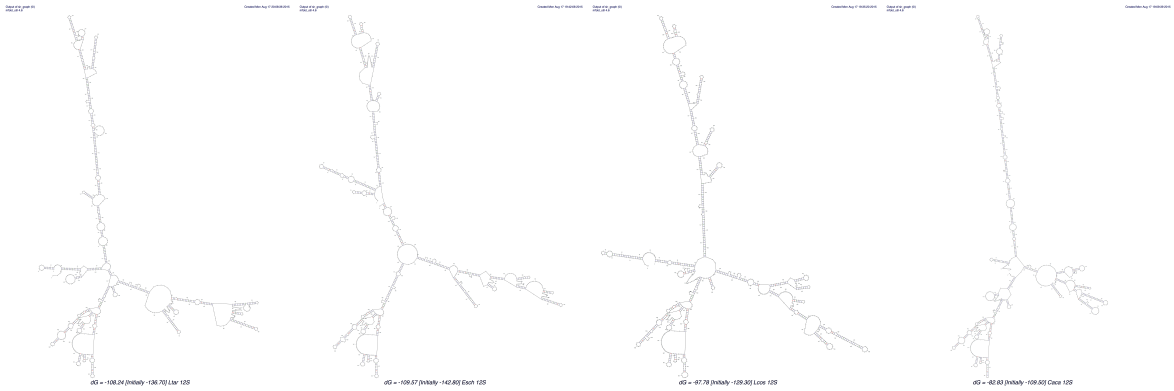
The matrix above shows the percent identities of the SSU rDNA and GAPDH nucleotide sequences among the organisms in this report. The values in the parantheses are the identities of GAPDH nucleotide sequences and outside of it are of SSU rDNA sequences. Identities of Leishmaninae family members to each other is shaded in green, while those of the SHTs to each other is shaded in yellow. Identities of P. ser to all other trypanosomatids is shown in gray.

Supplementary Figures for Chapter 3

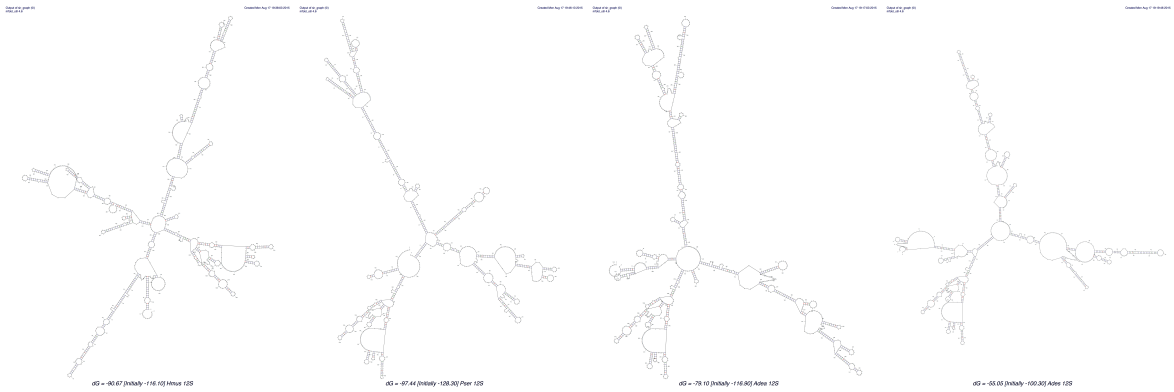


Supplementary Figure 3.1. Binary approach to quantitate editing events: Shown above is the inferred editing in the 5' end of the CYb transcript for all the organisms reported (first 75 nts of the coding sequence are shown). Abbreviations for the names of the organisms are as mentioned in the text. All colored bases are maxicircle-encoded. Reds in the alignment are As, yellows are Gs, greens are Cs and blues are Ts. U insertions are indicated as "U" against a white background in the alignment. Reference is the cDNA of *L. tar* (RNA_Ltar_C).

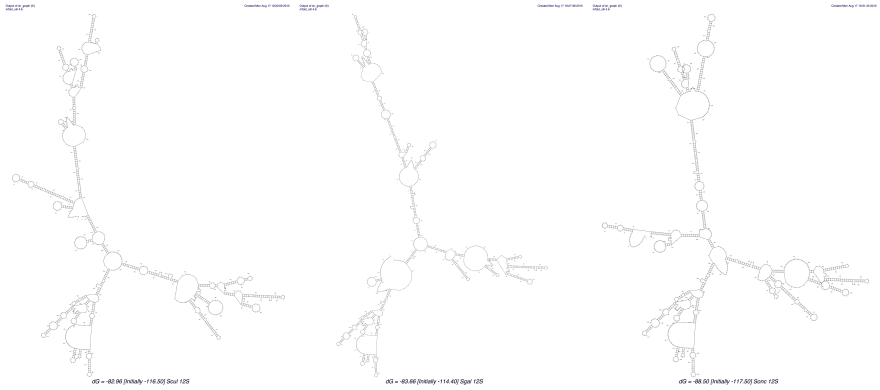
Any single U insertion or a stretch of Us inserted is represented as 1 in the binary matrix below, in accordance with the molecular mechanism of editing (i.e., at the site of editing, the Terminal Uridylyl Transferase enzyme adds a string of Us after the endonuclease nick and before the exonuclease chews off the excess Us). Therefore, regardless of the number of Us added at an editing locus, the event as the whole, is just counted as a single event and is scored as "1" (black boxes; each box represents an editing event at that locus). The corresponding absence of an event at the same locus is denoted as "0", as seen in the 13th and 14th columns of the binary matrix below (blue boxes), that correspond to the editing events at the 57th and 62nd columns above. Deletion events are also scored in the exact same fashion.



3.2A **3.2B** **3.2C** **3.2D**

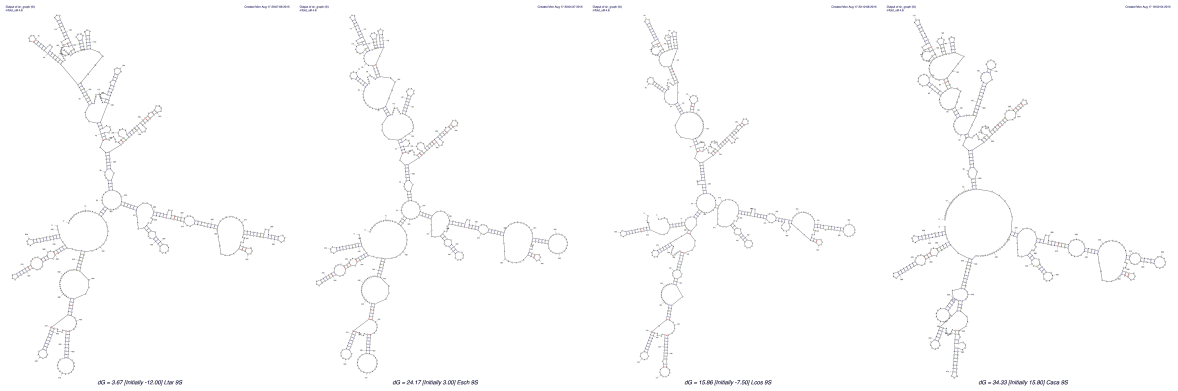


3.2E **3.2F** **3.2G** **3.2H**

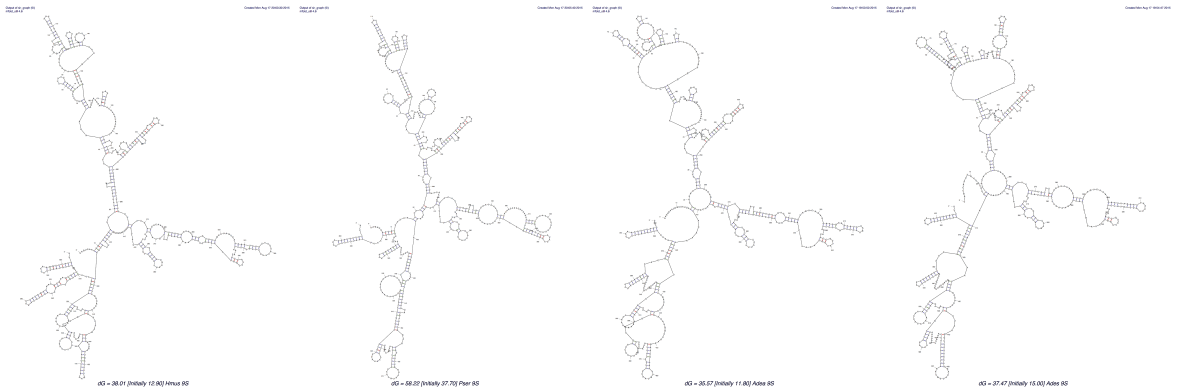


3.2I **3.2J** **3.2K**

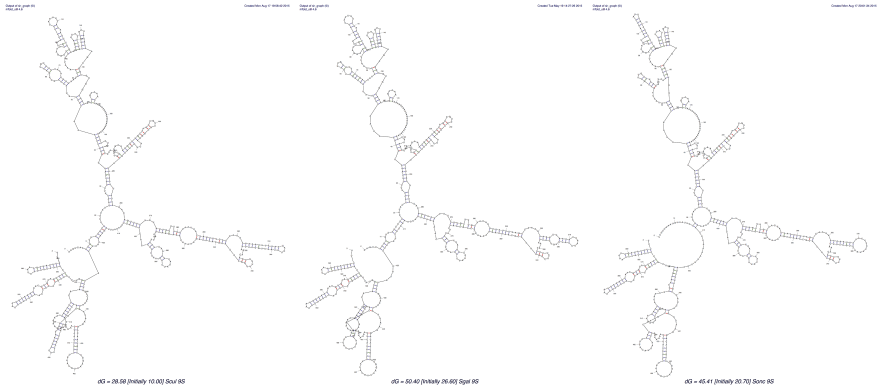
Supplementary Figure 3.2A. *L. tar* 12S folding; Supplementary Figure 3.2B. *E. sch* 12S folding; Supplementary Figure 3.2C. *L. cos* 12S folding; Supplementary Figure 3.2D. *C. aca* 12S folding; Supplementary Figure 3.2E. *H. mus* 12S folding; Supplementary Figure 3.2F. *P. ser* 12S folding; Supplementary Figure 3.2G. *A.dea* 12S folding; Supplementary Figure 3.2H. *A. des* 12S folding; Supplementary Figure 3.2I. *S. cul* 12S folding; Supplementary Figure 3.2J. *S. gal* 12S folding; Supplementary Figure 3.2K. *S. onc* 12S folding.



3.3A **3.3B** **3.3C** **3.3D**



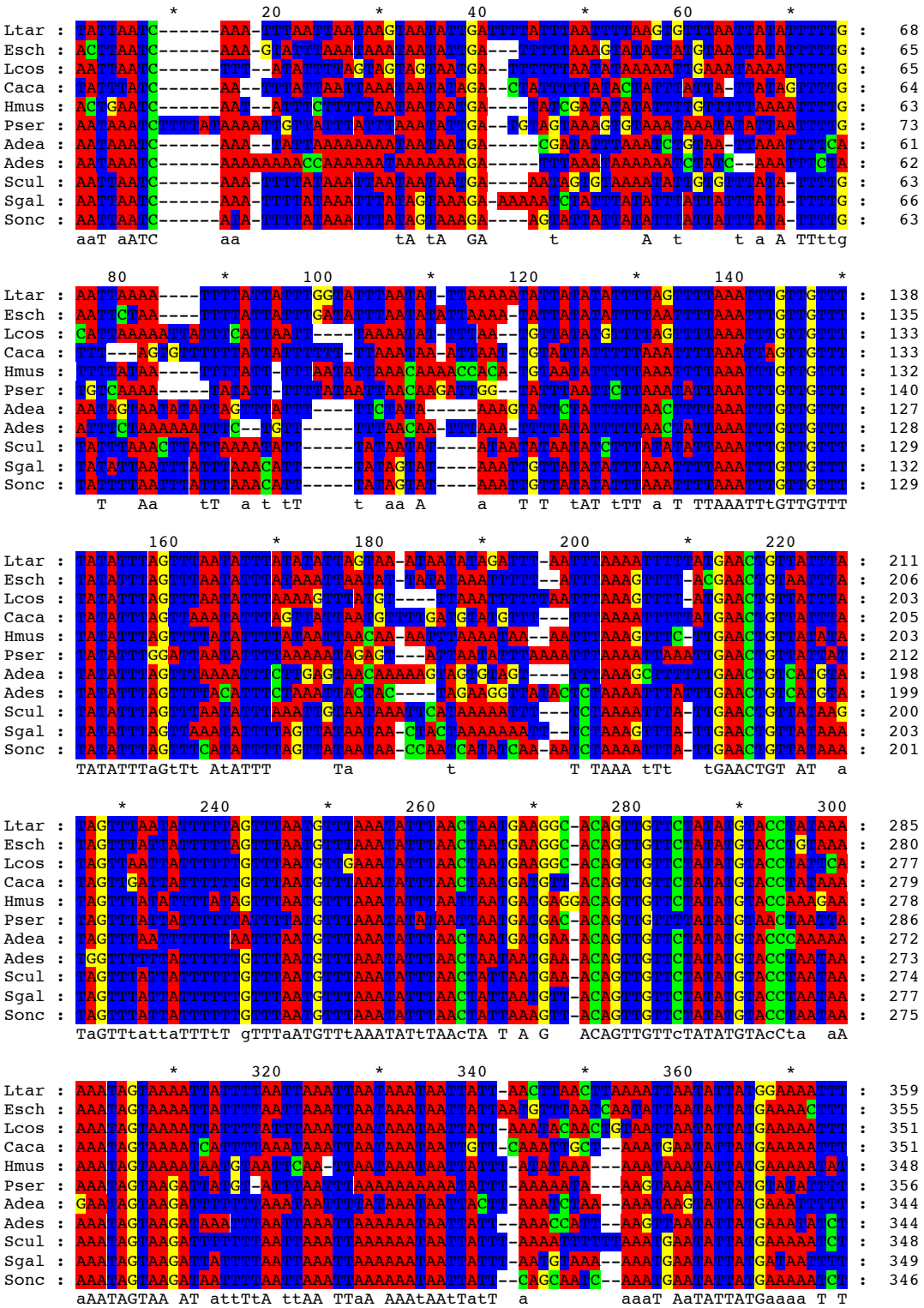
3.3E **3.3F** **3.3G** **3.3H**

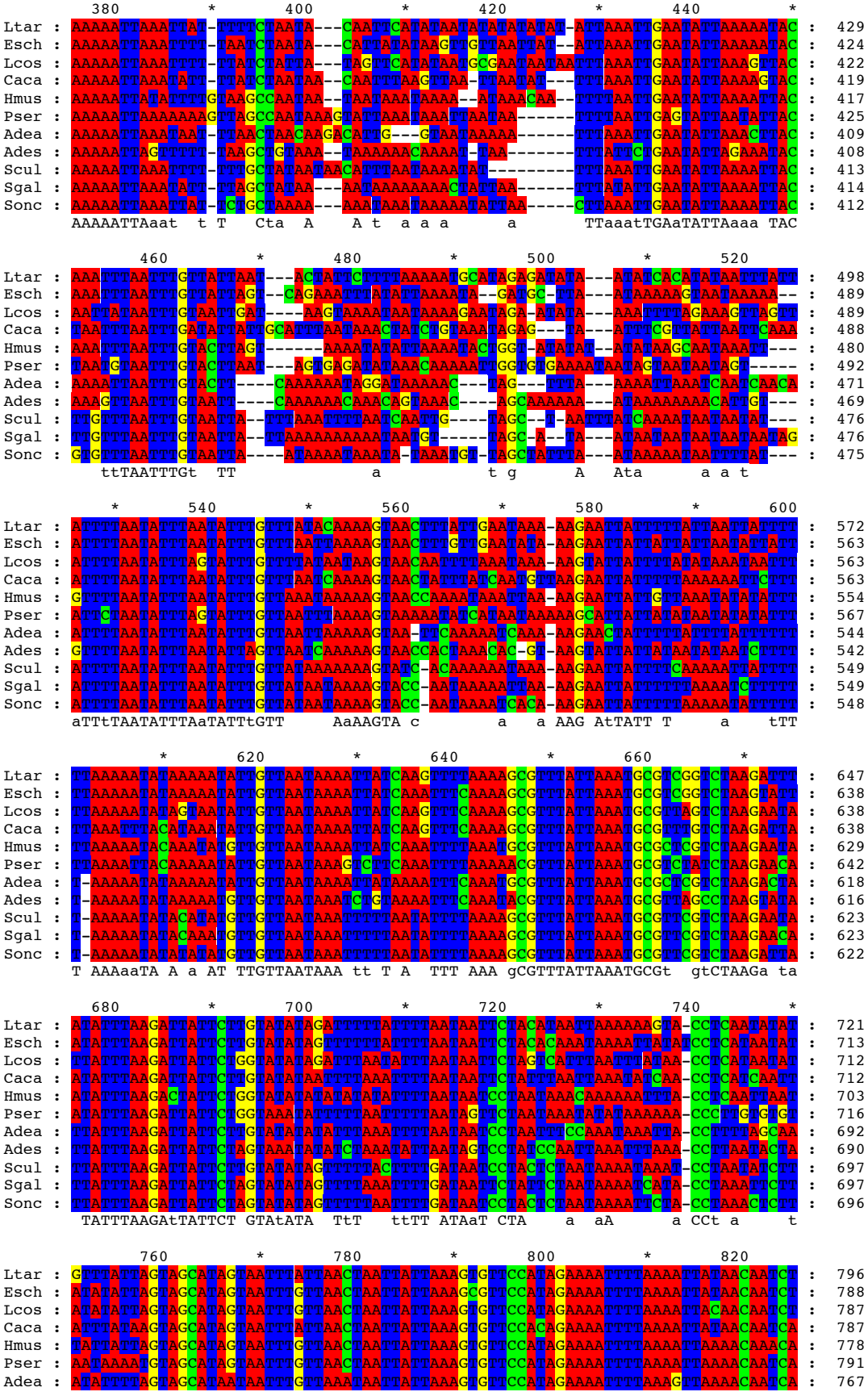


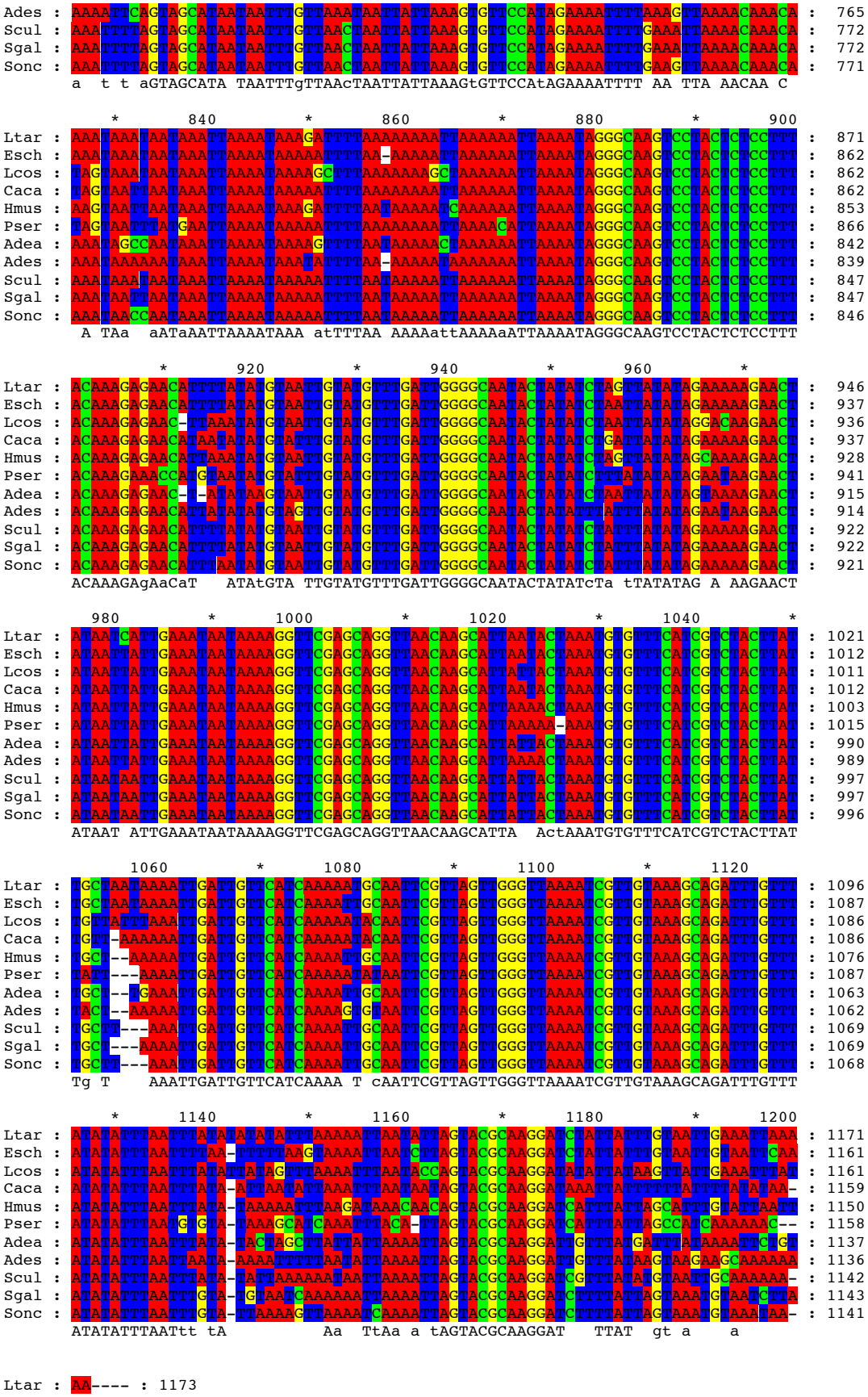
3.3I **3.3J** **3.3K**

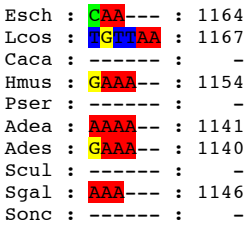
Supplementary Figure 3.3A. *L. tar* 9S folding; Supplementary Figure 3.3B. *E. sch* 9S folding; Supplementary Figure 3.3C. *L. cos* 9S folding; Supplementary Figure 3.3D. *C. aca* 9S folding; Supplementary Figure 3.3E. *H. mus* 9S folding; Supplementary Figure 3.3F. *P. ser* 9S folding; Supplementary Figure 3.3G. *A. dea* 9S folding; Supplementary Figure 3.3H. *A. des* 9S folding; Supplementary Figure 3.3I. *S. cul* 9S folding; Supplementary Figure 3.3J. *S. gal* 9S folding; Supplementary Figure 3.3K. *S. onc* 9S folding.

Supporting Information (SI) for Chapter 3





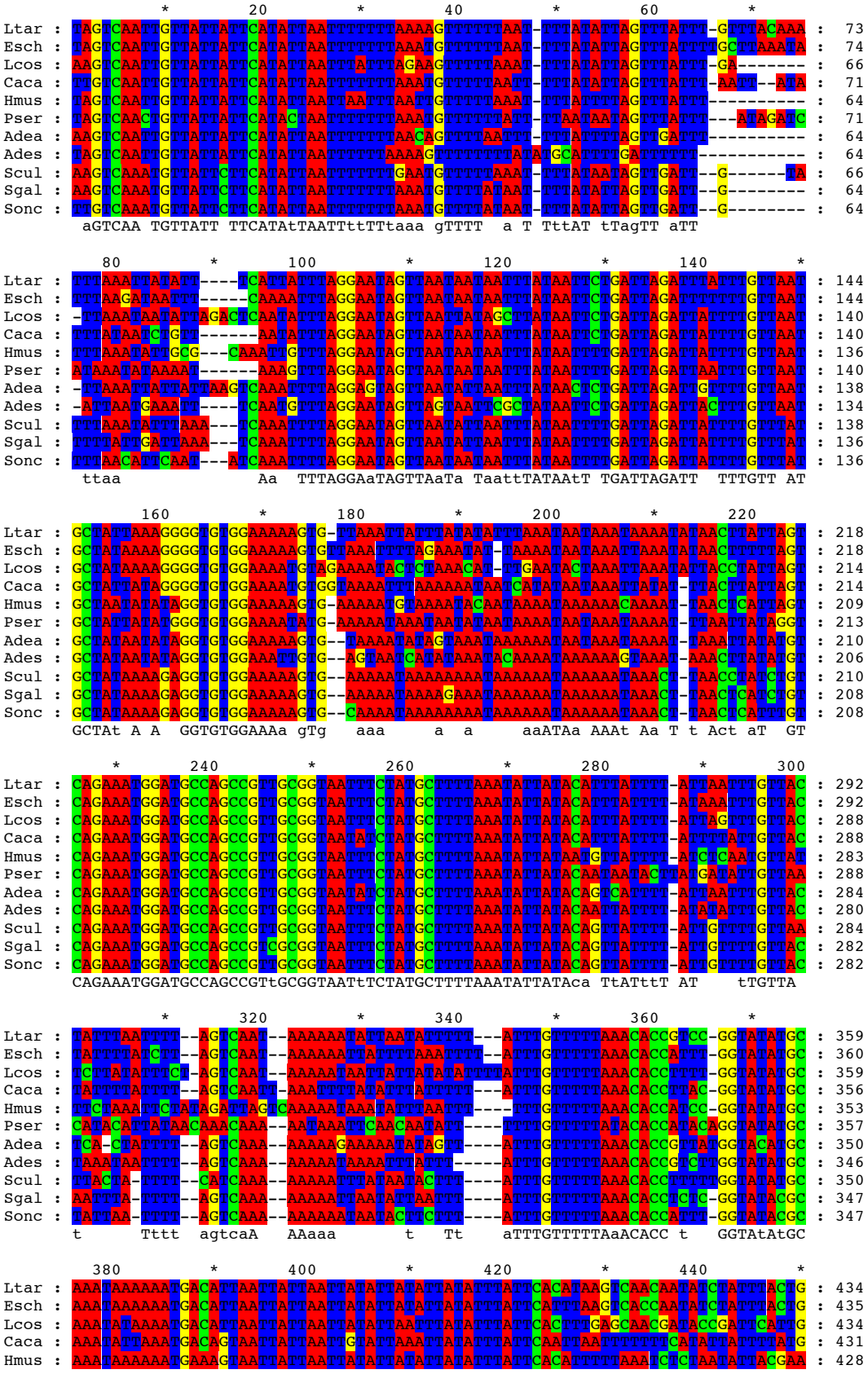


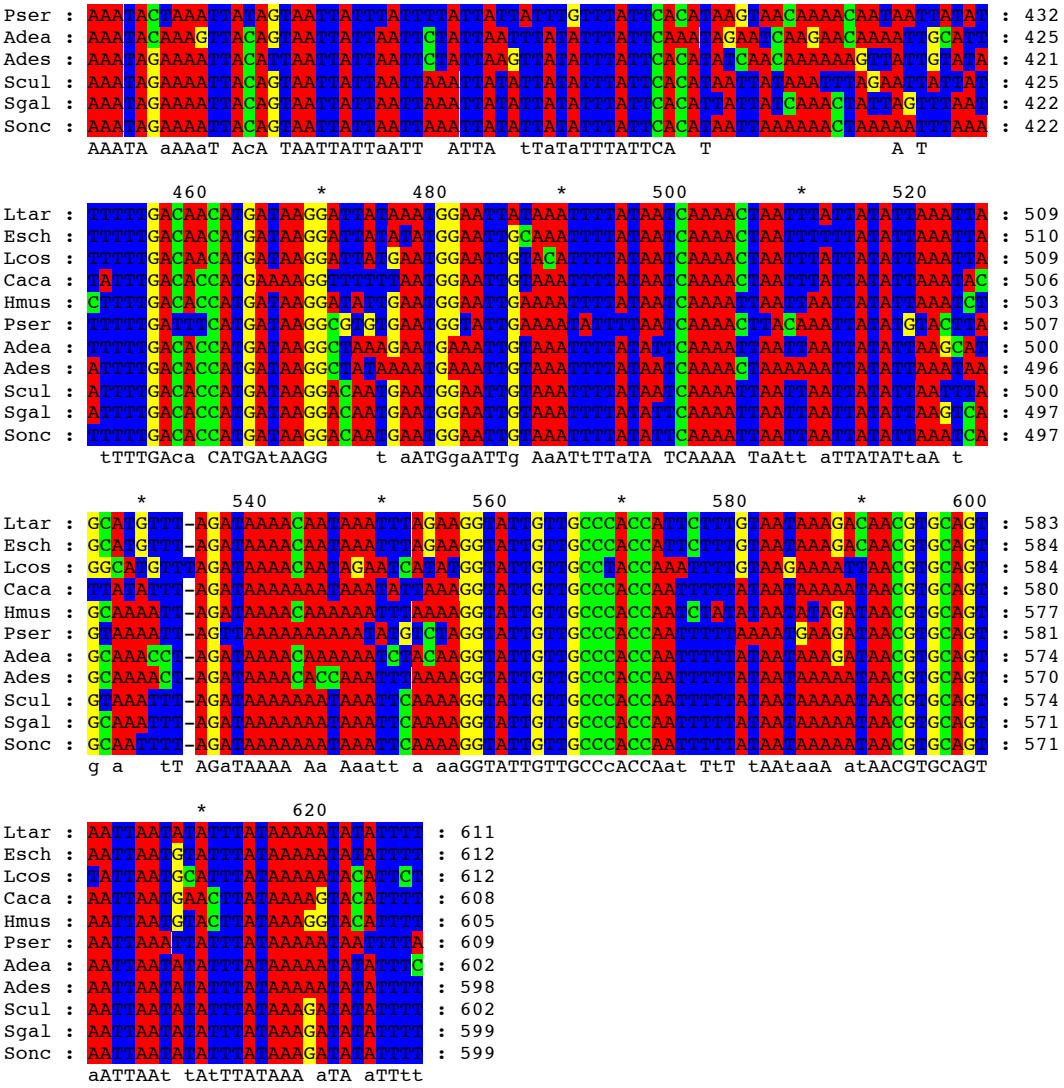


3.SI-S1a: GeneDoc 12S gene alignment: Abbreviations for the names of the organisms are as mentioned in the text. Reds in the alignments are As, yellows are Gs, greens are Cs and blues are Ts. Gaps in the alignments are indicated by '-'. Any consensus in each column is denoted as stretches of plain alphabets under *P. ser* (lower case alphabets mean lower confidence).

3.SI-S1b: 12S nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	89	**100**									
L. cos	85	84	**100**								
C. aca	85	84	82	**100**							
H. mus	82	83	81	81	**100**						
P. ser	77	77	78	78	81	**100**					
A. dea	82	81	80	80	81	76	**100**				
A. des	79	79	78	78	80	77	81	**100**			
S. cul	84	83	82	81	82	78	82	82	**100**		
S. gal	83	84	81	82	84	79	81	81	91	**100**	
S. onc	83	83	80	81	84	77	80	82	90	93	**100**

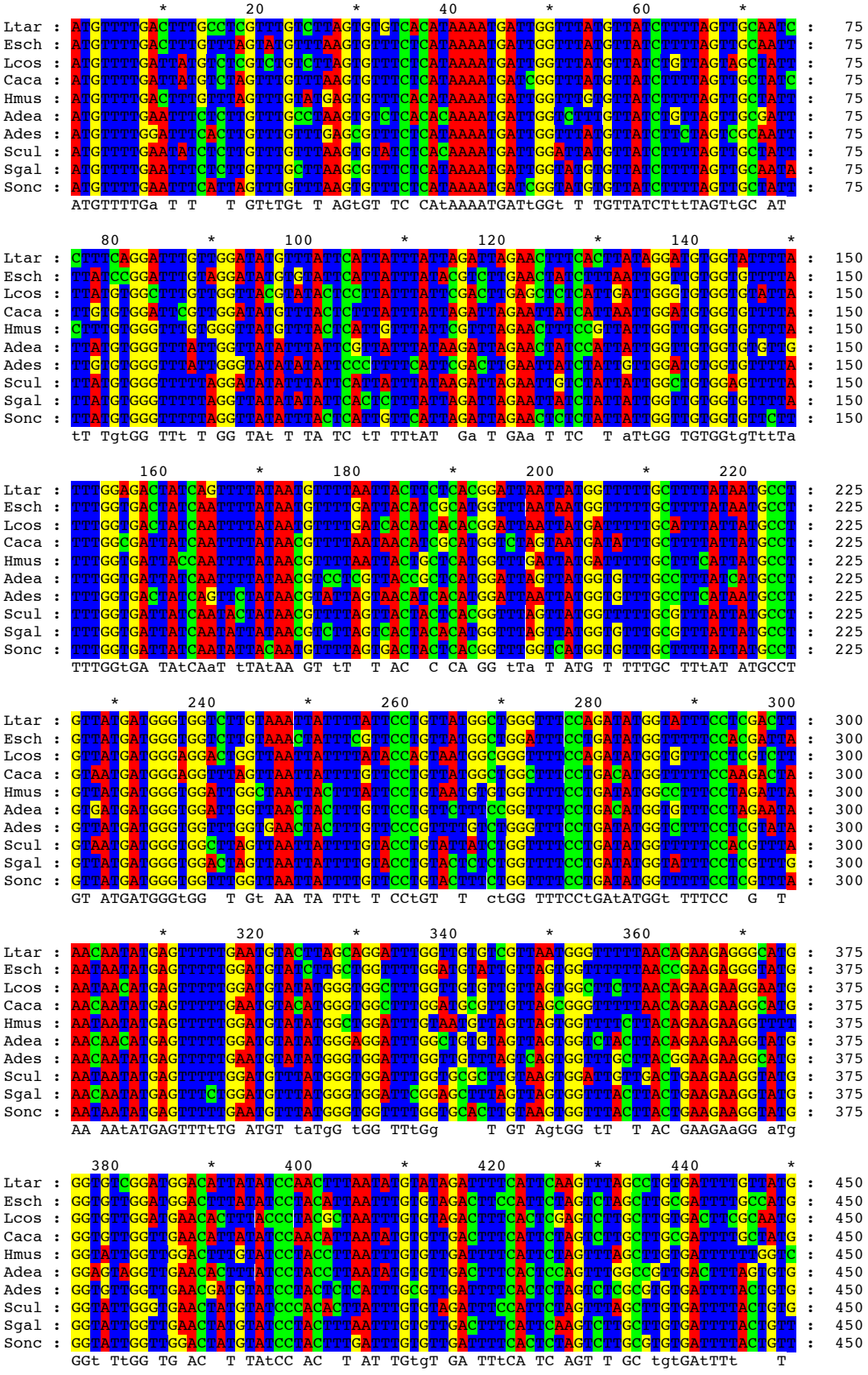


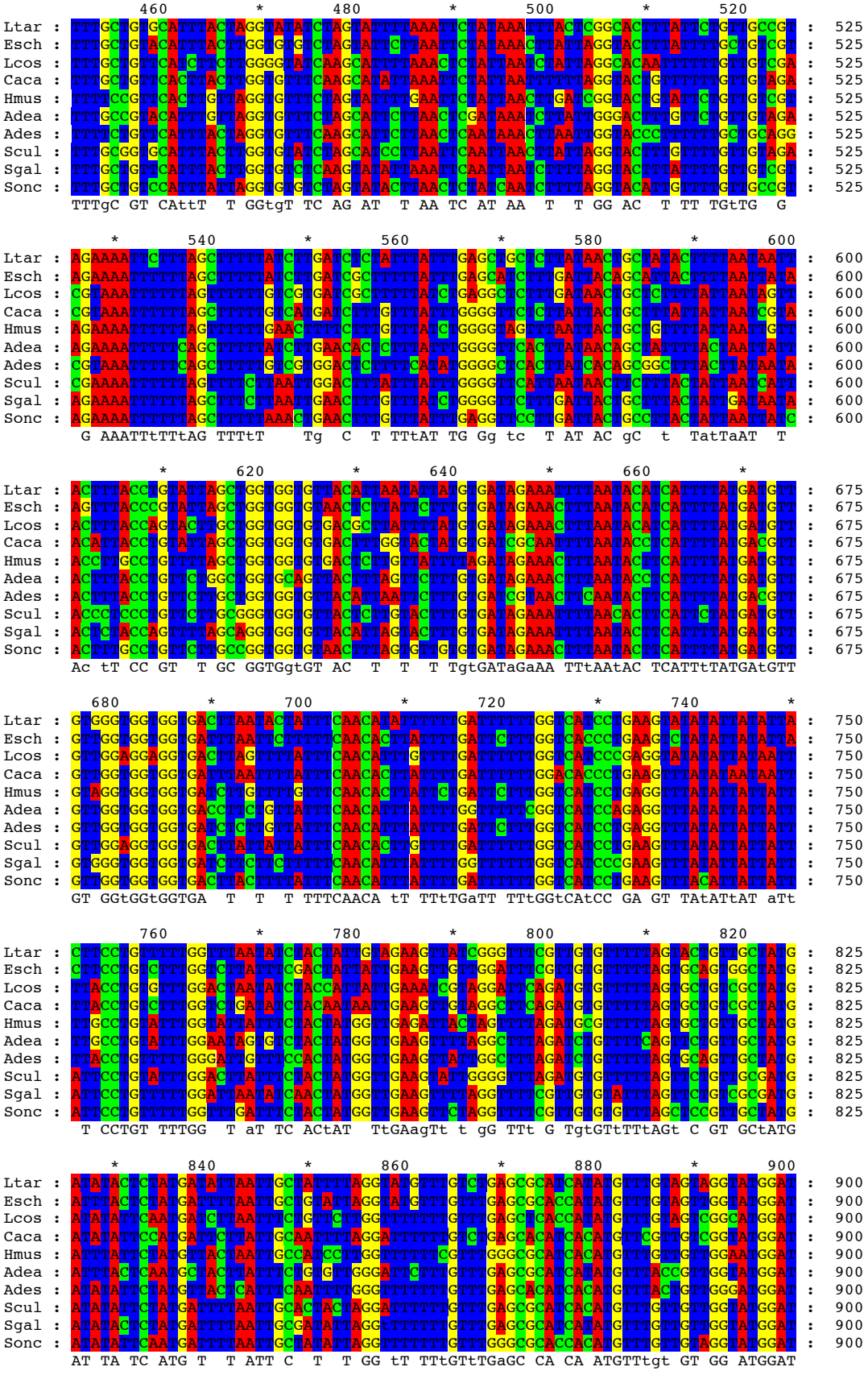


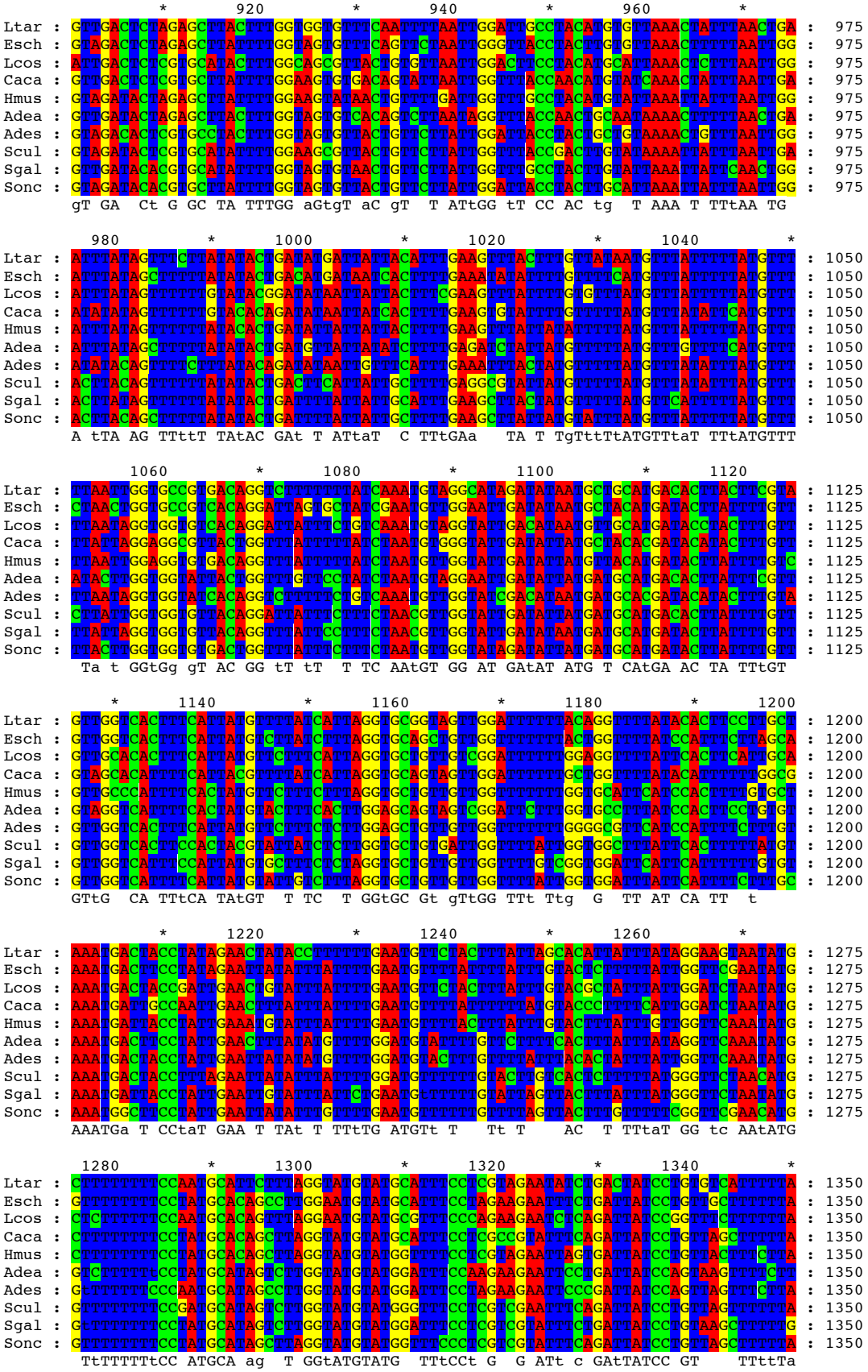
3.SI-S2a: GeneDoc 9S gene alignment: Abbreviations for the names of the organisms, color code, gaps and consensus for the figure are as mentioned in S1.

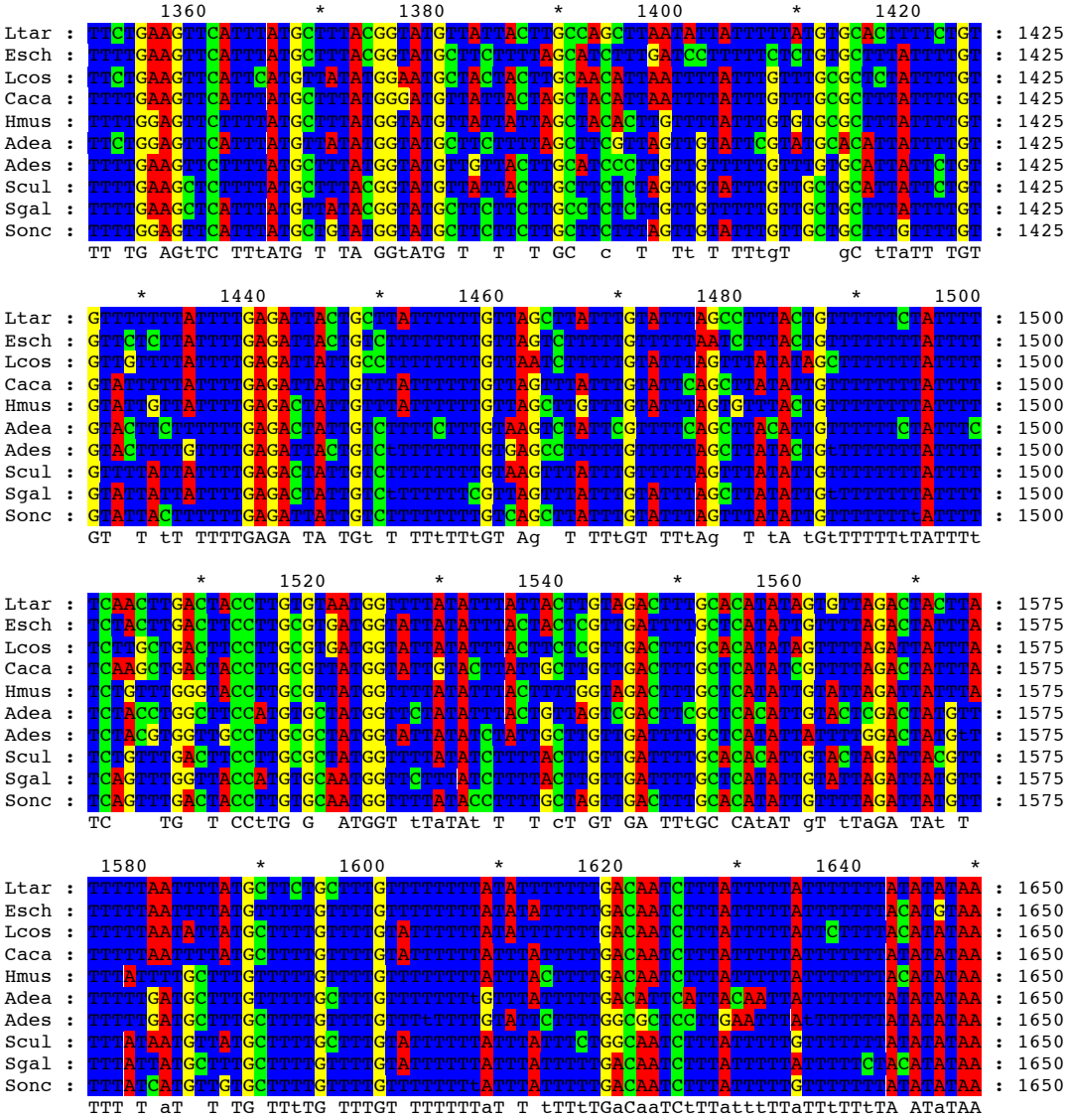
3.SI-S2b: 9S nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	93	**100**									
L. cos	86	86	**100**								
C. aca	86	85	83	**100**							
H. mus	83	82	79	82	**100**						
P. ser	79	78	74	78	76	**100**					
A. dea	82	81	79	81	80	77	**100**				
A. des	83	82	79	81	80	75	84	**100**			
S. cul	84	84	81	82	83	78	84	82	**100**		
S. gal	84	84	80	83	83	78	85	83	94	**100**	
S. onc	86	86	81	83	84	79	85	83	93	94	**100**

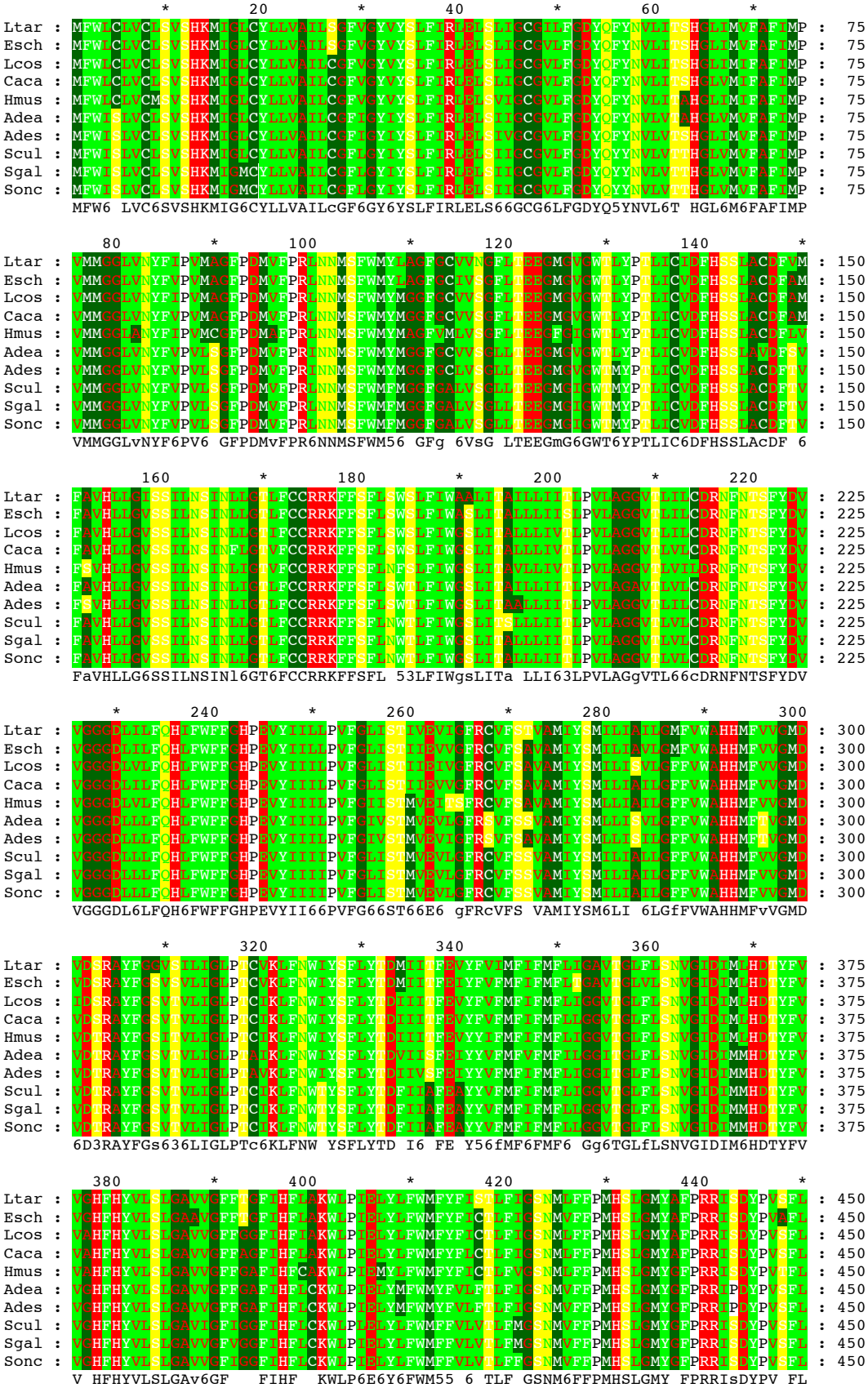


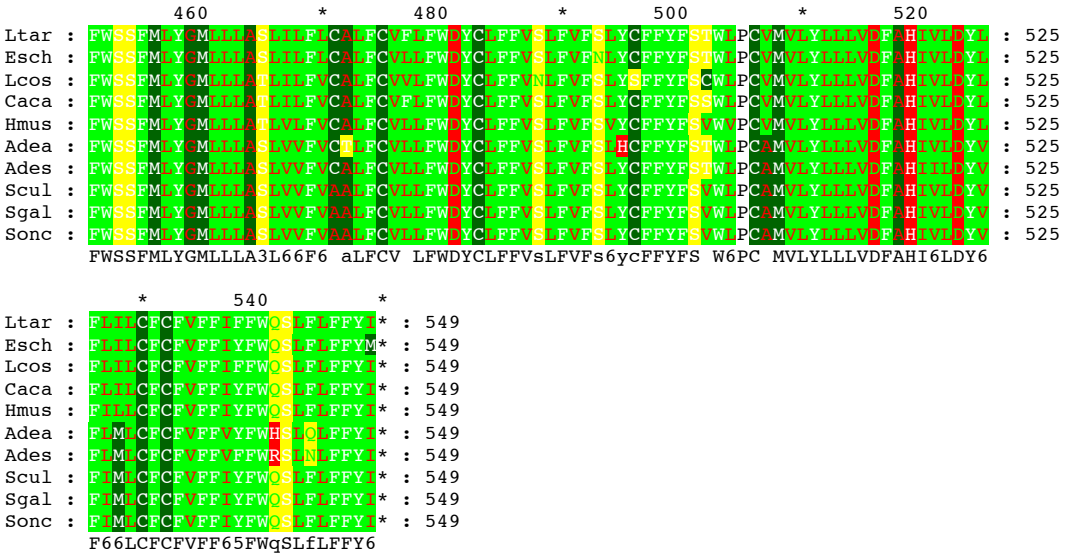

460 * 480 * 500 * 520
Ltar : 525
Esch : 525
Lcos : 525
Caca : 525
Hmus : 525
Adea : 525
Ades : 525
Scul : 525
Sgal : 525
Sonc : 525
TTTgC GT CAttT T GGtgT TC AG AT T AA TC AT AA T T GG AC T TT TGtTG G
* 540 * 560 * 580 * 600
Ltar : 600
Esch : 600
Lcos : 600
Caca : 600
Hmus : 600
Adea : 600
Ades : 600
Scul : 600
Sgal : 600
Sonc : 600
G AAATTtTTtAG TTTtT Tg C T TTtAT TG Gg tc T AT AC gC t TatTaAT T
* 620 * 640 * 660 *
Ltar : 675
Esch : 675
Lcos : 675
Caca : 675
Hmus : 675
Adea : 675
Ades : 675
Scul : 675
Sgal : 675
Sonc : 675
Ac tT CC GT T GC GGTGgtGT AC T T T TgtGATaGaAA TTtAAtAC TCATTtTATGAtGTT
680 * 700 * 720 * 740 *
Ltar : 750
Esch : 750
Lcos : 750
Caca : 750
Hmus : 750
Adea : 750
Ades : 750
Scul : 750
Sgal : 750
Sonc : 750
GT GGtGGtGGTGA T T T TTTCAACA tT TTtTGaTT TTtGGtCAtCC GA GT TAtATtAT aTt
760 * 780 * 800 * 820
Ltar : 825
Esch : 825
Lcos : 825
Caca : 825
Hmus : 825
Adea : 825
Ades : 825
Scul : 825
Sgal : 825
Sonc : 825
T CCTGT TTTGG T aT TC ACtAT TtGAagTt t gG TTt G TgtGTtTTtAGt C GT GCtATG
* 840 * 860 * 880 * 900
Ltar : 900
Esch : 900
Lcos : 900
Caca : 900
Hmus : 900
Adea : 900
Ades : 900
Scul : 900
Sgal : 900
Sonc : 900
AT TA TC ATG T T ATT C T T GG tT TTtGTtTGaGC CA CA ATGTTtgt GT GG ATGGAT





3.SI-S3a: GeneDoc COI nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1.





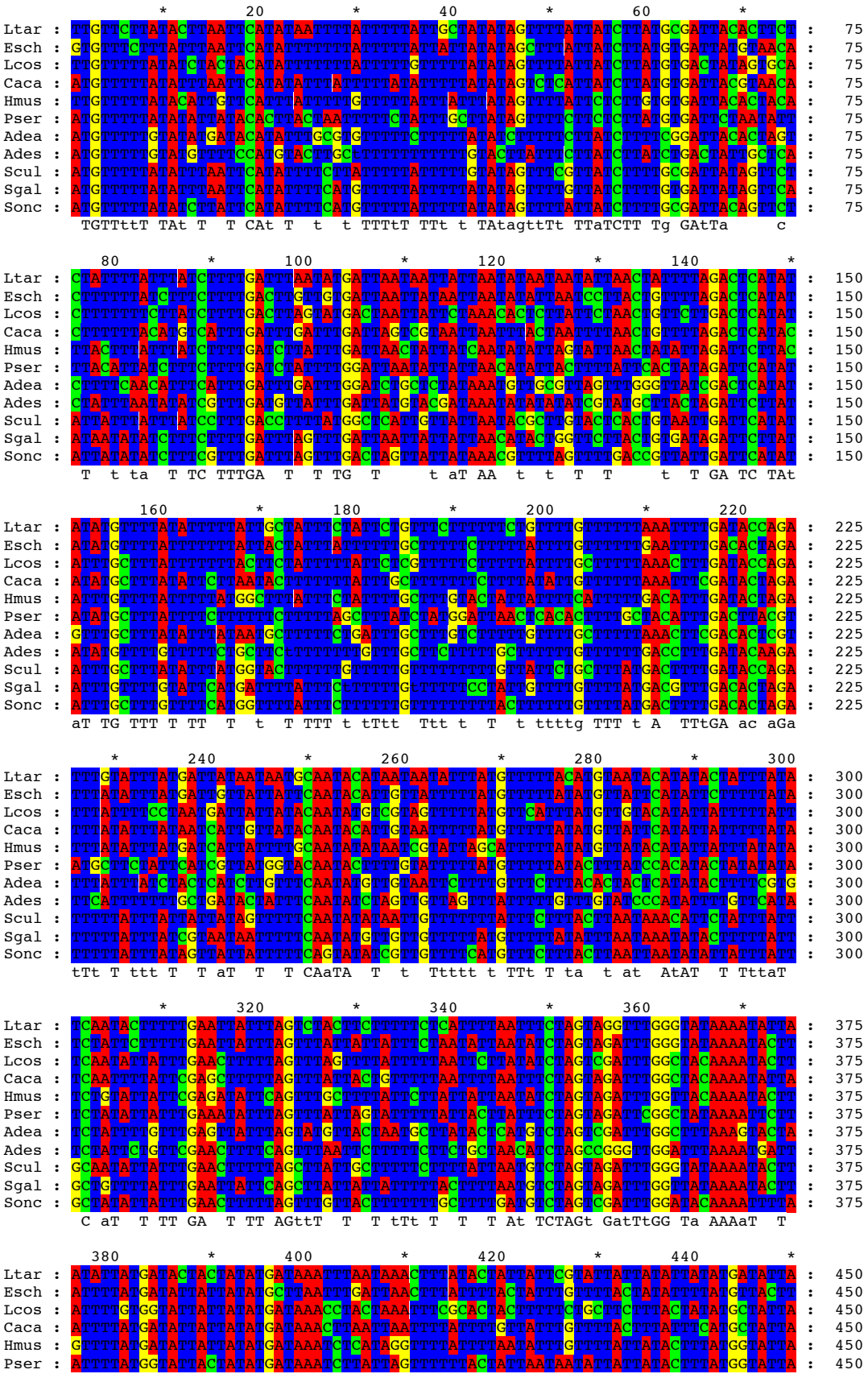
3.SI-S3b: GeneDoc COI protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment.

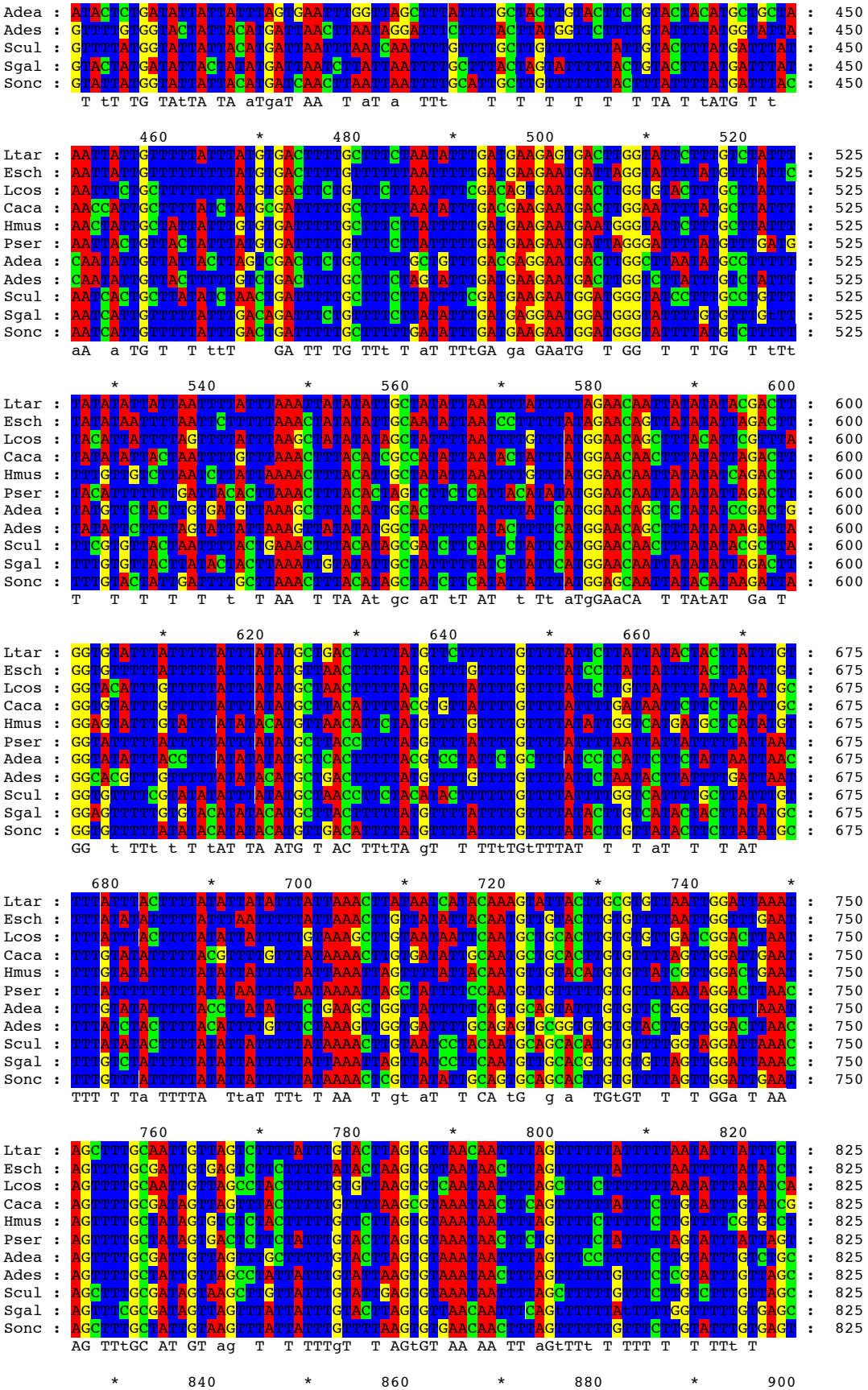
3.SI-S3c: COI nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

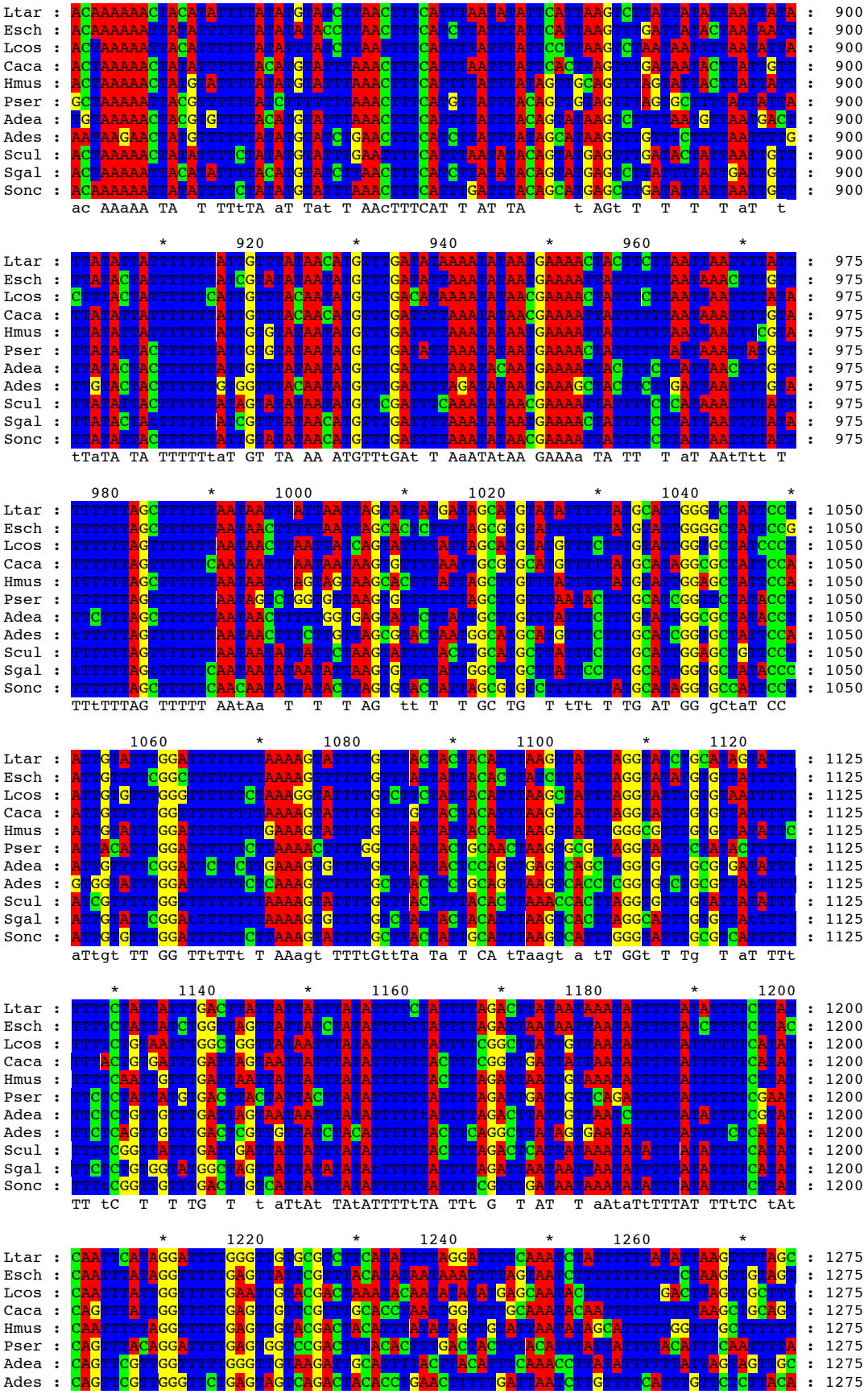
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	83	**100**								
L. cos	81	82	**100**							
C. aca	82	82	83	**100**						
H. mus	79	81	79	81	**100**					
A. dea	76	77	76	77	78	**100**				
A. des	77	79	78	79	79	81	**100**			
S. cul	78	80	78	81	80	79	80	**100**		
S. gal	78	80	78	81	82	80	81	86	**100**	
S. onc	78	81	78	81	82	80	80	86	88	**100**

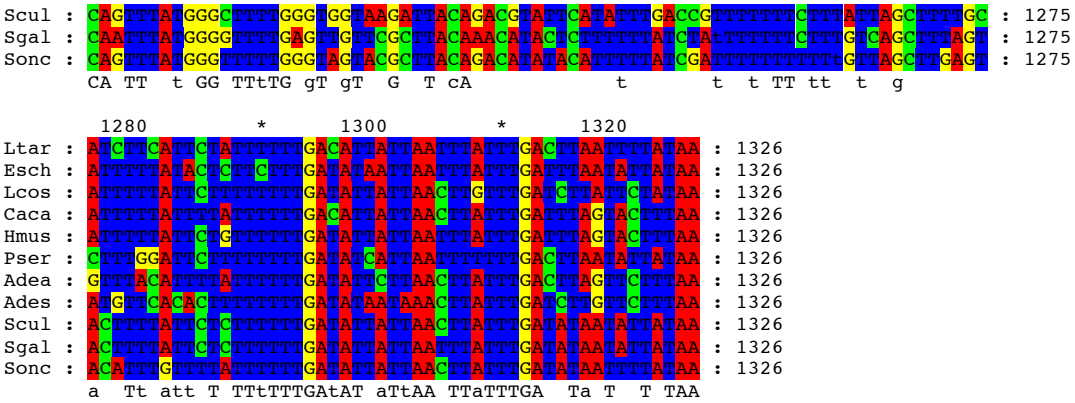
3.SI-S3d: COI protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	95	**100**								
L. cos	92	93	**100**							
C. aca	93	94	97	**100**						
H. mus	87	87	91	91	**100**					
A. dea	85	86	86	87	85	**100**				
A. des	86	86	87	87	86	95	**100**			
S. cul	86	87	87	89	87	91	91	**100**		
S. gal	86	87	87	89	87	91	91	99	**100**	
S. onc	86	87	87	89	87	91	91	99	100	**100**

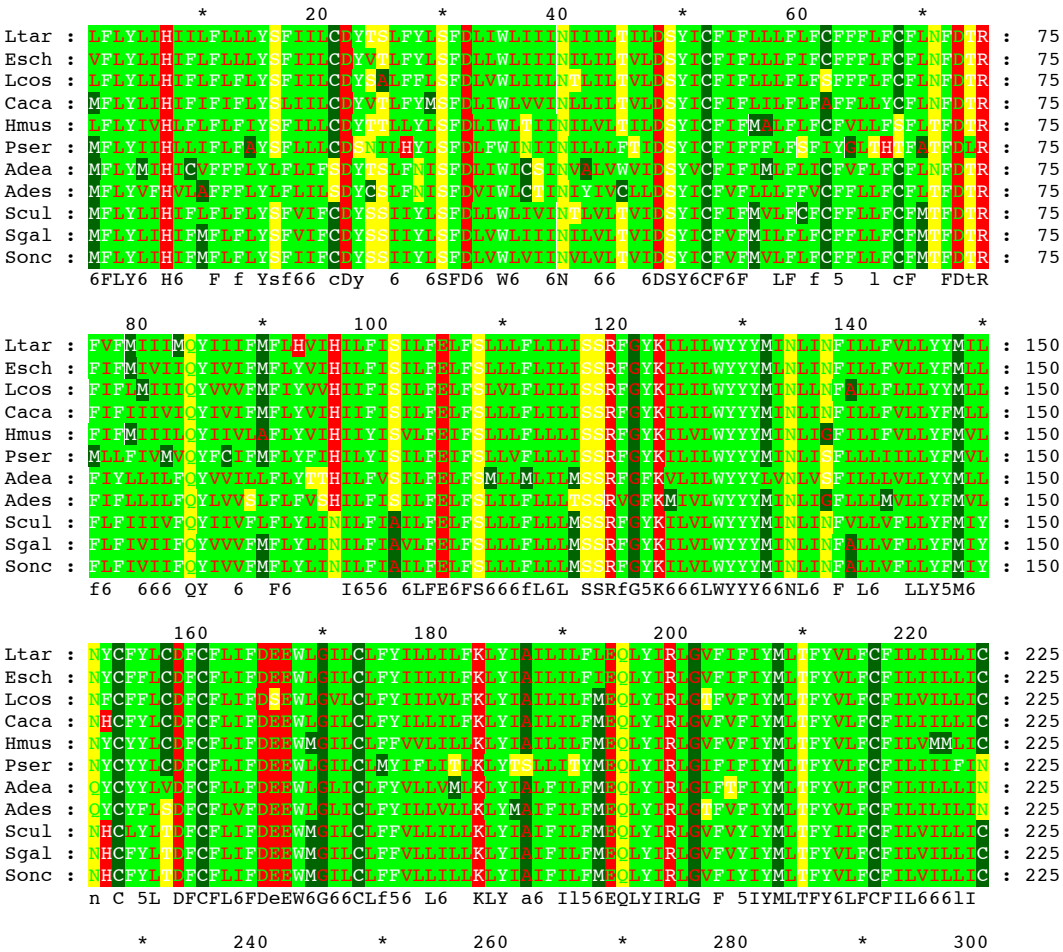

Ltar
Esch
Lcos
Caca
Hmus
Pser
Adea
Ades
Scul
Sgal
Sonc

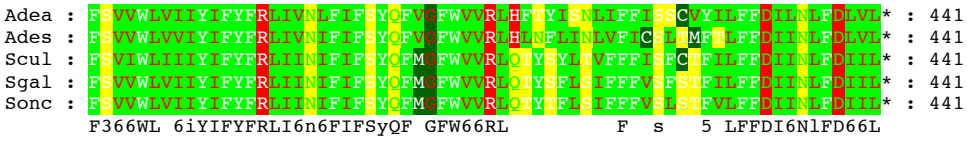

Adea : ATACTCTGATATTATTATTTAGTGAATTTGGTTAGCTTTATTTTGCTACTTGTACTTCTGTACTACATGCTGCTA : 450
Ades : GTTTTGTGGTACTATTACATGATTAACTTAATAGGATTTCTTTTACTTATGGTTCTTTTGTATTTTATGGTATTA : 450
Scul : GTTTTATGGTATTATTACATGATTAATTTAATCAATTTTGTTTTGCTTGTTTTTTTATTGTACTTTATGATTTAT : 450
Sgal : GTACTATGATATTACTATATGATTAATCTTATTAATTTTGCTTTACTAGTATTTTTACTGTACTTTATGATTTAT : 450
Sonc : GTATTATGGTATTATTACATGATCAACTTAATTAATTTTGCATTGCTTGTTTTTTTACTTTATTTTATGATTTAC : 450
T tT TG TAtTA TA aTgaT AA T aT a TTt T T T T T T TA T tATG T t
460 * 480 * 500 * 520
Ltar : AATTATTGTTTTTATTTATGTGACTTTTGCTTTCTAATATTTGATGAAGAGTGACTTGGTATTCTTTGTCTATTT : 525
Esch : AATTATTGTTTTTTTTTATGTGACTTTTGTTTTTTAATTTTTGATGAAGAATGATTAGGTATTTTATGTTTATTC : 525
Lcos : AATTTCTGCTTTTTTTTTATGTGACTTCTGTTTCTTAATTTTCGACAGTGAATGACTTGGTGTACTTTGCTTATTT : 525
Caca : AACCATTGCTTTTATCTATGCGATTTTTGCTTTTTTAATATTTGACGAAGAATGACTTGGAATTTTATGCTTATTT : 525
Hmus : AACTATTGCTATTATTTGTGTGATTTTTGCTTTCTTATTTTTGATGAAGAATGAATGGGTATTCTTTGCTTATTT : 525
Pser : AATTACTGTTACTATTTATGTGATTTTTGTTTTCTTATTTTTGATGAAGAATGATTAGGGATTTTATGTTTGATG : 525
Adea : CAATATTGTTATTACTTAGTCGACTTCTGCTTTTTTGCTGTTTTGACGAGGAATGACTTGGCTTTAATATGCCTTTTT : 525
Ades : CAATATTGTTACTTTTTGTCTGACTTTTGCTTTCTAGTATTTGATGAAGAATGACTTGGTCTTATTTGTCTATTT : 525
Scul : AATCACTGCTTATATCTAACTGATTTTTGCTTTCTTATTTTCGATGAAGAATGGATGGGTATCCTTTGCCTGTTT : 525
Sgal : AATCATTGTTTTTATTTGACAGATTTCTGTTTTCTTATATTTGATGAGGAATGGATGGGTATTTTGTGTTTGTTT : 525
Sonc : AATCATTGTTTTTATTTGACTGATTTTTGCTTTTTGATATTTGATGAAGAATGGATGGGTATTTTATGTCTTTTT : 525
aA a TG T T ttT GA TT TG TTt T aT TTtGA ga GAaTG T GG T T TG T tTt
* 540 * 560 * 580 * 600
Ltar : TATATATTATTAATTTTATTTAAATTATATATTGCTATATTAATTTTATTTTTAGAACAATTATATATACGACTT : 600
Esch : TATATAATTTTAATTCTTTTTAAACTATATATTGCAATATTAATTCCTTTTTATAGAACAGTTATATATTAGACTT : 600
Lcos : TACATTATTTTAGTTTTATTTAAGCTATATATAGCTATTTTAATTTTGTTTTATGGAACAGCTTTACATTCGTTTA : 600
Caca : TATATATTACTAATTTTGTTTAAACTTTACATCGCCATATTAATACTATTTATTGGAACAACTTTATATTAGACTT : 600
Hmus : TTTGTTGTCTTAATCTTATTAAAACTTTACATTGCTATATTAATTTTGTTTATGGAACAATTATATATCAGACTT : 600
Pser : TACATTTTTTTGATTACACTTAAACTTTACACTAGTCTTCTCATTACATATATGGAACAATTATATATTAGACTT : 600
Adea : TATGTTCTACTTGTGATGTTAAAGCTTTACATTGCACTTTTTATTTATTCATGGAACAGCTCTATATCCGACTG : 600
Ades : TATATTCTTTTAGTATTATTAAAGCTTATATATGGCTATTTTTATACTTTTCATGGAACAGCTTTATATAAGATTA : 600
Scul : TTCGTGTTACTAATTTTACTGAAACTTTACATAGCGATCTTCATTCTATTCATGGAACAACTTTATATACGCTTA : 600
Sgal : TTTGTGTTACTTATACTACTTAAATTGTATATTGCTATTTTTATCTTATTCATGGAACAATTATATATTAGACTT : 600
Sonc : TTTGTACTATTGATTTTGCTTAAACTTTACATAGCTATCTTCATATTATTTATGGAGCAATTATACATAAGATTA : 600
T T T T t T AA T TA At gc aT tT AT t Tt aTgGAaCA T TAtAT Ga T
* 620 * 640 * 660 *
Ltar : GGTGTATTTATTTTTATTTATATGCTGACTTTTTATGTTCTTTTTTGTTTTATTCTTATTATACTACTTATTTGT : 675
Esch : GGTGTTTTTATTTTTATTTATATGCTTAACTTTTTATGTTTTGTTTTGTTTTATCCTTATTATTTTACTTATTTGT : 675
Lcos : GGTACATTTGTTTTTATTTATATGCTAACTTTTTATGTTTTATTTTGTTTTATTCTTGTTATTTTATTTAATATGC : 675
Caca : GGTGTATTTGTTTTTATTTATATGCTTACATTTTACGTGTTATTTTGTTTTTATTTTGATAATTCTTCTTATTTGC : 675
Hmus : GGAGTATTTGTATTTTATATACATGTTAACATTCTATGTTTTGTTTTGTTTTATATTGGTCATGATGCTCATATGT : 675
Pser : GGTATTTTTATTTTTATTTATATGCTTACCTTTTATGTTTTATTTTGTTTTATTTTAATTATTATTTTTATTAAT : 675
Adea : GGTATATTTACCTTTATATATATGCTCACTTTTTACGTCCTATTCTGCTTTATCCTCATTCTTCTATTAATTAAC : 675
Ades : GGCACGTTTGTTTTTATATACATGCTGACTTTTTATGTTTTGTTTTTATTCTAATACTTTATTTTGATTAATT : 675
Scul : GGTGTTTTCGTTATATATTTATATGCTAACCTTCTACATACTTTTTTGTTTTATTTTGGTCATTTTGCTTATTTGT : 675
Sgal : GGAGTTTTTGTGTACATATACATGCTTACTTTTTATGTTTTATTTTGTTTTATACTTGTCATACTACTTATATGC : 675
Sonc : GGTGTTTTTATATACATATACATGTTGACATTTTATGTTTTATTTTGTTTTATACTTGTTATACTTCTTATATGC : 675
GG t TTt t T tAT TA ATG T AC TTtTA gT T TTtTGtTTTAT T aT T AT
680 * 700 * 720 * 740 *
Ltar : TTTATTTACTTTTATATTATATTTATTAAACTTATAATCATACAAAGTATTACTTGCGTGTTAATTGGATTAAAT : 750
Esch : TTTATATATTTTTATTTAATTTTTATTAAACTTGTTATATTACAATGTTGTACTTGTGTTTTAATTGGTTTGAAT : 750
Lcos : TTTATTTACTTTTATTATTATTTTGTAAAGCTTGTAATAATTGCAATGCTGCACTTGTGTGTTGATCGGACTTAAT : 750
Caca : TTTGTATATTTTTACGTTTGTTTATAAAACTTGTGATATTGCAATGCTGCACTTGTGTTTAGTTGGATTGAAT : 750
Hmus : TTTGTATATTTTTATATTATTTTTATTAAATTAGTTTTATTACAATGTTGTACATGTGTTATCGTTGGACTGAAT : 750
Pser : TTTATTTTTTTTTATATAATTTTAATAAAATTAGCTATTTTCCAATGTTGTTTTTGTGTTTTAATAGGACTTAAC : 750
Adea : TTTGTATATTTTTACCTTATATTTCTGAAGCTGGTTATTTTTCACTGCAGTATTTGTGTTCTGGTTGGTTTAAAT : 750
Ades : TTTATCTACTTTTACATTTGTTTCTAAAGTTGGTGATTTTGCACAGTGCGGTGTGTTACTTGTTGGACTTAAC : 750
Scul : TTTATATACTTTTATATTATTTTTATAAAACTTGTAATCCTACAATGCAGCACATGTGTTTTGGTAGGATTAAAC : 750
Sgal : TTTGTCTATTTTTATATTATTTTTATTAAATTAGTTATCCTTCAATGTTGCACGTGTGTTAGTTGGATTAAAC : 750
Sonc : TTTGTTATTTTTATATTATTTTTATAAAACTCGTTATATTGCACTGCAGCACTTGTGTTTTAGTTGGATTGAAT : 750
TTT T Ta TTTTA TtaT TTt T AA T gt aT T CA tG g a TGtGT T T GGa T AA
760 * 780 * 800 * 820
Ltar : AGCTTTGCAATTGTTAGCTTTTATTTGTACTTAGTGTTAACAATTTTAGTTTTTATTTTTAATATTTATTTCT : 825
Esch : AGTTTTGCGATTGTGAGCTTCTTTTTATACTAAGTGTTAATAACTTTAGTTTTTATTTTTAATTTTATATCT : 825
Lcos : AGTTTTGCAATTGTTAGCCTACTTTTTGTGTTAAGTGTCAATAATTTTAGCTTTCTTTTTTAATATTTATATCA : 825
Caca : AGTTTTGCGATAGTTAGTTTACTTTTTGTTTTAAGCGTAAATAACTTCAGTTTTTATTTCTTGTTATTTGTATCG : 825
Hmus : AGTTTTGCTATAGTGACTCTTCTATTTGTACTTAGTGTAAATAACTTTAGTTTCTTTTTCTTGTTTTCGTGTCT : 825
Pser : AGTTTTGCTATAGTGACTCTTCTATTTGTACTTAGTGTAAATAACTTCTGTTTTCTATTTTTAGTATTTATTAGT : 825
Adea : AGTTTTGCGATTGTTAGTTTGCTTTTTGTACTTAGTGTAAATAATTTTAGTTTCCTTTTTCTTGTATTTGTCTGC : 825
Ades : AGTTTTGCTAATTGTTAGCCTATTTTTGTATTTAAGTGTAAATAACTTTAGTTTTTGTTTCTTGTATTTGTTAGC : 825
Scul : AGCTTTGCGATAGTAAGCTTGTTATTTGTATTGAGTGTAAATAATTTTAGCTTTTTGTTTCTTGTCTTTGTTAGC : 825
Sgal : AGTTTCGCGATAGTTAGTTTATTATTTGTACTTAGTGTAACAATTTCAGTTTTTATTTTGGTTTTGTGAGC : 825
Sonc : AGCTTTGCTATTGTAAGTTTATTATTTGTTTTAAGTGTGAACAACTTTAGTTTTTGTTTCTTGTATTTGTGAGT : 825
AG TTtGC AT GT ag T T TTTgT T AGtGT AA AA TT aGtTTt T TTT T T TTt T
* 840 * 860 * 880 * 900


Ltar : ACAAAAAACTACATATTTTATATGCTATCTTAACTTTCATTTAATATATTCATTAAGTCTTATTATATTAATTATA : 900
Esch : ACAAAAAATTATATTTTTATATGCTATCCTTAACTTTCATCTTATTTATTCATTAAGTTTGATTATACTAATTATT : 900
Lcos : ACTAAAAATTACATTTTTTATATTTATCTTAATTTTCATTTTATTTATTCCTTAAGTCTAATAATTTTAATTATTA : 900
Caca : ACTAAAAACTATATTTTTTACATGTATTTAAACTTTCATTTAATTTATTCACTTAGTTTGATAATACTTATTGTT : 900
Hmus : ACTAAAAACTATGTATTTTATATGCTATTTAAACTTTCATTTTATTTATAGTTGCAGTTTAGTATTACTTATTATT : 900
Pser : GCTAAAAATTACGTTTTTTATCTTTTTTTAAACTTTCATGTTATTTACAGTTGTAGTTTAGTGCTTTTATTATTA : 900
Adea : ACTAAAAACTACGTGTTTTACATGTATTTAAACTTTCATTTTATTTACAGTATAAGTCTTTTAATGTTAATGACT : 900
Ades : AATAAGAACTATGTTTTTTATATGCTATCTGAACTTTCATCTTATTTATAGCATAAGTTTGTTCTTTTAATTTTG : 900
Scul : ACTAAAAACTATATTTTCTATATGCTATTTGAATTTTCATTTAATATACAGTATGAGTTTGATACTATTAATTGTT : 900
Sgal : ACTAAAAATTACATATTTTACATGCTATCTTAACTTTCATCTTATATACAGTATGAGTCTTATTTTATTGATTGTT : 900
Sonc : ACAAAAAATTATATTTTCTATATGCTATTTAAACTTTCATTTGATTTACAGCATGAGCTTTGATTATTAATTGTT : 900
ac AAaAA TA T TTtTA aT Tat T AAcTTTCAT T AT TA t AGt T T T T aT t
920 940 960
Ltar : TTATATTATTTTTTTATTGTTTATTAACATGTTTGATATTAAAATATTAATGAAAACTACTTCTTAATTAATTTATT : 975
Esch : TTATACTATTTTTTTATCGTATTATAATATGTTTGATATTAAATATTAATGAAAATTATTTTTTAATAAACTTTGTT : 975
Lcos : CTTTACTATTTTTTCATTGTTTACAATATGTTTGACATTAAAATATTAACGAAAACTATTTCTTAATTAATTTTATA : 975
Caca : TTATATTATTTTTTTATTGTTTACAACATGTTGATTTTAAATATTAACGAAAATTATTTTTTAATAAATTTTGTA : 975
Hmus : TTATATTATTTTTTTATTGTTGTTATAATATGTTTGATTTTAAATTATTAATGAAAATTATTTTTTAATTAATTTCGTA : 975
Pser : TTATATTACTTTTTTTATTGTTGTTATAATATGTTTGATTTAAATTATTAATGAAAACTATTTTTTATTAAATTATGTT : 975
Adea : TTATACTACTTTTTTATTGTTTATTAATATGTTTGATTTTAAATTACAATGAAAATTACTTTCTTATTAACTTTGTT : 975
Ades : TTGTACTACTTTTTTGTTGGTTTACAATATGTTTGATTTTAGATTATTAATGAAAGCTACTTCTTGATTAATTTTGTA : 975
Scul : TTATATTACTTTTTTTATAGTATTATAATATGTTCGATTTCAAATTATTAACGAAAATTATTTTCTCATAAATTTTATT : 975
Sgal : TTATACTATTTTTTTATCGTTTATTAACATGTTTGATTTTAAATTATTAATGAAAACTATTTTCTTATTAATTTTATA : 975
Sonc : TTATATTACTTTTTTTATTGTTATTAACACGTTTGATTTTAAATTATTAACGAAAATTATTTTCTTATTAATTTTATT : 975
tTaTA TA TTTTTtaT GT TA AA ATGTTTtGAt T AaATAtAA GAAAa TA TT T aT AAtTtt T
980 1000 1020 1040
Ltar : TTTTTTAGCTTTTTTAATAATTTATTAATTAGTATTATGATAGCATGTATATTTTATGCATTGGGTCTATTCCT : 1050
Esch : TTTTTTAGCTTTTTTAATAACTTTTAATTAGCACTCTTTTAGCGTGTATTTTTTATGTATTGGGGCTATTCCG : 1050
Lcos : TTTTTTAGTTTTTTTAATAACTTAATTATCAGTATTTTATTAGCATGTATGCTTCTTTGTATTGGTGCTATCCCT : 1050
Caca : TTTTTTAGTTTTTTCAATAATTTAATAATAAGTGTTTTAATTGCGTGCAGTTTTTATGCATAGGCGCTATTCCA : 1050
Hmus : TTTTTTAGCTTTTTTAATAATTTAGTAGTAAGCACTTTATTAGCTTGTTTATTTTATGTATTGGAGCTATTCCA : 1050
Pser : TTTTTTAGTTTTTTTAATAGTCTGGTGTTAAGTGTTTTTTAGCTTGTTTAATACTTTGCATCGGTTCTATACCT : 1050
Adea : TTCTTTAGCTTTTTTAATAACTTTTGGTGAGTATCTTTATTGCTTGTTTATTTCTTTGTATTGGCGCTATACCT : 1050
Ades : tTTTTTAGTTTTTTTAATAACTTTCTTGTTAGCGTACTAATGGCATGCAAGTTTCTTTGCATTGGGTGCTATTCCA : 1050
Scul : TTTTTTAGTTTTTTTAATAATATTATTCTAAGTATTTTACTTGCATGCTTATTTCTTTGCATTGGAGCTGTTCCT : 1050
Sgal : tTTTTTAGTTTTTTCAATAATATAATATTAAGTGTTTTATTGGCTTGCTTATTCCTTTGCATTGGTGCTAATACCC : 1050
Sonc : TTTTTTAGCTTTTTCAACAATATTATACTTAGTGTACTATTAGCGTGTCTTTTTTTATGCATAGGTGCCATTCCT : 1050
TTtTTTAG TTTTT AAtAa T T T AG tt T T GC TG T tTt T TG AT GG gCtaT CC
1060 1080 1100 1120
Ltar : ATTGTATTTGGATTTTTTTTAAAAGTATTTTGTTTACTACTACATTTAAGTTATTTAGGTATCTGCATAGTATTT : 1125
Esch : ATTGTTTTCGGCTTTTTTTTAAAAGTTTTTTGTTTATTATTACACTTATCTTATTTAGGTATATGTGTTATTTTT : 1125
Lcos : ATTGTGTTTGGGTTTTTTCTAAAGGTATTTTGTCTTCTATTACATTTAAGCTATTTAGGTATTTGTGTAATTTTT : 1125
Caca : ATTGTTTTTGGTTTTTTTTTTAAAAGTATTTTGTTTGTTACTACATTTAAGTTATTTAGGTATTTGTGTATTTTT : 1125
Hmus : ATTGTATTTGGATTTTTTTTGAAAGTATTTTGTTTATTATTACATTTAAGTTATTTGGGCGTTTGTGTATTATTC : 1125
Pser : ATTACATTTGGATTTTTTTCTTAAAACTTTTGGTTTATTACTGCAACTAAGTGCGTTAGGTATTTCTATACTTTTT : 1125
Adea : ATTGTTTTCGGATTCTTCTTGAAAGTGTTTTGTTTATTACTCCACTTGAGTCAGCTTTGGTGTTTGCGTGATATTT : 1125
Ades : GTGGTATTTGGATTTTTTCTCAAAGTTTTTTGCTTACTTCTGCACTTAAGTCACCCTCGGTGTCTGCGTTATTTT : 1125
Scul : ATCGTTTTTGGTTTTTTTTTAAAAGTATTTTGTTTACTTTACACTTAAACCACTTAGGTGTTTGTATTATATTT : 1125
Sgal : ATTGTATTCGGATTTTTTTTAAAAGTGTTTTGTCTATTACTACATTTAAGTCACTTAGGCATTTGTGTTACTTTT : 1125
Sonc : ATTGTGTTTGGATTTTTTCTTAAAGTATTTTGCTTACTATTGCATTTAAGTCATTTGGGTATTTGCGTCATTTTT : 1125
aTtgt TT GG TTtTTt T AAagt TTTtGttTa Ta T CA tTaagt a tT GGt T Tg T aT TTt
1140 1160 1180 1200
Ltar : TTTTCTATTATTTGACTTATTATTATTTATATTTCTATTTTAGACTTATTAATAAATATTTTTATATTTCTTAT : 1200
Esch : TTTTCTATTATTCTGGTTAGTTATTCTATATTTTTATTTTAGATTAATAATTATATTTTATCTTTCTTAC : 1200
Lcos : TTTTCTGTAATTTGGCTGGTTATTAATTTATATTTTTTATTTTCGGCTTATTGTTAATATTTTTATTTTTTCAAAT : 1200
Caca : TTTACTGTGATTTGATTAGTAATTATTTATATTTTTTACTTTCGGTTGATTATTAATATTTTTATTTTTTCAAAT : 1200
Hmus : TTTTCAATTGTTTGATTAATTATTATTTATATTTTTTACTTTAGATTAATTGTAAATATTTTTATTTTTTCTTAT : 1200
Pser : TTCTCTATTATCTGACTTACTATTACTTATATTTTTTATTTTAGATTGATTGTTCAGATTTTTATTTTTTCGAAT : 1200
Adea : TTCTCTGTTGTTTGATTAGTAATAATTTATATTTTTATTTTAGACTTATTGTTAATCTTTTTATATTTTCGTAT : 1200
Ades : TTCTCAGTTGTTTGACTCGTTGTTATCTACATTTTTTACTTCAGGCTTATAGTGAATATTTTTATTTTCTCAAAT : 1200
Scul : TTTTCGGTTATTTGATTGATTATTATTTATATTTTTTACTTTAGACTCATTATTAAATTATATTTTATATTTTCAAAT : 1200
Sgal : TTCTCTGTGGCTATGGCTAGTTATTATATTATATTTTTATTTTAGATTAATTAATTATATTTTTATATTTTCAAAT : 1200
Sonc : TTTtCGGTTGTTTGACTTGTCATTATTTATATTTTTTATTTTCGTTTGATTAATTAAATTATTTTATATTTTCTTAT : 1200
TT tC T T TG T t aTtAt TAtATTTTtTA TTt G T AT T aAtaTtTTTAT TTtTC tAt
1220 1240 1260
Ltar : CAATTCATTAGGATTTTGGGTTGTGCGTCTTCATATTTTAGGATTTTCAAATCTATTTTTTATATTAAGTTTTAGC : 1275
Esch : CAATTTATTAGGTTTTTGAGTTATTCGTTTACATATAAATAAATTTTAGTTAATCTTTTTTTTTTCTAAGTTGTTAGT : 1275
Lcos : CAATTTATTGGTTTTTGAATTGTACGACTAAATACAATATATATGAGCAATACTTTTTTTTGACTTAGTTGCTTT : 1275
Caca : CAGTTTATTGGTTTTTGAGTTGTTCGTTTGCACCTAATTGGTTTTGCAAATACAATTTTTTTTTTTAAGCTGCAGT : 1275
Hmus : CAATTTTTAGGTTTTTGAGTTGTACGACTACATTTATATAGTTGTTATTAATATAGCATTTTTGGTTTGCTTTTTT : 1275
Pser : CAGTTTACAGGATTTTGAGTTGGTCCGACTTTACACTTTGACTACTTTACATTTATTATTTTACATTTCAATTTTA : 1275
Adea : CAGTTCGTTGGTTTTGGGTTGTAAGATTGCATTTTACTTTACATTTCAAACCTTATATTTTTTATTAGTAGTTGC : 1275
Ades : CAGTTCGTTGGGTTTCTGAGTAGTCAGACTACACCTGAACTTTTTGATTAATCTTGTTTTCATTTGTTCTTTACA : 1275



3.SI-S4a: GeneDoc MURF1 / ND2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1.





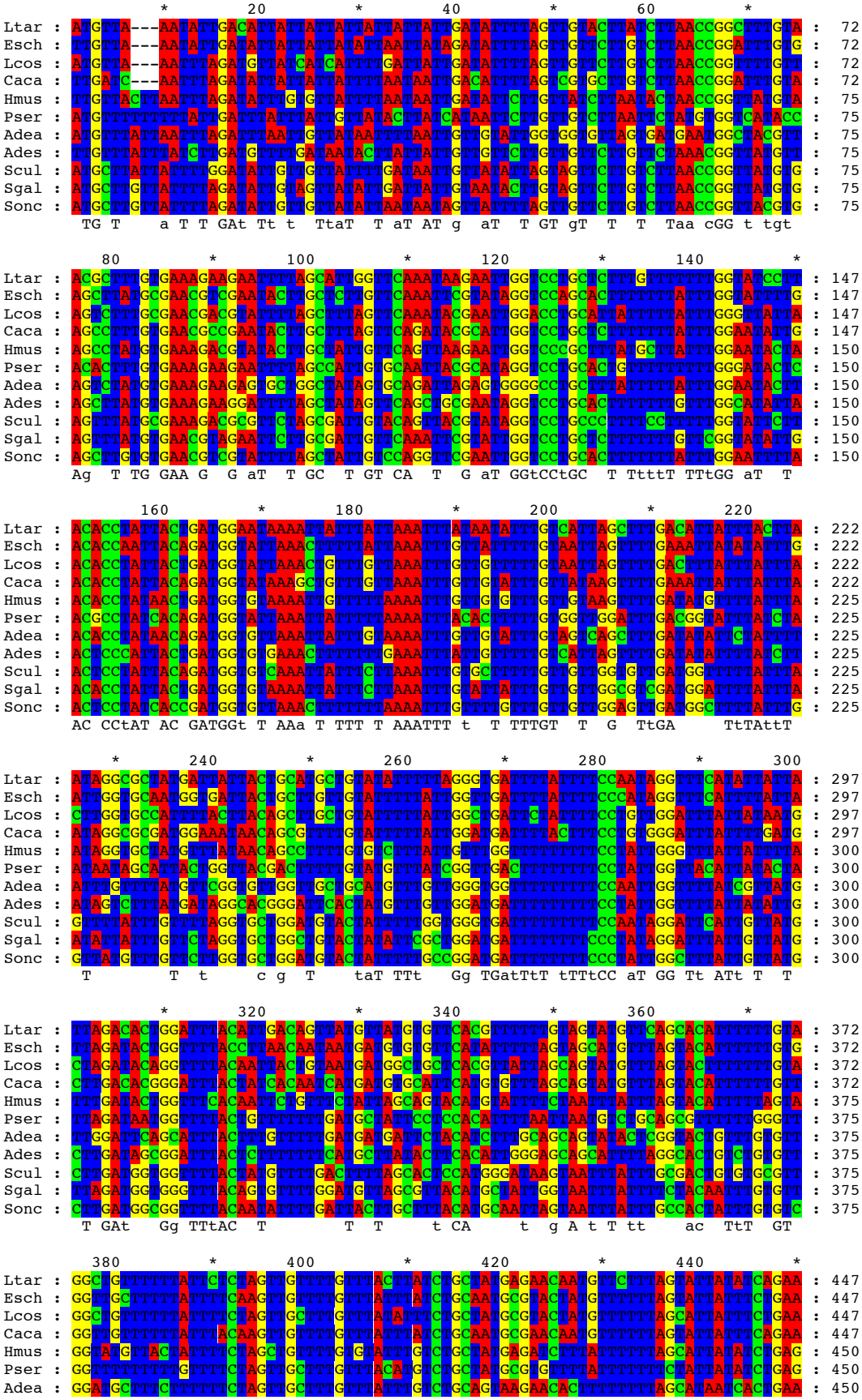
3.SI-S4b: GeneDoc MURF1 / ND2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment.

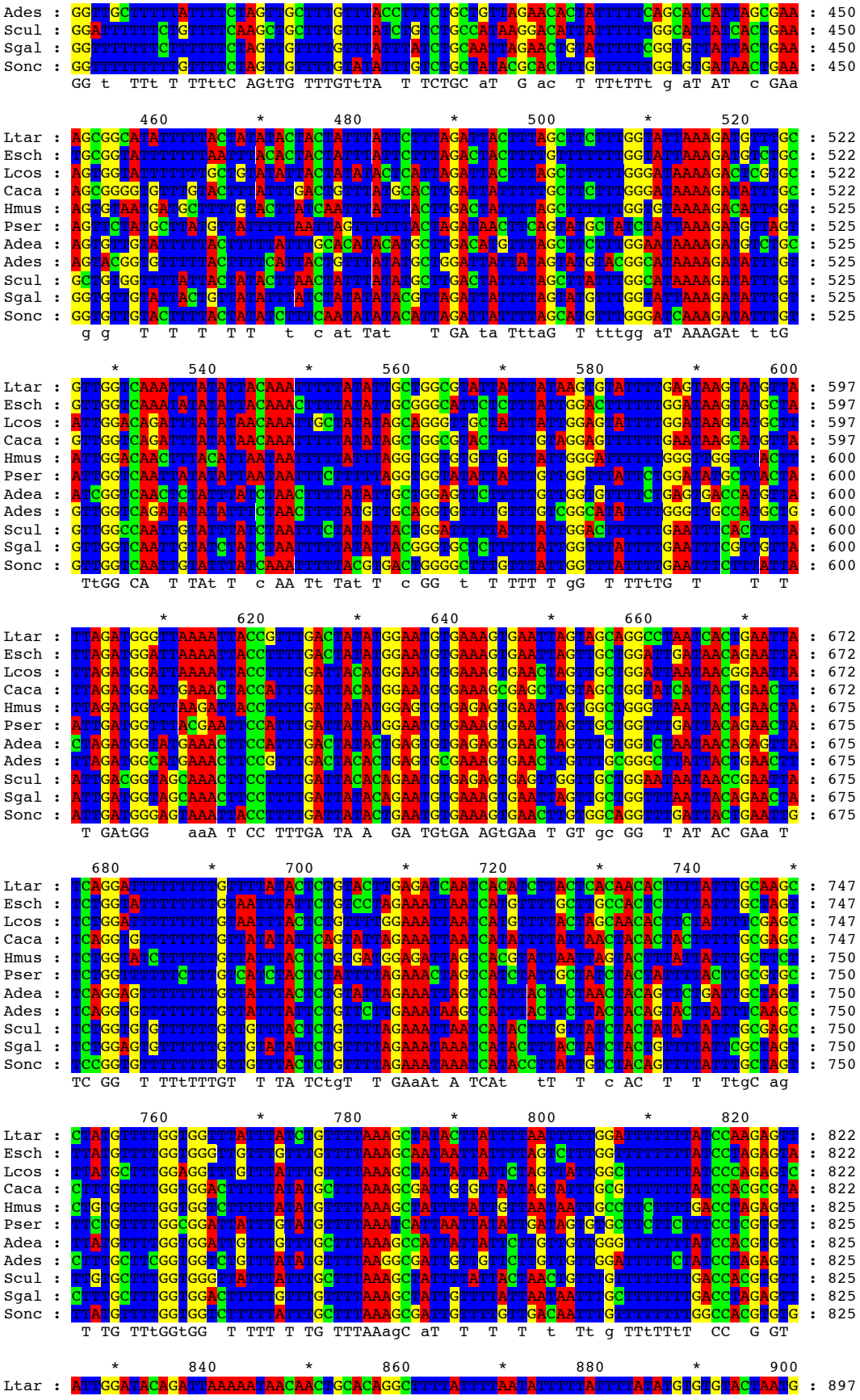
3.SI-S4c: MURF1 nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

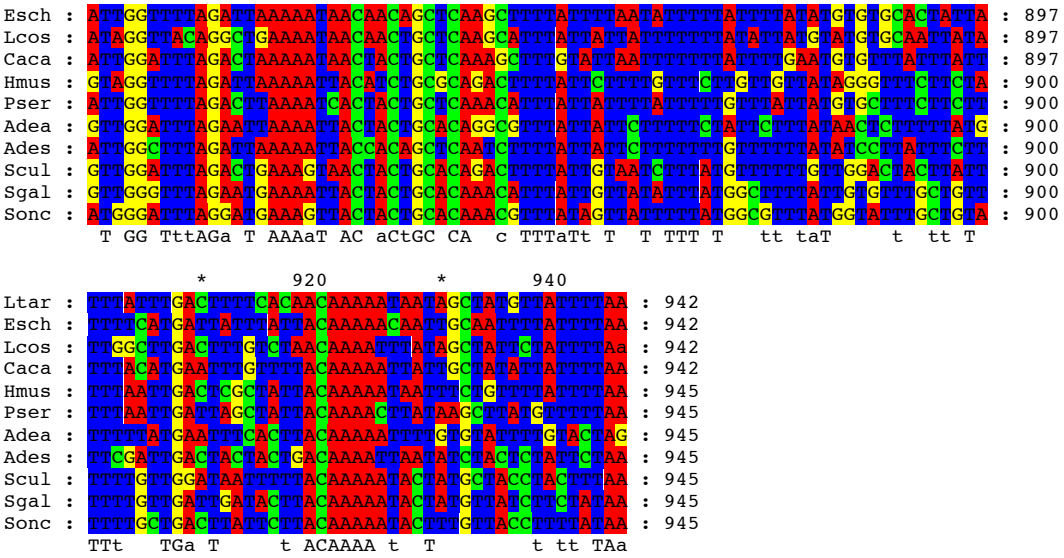
	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	82	**100**									
L. cos	77	78	**100**								
C. aca	79	81	77	**100**							
H. mus	77	78	74	78	**100**						
P. ser	71	72	69	70	73	**100**					
A. dea	71	71	70	73	70	66	**100**				
A. des	71	70	71	70	71	66	70	**100**			
S. cul	75	74	74	76	75	69	70	70	**100**		
S. gal	76	78	73	77	78	72	71	70	80	**100**	
S. onc	76	75	73	78	76	70	71	70	81	83	**100**

3.SI-S4d: MURF1 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

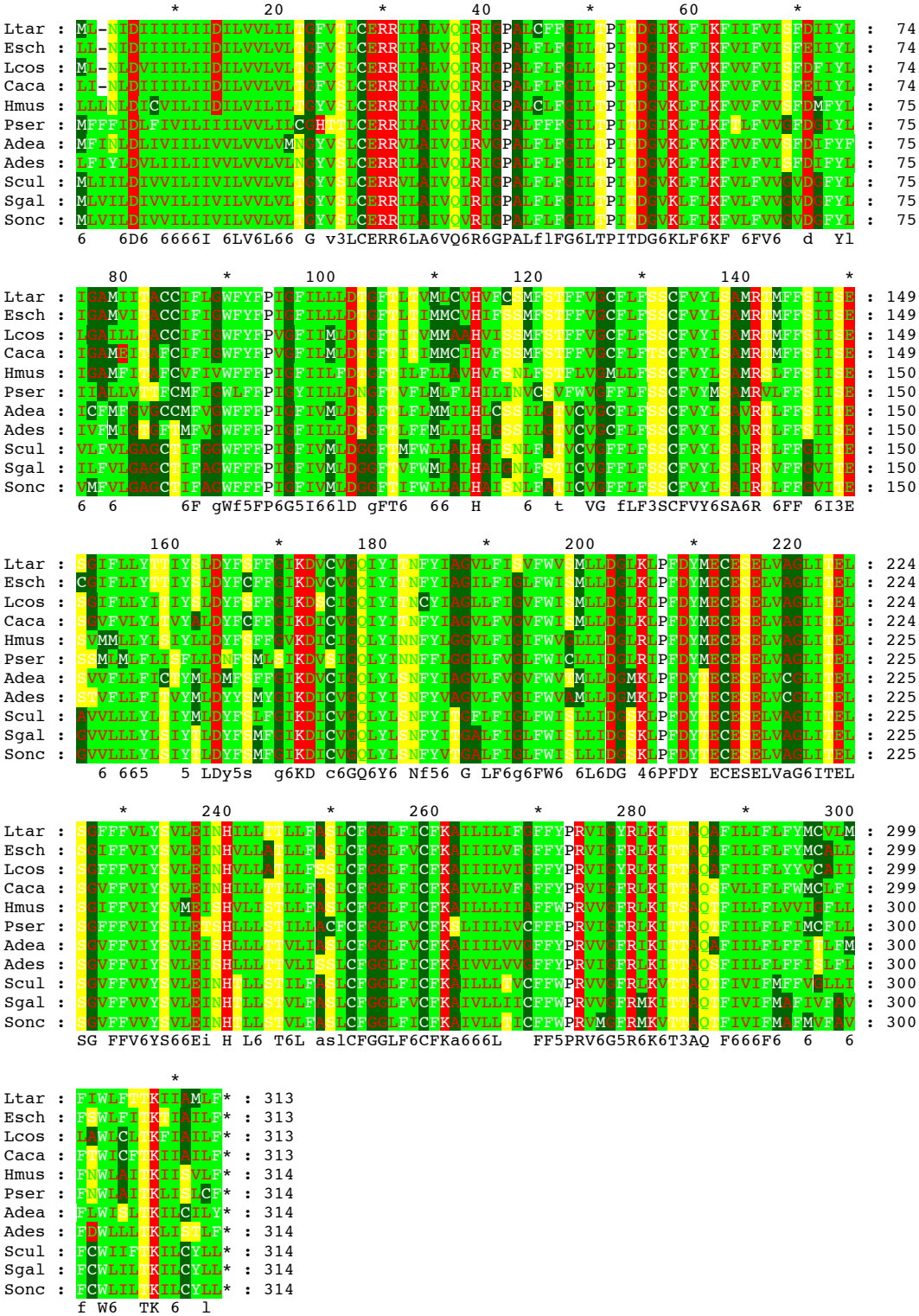
	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	91	**100**									
L. cos	85	88	**100**								
C. aca	86	89	85	**100**							
H. mus	80	81	78	81	**100**						
P. ser	71	71	67	69	71	**100**					
A. dea	73	73	72	75	72	65	**100**				
A. des	73	73	73	76	74	64	78	**100**			
S. cul	80	79	78	81	80	68	72	72	**100**		
S. gal	80	81	80	83	81	69	73	73	93	**100**	
S. onc	80	81	79	83	80	69	73	73	93	98	**100**


Ltar
Esch
Lcos
Caca
Hmus
Pser
Adea
Ades
Scul
Sgal
Sonc


Ades : GGTTGCTTTTTATTTCTAGTTGCTTTGTTTACCTTTCTGCTGTTAGAACACTATTTTTCAGCATCATTAGCGAA : 450
Scul : GGATTTTTCTGTTTCAAGCTGCTTTGTTTATCTGTCTGCCATAAGGACATTATTTTTGGCATTATCACTGAA : 450
Sgal : GGTTTTTTCTTTTTCTAGTTGTTTGTTTATTTATCTGCAATTAGAACTGTATTTTTCGGTGTTATTACTGAA : 450
Sonc : GGTTTTTTTTTGTTTCTAGTTGTTTGTATATTTGTCTGCTATACGCACTTTGTTTTTTGGTGTGATAACTGAA : 450
GG t TTt T TTttC AGtTG TTTGTtTA T TCTGC aT G ac T TTtTTt g aT AT c GAa
460 * 480 * 500 * 520
Ltar : AGCGGCATATTTTTACTATATACTACTATTTATTCTTTAGATTACTTTAGCTCTTTGGTATTAAAGATGTTTGC : 522
Esch : AGCGGTATTTTTTTAATTTACACTACTATTTATTCTTTAGACTACTTTGTTTTTTGGTATTAAAGATGTCTGC : 522
Lcos : AGTGGTATTTTTTTGCTGTATATTACTATATACTCATTAGATTACTTTAGCTTTTTGGGATAAAAGACTCGTGC : 522
Caca : AGCGGGGTGTTTGTACTTTATTTGACTGTTTATGCACTTGATTATTTTTGCTTCTTTGGGATAAAAGATATTTGC : 522
Hmus : AGTGTAAAGATGCTTTTGTACTTATCAATTTATTTACTTGACTATTTTAGCTTTTTGGTGTAAAAGACATTTGT : 525
Pser : AGTCTATAGCTTATGTTATTTTTAATTAGTTTTTACTAGATAACTTCAGTATGCTATCTATTAAAGATGTTAGT : 525
Adea : AGTGTTGTATTTTTACTTTTTATTTGCACATACATGCTTGACATGTTTAGCTTCTTTGGAATAAAAGATGTCTGC : 525
Ades : AGTACGGTGTTTTTACTTTTCATTACTGTTTATATGCTTGGATTATTATAGTATGTACGGCATAAAAGATATTTGT : 525
Scul : GCTGTGGTTTTATTACTATACTTAACTATTTATATGCTTGACTATTTTAGCTTTATTTGGCATAAAAGATATTTGT : 525
Sgal : GGTGTTGTATTACTGTTATATTTATCTATATATACGTTAGATTATTTTAGTATGTTTGGTATTAAAGATATTTGT : 525
Sonc : GGTGTTGTACTTTTACTATATCTTTCAATATATACATTAGATTATTTTAGCATGTTTGGGATCAAAGATATTTGT : 525
g g T T T T T t c at Tat T GA ta TttaG T tttgg aT AAAGAt t tG
* 540 * 560 * 580 * 600
Ltar : GTTGGTCAAATTTATATTACAAATTTTTATATTGCTGGCGTATTATTTATTAAGTTGTATTTTGAGTAAGTATTGTTA : 597
Esch : GTTGGTCAAATATATATTACAAACTTTTATATTGCGGGCATTCTCTTTATTGGACTTTTTTGGATAAGTATTGCTA : 597
Lcos : ATTGGACAGATTTATATTAACAAATTTGCTATATAGCAGGGTTTGCTATTTATTTGGAGTATTTTGGATAAGTATTGCTT : 597
Caca : GTTGGTCAGATTTATATTAACAAATTTTTATATAGCTGGCGTACTTTTGTTAGGAGTTTTTTGGAATAAGCATGGTTA : 597
Hmus : ATTGGACAACTTTACATTAATTAATTTTTATTTAGGTGGTGTGTTGTTTATTGGGATTTTTTGGGTTGGTTTACTT : 600
Pser : ATTGGTCAATTATATATTAATTAATTTCTTTTTAGGTGGTATATTATTTGTTGGTTTATTCTGGATATGCTTACTTA : 600
Adea : ATCGGTCAACTCTATTTATCTTAACTTTTATATTGCTGGAGTTCTTTTTGTTGGTGTTTCTGGAGTGACCATGTTA : 600
Ades : GTTGGTCAGATATATATTCTAACTTTTATGTTGCAGGTGTTTGTTTGTTGTCGGCATATTTTGGGTTGCCATGCTG : 600
Scul : GTTGGCCAATTGTATTTATCTAATTTCTATATTACTGGATTTTATTTATTGGACTTTTTTGAATTTCACTTTTTA : 600
Sgal : GTTGGTCAATTGTATCTATCTAATTTTTATATTACGGGTGCTCTTTTTATTGGTTTATTTTGAATTTCGTTGCTA : 600
Sonc : GTTGGTCAATTGTATTTATCAAATTTTTACGTGACTGGGGCTTTGTTTATTGGTTTATTTTGAATTTCTTTATTA : 600
TtGG CA T TAt T c AA Tt Tat T c GG t T TTT T gG T TTtTG T T T
* 620 * 640 * 660 *
Ltar : TTAGATGGCTTAAAATTACCGTTTGACTATATGGAATGTGAAAGTGAATTACTAGCAGGCCTAATCACTGAATTA : 672
Esch : TTAGATGGATTAAAATTACCTTTTGACTATATGGAATGTGAAAGTGAATTACTTGCTGGATTGATAACAGAATTA : 672
Lcos : TTAGATGGATTAAAATTACCTTTTGATTACATGGAATGTGAAAGTGAACTACTTGCTGGATTAATAACGGAATTA : 672
Caca : TTAGATGGATTGAAACTACCATTTGATTACATGGAAATGTGAAAGCGAGCTTGCTAGCTGGTATCATTACTGAACTT : 672
Hmus : TTAGATGGTTTAAGATTACCTTTTGATTATATGGAGTGTGAGAGTGAATTAGTGGCTGGGTTAATTACTGAACTA : 675
Pser : ATTGATGGTTTACGAATTCCATTTGATTATATGGAATGTGAAAGTGAATTAGTTGCTGGTTTGATTACAGAACTA : 675
Adea : CTAGATGGTATGAAACTTCCATTTGACTATACTGAGTGTGAGAGTGAACTACTTTGTGGTCTAATAACAGAGTTA : 675
Ades : TTAGATGGCATGAAACTTCCGTTTGACTACACTGAGTGCGAAAGTGAACTTGCTTTGCGGGCTTATTACTGAACTT : 675
Scul : ATTGACGGTAGCAAACTTCCTTTTGATTATACACAGAATGTGAGAGTGAGTTGGTTGCTGGAATAATAACCGAATTA : 675
Sgal : ATTGATGGTAGCAAACTTCCTTTTGATTATACAGAATGTGAAAGTGAATTAGTTGCTGGTTTAATTACAGAACTA : 675
Sonc : ATTGATGGGAGTAAATTACCTTTTGATTATACTGAATGTGAAAGTGAACTTGTGGCAGGTTTGATTACTGAATTG : 675
T GAtGG aaA T CC TTTGA TA A GA TGtGA AGtGAa T GT gc GG T AT AC GAa T
680 * 700 * 720 * 740 *
Ltar : TCAGGATTTTTTTTTGTTTTATACTCTGTACTTGAGATCAATCACATCTTACTCACAACACTTTTATTTGCAAGC : 747
Esch : TCTGGTATTTTTTTTGTAATTTATTCTGTCCTAGAAATTAATCATGTTTTGCTTGCCACTCTTTTATTTGCTAGT : 747
Lcos : TCTGGATTTTTTTTTGTAATTTACTCTGTTTTGGAAATTAATCATGTTTTACTAGCAACACTTCTATTTTCGAGC : 747
Caca : TCAGGTGTTTTTTTTGTTATATTATTCAGTATTAGAAATTAATCATGTTTTATTAACTACACTATTTTGTCGAGC : 747
Hmus : TCTGGTATCTTTTTTGTTATTTACTCTGTGATGGAGATTAGTCACGTATTTAATTAGTACTTTATTATTGCTTCT : 750
Pser : TCTGGTTTTTTCTTTGTCATCTACTCTATTTTAGAAACTAGTCATCTATTGCTATCTACTATTTTACTTGCGAGC : 750
Adea : TCAGGAGTTTTTTTTGTTATTTACTCTGTTATTAGAAATTAGTCATTTACTTCTAACTACAGTTCTGATTGCTAGT : 750
Ades : TCAGGTGTTTTTTTTGTTATTTATTCTGTTCTTGAAATTAAGTCATTTACTTCTTACTACAGTACTTATTTCAAGC : 750
Scul : TCTGGTGTGTTTTTTGTTATTTACTCTGTTTTAGAAATTAATCATACTTTGCTTATCTACTATATTATTGCGAGC : 750
Sgal : TCTGGAGTGTTTTTTGTTGTTATATTCTGTTTTAGAAATTAAATCATACTTTACTTATCTACTGTTTTATTCGCTAGT : 750
Sonc : TCCGGTGTTTTTTTTGTTGTTTACTCTGTTTTAGAAATTAAATCATACCTTATTGTCTACAGTTTTATTTGCTAGT : 750
TC GG T TTtTTTGT T TA TCtgT T GAaAt A TCAt tT T c AC T T TtgC ag
760 * 780 * 800 * 820
Ltar : CTATGTTTTGGTTGGTTTATTTATCTGTTTTAAAGCTTATACTTATTTTAATTTTTGGATTTTTTTATCCAAGAGTT : 822
Esch : TTATGTTTTGGTTGGGTTGTTTGTTTGTTTTAAAGCAAATAATTATTTTAGTCTTTGGTTTTTTTTATCCTAGAGTA : 822
Lcos : TTATGCTTTGGAGGTTTGTTTATTTGTTTTAAAGCTTATTATTATTCTAGTTATTGGCTTTTTTTATCCCAGAGTC : 822
Caca : CTTTGTTTTGGTTGGACTTTTTATATGCTTTAAAGCGAATTGTGTTATTAGTATTGCGTTTTTTTATCCACGCGTA : 822
Hmus : CTGTGTTTTGGTTGGTCTTTTTATATGTTTTAAAGCTTATTTTATTGCTTAATAATTGCCTTCTTTTGACCTTAGACTT : 825
Pser : TTCTGTTTTGGCGGATTATTTGTATGTTTTAAATCATTAATTATATTGATACGTGGCTTCTTCTTTCCTCGTGTT : 825
Adea : TTATGTTTTGGTGGATTGTTTGTTTGCTTTAAAGCCATTATTATTCTTGTTGTTGGGTTTTTTTATCCACGTGTT : 825
Ades : CTTTGCTTTCGGTTGGTCTGTTTATATGTTTTAAGGCGAATTGTTGTTCTTGTTGCTTGGATTTTTCTATCCTAGAGTT : 825
Scul : TTGTGCTTTGGTTGGGTTATTTATTGCTTTAAAGCTTATTTTATTACTAACTGCTTGTTTTTTTGACCACGTGTT : 825
Sgal : CTTTGCTTTGGTTGGACTTTTTGTTGTTTTAAAGCTTATTGTTTTATTAATAATTTGCTTTTTTTGACCTTAGAGTT : 825
Sonc : TTATGTTTTGGTTGGTCTTTTTATTTGCTTTAAAGCGATTGTTTTGCTTGACAATTTGTTTTTTTTGGCCACGTGTG : 825
T TG TTtGGtGG T TTT T TG TTTAAagC aT T T T t Tt g TTtTTtT CC G GT
* 840 * 860 * 880 * 900
Ltar : ATTGGATACAGATTAAAAATAACAACTGCACAGGCTTTTATTTTAATATTTTTATTTTATATGTGTGTACTAATG : 897



3.SI-S5a: GeneDoc ND1 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. The absence of the third codons in *L. tar, C. aca, L. cos* and *E. sch* are shown as gaps in the alignment.



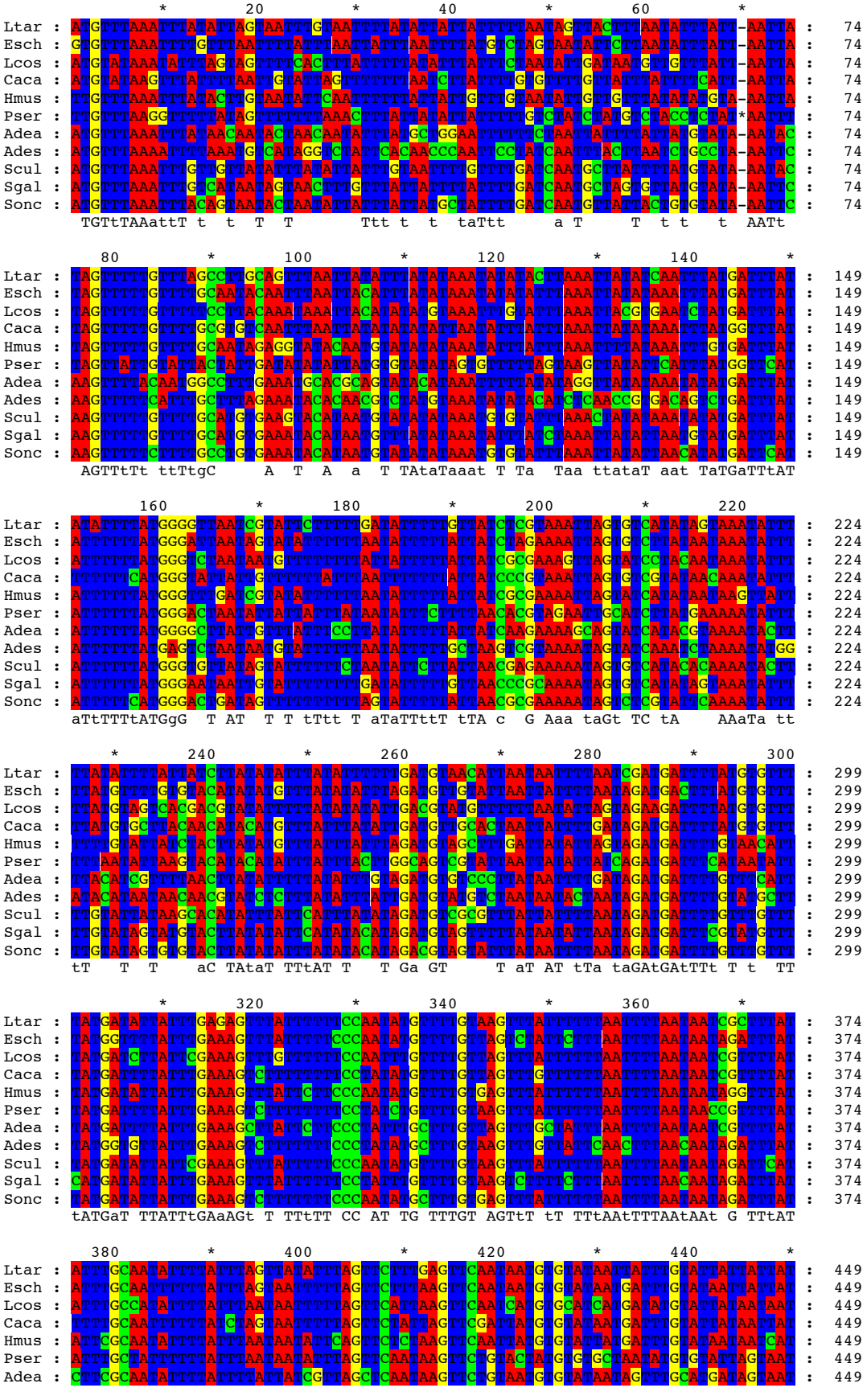
3.SI-S5b: GeneDoc ND1 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment.

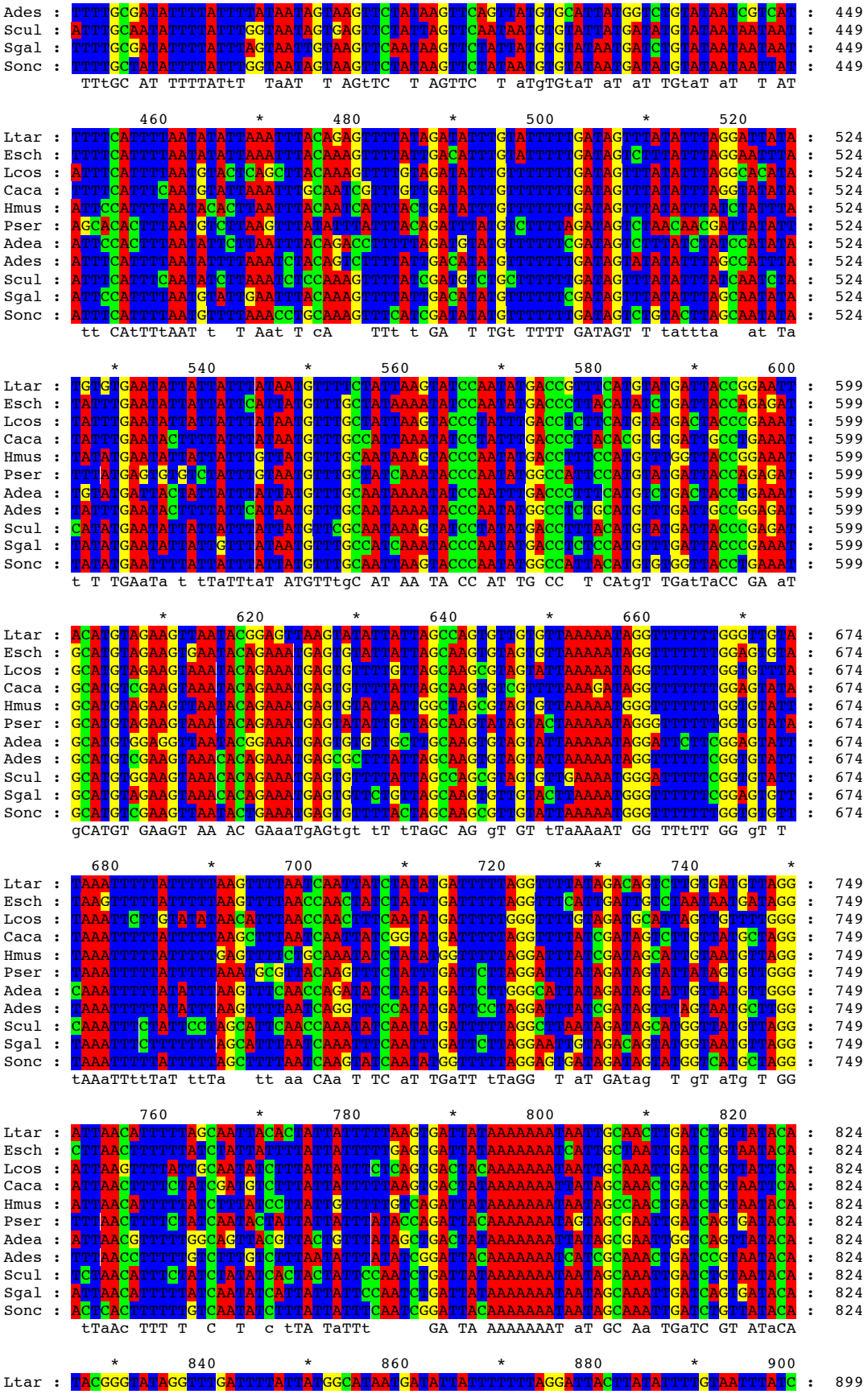
3.SI-S5c: ND1 nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

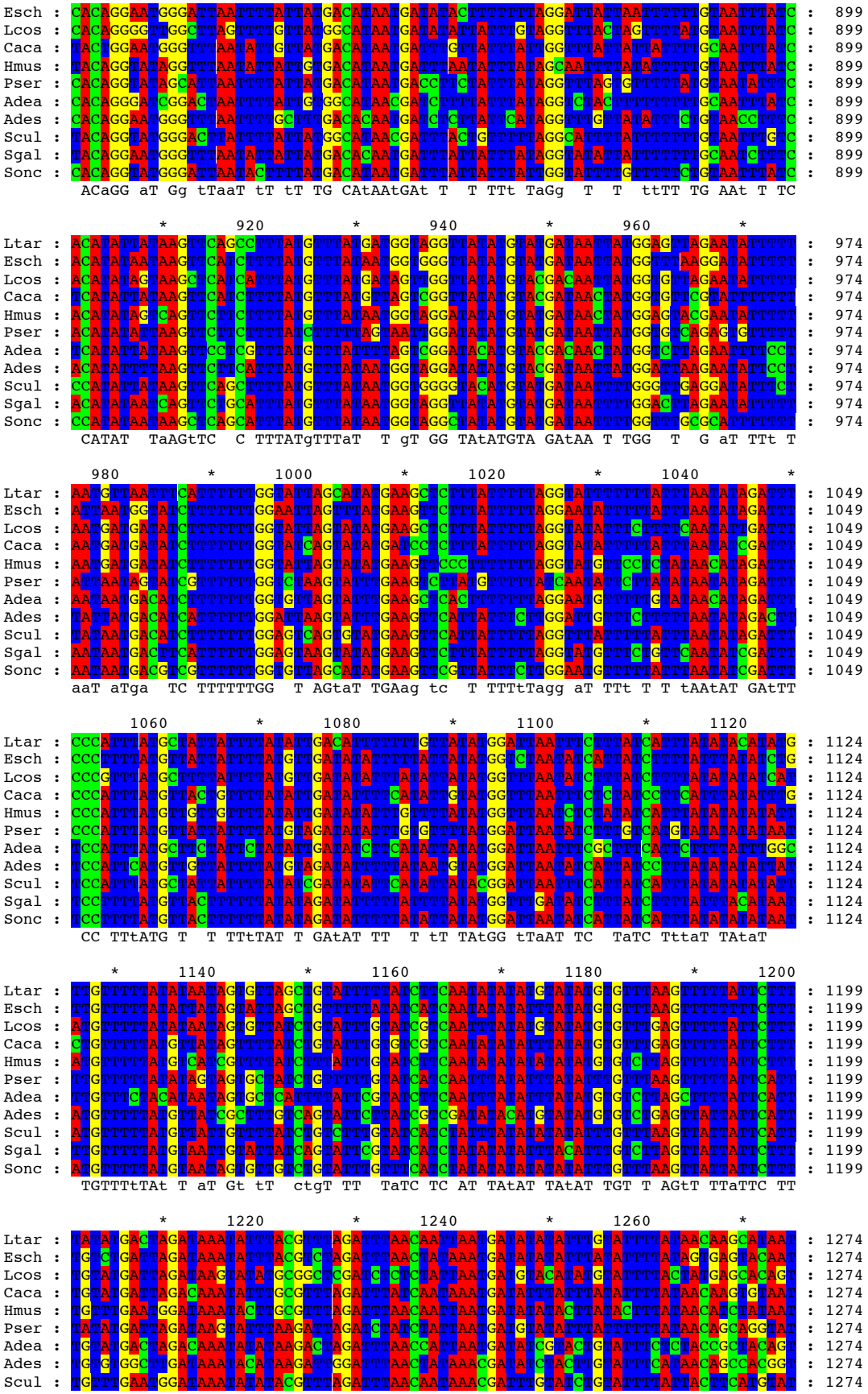
	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	81	**100**									
L. cos	78	81	**100**								
C. aca	78	79	78	**100**							
H. mus	72	72	72	72	**100**						
P. ser	68	69	68	66	69	**100**					
A. dea	70	69	70	70	67	65	**100**				
A. des	70	70	71	71	68	67	73	**100**			
S. cul	68	69	69	69	69	67	71	69	**100**		
S. gal	69	70	69	70	70	68	71	70	81	**100**	
S. onc	67	69	70	70	70	66	69	69	79	83	**100**

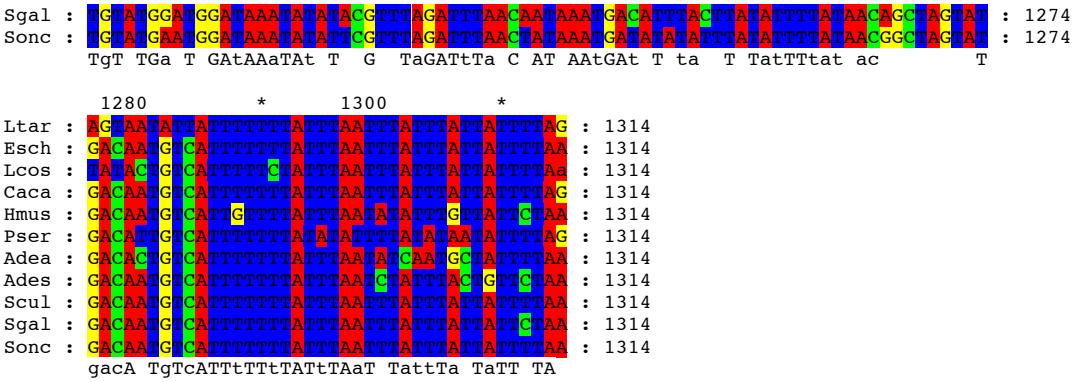
3.SI-S5d: ND1 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	89	**100**									
L. cos	83	85	**100**								
C. aca	83	87	83	**100**							
H. mus	73	73	71	73	**100**						
P. ser	65	65	62	61	67	**100**					
A. dea	68	69	69	70	66	61	**100**				
A. des	71	71	70	73	70	65	81	**100**			
S. cul	66	67	67	69	70	61	72	70	**100**		
S. gal	67	68	68	67	70	63	71	69	89	**100**	
S. onc	65	66	67	67	70	61	69	69	91	96	**100**

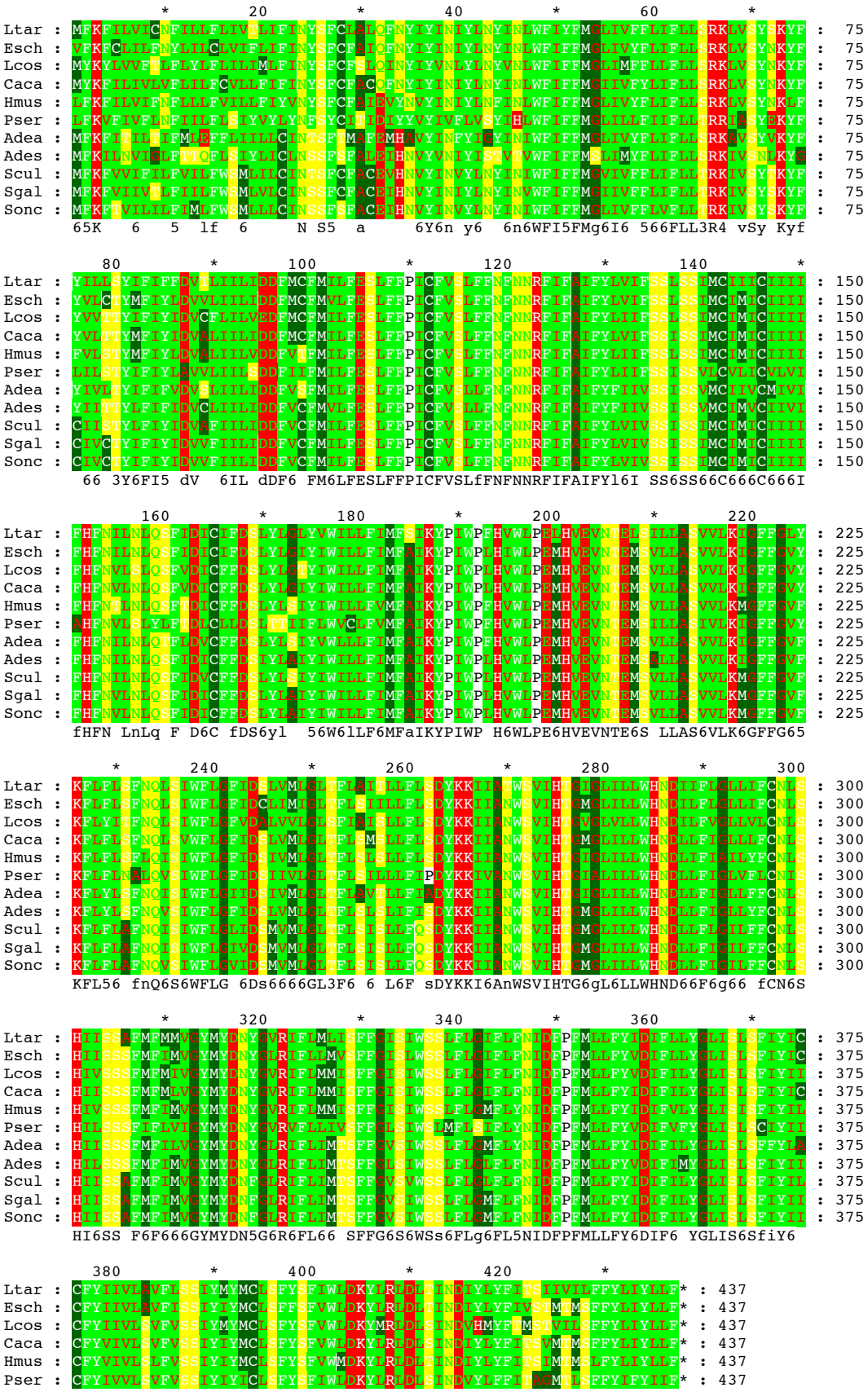


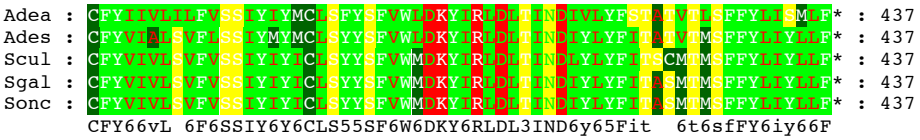

Ades : 449
Scul : 449
Sgal : 449
Sonc : 449
TTtGC AT TTTTATtT TaAT T AGtTC T AGTTC T aTgTGtaT aT aT TGtaT aT T AT
460 * 480 * 500 * 520
Ltar : 524
Esch : 524
Lcos : 524
Caca : 524
Hmus : 524
Pser : 524
Adea : 524
Ades : 524
Scul : 524
Sgal : 524
Sonc : 524
tt CAtTTtAAT t T Aat T cA TTt t GA T TGt TTTT GATAGT T tattta at Ta
* 540 * 560 * 580 * 600
Ltar : 599
Esch : 599
Lcos : 599
Caca : 599
Hmus : 599
Pser : 599
Adea : 599
Ades : 599
Scul : 599
Sgal : 599
Sonc : 599
t T TGAaTa t tTaTTtaT ATGTTtgC AT AA TA CC AT TG CC T CAtgT TGatTaCC GA aT
* 620 * 640 * 660 *
Ltar : 674
Esch : 674
Lcos : 674
Caca : 674
Hmus : 674
Pser : 674
Adea : 674
Ades : 674
Scul : 674
Sgal : 674
Sonc : 674
gCATGT GAaGT AA AC GAaaTgAGtgt tT tTaGC AG gT GT tTaAAaAT GG TTtTT GG gT T
680 * 700 * 720 * 740 *
Ltar : 749
Esch : 749
Lcos : 749
Caca : 749
Hmus : 749
Pser : 749
Adea : 749
Ades : 749
Scul : 749
Sgal : 749
Sonc : 749
tAAaTTttTaT ttTa tt aa CAa T TC aT TGaTT tTaGG T aT GAtag T gT aTg T GG
760 * 780 * 800 * 820
Ltar : 824
Esch : 824
Lcos : 824
Caca : 824
Hmus : 824
Pser : 824
Adea : 824
Ades : 824
Scul : 824
Sgal : 824
Sonc : 824
tTaAc TTT T C T c tTA TaTTt GA TA AAAAAAAT aT GC Aa TGaTC GT ATaCA
* 840 * 860 * 880 * 900
Ltar : 899





3.SI-S6a: GeneDoc ND4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. The T residue of *P. ser* at position 70 that is causing a frame-shift and should have been removed is denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position to maintain the reading frame.





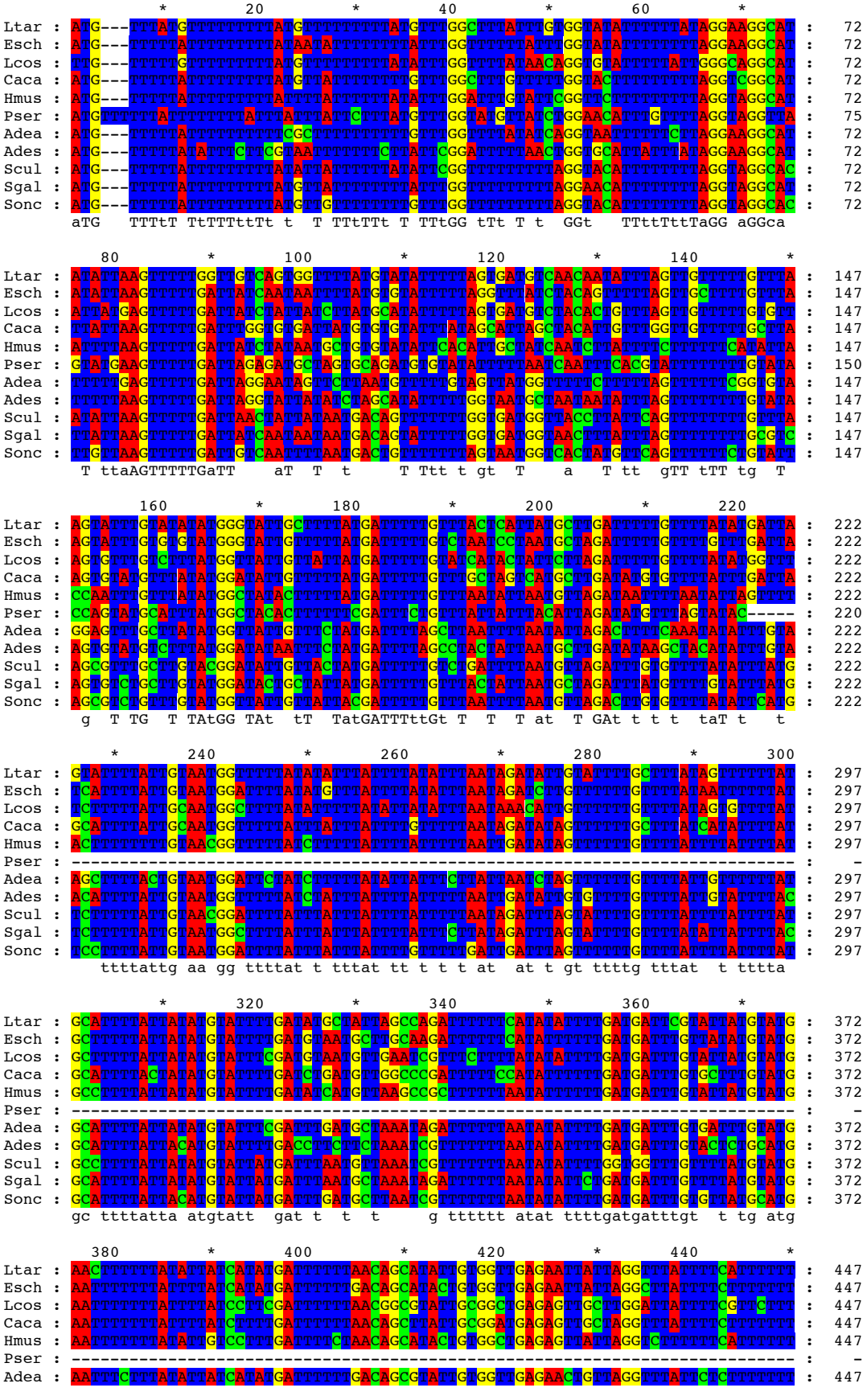
3.SI-S6b: GeneDoc ND4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment.

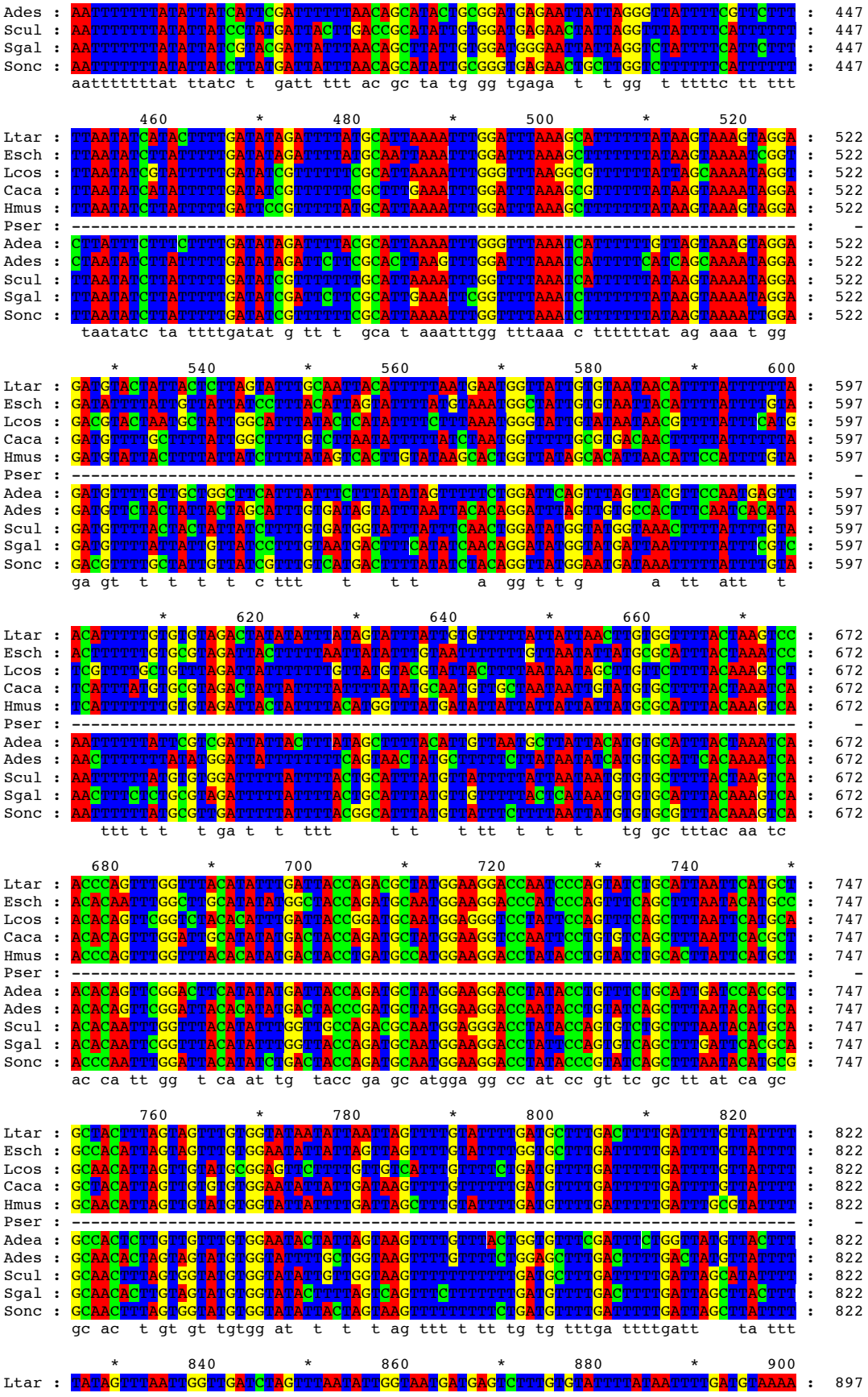
3.SI-S6c: ND4 nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	85	**100**									
L. cos	80	81	**100**								
C. aca	82	84	80	**100**							
H. mus	81	82	78	80	**100**						
P. ser	34	34	32	34	31	**100**					
A. dea	75	74	74	75	75	30	**100**				
A. des	73	75	74	74	76	31	73	**100**			
S. cul	78	80	77	78	80	31	75	75	**100**		
S. gal	79	80	78	79	81	32	75	78	83	**100**	
S. onc	79	81	78	80	80	31	76	78	84	86	**100**

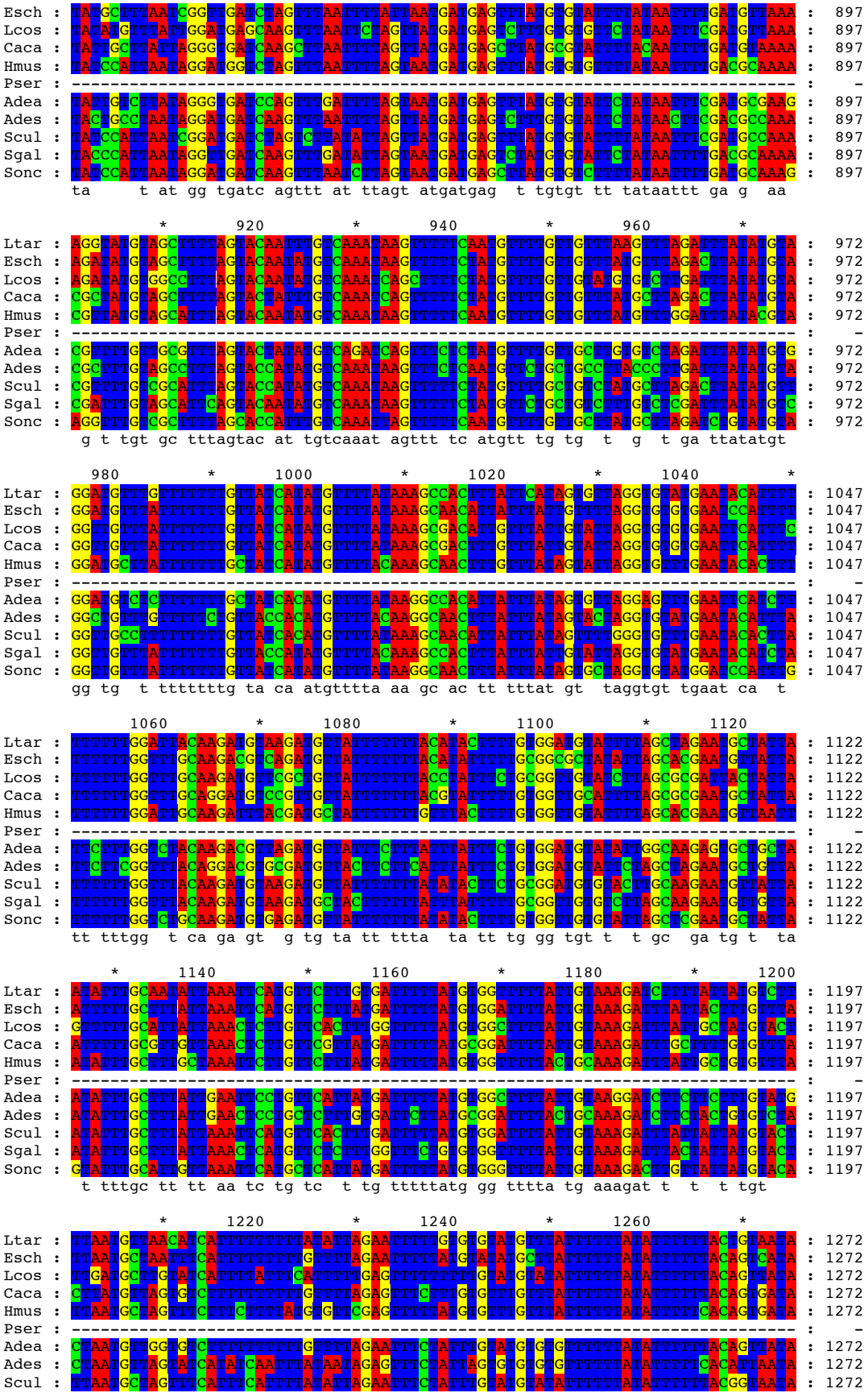
3.SI-S6d: ND4 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

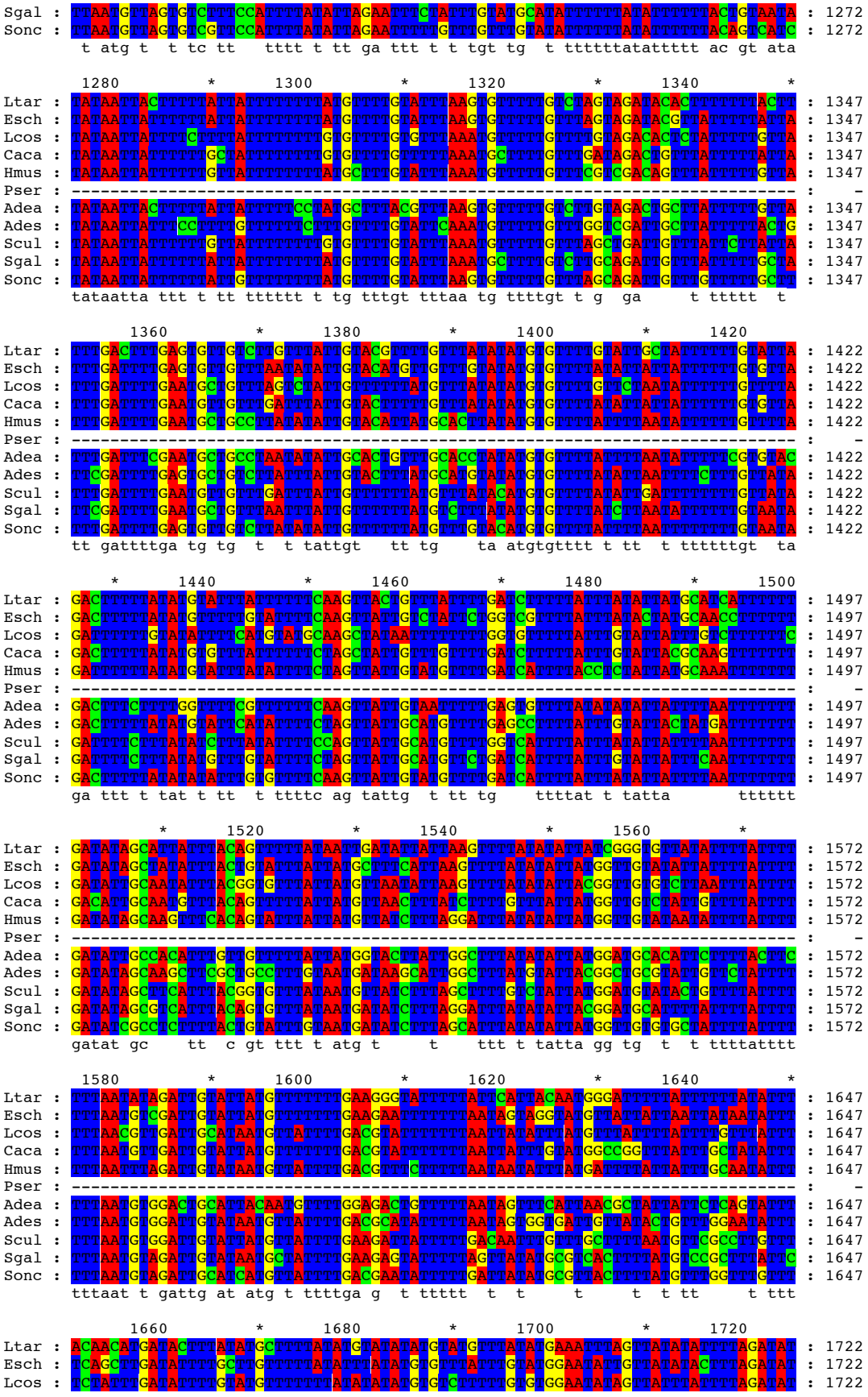
	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	86	**100**									
L. cos	80	81	**100**								
C. aca	86	89	85	**100**							
H. mus	81	83	78	86	**100**						
P. ser	68	70	68	70	71	**100**					
A. dea	77	77	74	78	78	68	**100**				
A. des	76	79	77	80	78	68	81	**100**			
S. cul	78	81	77	84	83	68	81	82	**100**		
S. gal	79	82	78	86	83	70	81	83	94	**100**	
S. onc	79	82	78	84	82	70	81	84	94	97	**100**

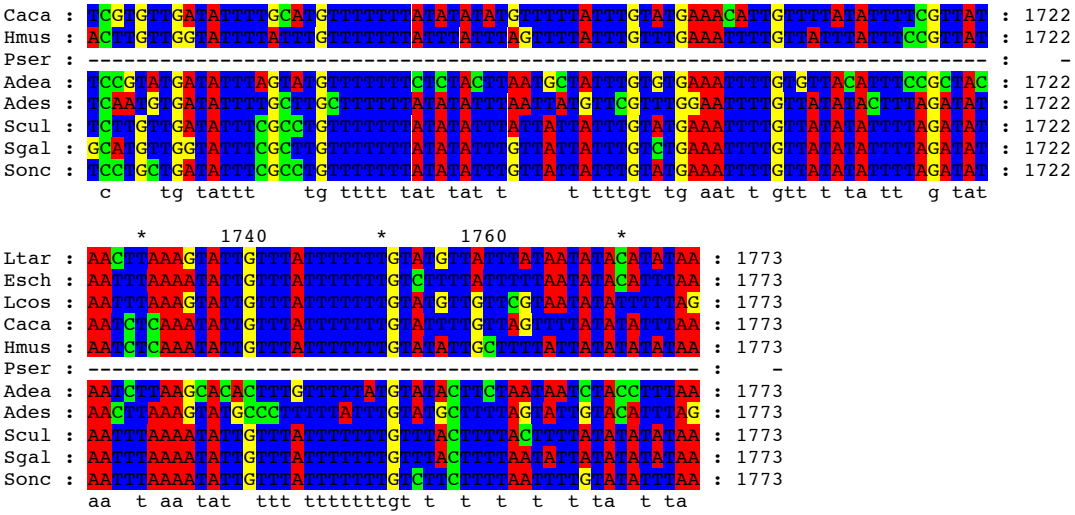




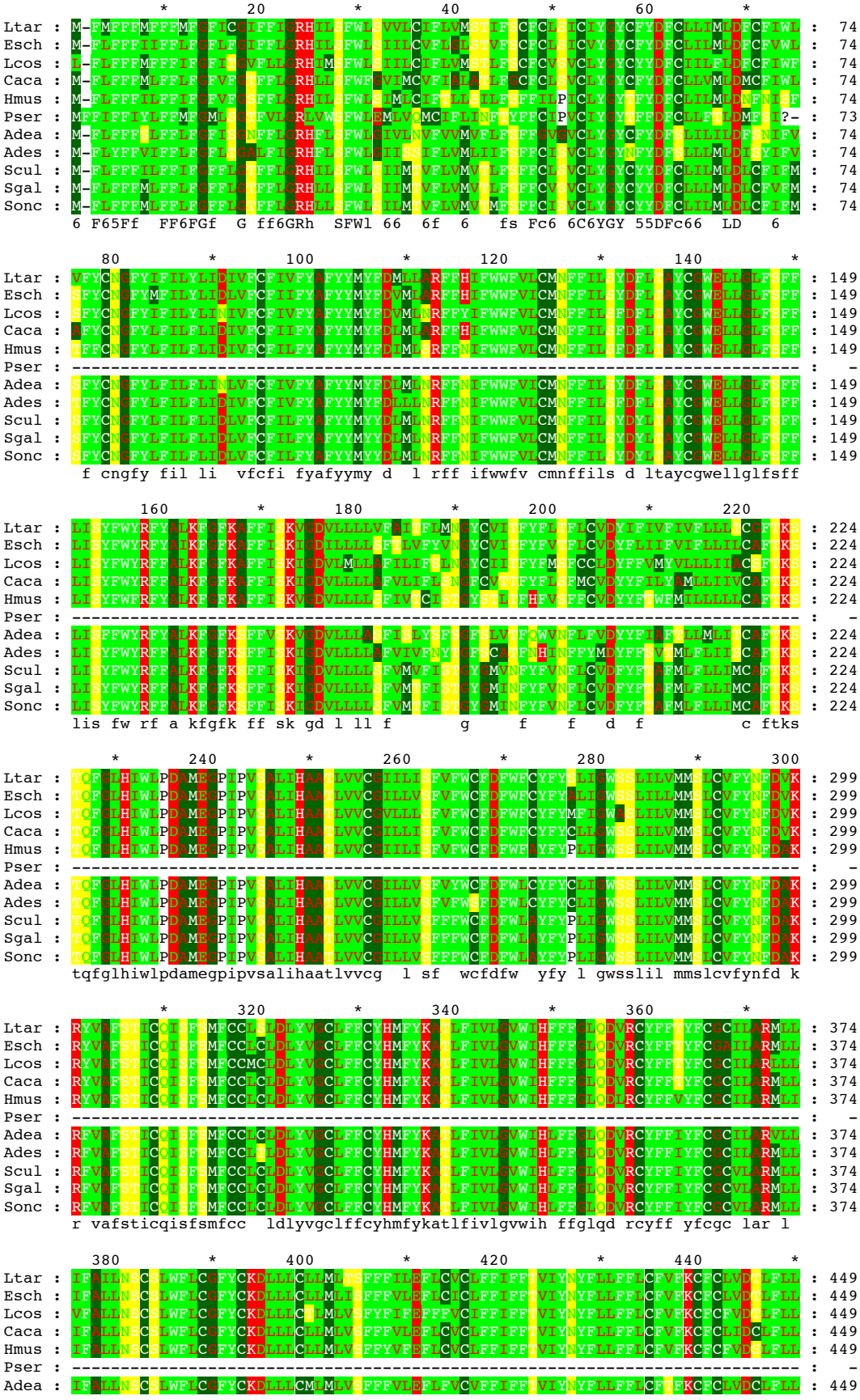
122

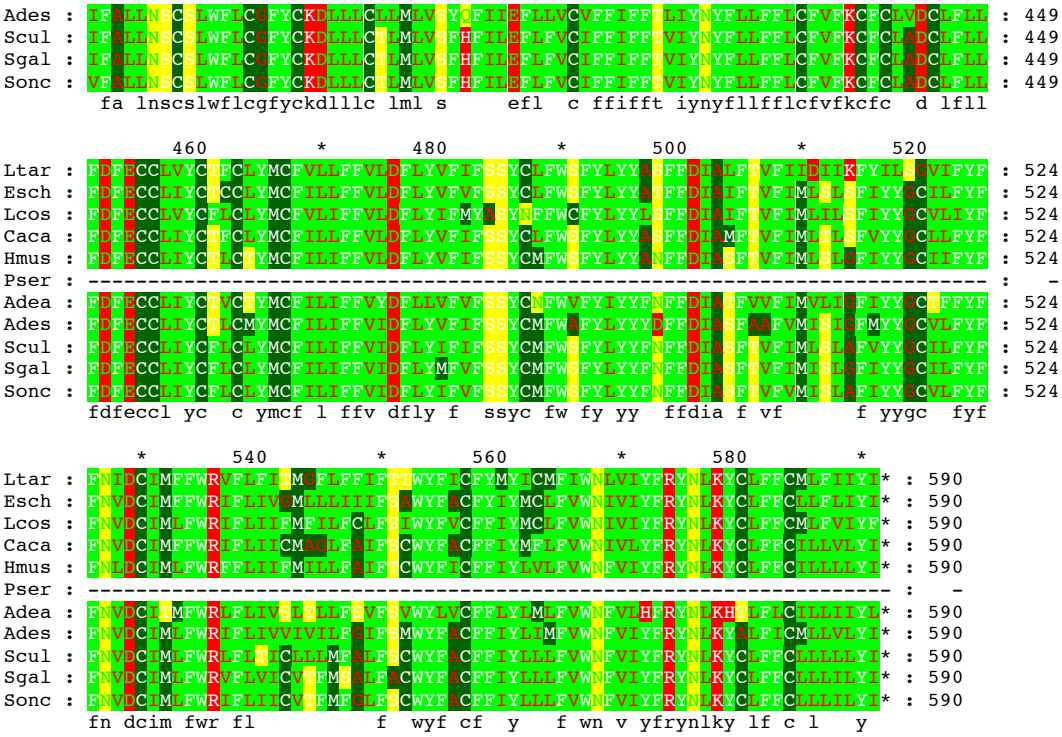

Esch : 897
Lcos : 897
Caca : 897
Hmus : 897
Pser : -
Adea : 897
Ades : 897
Scul : 897
Sgal : 897
Sonc : 897
ta t at gg tgatc agttt at ttagt atgatgag t tgtgt tt tataattt ga g aa
* 920 * 940 * 960 *
Ltar : 972
Esch : 972
Lcos : 972
Caca : 972
Hmus : 972
Pser : -
Adea : 972
Ades : 972
Scul : 972
Sgal : 972
Sonc : 972
g t tgt gc tttagtac at tgtcaaat agttt tc atgtt tg tg t g t ga ttatatgt
980 * 1000 * 1020 * 1040 *
Ltar : 1047
Esch : 1047
Lcos : 1047
Caca : 1047
Hmus : 1047
Pser : -
Adea : 1047
Ades : 1047
Scul : 1047
Sgal : 1047
Sonc : 1047
gg tg t tttttttg ta ca atgtttta aa gc ac tt tttat gt taggtgt tgaat ca t
1060 * 1080 * 1100 * 1120
Ltar : 1122
Esch : 1122
Lcos : 1122
Caca : 1122
Hmus : 1122
Pser : -
Adea : 1122
Ades : 1122
Scul : 1122
Sgal : 1122
Sonc : 1122
tt tttgg t ca ga gt g tg ta tt ttta ta tt tg gg tgt t t gc ga tg t ta
* 1140 * 1160 * 1180 * 1200
Ltar : 1197
Esch : 1197
Lcos : 1197
Caca : 1197
Hmus : 1197
Pser : -
Adea : 1197
Ades : 1197
Scul : 1197
Sgal : 1197
Sonc : 1197
t tttgc tt tt aa tc tg tc t tg ttttatg gg tttta tg aaagat t t t tgt
* 1220 * 1240 * 1260 *
Ltar : 1272
Esch : 1272
Lcos : 1272
Caca : 1272
Hmus : 1272
Pser : -
Adea : 1272
Ades : 1272
Scul : 1272





3.SI-S7a: GeneDoc ND5 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. A gap has been incorporated for the other organisms where an additional codon is encoded for *P. ser.*





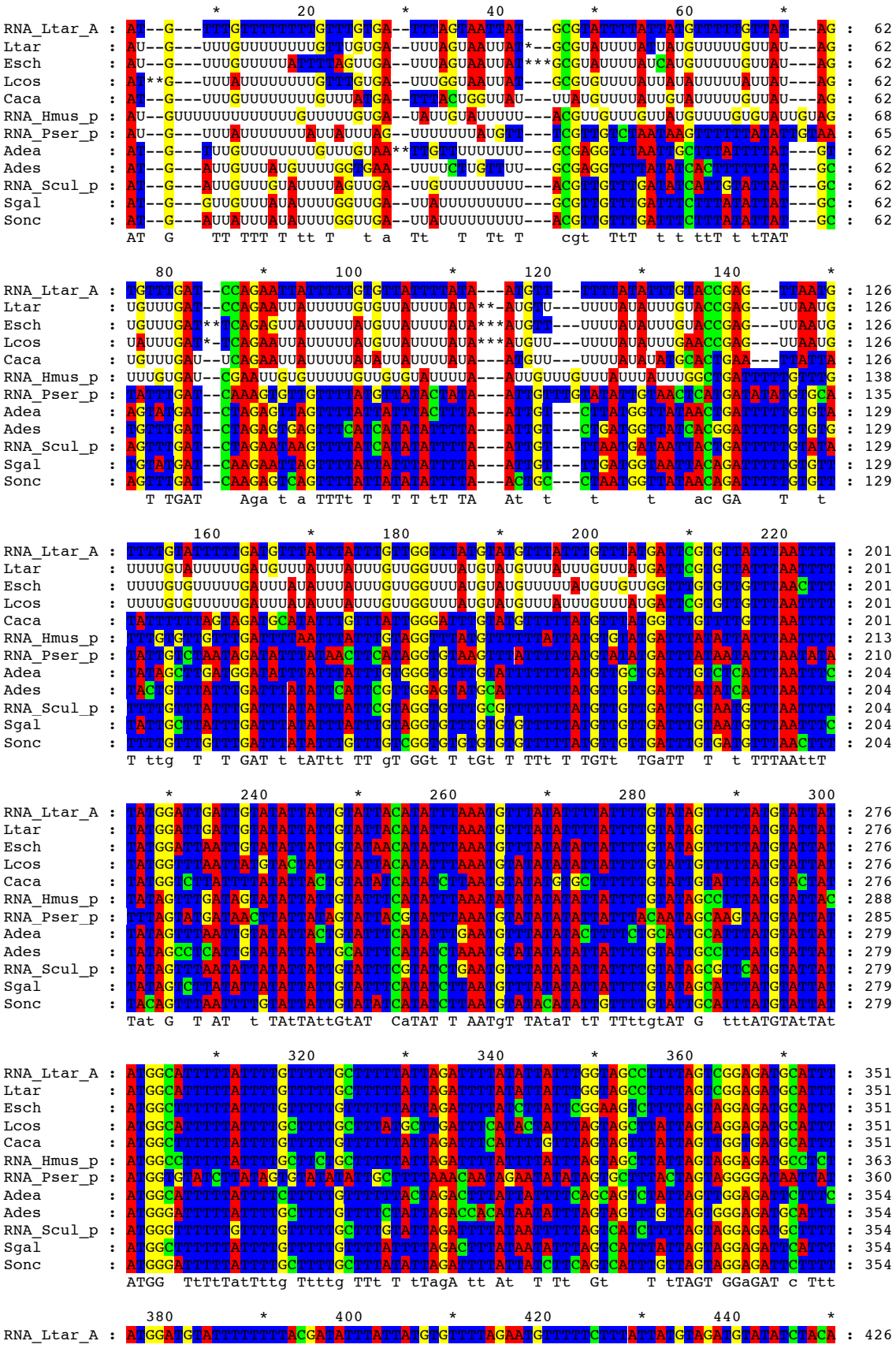
3.SI-S7b: GeneDoc ND5 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. ? in *P. ser*'s alignment refers to incomplete sequence or missing information.

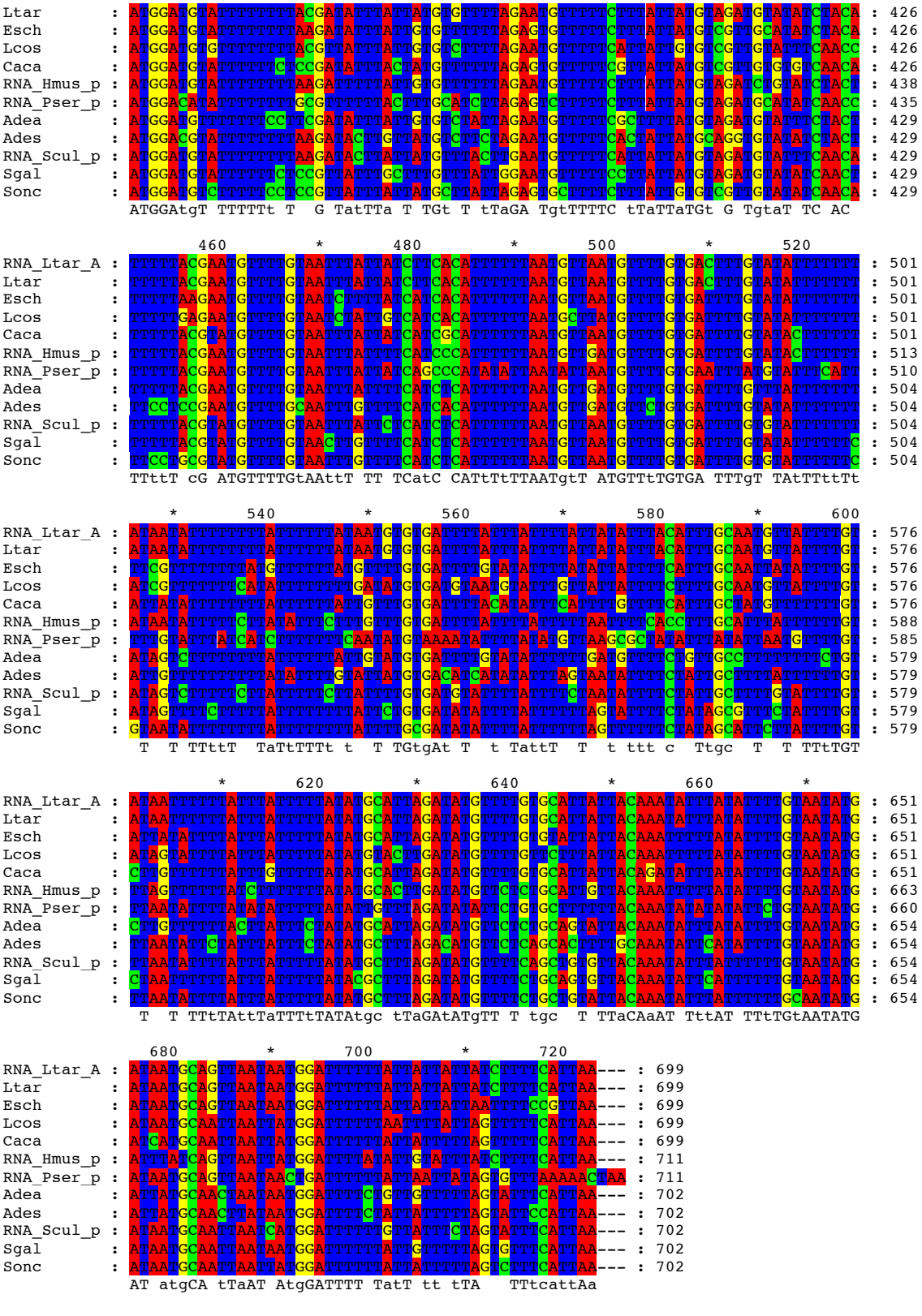
3.SI-S7c: ND5 nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	84	**100**									
L. cos	78	80	**100**								
C. aca	80	83	80	**100**							
H. mus	78	81	77	81	**100**						
P. ser	61	61	58	64	67	**100**					
A. dea	73	76	73	74	74	62	**100**				
A. des	75	75	74	76	75	59	74	**100**			
S. cul	78	81	78	79	80	64	75	77	**100**		
S. gal	78	80	78	79	79	65	76	78	87	**100**	
S. onc	78	80	77	80	79	65	76	77	87	85	**100**

3.SI-S7d: ND5 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

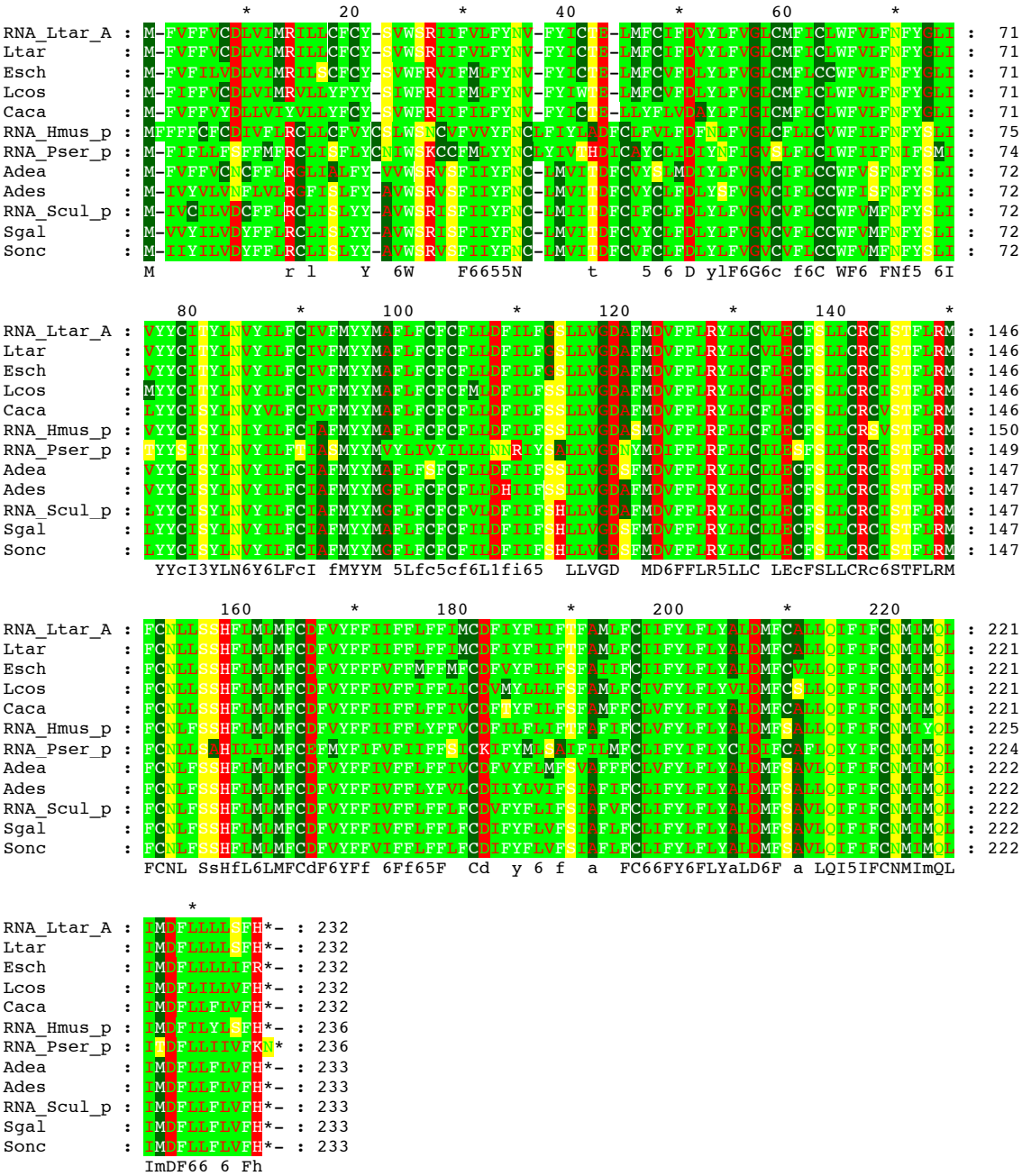
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	85	**100**									
L. cos	81	83	**100**								
C. aca	84	87	83	**100**							
H. mus	81	83	79	84	**100**						
P. ser	46	44	40	51	50	**100**					
A. dea	75	79	76	79	80	46	**100**				
A. des	78	79	78	81	80	44	81	**100**			
S. cul	78	83	81	84	84	47	82	83	**100**		
S. gal	79	82	80	83	83	47	82	83	96	**100**	
S. onc	78	82	81	83	83	50	82	84	96	97	**100**





3.SI-S8a: GeneDoc ATPase 6 / MURF4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position to

maintain the frame. References are cDNA of *L. tar* and RNA sequences of *S. cul*, *H. mus* and *P. ser*.



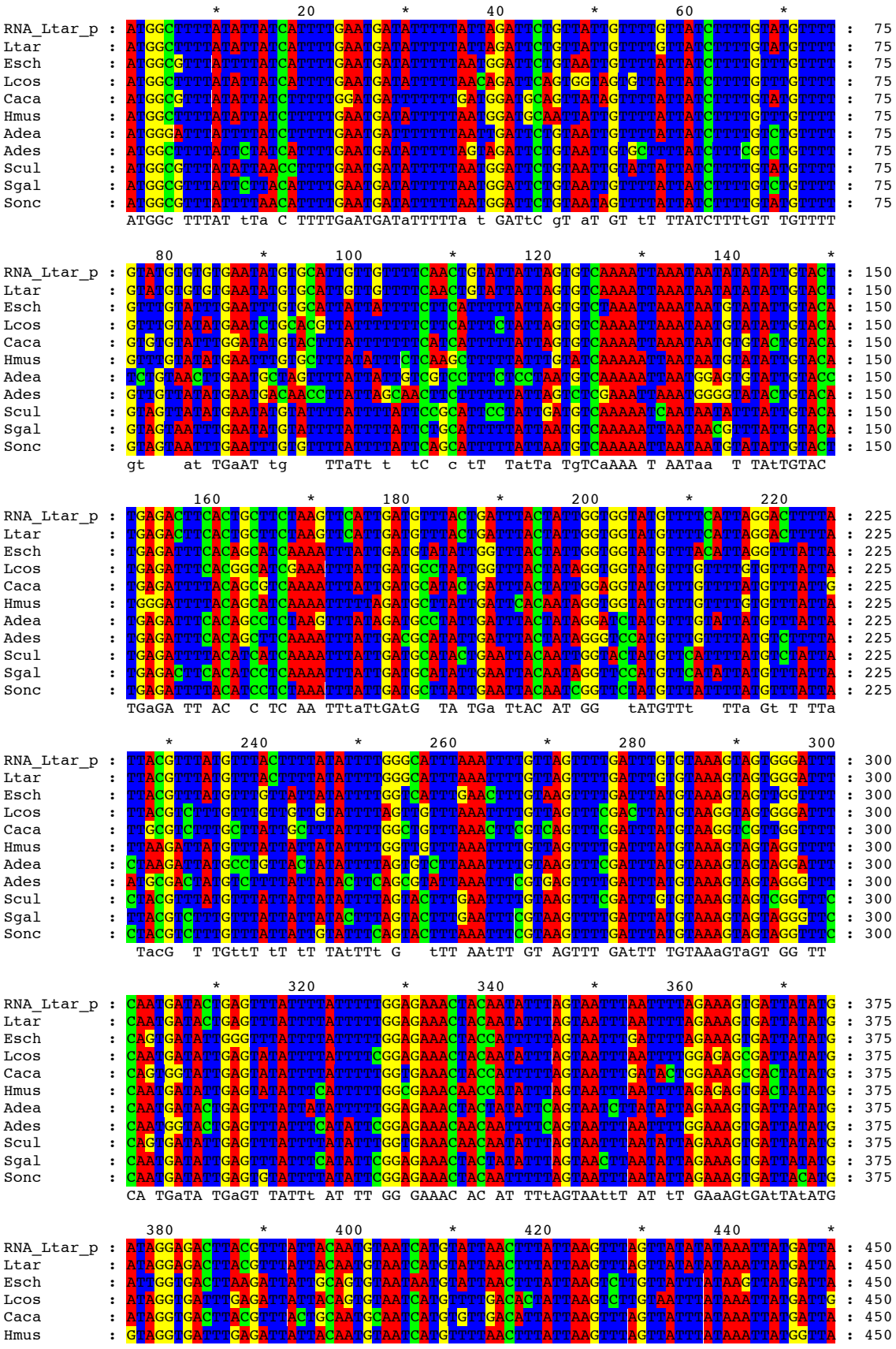
3.SI-S8b: GeneDoc ATPase 6 / MURF4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated and shown.

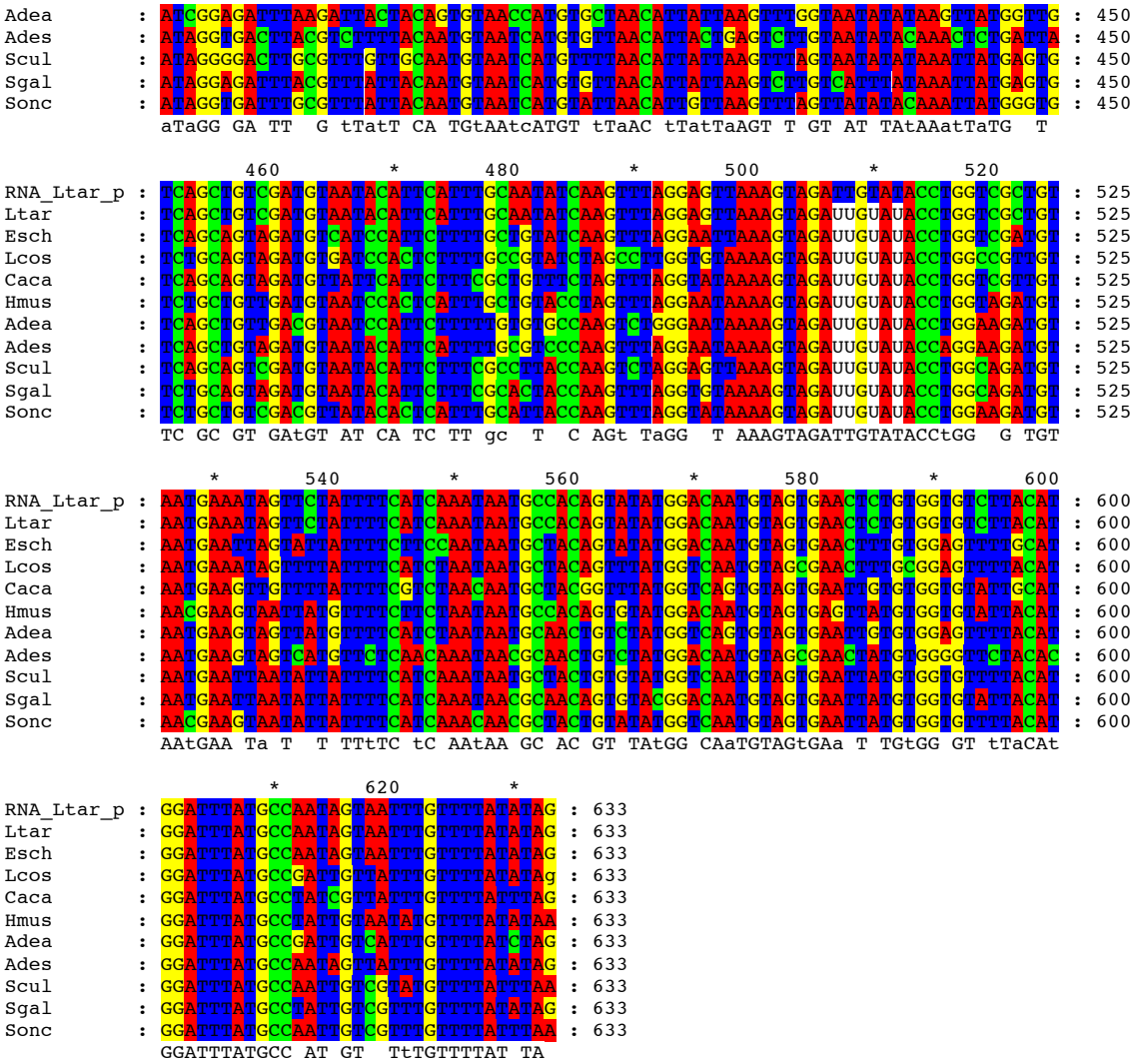
3.SI-S8c: A6 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	91	**100**									
L. cos	88	87	**100**								
C. aca	85	84	82	**100**							
H. mus	82	78	78	77	**100**						
P. ser	70	68	68	66	69	**100**					
A. dea	79	77	76	79	78	68	**100**				
A. des	77	77	75	75	76	68	80	**100**			
S. cul	80	80	78	77	78	71	81	83	**100**		
S. gal	79	79	77	78	78	71	83	82	88	**100**	
S. onc	77	78	78	77	76	69	80	80	86	88	**100**

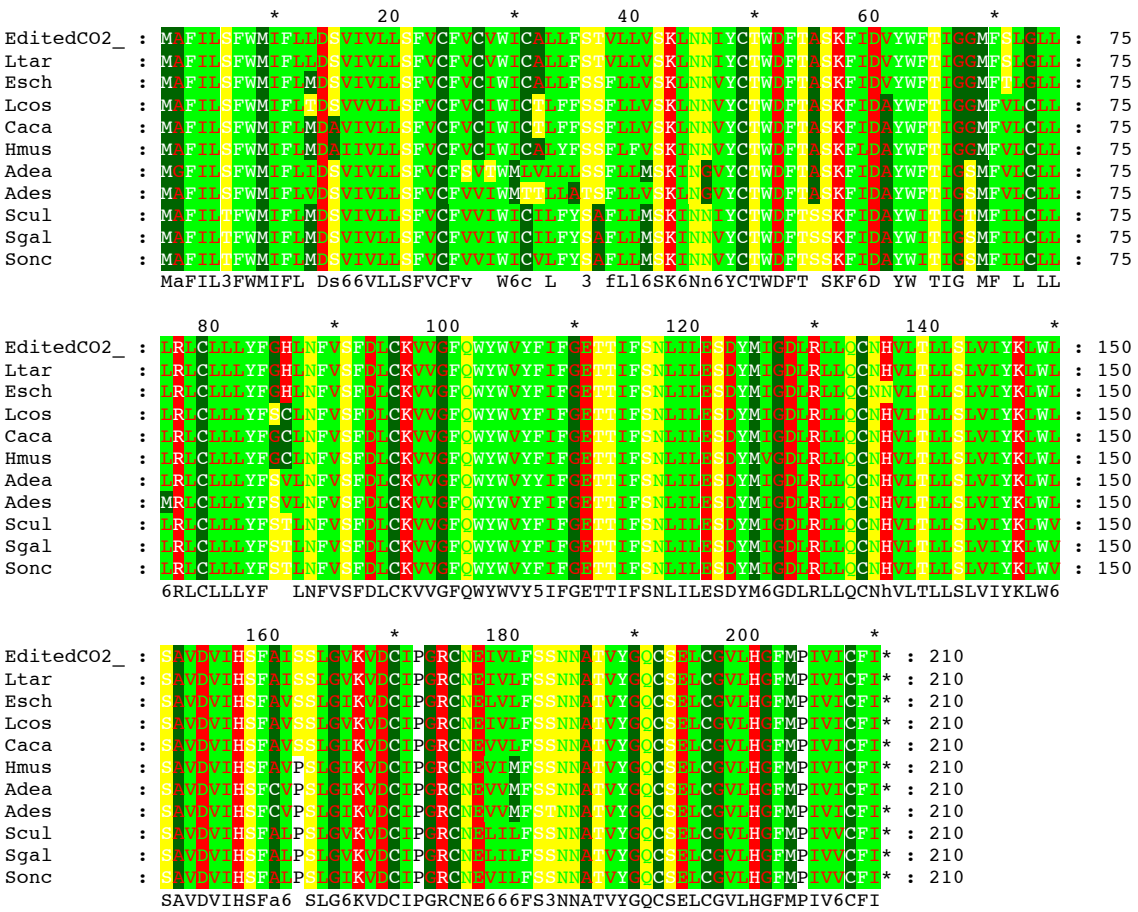
3.SI-S8d: A6 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	89	**100**									
L. cos	88	86	**100**								
C. aca	86	83	83	**100**							
H. mus	76	72	70	73	**100**						
P. ser	55	56	56	55	55	**100**					
A. dea	75	73	72	77	75	59	**100**				
A. des	72	73	70	72	73	59	86	**100**			
S. cul	74	75	74	74	74	59	86	87	**100**		
S. gal	73	74	73	74	73	61	88	88	95	**100**	
S. onc	72	74	72	72	73	60	86	87	94	97	**100**


RNA_Ltar_p
Ltar
Esch
Lcos
Caca
Hmus
Adea
Ades
Scul
Sgal
Sonc



3.SI-S9a: GeneDoc COII nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Reference is cDNA of *L. tar*.



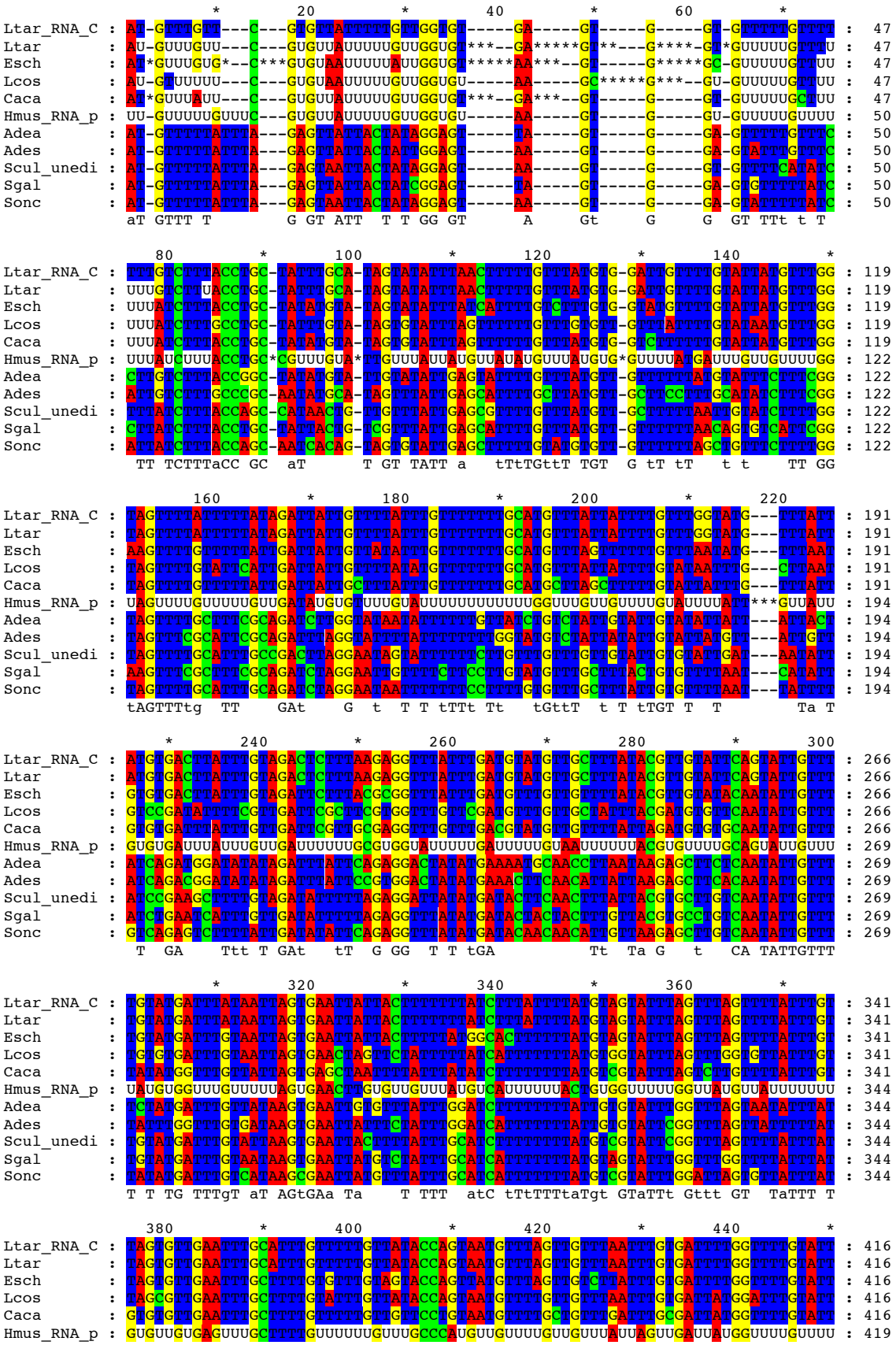
3.SI-S9b: GeneDoc COII protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

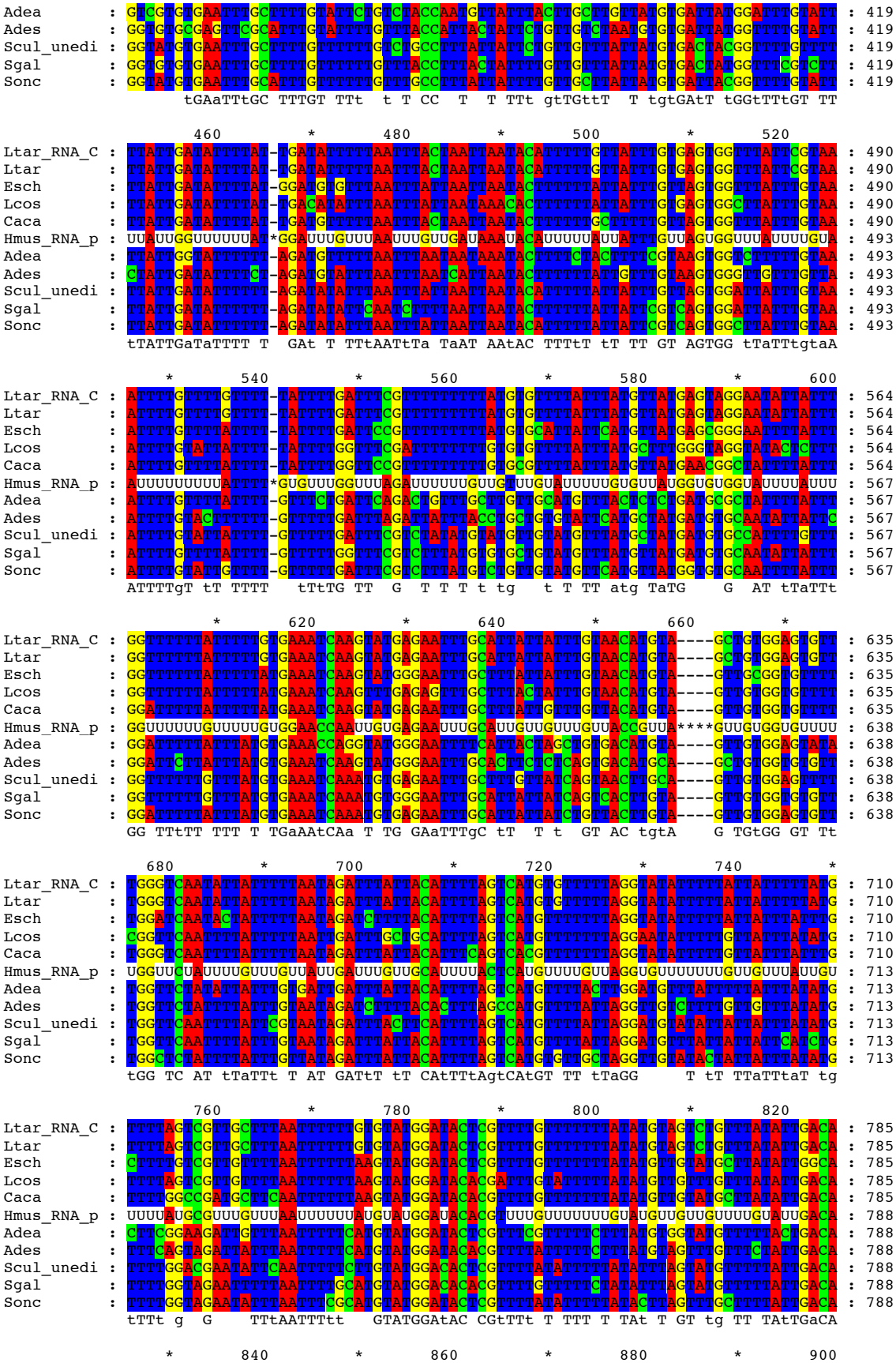
3.SI-S9c: COII edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	87	**100**								
L. cos	83	86	**100**							
C. aca	81	85	84	**100**						
H. mus	83	85	84	83	**100**					
A. dea	79	81	80	78	81	**100**				
A. des	80	80	79	77	78	80	**100**			
S. cul	83	84	82	82	83	80	79	**100**		
S. gal	82	84	82	81	82	81	82	89	**100**	
S. onc	82	83	82	82	84	82	80	89	90	**100**

3.SI-S9d: COII protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	95	**100**								
L. cos	93	94	**100**							
C. aca	93	96	97	**100**						
H. mus	90	92	92	95	**100**					
A. dea	87	89	90	90	89	**100**				
A. des	88	90	91	92	90	94	**100**			
S. cul	88	89	90	90	88	87	87	**100**		
S. gal	88	89	90	90	89	88	88	99	**100**	
S. onc	87	89	90	91	90	90	89	98	99	**100**





Adea : GTCGTGTGAATTTGCTTTTGTATTCTGTCTACCAATGTTATTTACTTGCTTGTTATGTGATTATGGATTTGTATT : 419
Ades : GGTGTGCGAGTTCGCATTTGTATTTTGTTACCATTACTATTCTGTTGTCTAATGTGTGATTATGGTTTTGTATT : 419
Scul_unedi : GGTATGTGAATTTGCTTTTGTTTTTTGTCTGCCTTTATTATTCTGTGTTTATTATGTGACTACGGTTTTGTTTT : 419
Sgal : GGTGTGTGAATTTGCTTTTGTTTTTTGTTTACCTTTACTATTTCGTGTTTATTATGTGACTATGGTTTCGTCTT : 419
Sonc : GGTATGTGAATTTGCATTTGTTTTTTGTTTGCCTTTATTATTTCGTTGCTTATTATGTGATTACGCTTTTGTATT : 419
tGAaTTtGC TTTGT TTt t T CC T T TTt gtTGttT T tgtGAtT tGGtTTtGT TT

460 * 480 * 500 * 520

Ltar_RNA_C : TTATTGATATTTTAT-TGATATTTTTAATTTACTAATTAATACATTTTTGTTATTTGTGAGTGGTTTATTCGTAA : 490
Ltar : TTATTGATATTTTAT-TGATATTTTTAATTTACTAATTAATACATTTTTGTTATTTGTGAGTGGTTTATTCGTAA : 490
Esch : TTATTGATATTTTAT-GGATTGTGTTTAATTTATTAATTAATACTTTTTTATTATTTGTTAGTGGTTTATTTGTAA : 490
Lcos : TTATTGATATTTTAT-TGACATATTTAATTTATTAATAAACACTTTTTTATTATTTGTGAGTGGCTTATTTGTAA : 490
Caca : TTATTGATATTTTAT-TGATGTTTTTAATTTACTAATTAATACTTTTTTGCTTTTTGTTAGTGGTTTATTTGTAA : 490
Hmus_RNA_p : UUAUUGGUUUUUUAT*GGAUUUGUUUAAUUUGUUGAUAAAUACAUUUUUAUUATTTGUUAGUGGUUUAUUUUGUA : 493
Adea : TTATTGGTATTTTTT-AGATTGTTTTAATTTAATAATAAATACTTTTCTACTTTCGTTAAGTGGTCTTTTGTTAA : 493
Ades : CTATTGATATTTTCT-AGATGTATTTAATTTAATCATTAATACTTTTTTATTGTTTGTAAGTGGGTTGTTTGTTA : 493
Scul_unedi : TTATTGATATTTTTT-AGATATATTTAATTTATTAATTAATACATTTTTATTATTTGTTAGTGGATTATTTGTAA : 493
Sgal : TTATTGATATTTTTT-AGATATATTTCAATCTTTTAATTAATACTTTTTTATTATTCGTCAGTGGATTATTTGTAA : 493
Sonc : TTATTGATATTTTTT-AGATATATTTAATTTATTAATTAATACATTTTTATTATTCGTCAGTGGCTTATTTGTAA : 493
tTATTGaTaTTTT T GAt T TTtAATTTa TaAT AAtAC TTTtT tT TT GT AGTGG tTaTTtgtaA

* 540 * 560 * 580 * 600

Ltar_RNA_C : ATTTTGTTTTGTTTT-TATTTTGATTTCGTTTTTTTTTATGTGTTTTATTTATGTTATGAGTAGGAATATTATTT : 564
Ltar : ATTTTGTTTTGTTTT-TATTTTGATTTCGTTTTTTTTTATGTGTTTTATTTATGTTATGAGTAGGAATATTATTT : 564
Esch : ATTTTGTTTTATTTT-TATTTTGATTCCGTTTTTTTTTATGTGCATTATTTCATGTTATGAGCGGGAATTTTATTT : 564
Lcos : ATTTTGTATTATTTT-TATTTTGGTTTCGATTTTTTTTTGTGTGTTTTATTTATGCTTTGGGTAGGTTATACTCTTT : 564
Caca : ATTTTGTTTTATTTT-TATTTTGGTTCCGTTTTTTTTTGTGCGTTTTATTTATGTTATGAACGGCTATTTTATTT : 564
Hmus_RNA_p : AUUUUUUUUUATTTT*GUGUUUGGUUUAGAUUUUUUGUUGTUUGUAUUUUUGUGUUAUGGUGUGGUAUUUUAUUU : 567
Adea : ATTTTGTTTTATTTT-GTTTCTGATTCAGACTGTTTGCTTGTTGCATGTTTACTCTCTGATGCGCTATTTTATTT : 567
Ades : ATTTTGTACTTTTTT-GTTTTTGATTTAGATTATTTACCTGCTGTGTATTCATGCTATGATGTGCAATTATTATTC : 567
Scul_unedi : ATTTTGTATTATTTT-GTTTTTGATTTCGTCTATATGTATGTTGTATGTTTATGCTATGATGTGCCATTTTGTTT : 567
Sgal : ATTTTGTTTTATTTT-GTTTTTGGTTTCGTCTTTATGTGTGCTGTATGTTTATGTTATGATGTGCAATATTATTT : 567
Sonc : ATTTTGTATTGTTTT-GTTTTTGATTTCGTCTTTATGTGCTGTTGTATGTTCATGTTATGGTGTGCAATTTTATTT : 567
ATTTTgT tT TTTT tTtTG TT G T T T t tg t T TT atg TaTG G AT tTaTTt

* 620 * 640 * 660 *

Ltar_RNA_C : GGTTTTTTATTTTTGTGAAATCAAGTATGAGAATTTGCATTATTATTATTGTAACATGTA----GCTGTGGAGAGTGTT : 635
Ltar : GGTTTTTTATTTTTGTGAAATCAAGTATGAGAATTTGCATTATTATTATTGTAACATGTA----GCTGTGGAGAGTGTT : 635
Esch : GGTTTTTTATTTTTATGAAATCAAGTATGGGAATTTGCTTTATTATTTTGTAACATGTA----GTTGGCGGTGTTTTT : 635
Lcos : GGTTTTTTATTTTTATGAAATCAAGTTTGAGAGTTTGCTTTACTATTTGTAACATGTA----GTTGTGGTGTTTT : 635
Caca : GGATTTTTATTTTTATGAAATCAAGTATGAGAATTTGCTTTATGTGTTGTTACATGTA----GTTGTGGTGTTTT : 635
Hmus_RNA_p : GGUUUUUUGUUUUUGUGGAACCAAUUGUGAGAAUUUGCAUUGUUGUUUGUUACCGUUA****GUUGUGGUGUUUU : 638
Adea : GGATTTTTATTTATGTGAAACCAGGTATGGGAATTTGCATTACTAGCTGTGACATGTA----GTTGTGGAGTTATA : 638
Ades : GGATTCTTATTTATGTGAAATCAAGTATGGGAATTTGCACTTCTCCTCAGTGACATGCA----GCTGTGGTGTGTT : 638
Scul_unedi : GGTTTTTTGTTTATGTGAAATCAAATGTGAGAATTTGCTTTGTTATCAGTAACTTGCA----GTTGTGGAGTTTT : 638
Sgal : GGTTTTTTGTTTATGTGAAATCAAATGTGGGAATTTGCATTATTATCAGTCACTTGTA----GTTGTGGTGTGTT : 638
Sonc : GGATTTTTATTTATGTGAAATCAAATGTGAGAATTTGCATTATTATCAGTTACTTGTA----GTTGTGGAGAGTT : 638
GG TTtTT TTT T TGaAAtCAa T TG GAaTTTgC tT T t GT AC tgtA G TGtGG GT Tt

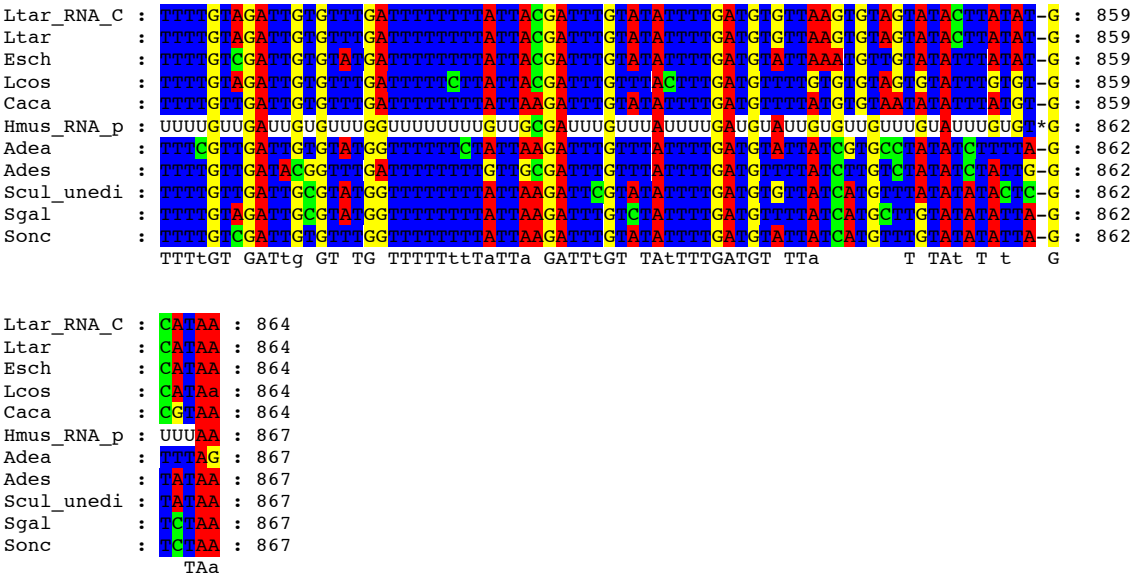
680 * 700 * 720 * 740 *

Ltar_RNA_C : TGGGTCAATATTATTTTTAATAGATTTATTACATTTTAGTCATGTGTTTTTAGGTATATTTTTATTATTTTTATG : 710
Ltar : TGGGTCAATATTATTTTTAATAGATTTATTACATTTTAGTCATGTGTTTTTAGGTATATTTTTATTATTTTTATG : 710
Esch : TGGATCAATACTATTTTTAATAGATCTTTTACATTTTAGTCATGTTTTTTTAGGTATATTTTTATTATTATTTG : 710
Lcos : CGGTTCAATTTTATTTTTAATTGATTTGCTGCATTTTAGTCATGTTTTTTTAGGAATATTTTTGTTATTTATTG : 710
Caca : TGGGTCAATTTTATTTTTAATAGATTTATTACATTTCAGTCACGTTTTTTTAGGTATATTTTTGTTATTTATTG : 710
Hmus_RNA_p : UGGUUCUAUUUUGUUUGUUAUUGAUUUGUUGCAUUUUACTCAUGUUUUGUUAGGUGUUUUUUUGUUGUUUAUUGU : 713
Adea : TGGTTCTATAATATTTGTGATTGATTTATTACATTTTAGTCATGTTTTACTTGGATGTTTATTTTTATTTATATG : 713
Ades : TGGTTCTATTTTATTTGTAATAGATCTTTTACACTTTAGTCCATGTTTTATTAGGTTGTCTTTTGTTGTTTATATG : 713
Scul_unedi : TGGTTCAATTTTATTTGTAATAGATTTACTTCATTTTAGTCATGTTTTATTAGGATGTTATATTATTATTTATATG : 713
Sgal : TGGTTCAATTTTATTTGTAATAGATTTATTACATTTTAGTCATGTTTTATTAGGATGTTTATTATTATTCATCTG : 713
Sonc : TGGTCTATTTTATTTGTTATAGATTTATTACATTTTAGTCATGTGTTGCTAGGTTGTTATACTATTATTTATATG : 713
tGG TC AT tTaTTt T AT GATtT tT CAtTTtAgtCAtGT TT tTaGG T tT TTaTTtaT tg

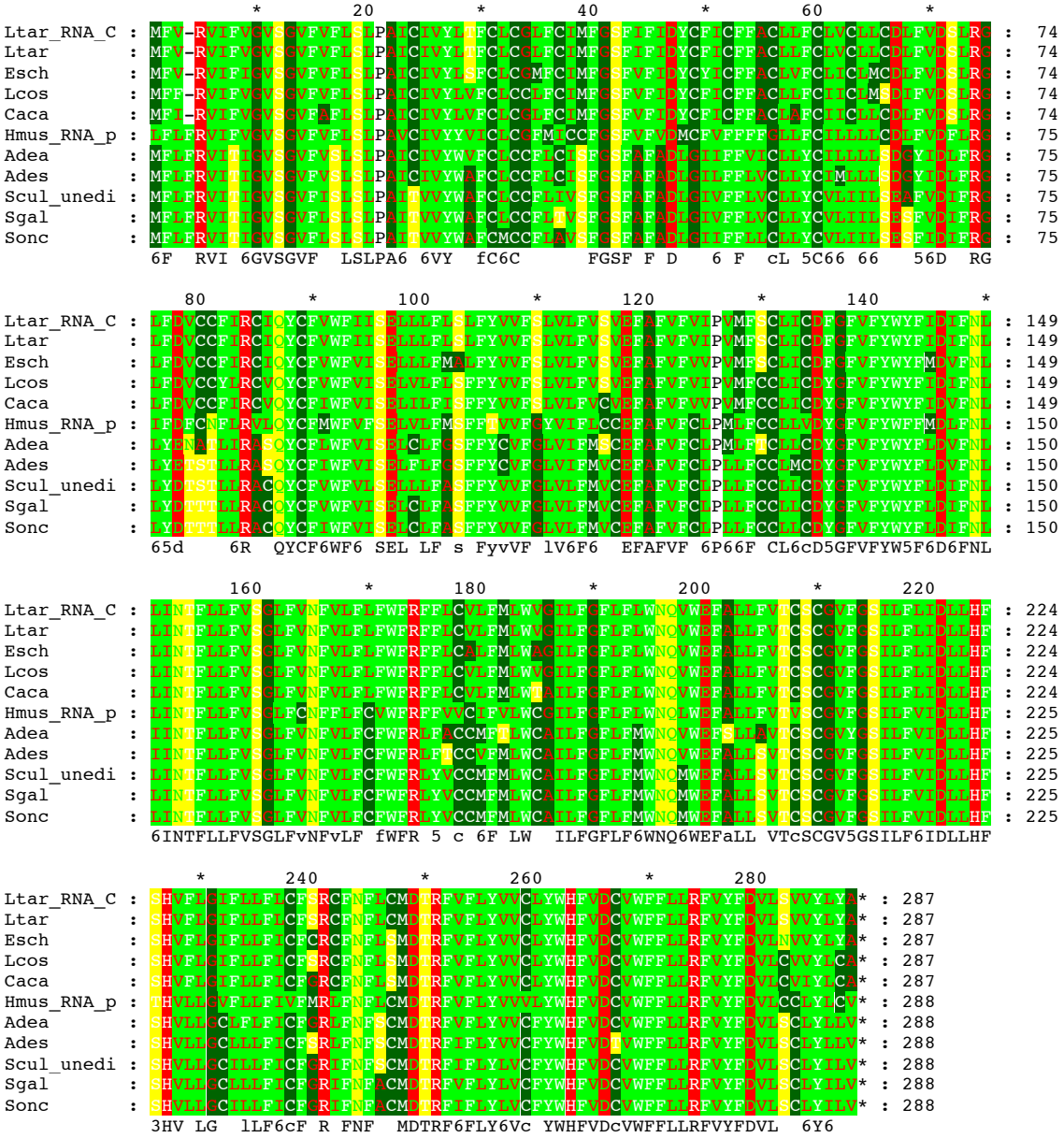
760 * 780 * 800 * 820

Ltar_RNA_C : TTTTAGTCGTTGCTTTAATTTTTTGTGTATGGATACTCGTTTTGTTTTTTTATATGTAGTCTGTTTATATTGACA : 785
Ltar : TTTTAGTCGTTGCTTTAATTTTTTGTGTATGGATACTCGTTTTGTTTTTTTATATGTAGTCTGTTTATATTGACA : 785
Esch : CTTTTGTCGTTGTTTTAATTTTTTAAGTATGGATACTCGTTTTGTTTTTTATATGTTGTATGCTTATATTGGCA : 785
Lcos : TTTTAGTCGTTGTTTTAATTTTTAAGTATGGATACACGATTTGTATTTTTATATGTTGTTTGTTTATATTGACA : 785
Caca : TTTTGGCCGATGCTTCAATTTTTTAAGTATGGATACACGTTTTGTTTTTTTATATGTTGTATGCTTATATTGACA : 785
Hmus_RNA_p : UUUUAUGCGUUUGUUUAAUUUUUUAUGUAUGGAUACACGUUUUGUUUUUUUGUAUGUUGUUGUUUUGUAUUGACA : 788
Adea : CTTCGGAAGATTGTTTAATTTTTCATGTATGGATACTCGTTTCGTTTTTCTTTATGTGCTATGTTTTTACTGACA : 788
Ades : TTTCAGTAGATTATTTAATTTTTCATGTATGGATACACGTTTTATTTTTCTTTATGTAGTTTGTTTCTATTGACA : 788
Scul_unedi : TTTTGGACGAATATTCAATTTTTCCTTGTATGGACACTCGTTTTATATTTTATATTTAGTATGTTTTTATTGACA : 788
Sgal : TTTTGGTAGAATTTTTAATTTTGCATGTATGGACACACGTTTTGTTTTTCTATATTTAGTATGTTTTTATTGACA : 788
Sonc : TTTTGGTAGAATATTTAATTTCGCATGTATGGATACTCGTTTTATATTTTTATACTTAGTTGCTTTTATTGACA : 788
tTTt g G TTtAATTTTt GTATGGAtAC CGtTTt T TTT T TAt T GT tg TT TAtTGaCA

* 840 * 860 * 880 * 900



3.SI-S10a: GeneDoc COIII nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNA of *L. tar* and RNA sequences of *S. cul* and *H. mus*.



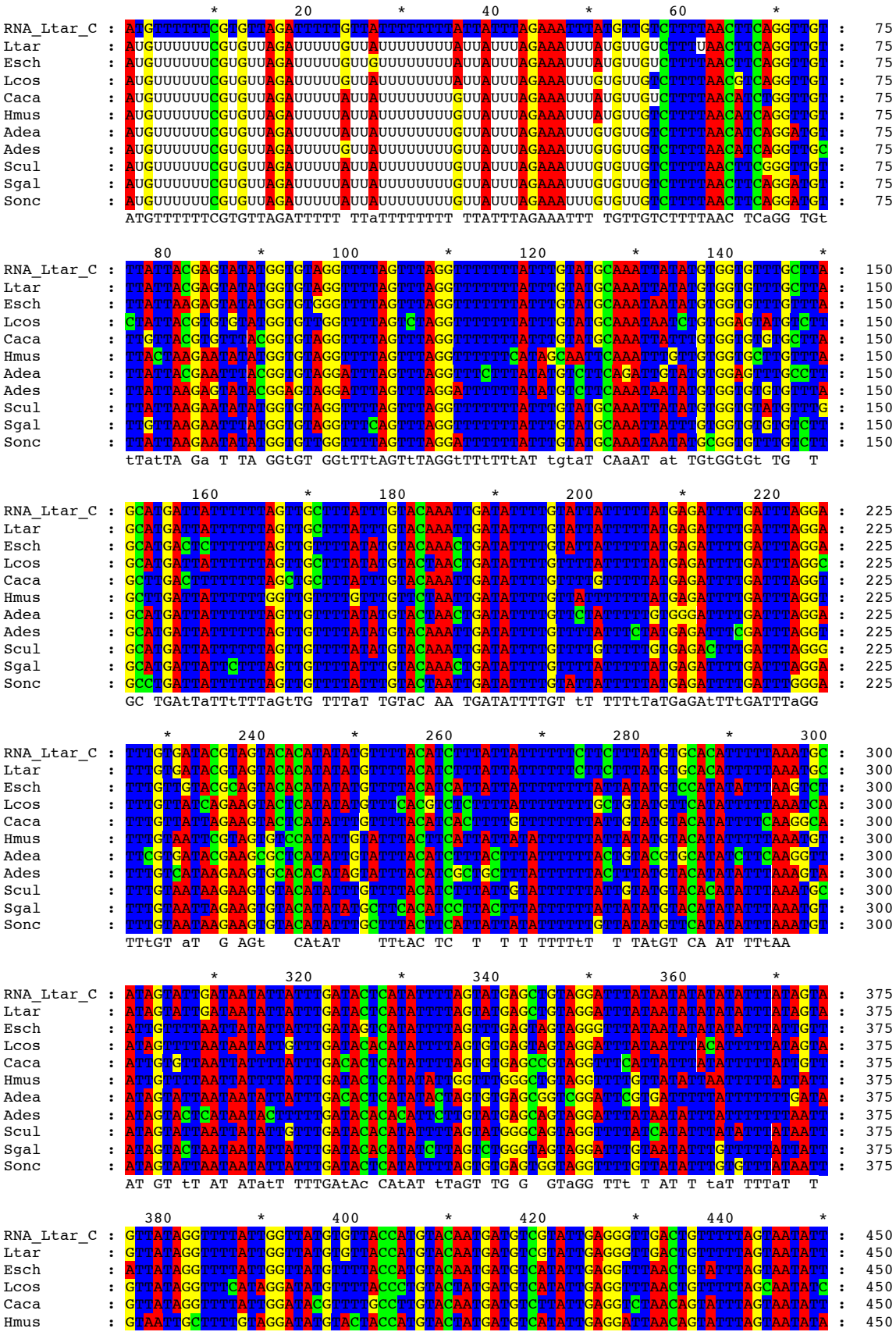
3.SI-S10b: GeneDoc COIII protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

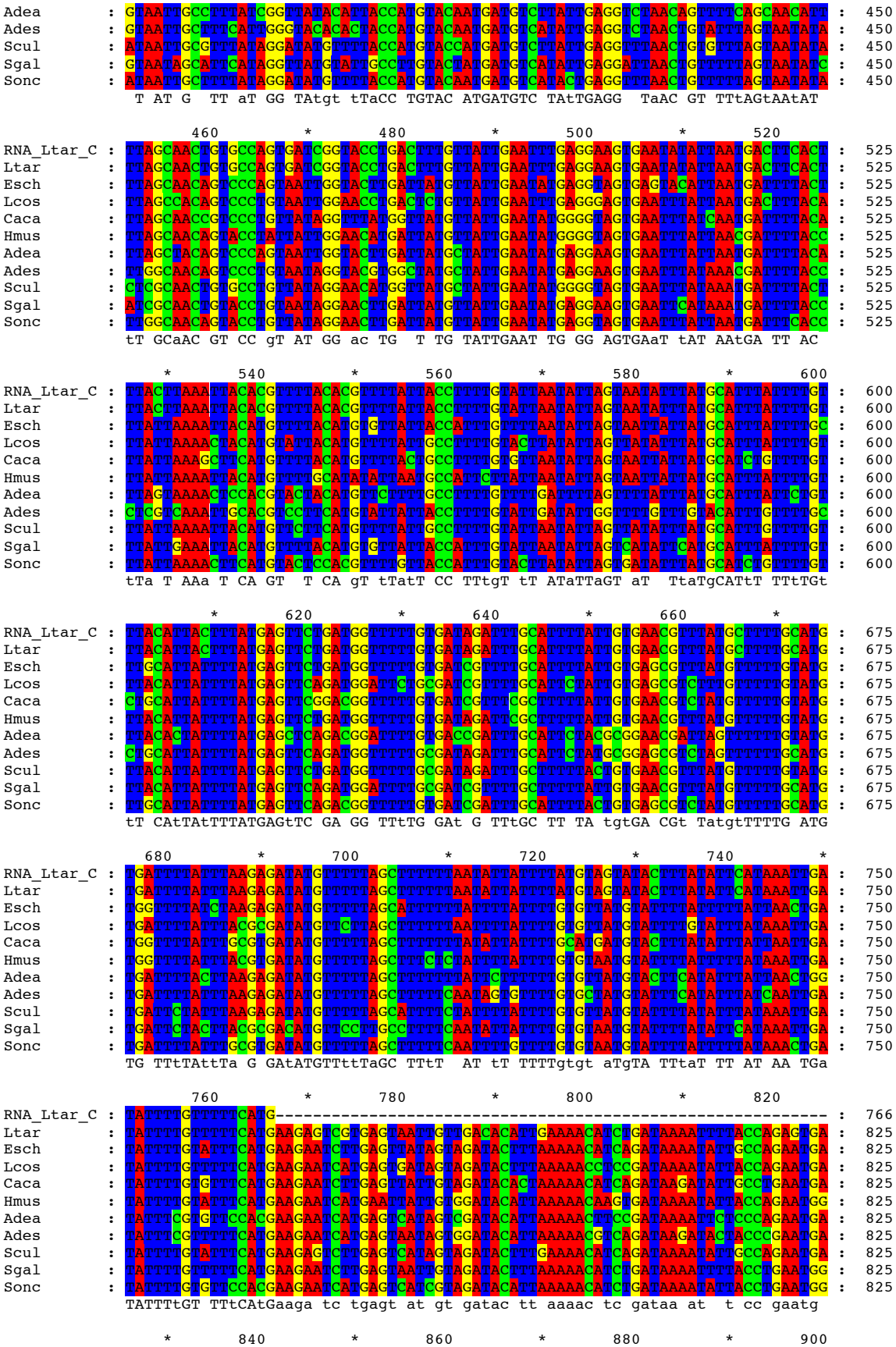
3.SI-S10c: COIII edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

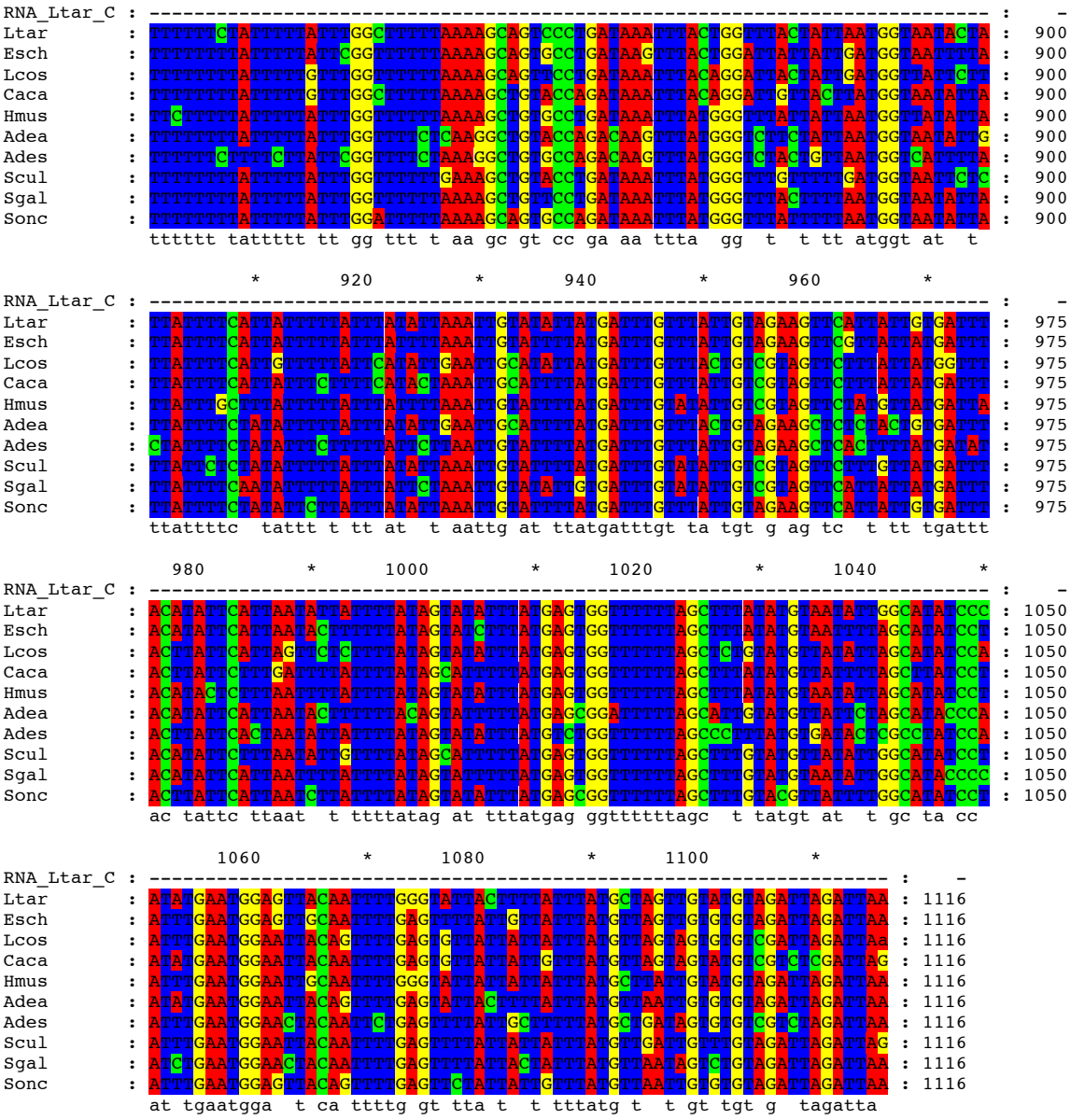
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	90	**100**								
L. cos	86	85	**100**							
C. aca	87	87	86	**100**						
H. mus	75	75	77	77	**100**					
A. dea	72	72	71	73	72	**100**				
A. des	73	72	73	72	72	81	**100**			
S. cul	74	73	74	73	74	79	79	**100**		
S. gal	75	74	73	74	74	80	79	87	**100**	
S. onc	74	73	72	74	73	80	79	87	88	**100**

3.SI-S10d: COIII protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

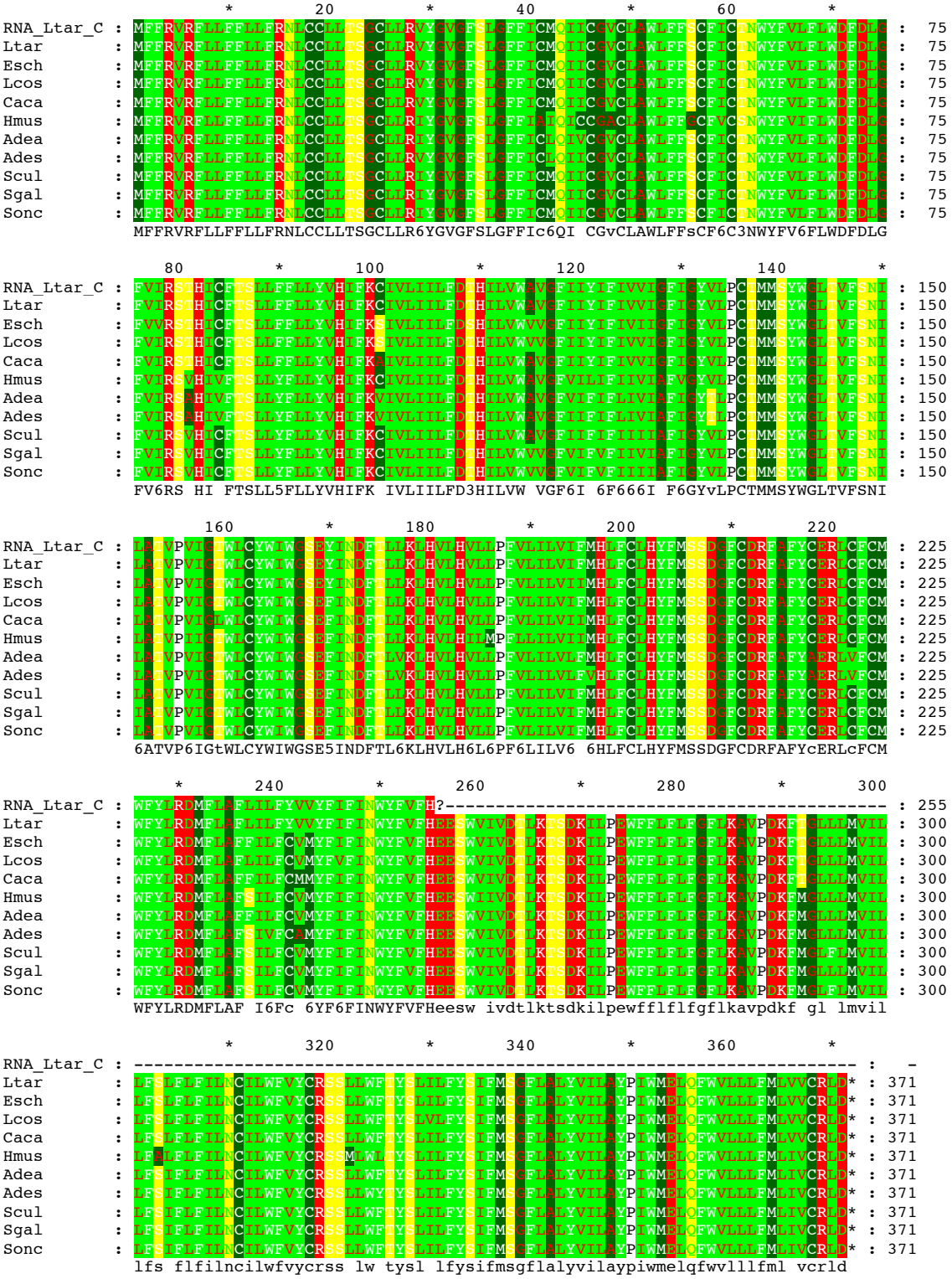
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	93	**100**								
L. cos	93	91	**100**							
C. aca	91	91	93	**100**						
H. mus	73	72	75	75	**100**					
A. dea	71	71	73	73	73	**100**				
A. des	71	71	74	73	72	92	**100**			
S. cul	70	69	72	71	72	84	88	**100**		
S. gal	71	70	72	71	72	85	87	97	**100**	
S. onc	70	69	71	71	70	85	87	96	97	**100**




Adea : 450
Ades : 450
Scul : 450
Sgal : 450
Sonc : 450
460 480 500 520
RNA_Ltar_C : 525
Ltar : 525
Esch : 525
Lcos : 525
Caca : 525
Hmus : 525
Adea : 525
Ades : 525
Scul : 525
Sgal : 525
Sonc : 525
540 560 580 600
RNA_Ltar_C : 600
Ltar : 600
Esch : 600
Lcos : 600
Caca : 600
Hmus : 600
Adea : 600
Ades : 600
Scul : 600
Sgal : 600
Sonc : 600
620 640 660
RNA_Ltar_C : 675
Ltar : 675
Esch : 675
Lcos : 675
Caca : 675
Hmus : 675
Adea : 675
Ades : 675
Scul : 675
Sgal : 675
Sonc : 675
680 700 720 740
RNA_Ltar_C : 750
Ltar : 750
Esch : 750
Lcos : 750
Caca : 750
Hmus : 750
Adea : 750
Ades : 750
Scul : 750
Sgal : 750
Sonc : 750
760 780 800 820
RNA_Ltar_C : 766
Ltar : 825
Esch : 825
Lcos : 825
Caca : 825
Hmus : 825
Adea : 825
Ades : 825
Scul : 825
Sgal : 825
Sonc : 825
840 860 880 900



3.SI-S11a: GeneDoc CYb nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Reference is cDNA of *L. tar.*



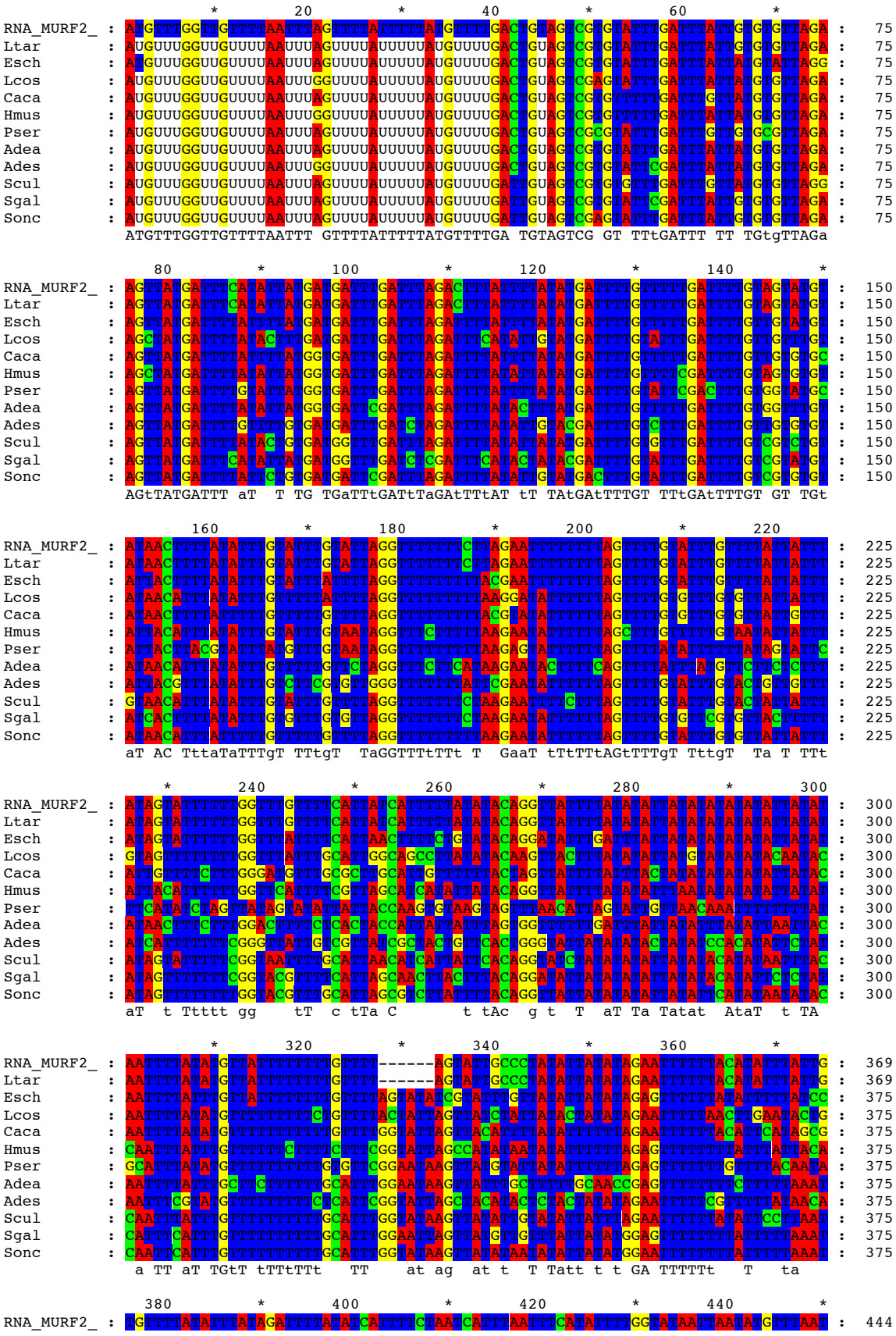
3.SI-S11b: GeneDoc CYb protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. ? in the alignment indicates missing information or incomplete sequence. References have also been translated.

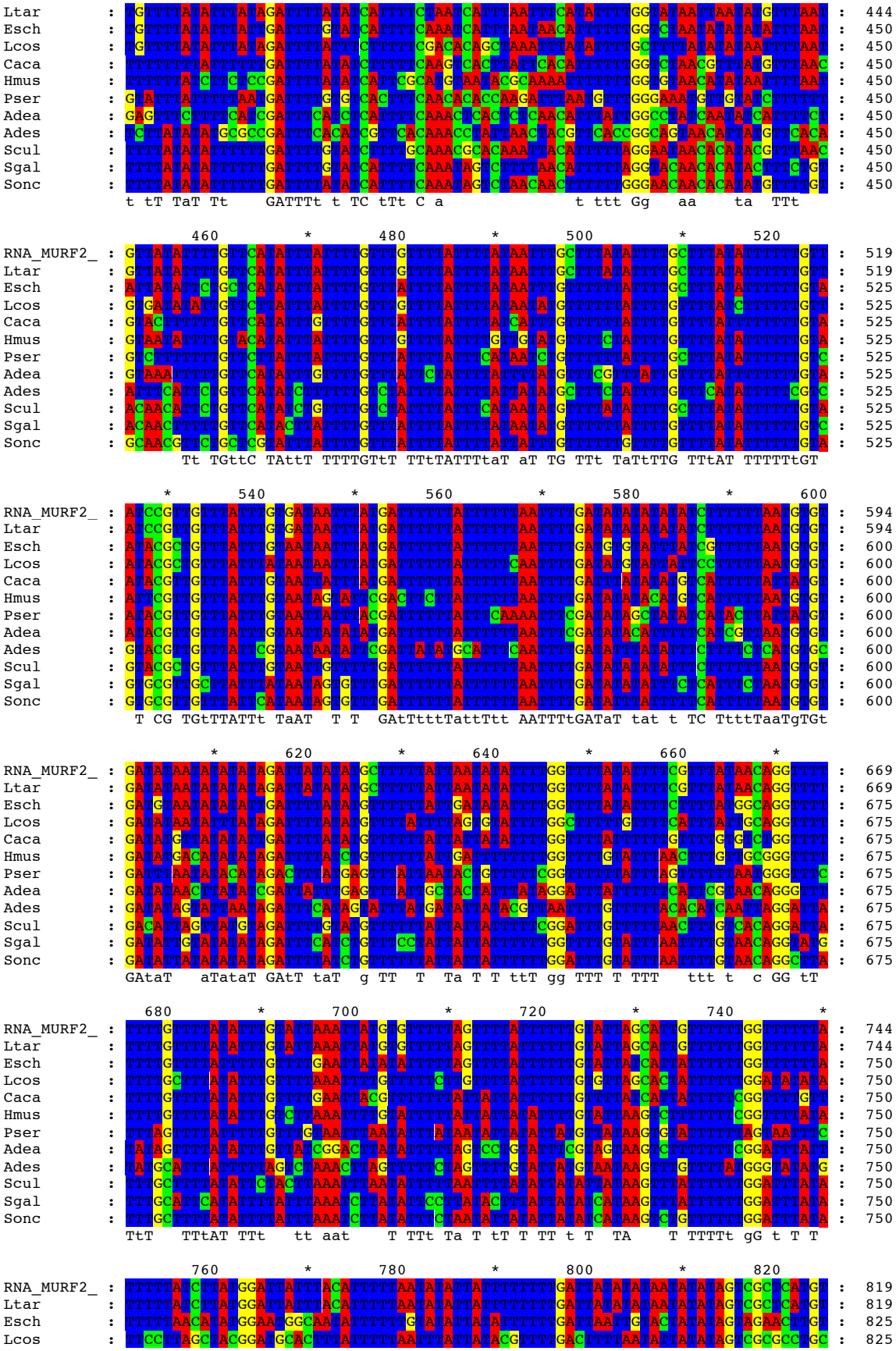
3.SI-S11c: CYb edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

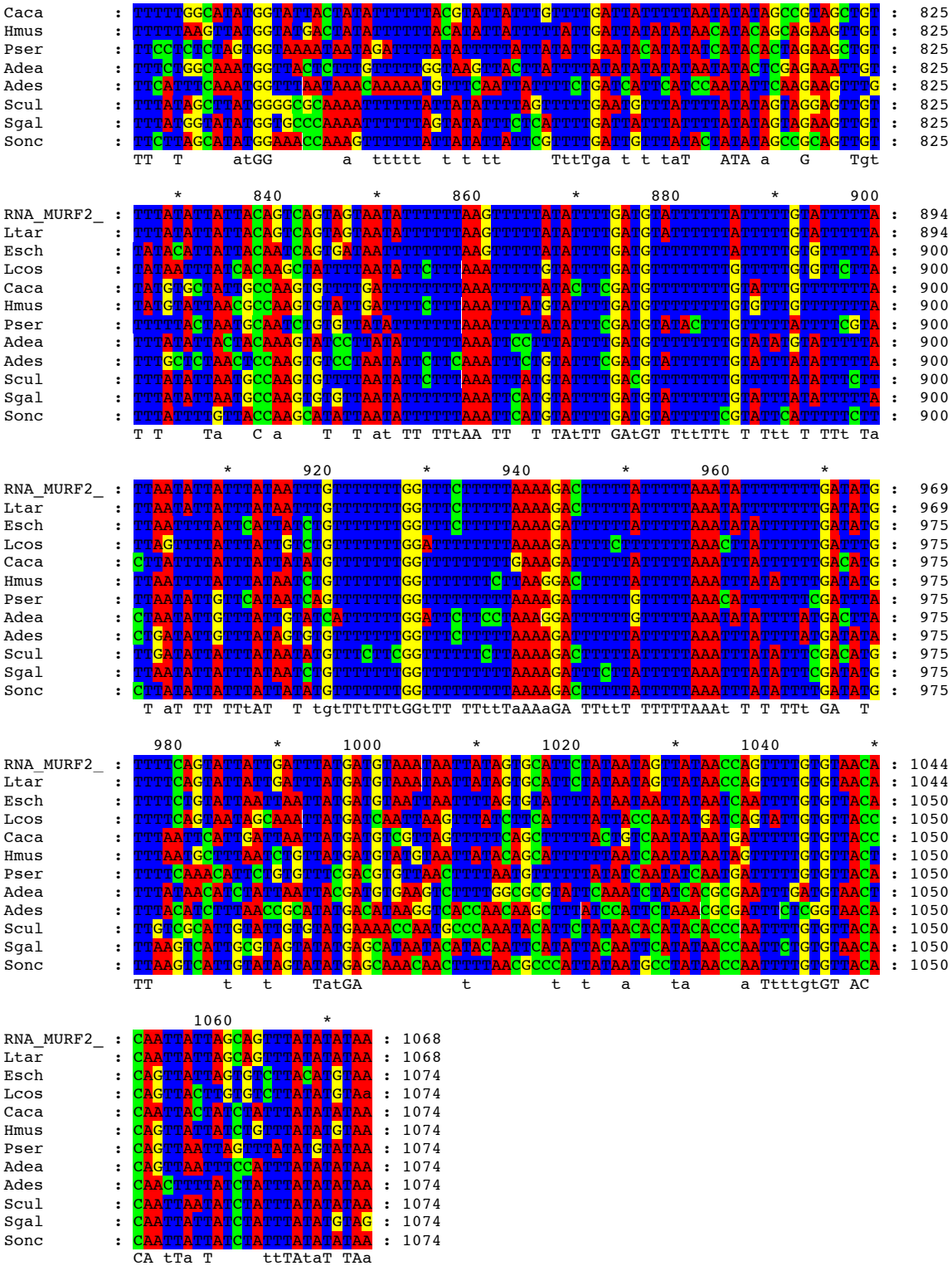
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	89	**100**								
L. cos	86	87	**100**							
C. aca	84	86	85	**100**						
H. mus	84	86	82	84	**100**					
A. dea	82	82	82	80	80	**100**				
A. des	81	82	81	80	80	82	**100**			
S. cul	86	88	85	85	87	82	83	**100**		
S. gal	86	86	86	83	86	81	82	89	**100**	
S. onc	86	87	86	84	86	83	82	88	88	**100**

3.SI-S11d: CYb protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

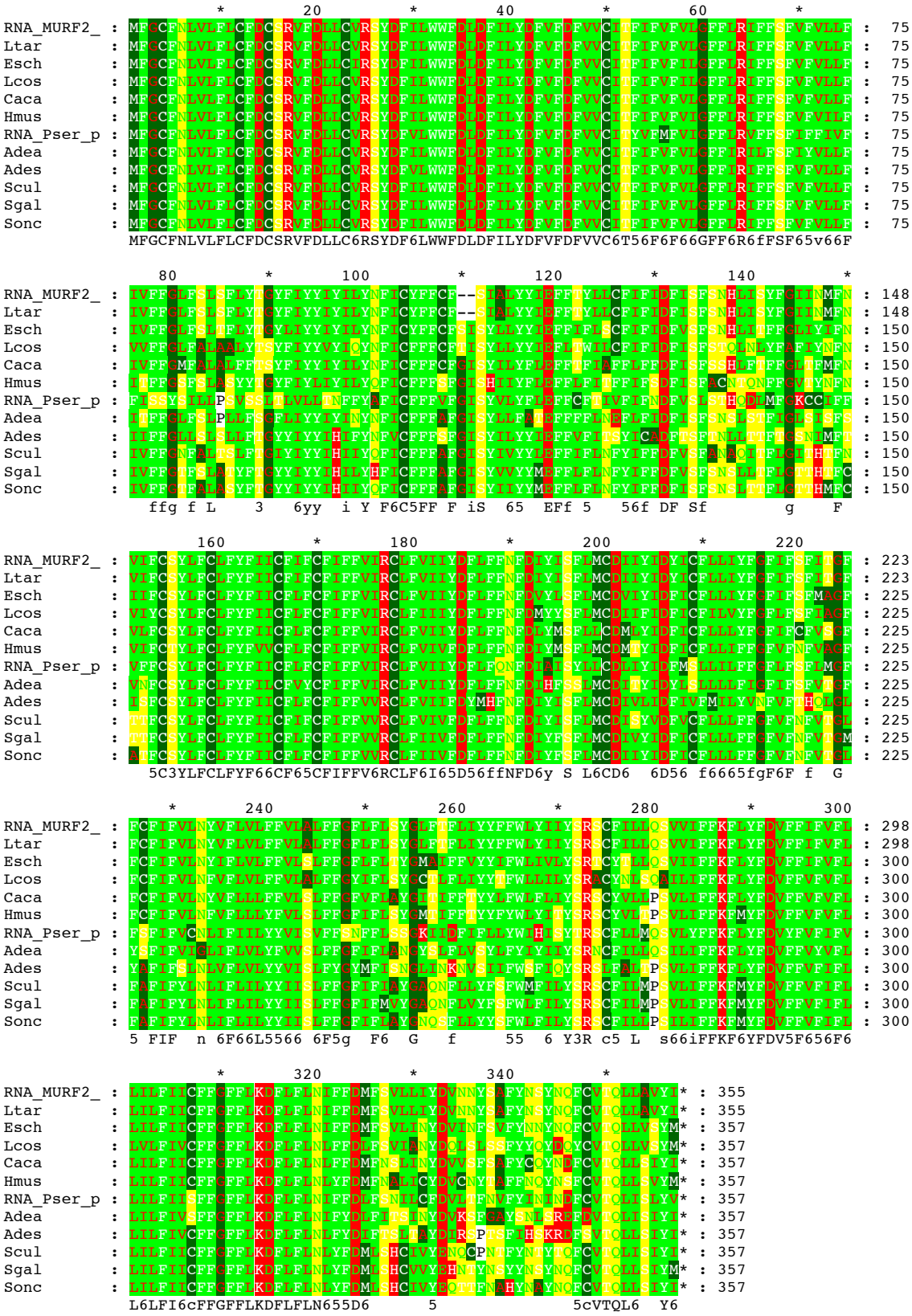
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**									
E. sch	98	**100**								
L. cos	98	98	**100**							
C. aca	98	98	98	**100**						
H. mus	91	91	91	92	**100**					
A. dea	94	93	94	94	92	**100**				
A. des	93	92	93	94	91	98	**100**			
S. cul	96	95	96	96	94	96	95	**100**		
S. gal	95	95	96	95	94	96	95	98	**100**	
S. onc	95	95	95	95	93	96	95	99	99	**100**







3.SI-S12a: GeneDoc MURF2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as “U” against a white background in the alignment. References are cDNA of *L. tar* and RNA of *P. ser*.



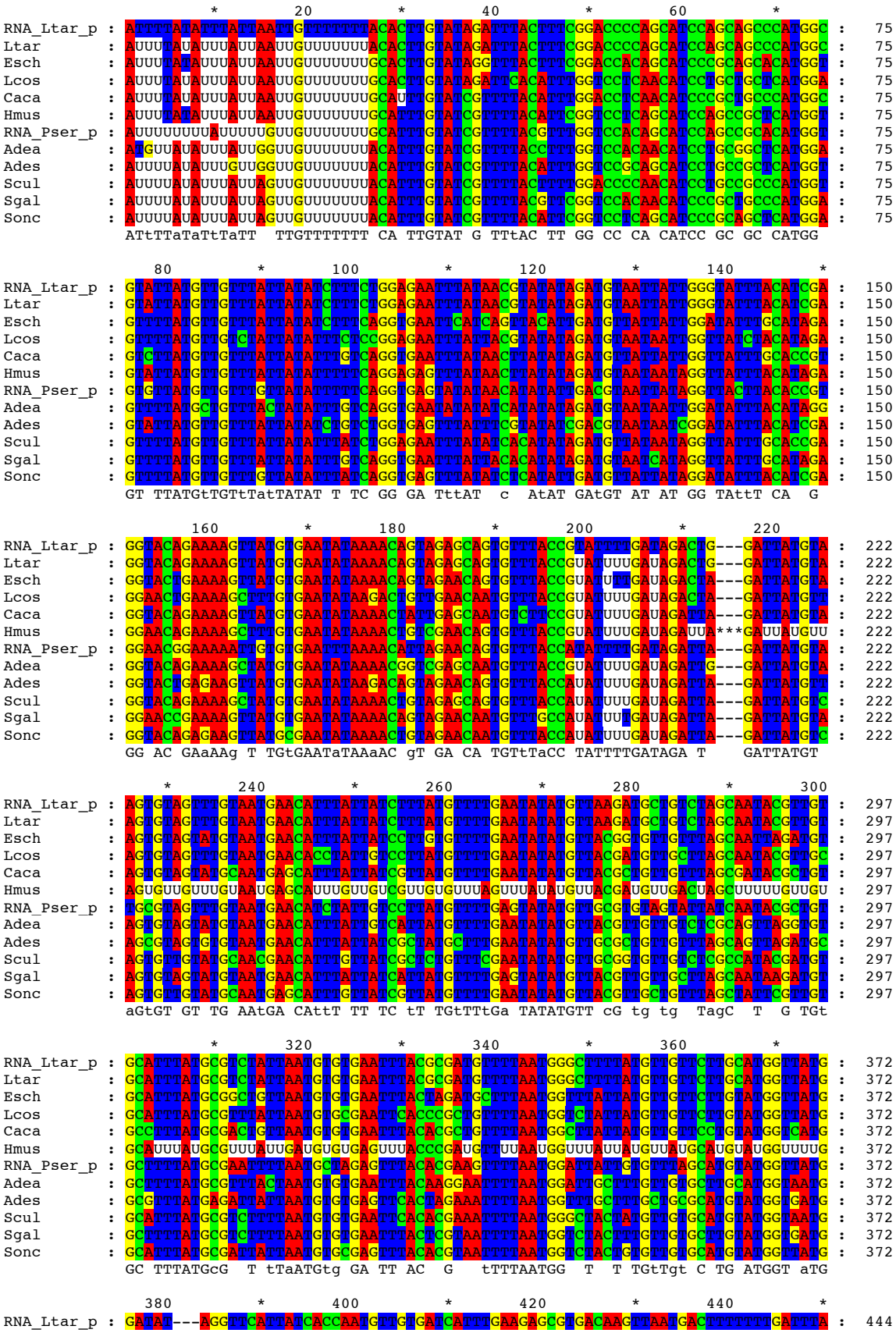
3.SI-S12b: GeneDoc MURF2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

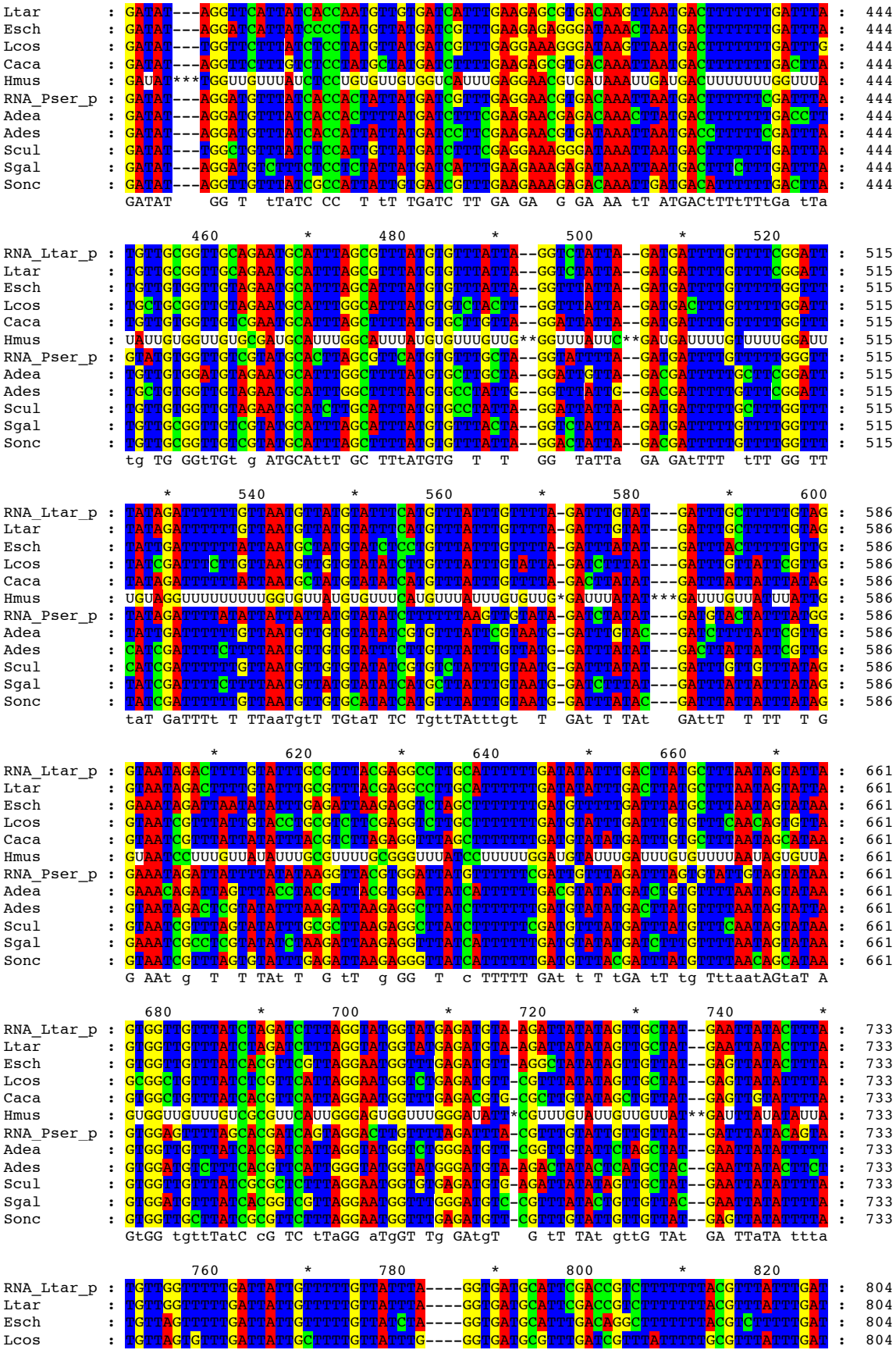
3.SI-S12c: MURF2 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

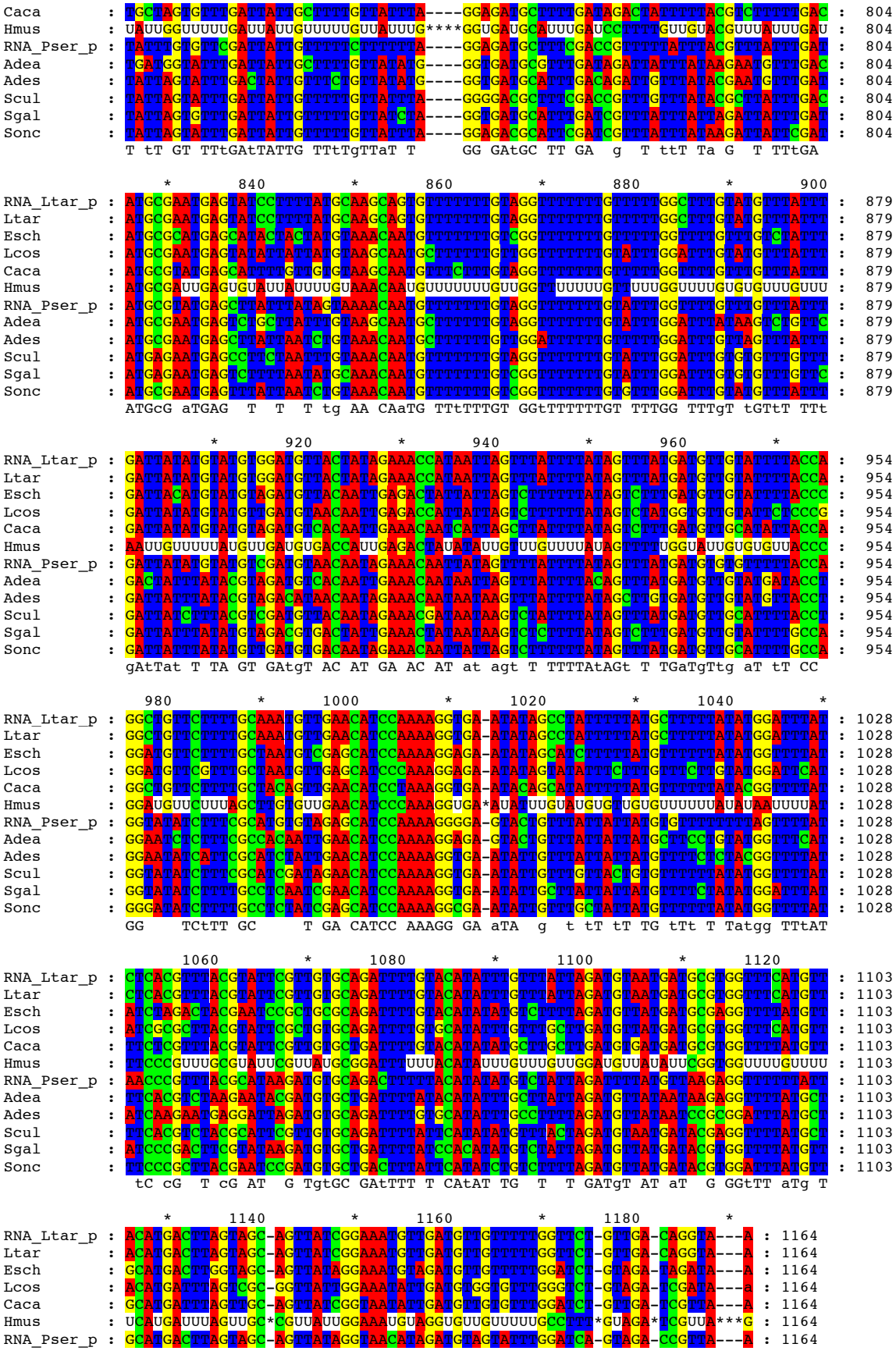
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	88	**100**									
L. cos	81	81	**100**								
C. aca	81	83	79	**100**							
H. mus	80	80	76	81	**100**						
P. ser	73	73	69	71	72	**100**					
A. dea	73	73	69	74	73	68	**100**				
A. des	71	70	68	70	70	64	67	**100**			
S. cul	78	77	74	76	77	70	71	71	**100**		
S. gal	78	77	73	76	77	70	72	72	83	**100**	
S. onc	79	79	76	78	79	69	73	72	84	87	**100**

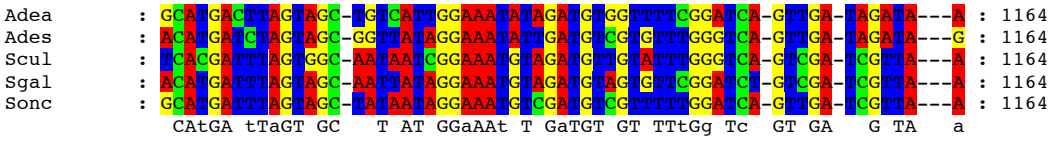
3.SI-S12d: MURF2 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	86	**100**									
L. cos	81	80	**100**								
C. aca	81	82	77	**100**							
H. mus	77	77	73	82	**100**						
P. ser	63	64	61	65	64	**100**					
A. dea	73	72	67	73	68	63	**100**				
A. des	69	68	66	71	67	59	67	**100**			
S. cul	73	73	67	75	74	62	69	71	**100**		
S. gal	75	74	70	75	76	62	70	72	89	**100**	
S. onc	75	74	71	78	77	62	71	72	89	92	**100**



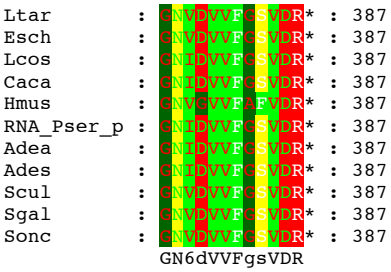






3.SI-S13a: GeneDoc ND7 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNA of *L. tar* and RNA sequence of *P. ser*.

RNA_Ltar_p
Ltar
Esch
Lcos
Caca
Hmus
RNA_Pser_p
Adea
Ades
Scul
Sgal
Sonc



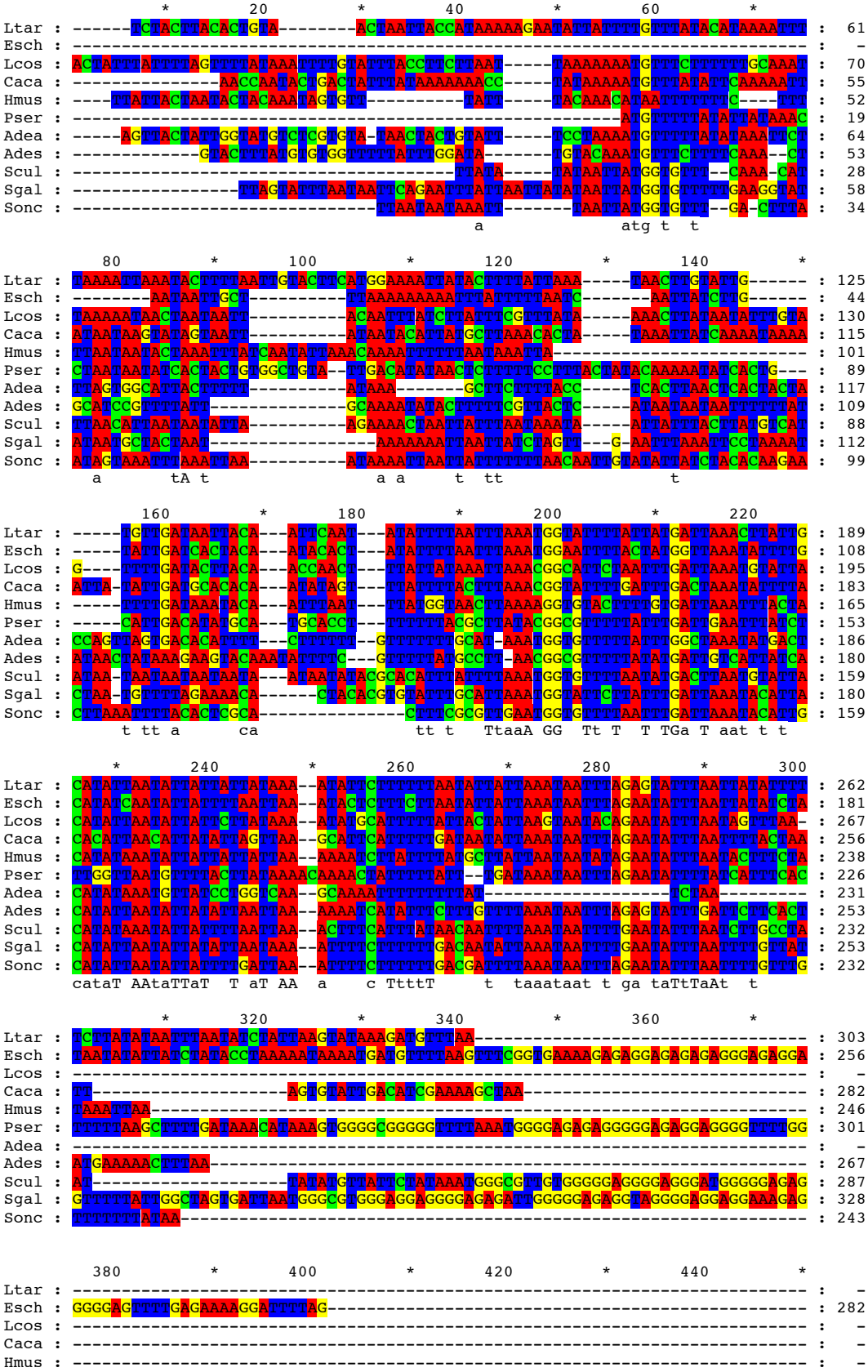
3.SI-S13b: GeneDoc ND7 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

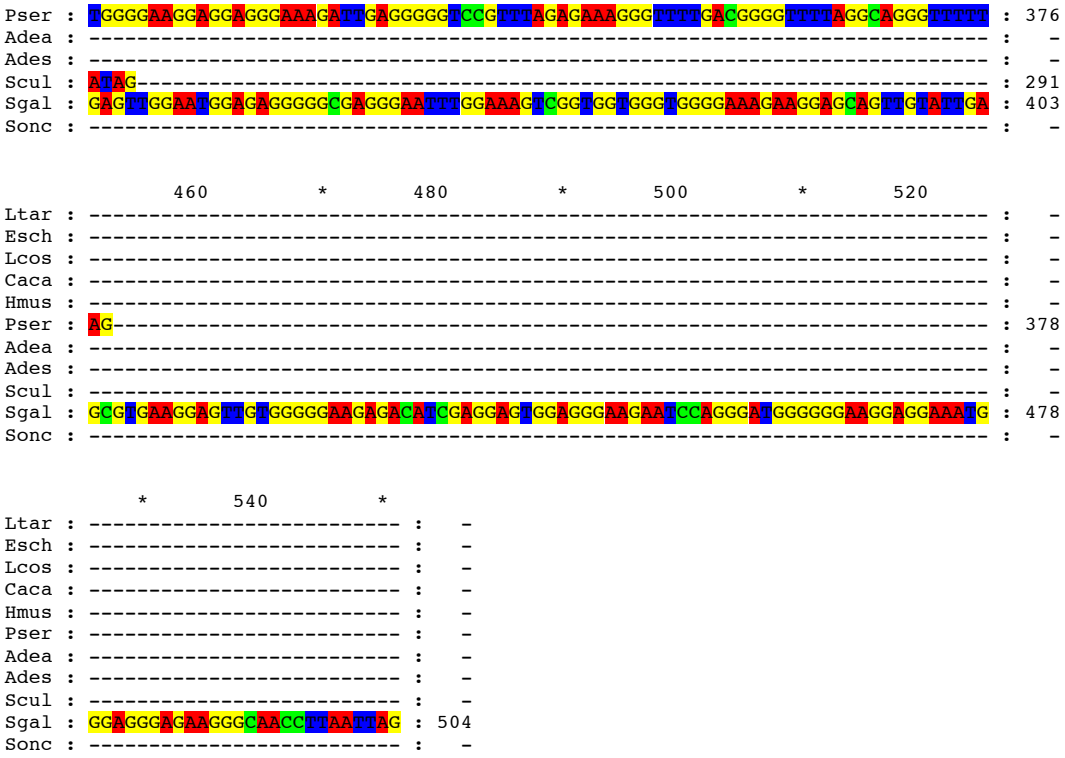
3.SI-S13c: ND7 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	87	**100**									
L. cos	85	84	**100**								
C. aca	85	86	84	**100**							
H. mus	81	80	82	81	**100**						
P. ser	77	78	75	77	75	**100**					
A. dea	80	79	79	80	76	75	**100**				
A. des	80	80	78	78	77	76	82	**100**			
S. cul	83	82	81	83	80	77	82	82	**100**		
S. gal	82	84	81	82	79	79	82	82	85	**100**	
S. onc	82	83	81	83	81	79	82	82	86	87	**100**

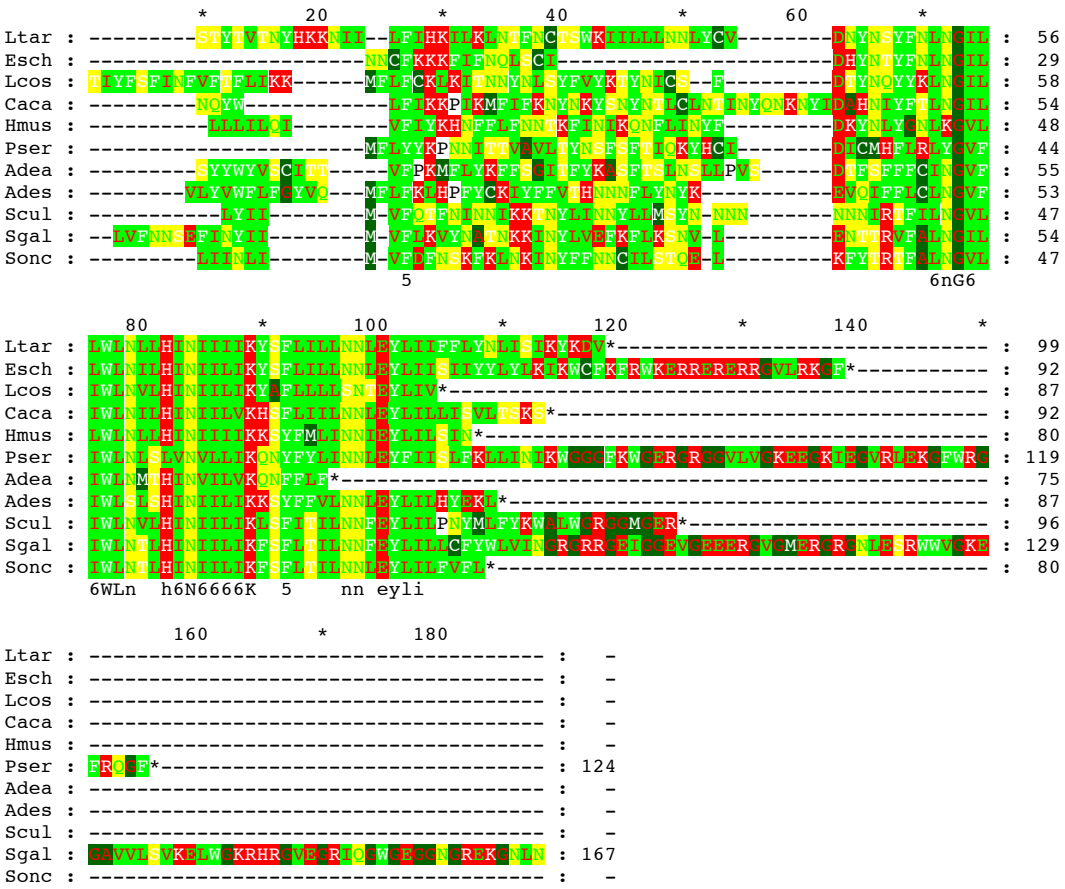
3.SI-S13d: ND7 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	99	**100**									
L. cos	99	99	**100**								
C. aca	98	98	98	**100**							
H. mus	83	83	84	83	**100**						
P. ser	81	80	81	81	74	**100**					
A. dea	90	91	90	91	80	80	**100**				
A. des	92	92	91	92	81	80	97	**100**			
S. cul	93	93	92	93	82	82	95	96	**100**		
S. gal	93	93	92	93	82	82	95	96	99	**100**	
S. onc	93	93	92	93	82	82	95	96	100	100	**100**


Ltar
Esch
Lcos
Caca
Hmus
Pser
Adea
Ades
Scul
Sgal
Sonc



3.SI-S14a: GeneDoc MURF5 nucleotide alignment: Abbreviations for the names of the organisms, color code, gaps and consensus for the figure are as mentioned in S1.



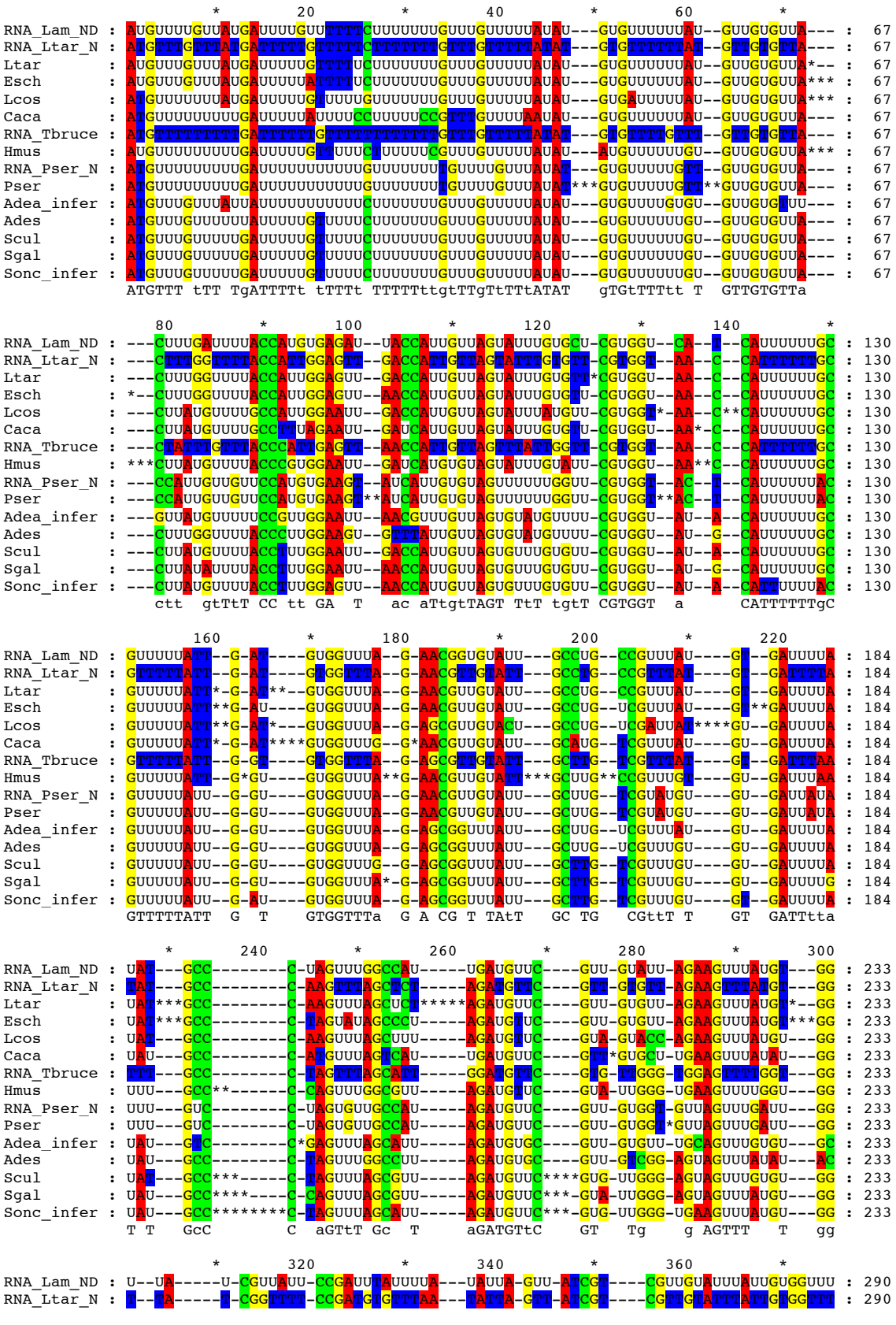
3.SI-S14b: GeneDoc MURF5 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment.

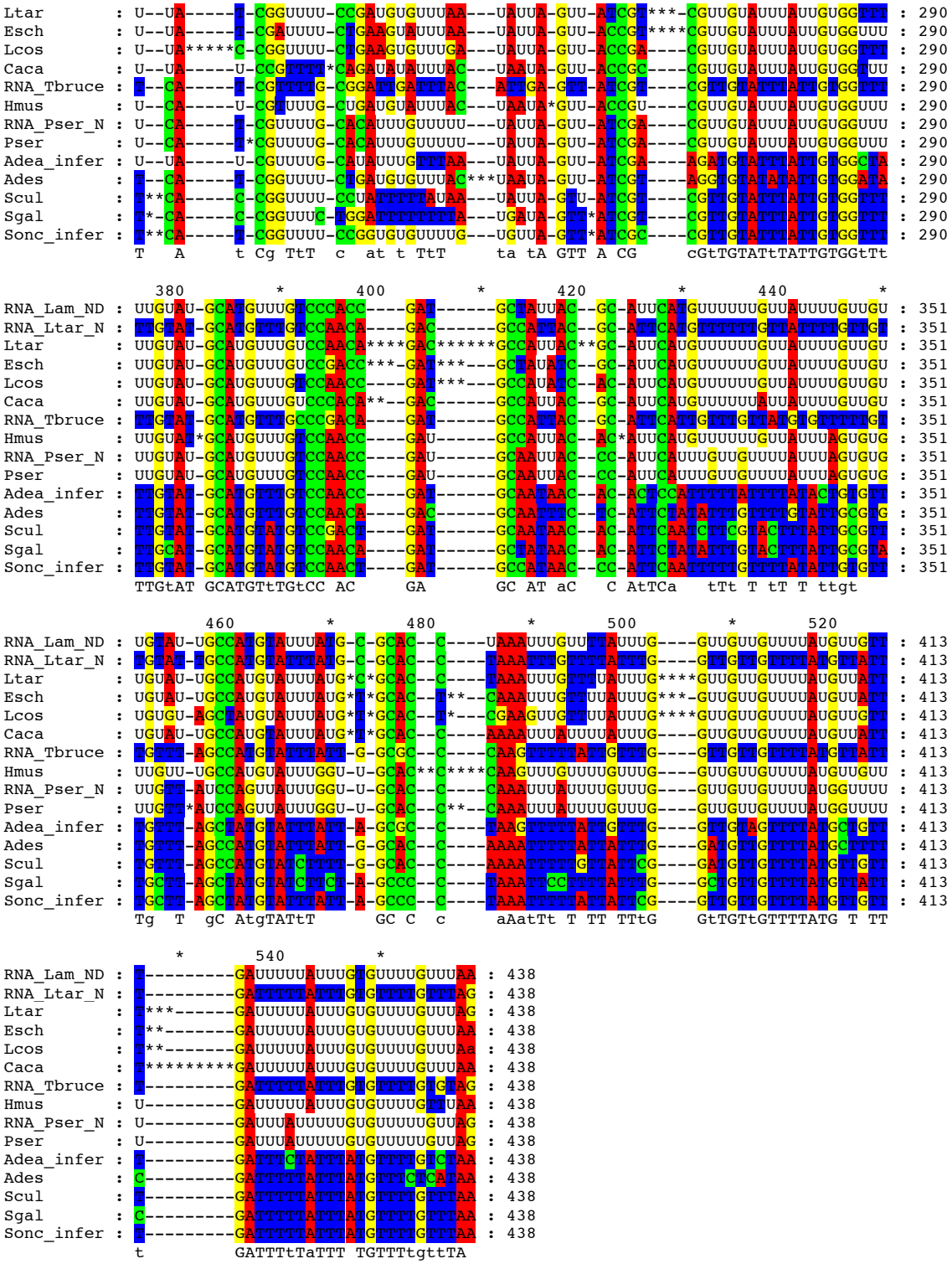
3.SI-S14c: MURF5 nucleotide percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	72	**100**									
L. cos	65	67	**100**								
C. aca	63	67	62	**100**							
H. mus	63	65	67	61	**100**						
P. ser	53	49	60	54	55	**100**					
A. dea	58	59	54	55	59	57	**100**				
A. des	56	61	62	60	62	59	52	**100**			
S. cul	62	62	61	62	65	47	53	61	**100**		
S. gal	57	55	59	60	62	44	53	55	64	**100**	
S. onc	62	65	58	62	65	53	56	55	67	72	**100**

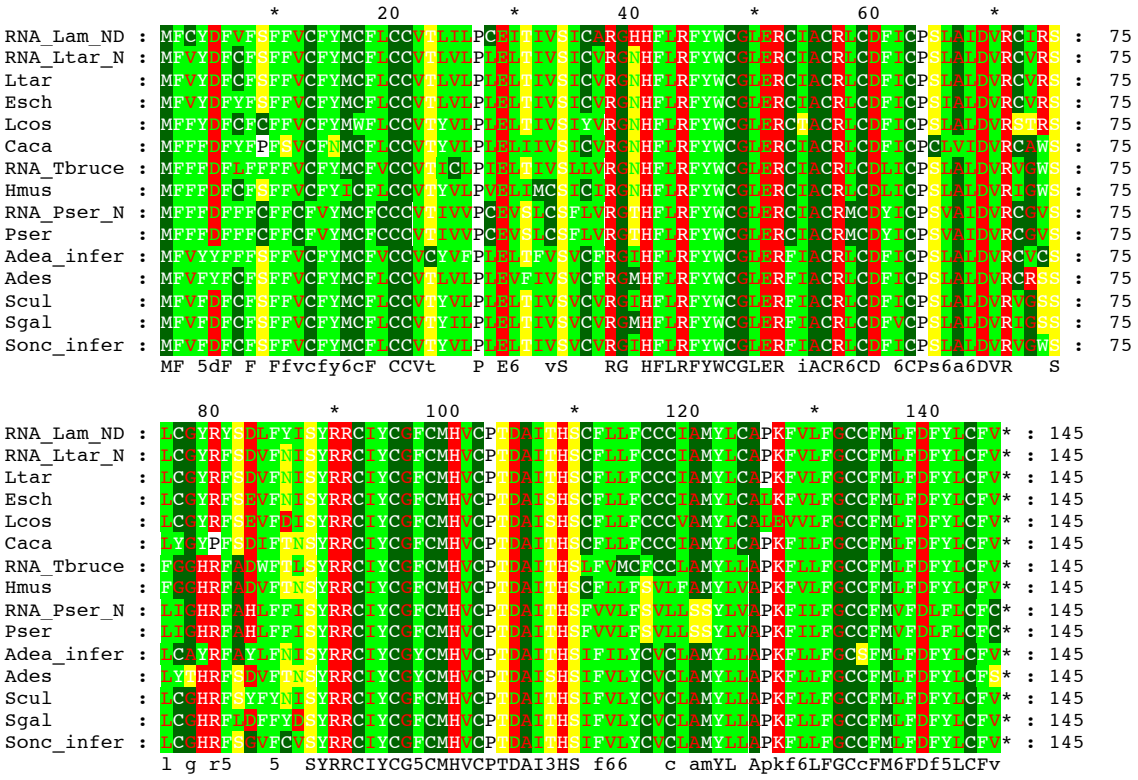
3.SI-S14d: MURF5 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	60	**100**									
L. cos	43	60	**100**								
C. aca	46	58	50	**100**							
H. mus	45	56	39	39	**100**						
P. ser	29	25	30	25	30	**100**					
A. dea	27	40	28	32	25	34	**100**				
A. des	31	38	31	39	35	39	31	**100**			
S. cul	33	40	42	43	39	25	26	41	**100**		
S. gal	38	41	40	43	35	21	30	35	52	**100**	
S. onc	39	51	43	43	39	26	32	39	55	60	**100**





3.SI-S15a: GeneDoc ND8 / G1 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNAs of *L. tar* and *T. bru* and RNA sequences of *P. ser, A. dea, S. onc* and *Lam.*



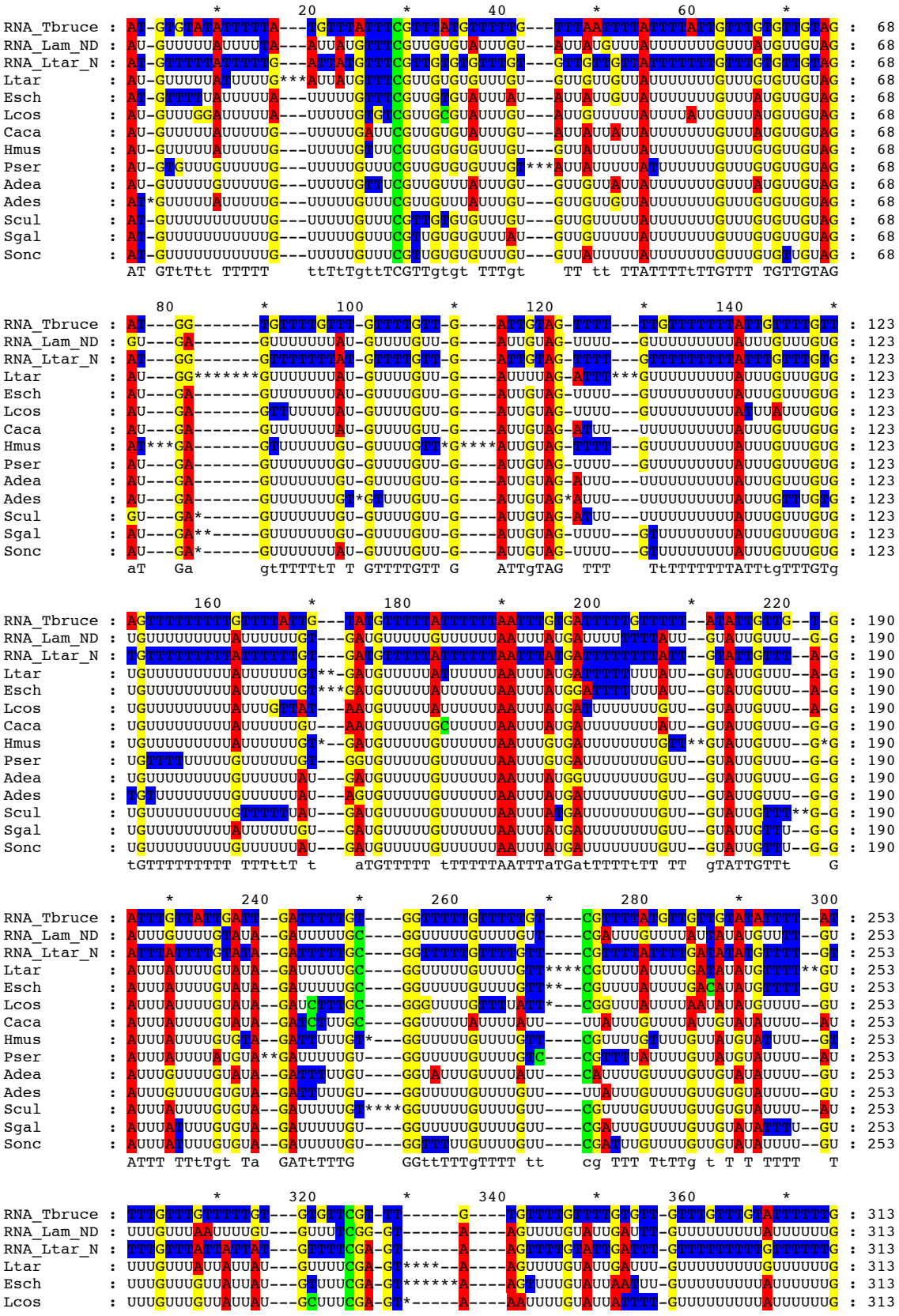
3.SI-S15b: GeneDoc ND8/ G1 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

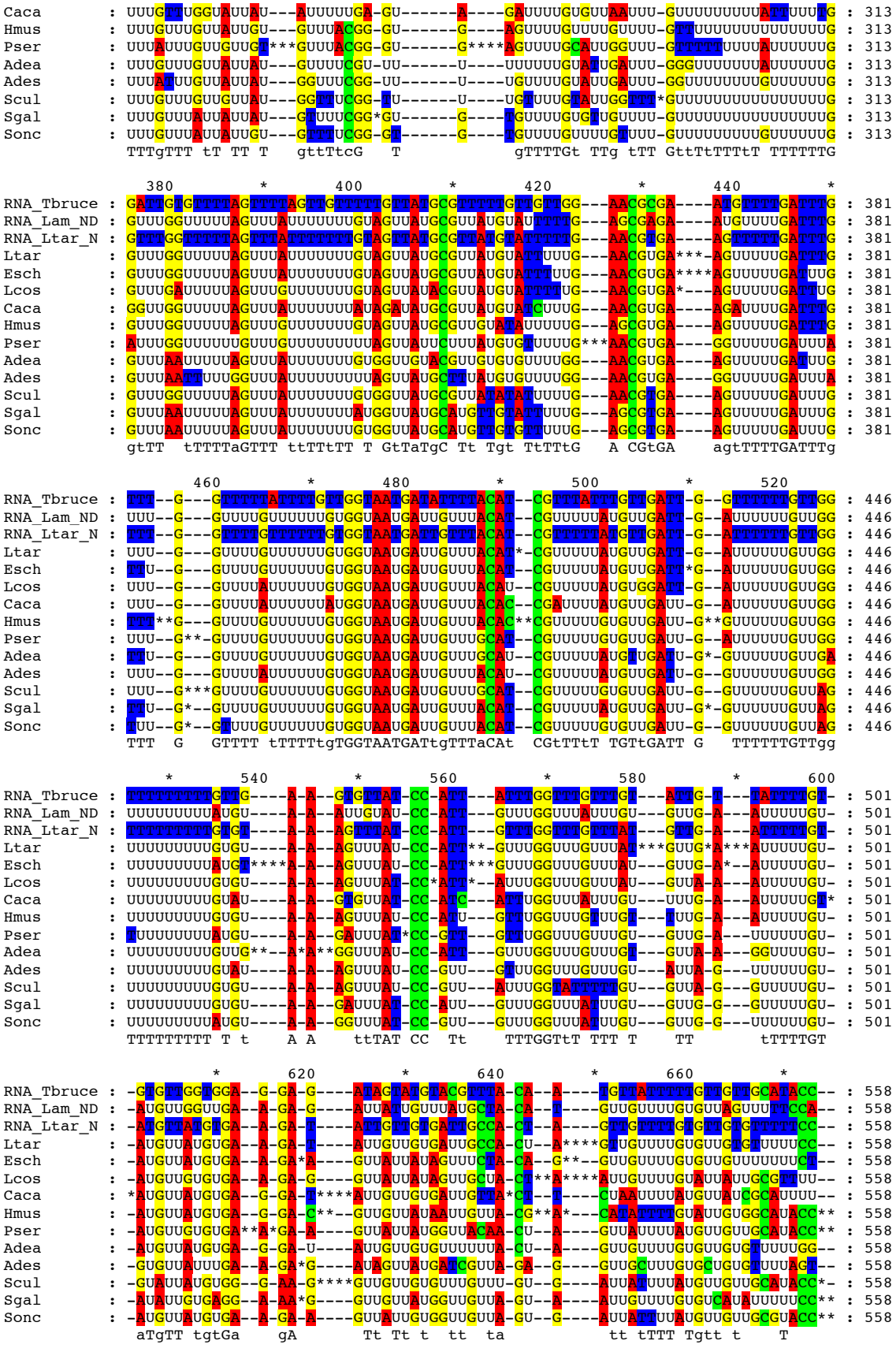
3.SI-S15c: ND8 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

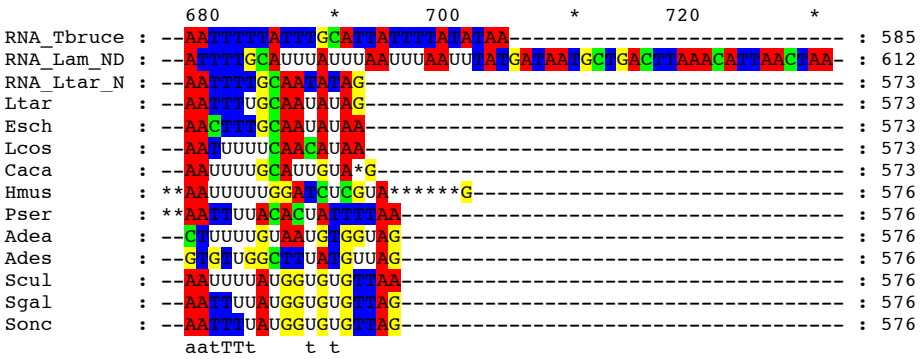
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	95	**100**									
L. cos	91	92	**100**								
C. aca	91	90	88	**100**							
H. mus	87	86	86	86	**100**						
P. ser	79	79	77	77	82	**100**					
A. dea	83	82	82	79	81	78	**100**				
A. des	83	82	80	81	82	77	84	**100**			
S. cul	85	84	84	83	84	77	85	86	**100**		
S. gal	84	83	83	81	83	76	83	85	92	**100**	
S. onc	87	86	86	84	85	79	87	85	92	90	**100**

3.SI-S15d: ND8 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

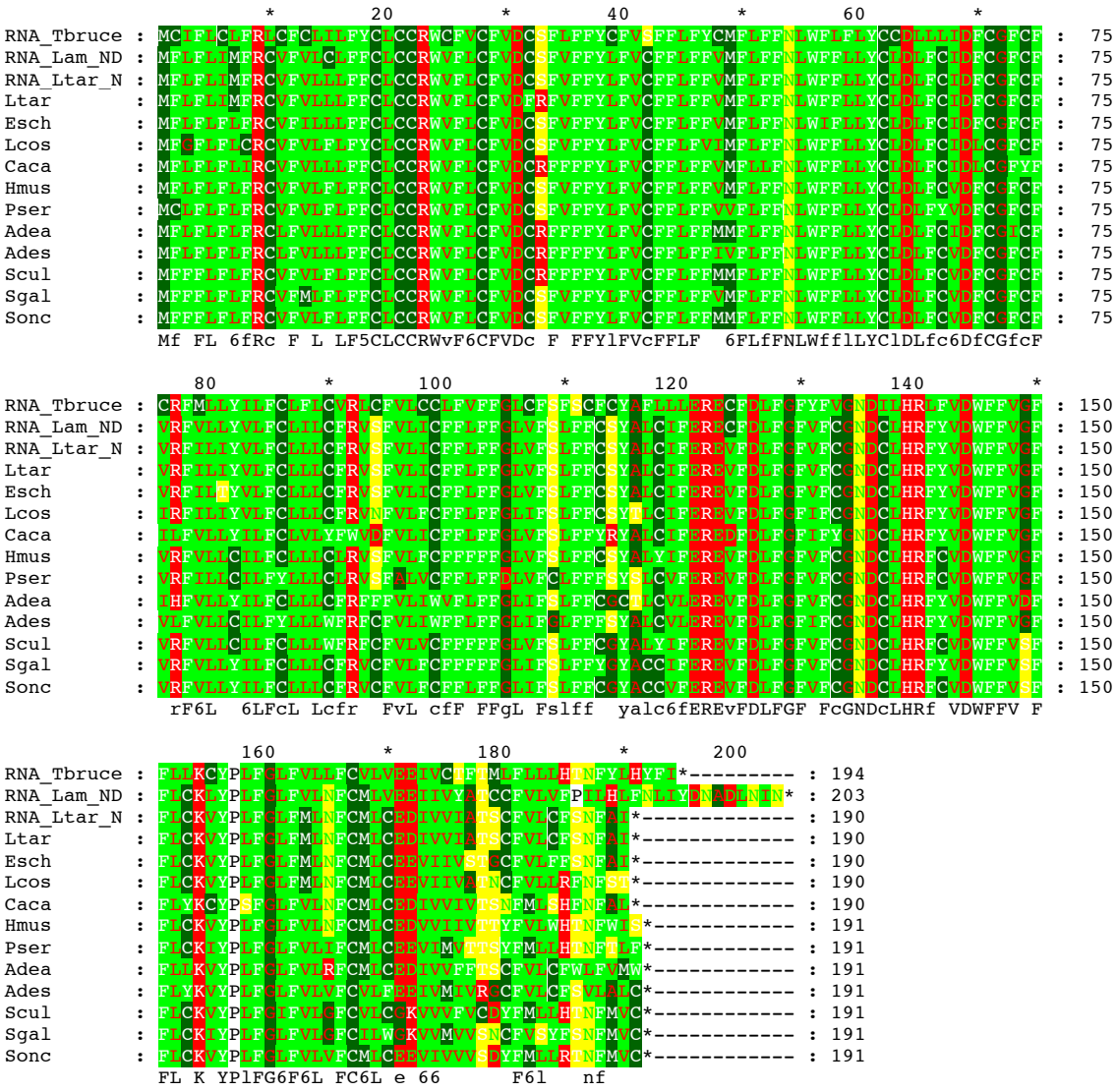
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	97	**100**									
L. cos	90	90	**100**								
C. aca	87	85	81	**100**							
H. mus	83	80	79	81	**100**						
P. ser	69	68	66	68	72	**100**					
A. dea	83	81	75	74	74	67	**100**				
A. des	83	81	76	78	78	68	83	**100**			
S. cul	87	84	81	79	80	70	87	88	**100**		
S. gal	85	81	79	79	81	69	83	88	94	**100**	
S. onc	87	84	81	81	82	70	86	88	96	93	**100**







3.SI-S16a: GeneDoc ND9 / G2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as “U” against a white background in the alignment. Deleted Ts have been denoted by “*” in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNAs of *L. tar* and *T. bru* and RNA sequence of *Lam.*



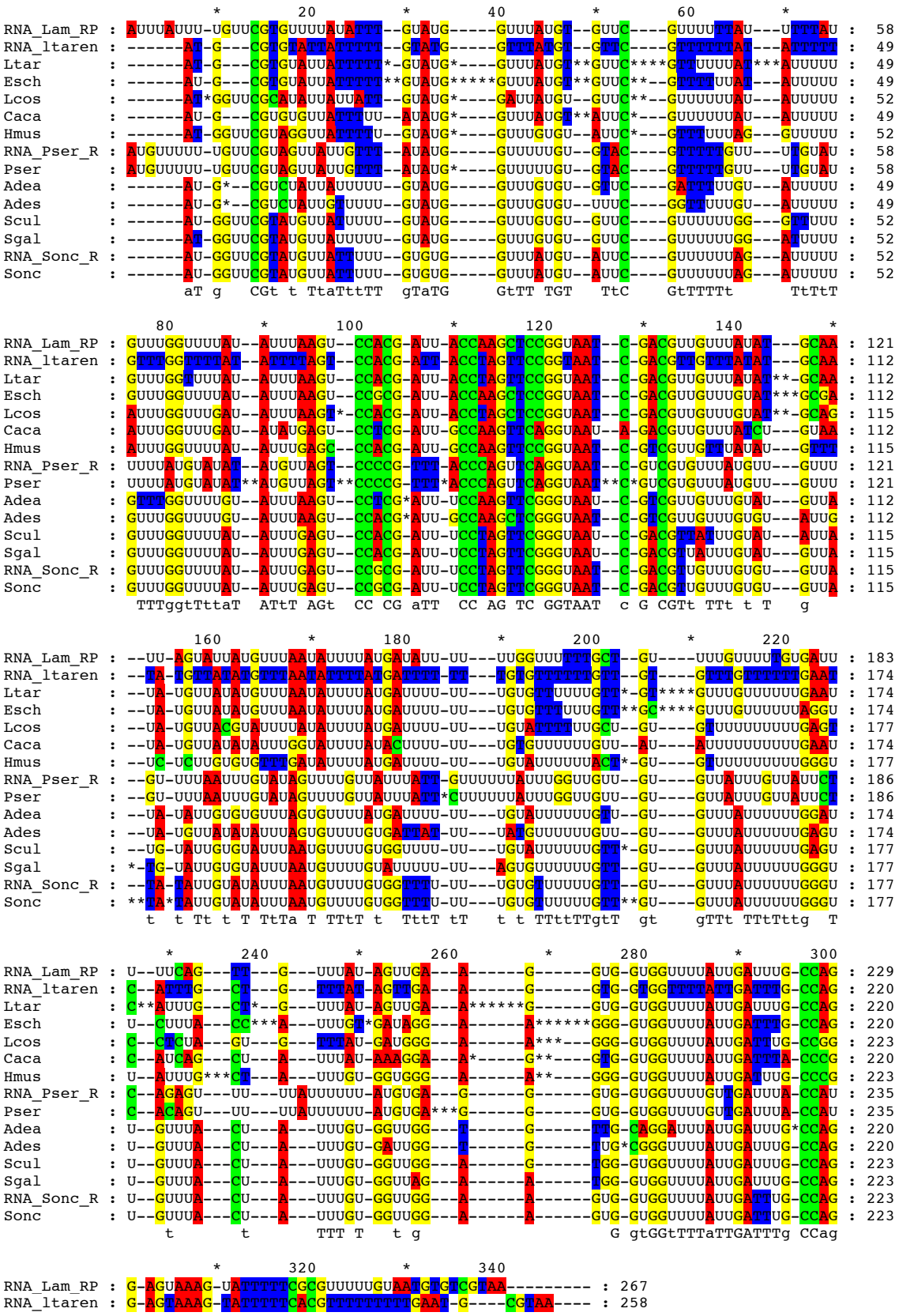
3.SI-S16b: GeneDoc ND9/ G2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

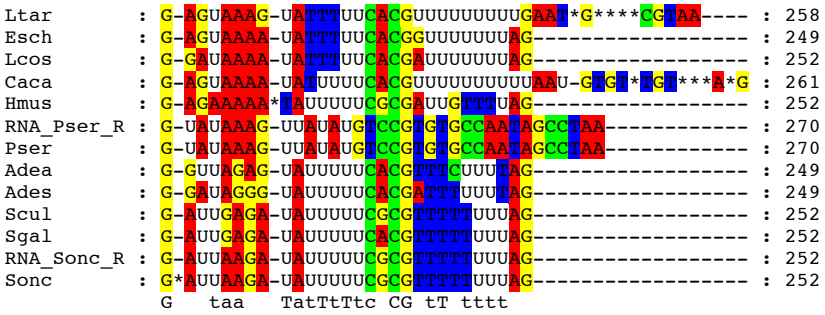
3.SI-S16c: ND9 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	94	**100**									
L. cos	90	92	**100**								
C. aca	88	87	86	**100**							
H. mus	90	89	87	86	**100**						
P. ser	88	88	85	83	90	**100**					
A. dea	89	87	85	85	87	84	**100**				
A. des	87	86	83	84	86	85	89	**100**			
S. cul	87	86	84	85	90	88	88	88	**100**		
S. gal	89	87	86	85	90	87	88	87	91	**100**	
S. onc	88	88	86	85	91	90	88	88	92	94	**100**

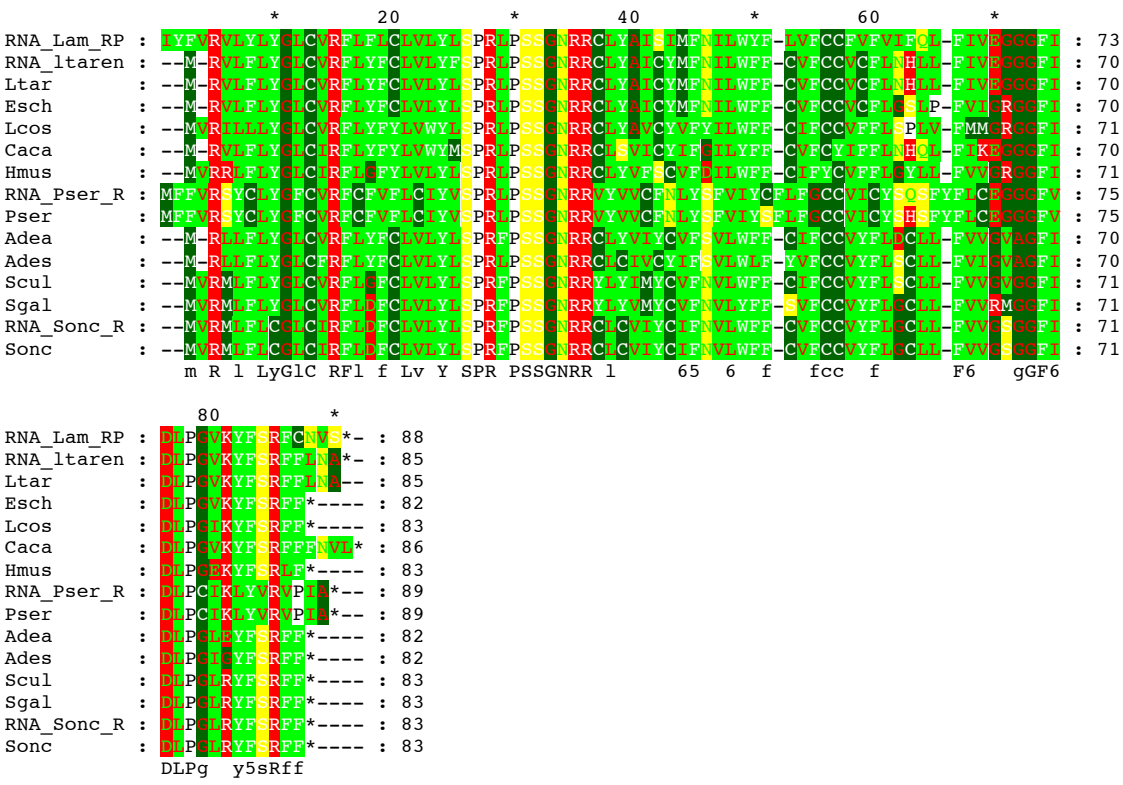
3.SI-S16d: ND9 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	*E. sch*	*L. cos*	*C. aca*	*H. mus*	*P. ser*	*A. dea*	*A. des*	*S. cul*	*S. gal*	*S. onc*
L. tar	**100**										
E. sch	92	**100**									
L. cos	85	88	**100**								
C. aca	83	79	78	**100**							
H. mus	87	87	83	81	**100**						
P. ser	80	83	79	75	86	**100**					
A. dea	83	80	78	79	79	74	**100**				
A. des	81	79	75	78	79	77	82	**100**			
S. cul	79	79	77	77	87	80	79	79	**100**		
S. gal	81	83	81	76	84	79	77	79	86	**100**	
S. onc	82	84	84	77	87	84	80	79	90	89	**100**





3.SI-S17a: GeneDoc RPS12 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNA of *L. tar* and RNA sequences of *Lam, S. onc* and *P. ser.*



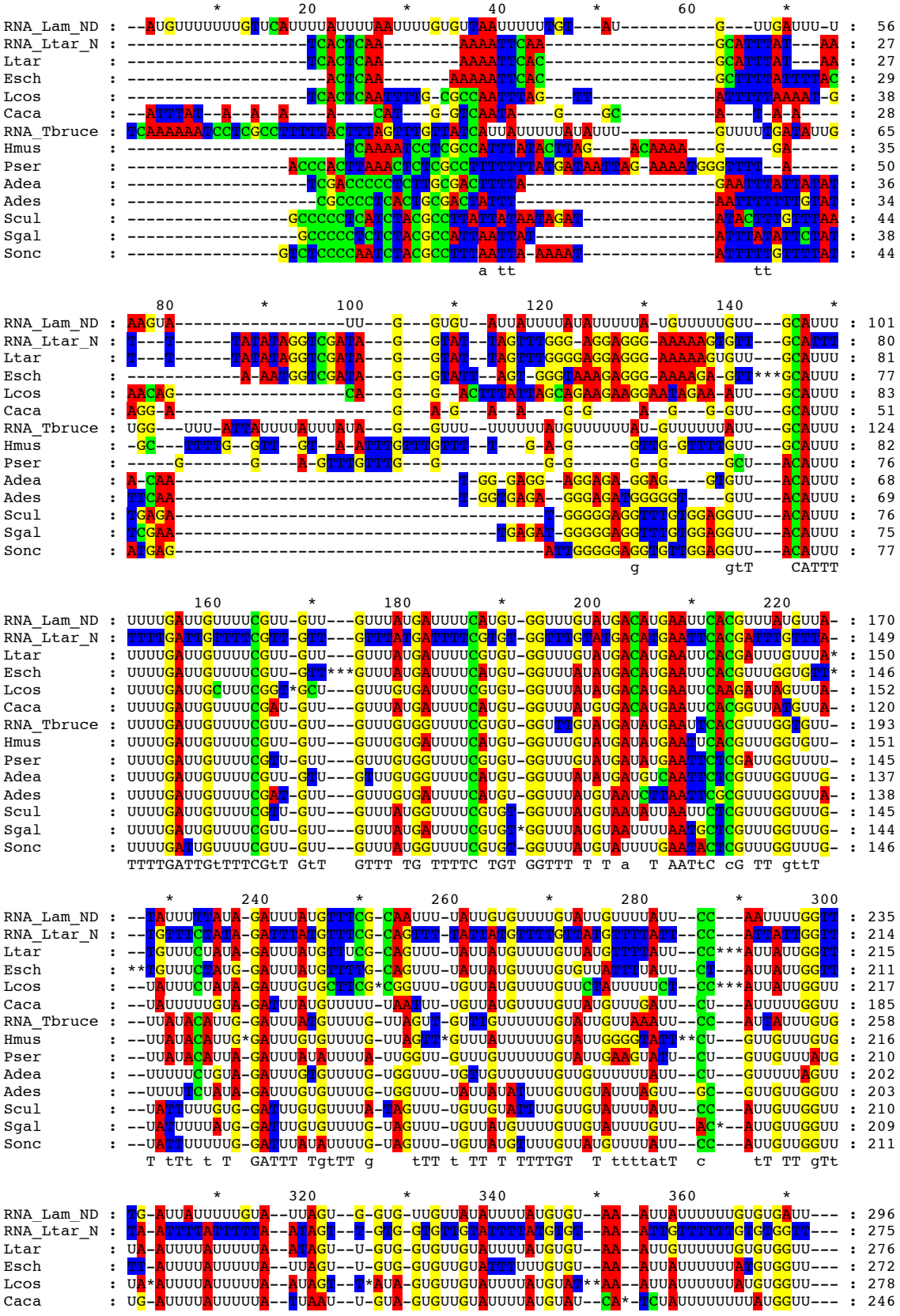
3.SI-S17b: GeneDoc RPS12 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

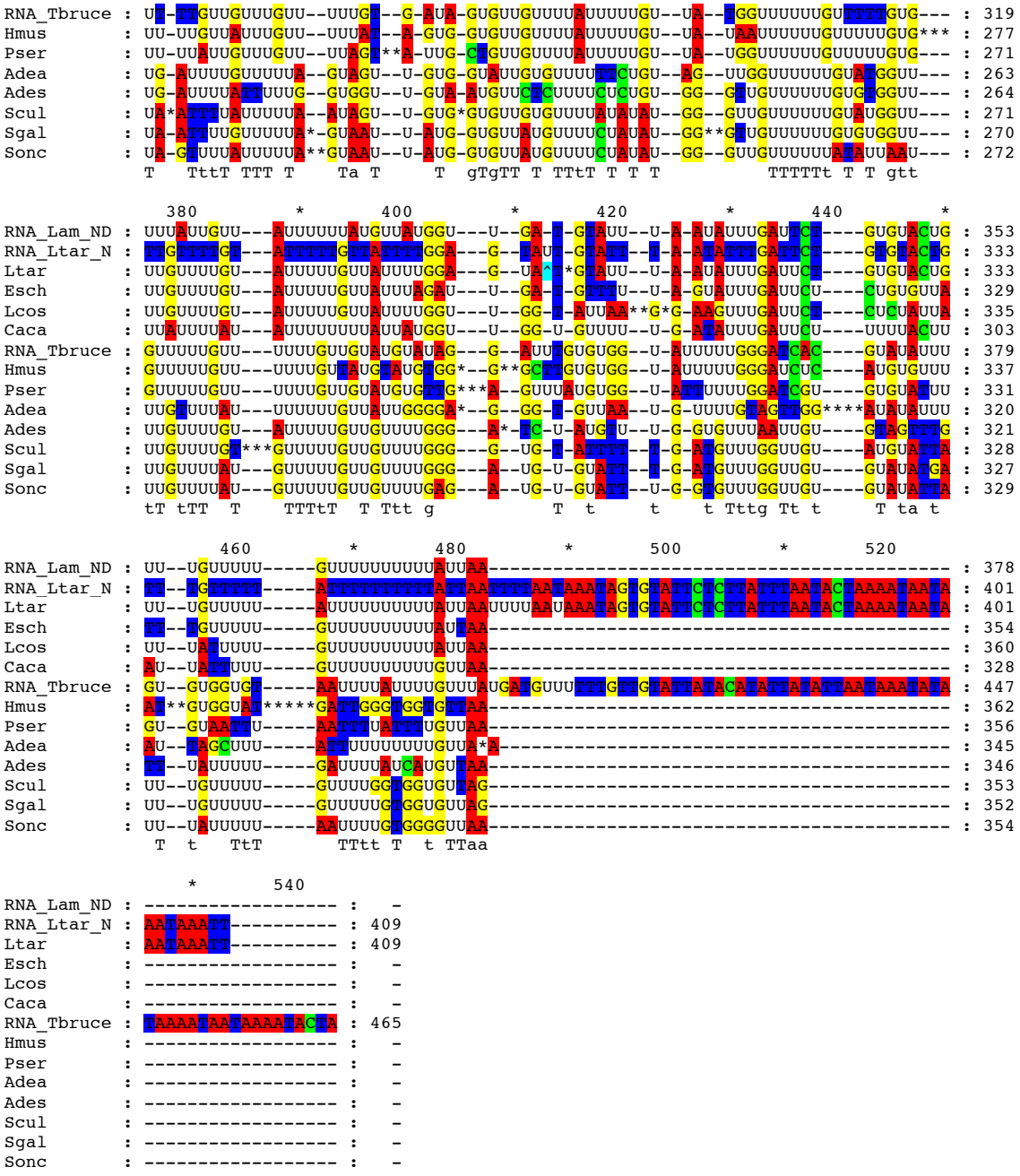
3.SI-S17c: RPS12 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	90	**100**									
L. cos	87	88	**100**								
C. aca	87	82	81	**100**							
H. mus	83	84	83	81	**100**						
P. ser	62	58	57	59	59	**100**					
A. dea	85	84	80	77	82	59	**100**				
A. des	82	84	81	77	78	57	90	**100**			
S. cul	84	84	82	78	86	59	88	85	**100**		
S. gal	84	85	81	78	85	58	87	84	96	**100**	
S. onc	86	88	82	81	86	59	87	85	93	93	**100**

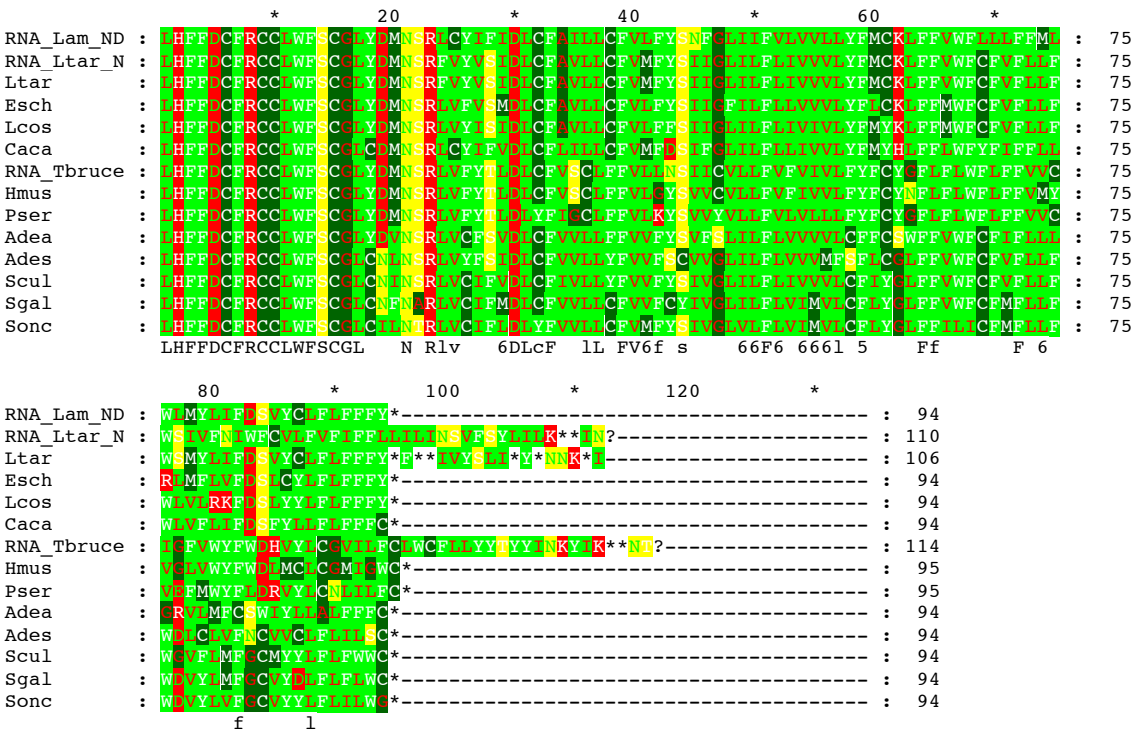
3.SI-S17d: RPS12 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	91	**100**									
L. cos	79	82	**100**								
C. aca	81	76	73	**100**							
H. mus	76	78	78	71	**100**						
P. ser	43	40	43	40	40	**100**					
A. dea	78	78	74	67	78	39	**100**				
A. des	76	78	74	68	70	40	82	**100**			
S. cul	77	77	75	63	78	40	89	78	**100**		
S. gal	76	76	69	66	75	40	84	73	90	**100**	
S. onc	77	78	69	70	77	37	85	78	87	88	**100**





3.SI-S18a: GeneDoc ND3 / G5 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. The discussed U residue has been denoted as '^' against a turquoise background. References are cDNA of *L. tar* and RNA sequences of *T. bru* and *Lam.*



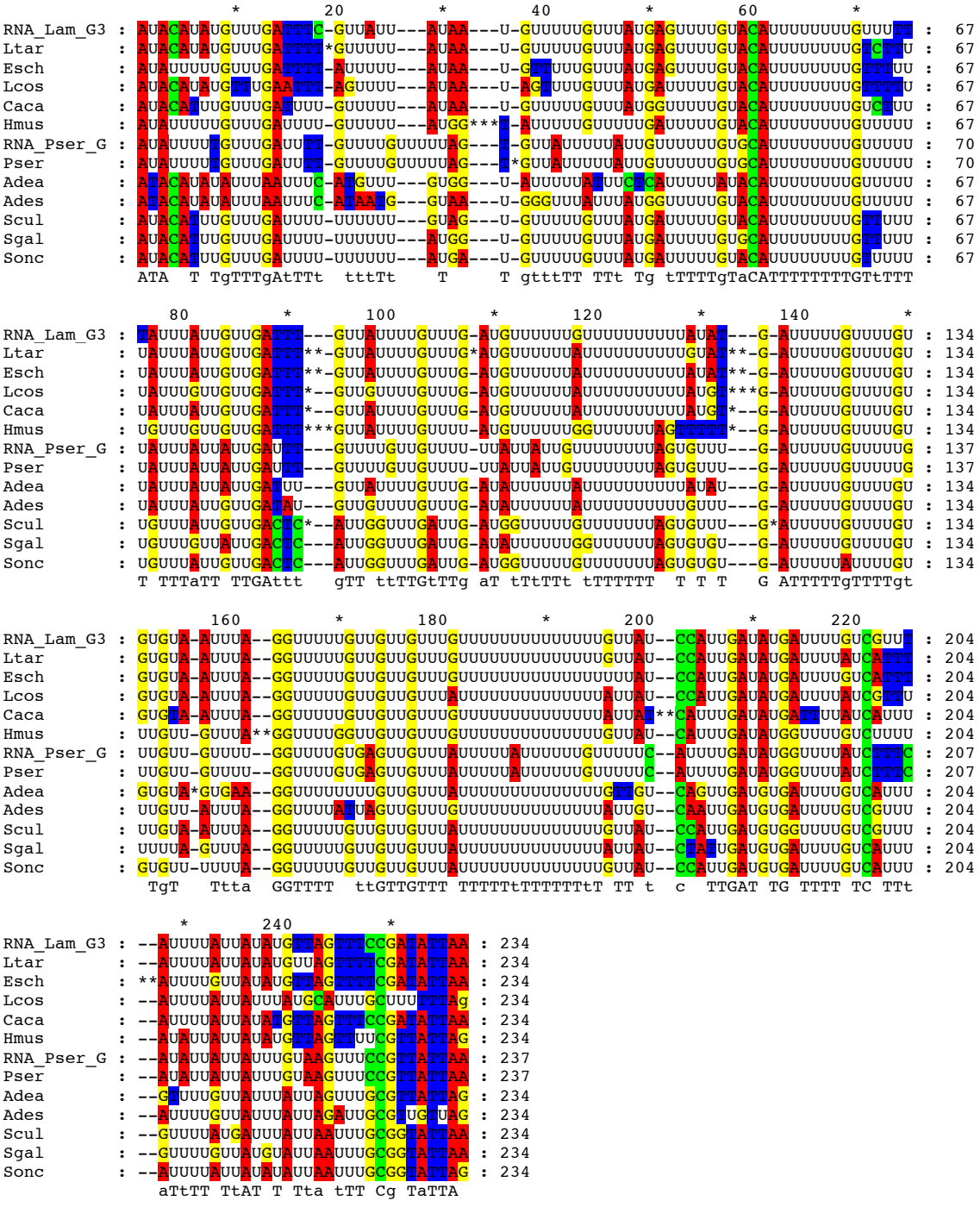
3.SI-S18b: GeneDoc ND3 / G5 protein alignment of the conserved motif: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated. ? denotes incomplete or missing information.

3.SI-S18c: ND3 edited transcript percent identity matrix (conserved motif): Abbreviations for the names of the organisms are as mentioned in the text.

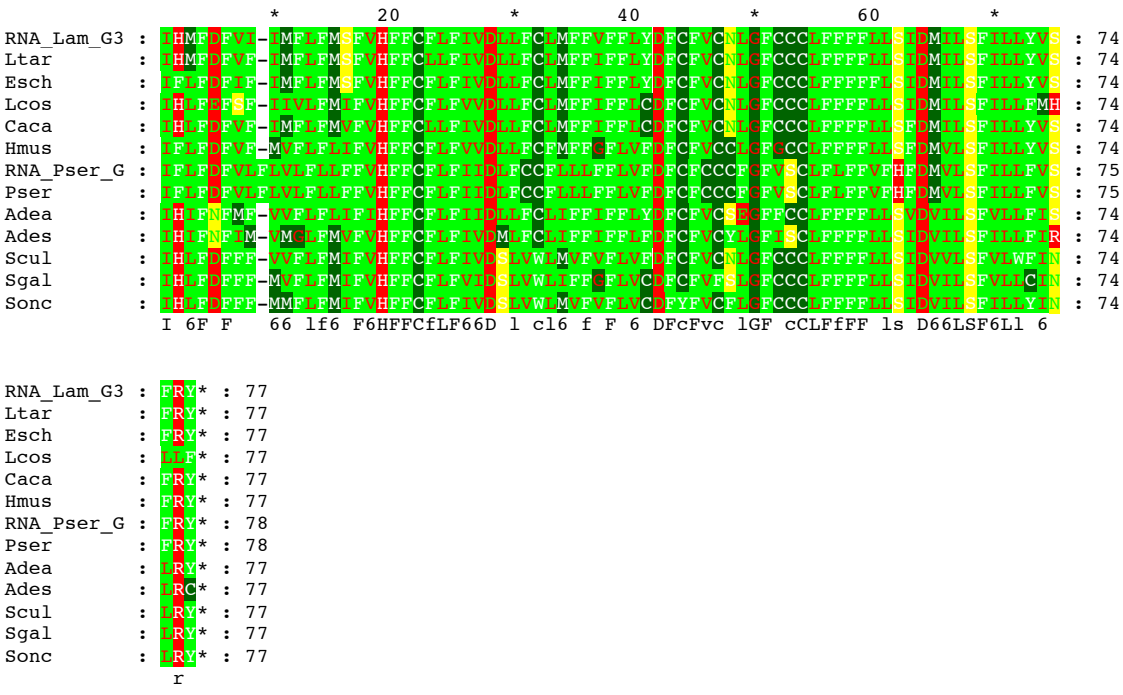
	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	90	**100**									
L. cos	86	85	**100**								
C. aca	84	83	82	**100**							
H. mus	65	69	61	66	**100**						
P. ser	67	68	62	65	84	**100**					
A. dea	78	78	76	78	69	71	**100**				
A. des	79	78	77	75	64	64	78	**100**			
S. cul	82	80	79	79	66	65	81	82	**100**		
S. gal	80	79	78	77	63	62	79	84	92	**100**	
S. onc	79	77	75	77	63	65	77	80	87	91	**100**

3.SI-S18d: ND3 protein percent identity matrix (conserved motif): Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	84	**100**									
L. cos	85	84	**100**								
C. aca	73	70	74	**100**							
H. mus	47	50	46	44	**100**						
P. ser	44	49	45	40	76	**100**					
A. dea	67	65	66	65	49	46	**100**				
A. des	69	65	64	59	45	41	64	**100**			
S. cul	71	69	72	70	45	44	68	72	**100**		
S. gal	67	64	67	67	41	37	64	73	82	**100**	
S. onc	64	62	64	63	45	44	57	68	74	83	**100**



3.SI-S19a: GeneDoc G3 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are RNA sequences of *P. ser* and *Lam.*



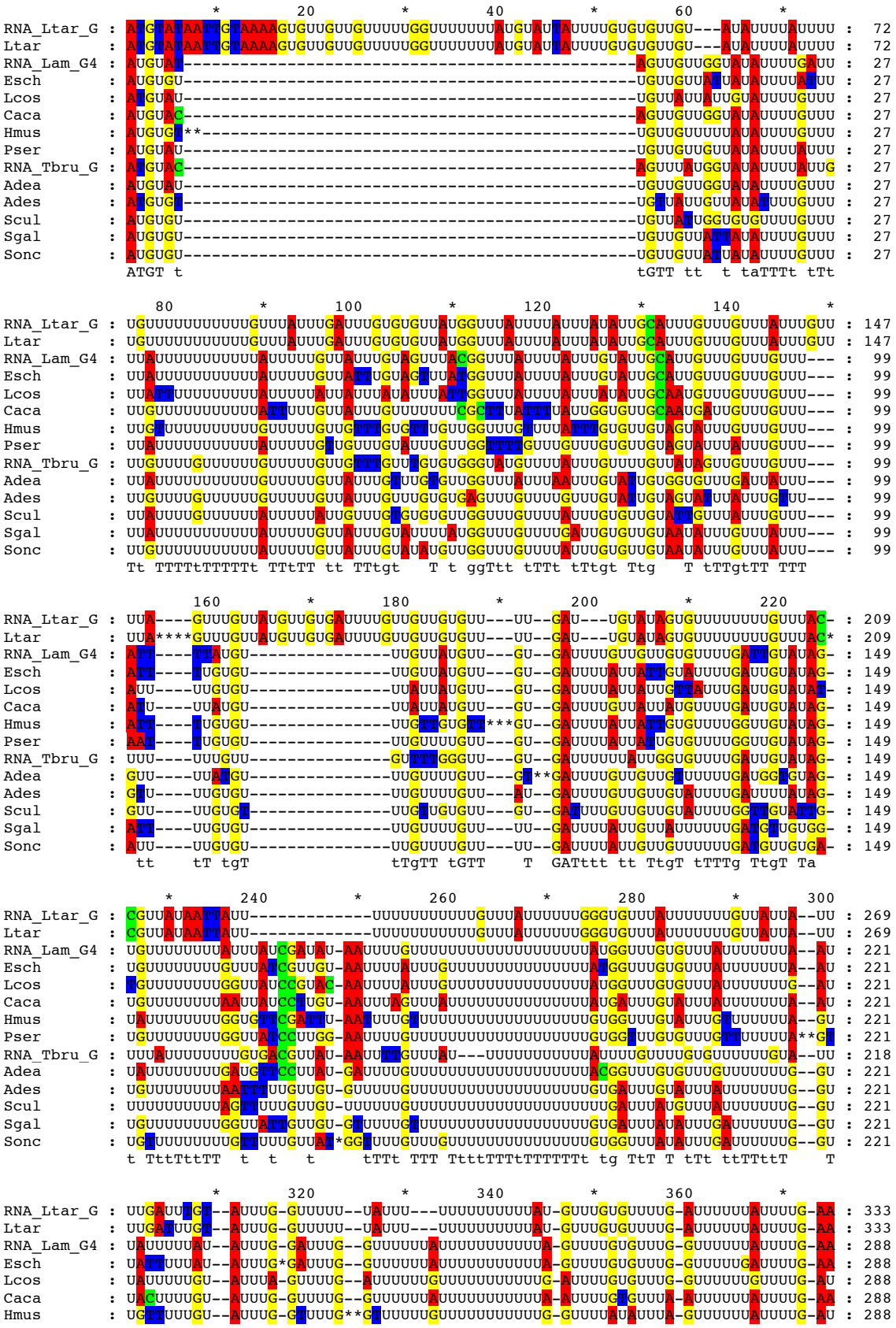
3.SI-S19b: GeneDoc G3 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

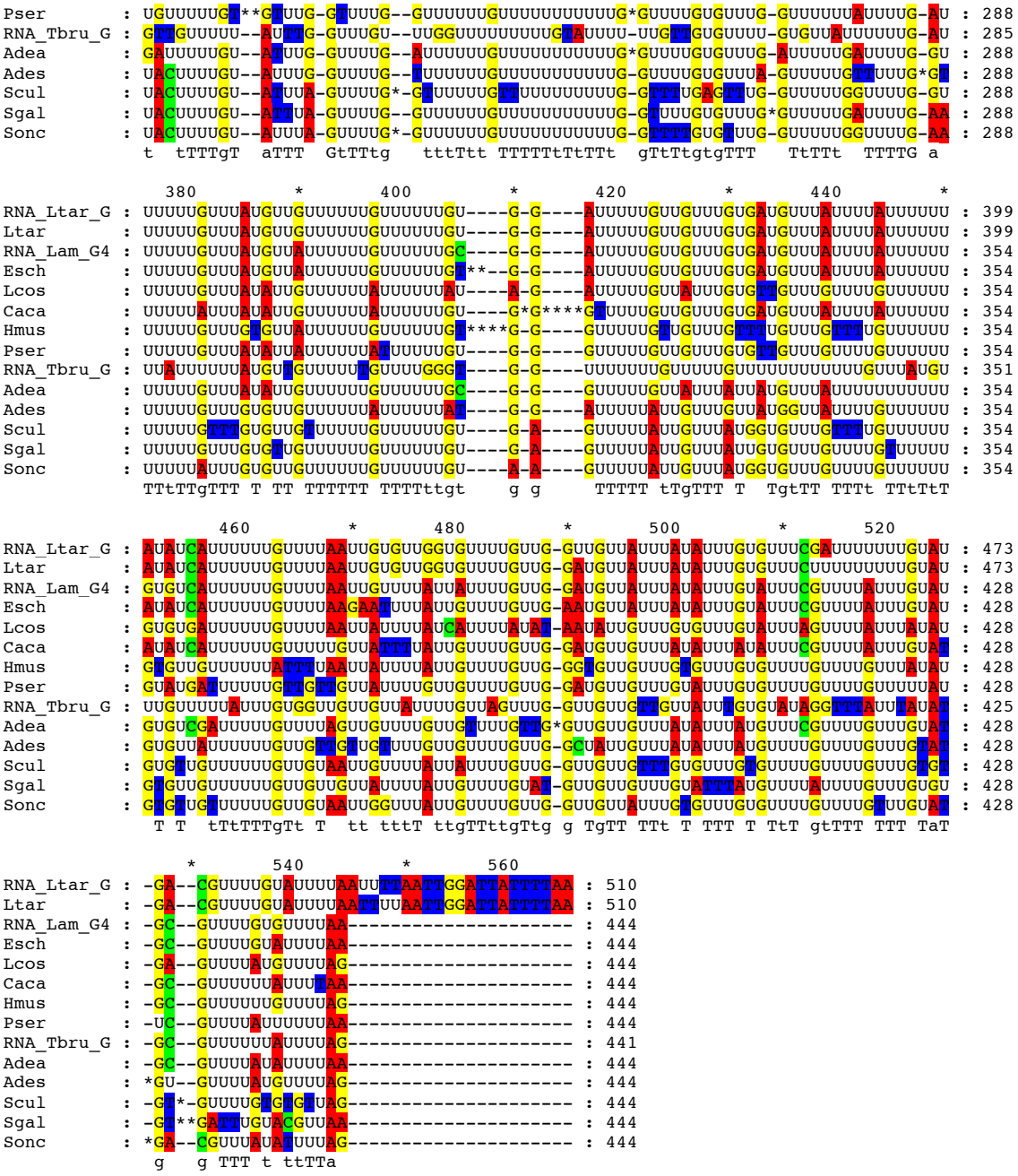
3.SI-S19c: G3 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	96	**100**									
L. cos	90	89	**100**								
C. aca	96	94	90	**100**							
H. mus	87	88	83	88	**100**						
P. ser	77	78	74	79	82	**100**					
A. dea	85	85	84	85	82	74	**100**				
A. des	84	84	85	85	80	75	87	**100**			
S. cul	87	86	85	85	85	76	82	82	**100**		
S. gal	84	85	84	85	85	76	83	80	92	**100**	
S. onc	88	87	86	87	86	74	82	81	94	92	**100**

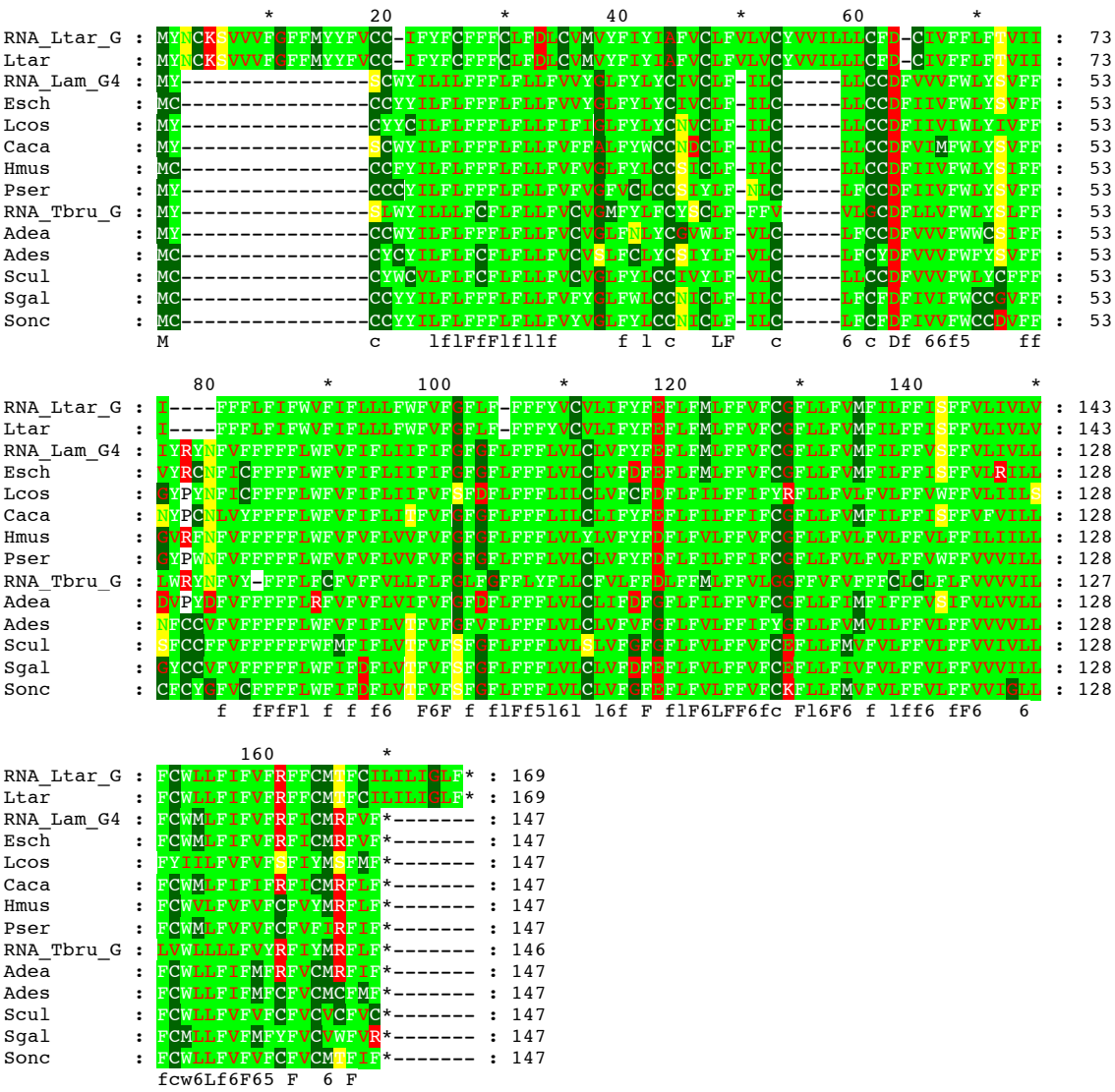
3.SI-S19d: G3 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	L. tar	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
L. tar	**100**										
E. sch	94	**100**									
L. cos	81	81	**100**								
C. aca	95	91	82	**100**							
H. mus	78	79	71	81	**100**						
P. ser	56	58	49	58	70	**100**					
A. dea	74	73	71	73	69	53	**100**				
A. des	74	74	74	74	64	51	77	**100**			
S. cul	71	71	71	71	73	55	71	70	**100**		
S. gal	69	69	71	70	73	51	74	68	84	**100**	
S. onc	75	75	73	77	73	51	68	70	88	86	**100**





3.SI-S20a: GeneDoc G4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are RNA sequences of *L. tar, T. bru* and *Lam.*



3.SI-S20b: GeneDoc G4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. * indicates stop codon in the alignment. References have also been translated.

3.SI-S20c: G4 edited transcript percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	Lam	***L. tar***	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
Lam	**100**											
L. tar	76	**100**										
E. sch	94	75	**100**									
L. cos	86	70	86	**100**								
C. aca	90	74	89	84	**100**							
H. mus	86	72	86	84	84	**100**						
P. ser	86	73	87	85	86	90	**100**					
A. dea	86	73	84	82	85	85	85	**100**				
A. des	83	70	82	81	84	86	86	86	**100**			
S. cul	84	70	82	82	80	87	84	84	88	**100**		
S. gal	82	69	83	81	82	85	85	84	86	88	**100**	
S. onc	82	71	83	83	81	87	84	84	86	88	92	**100**

3.SI-S20d: G4 protein percent identity matrix: Abbreviations for the names of the organisms are as mentioned in the text.

	Lam	***L. tar***	***E. sch***	***L. cos***	***C. aca***	***H. mus***	***P. ser***	***A. dea***	***A. des***	***S. cul***	***S. gal***	***S. onc***
Lam	**100**											
L. tar	62	**100**										
E. sch	90	61	**100**									
L. cos	73	49	76	**100**								
C. aca	82	56	81	71	**100**							
H. mus	78	54	78	73	73	**100**						
P. ser	74	51	74	73	75	84	**100**					
A. dea	78	56	73	64	71	72	73	**100**				
A. des	72	51	71	64	70	72	75	76	**100**			
S. cul	71	51	69	65	64	73	69	69	77	**100**		
S. gal	69	50	72	66	68	74	73	69	73	75	**100**	
S. onc	72	54	73	69	67	78	74	71	74	78	85	**100**

CHAPTER 4

The maxicircle genomes of several trypanosomatids of the *Trypanosoma* group (*T. cruzi* related and unrelated parasites)

Abstract

In this report, we have sequenced the maxcircle genomes of six new strains of *Trypanosoma cruzi* (*T. cruzi* CL, *T. cruzi* Tulahuen, *T. cruzi* M6241, *T. cruzi* CAN III, *T. cruzi* G and *T. cruzi* JR), representing several different *T. cruzi* DTUs, along with *T. serpentis*, *T. lewisi*, *T. conorhini*, *T. dionisii* and *T. erneyi*; and compared them to the previously reported maxicircle genomes of *T. brucei*, *T. cruzi* CL Brener, *T. cruzi* Sylvio X10 and *T. cruzi* Esmeraldo. The maxicircle sequences of *T. cruzi marinkellei* and *T. rangeli* available in the public sequence databases were included in this study for better representation of all the trypanosomal clades. The maxicircle genomes were characterized with respect to their sizes, gene-content and synteny. All the maxicircles contain a syntenic panel of 20 genes (excluding any gRNA genes) stretching from the two mitochondrial ribosomal RNA genes through the NADH dehydrogenase-5 gene. The RNA editing patterns of each of the edited transcripts and the identities of their predicted proteins were inferred. 13 of the 20 genes encoded in the maxcircle apparently undergo partial or pan-editing in these trypanosomatids. The total number of genes edited and their editing patterns are identical across all the organisms included in this report. Phylogenetic analyses on the nuclear and maxicircle datasets reveal differences in the relationships between the *T. cruzi* members,

but are congruent for the non-*T. cruzi* parasites. Maximum Likelihood (ML) analysis on the quantification of the inferred RNA editing patterns, using a binary value based strategy suggests that RNA editing has evolved in parallel with the nuclear genes on a broad scale, although subtle differences are again observed within the *T. cruzi* family.

Introduction:

The family Trypanosomatidae, belonging to the Order Kinetoplastida, is an extensive group of protists, which are of great significance, primarily owing to their basal placement in the eukaryotic evolutionary tree, and due to them being one of the most successful parasites ever known to mankind. These protozoans are parasites of several insects, invertebrates and almost all the members of Phylum Chordata, in addition to many plants (Vickerman, 1994). *Trypansoma cruzi* and *T. brucei*, the etiological agents of American trypanosomiasis (Chagas disease) and African sleeping sickness (Human African Trypanosomiasis or HAT), and *Leishmania* species, the etiological agent of Leishmaniasis, are three of the most extensively studied parasites of this family. Members of the *Trypanosoma* genus are all digenetic parasites, meaning they have two hosts, an animal host (vertebrates including aquatic animals, amphibians, reptiles, birds and mammals) and a vector host (including insects, arthropods and invertebrates) that transmits the disease between the animal hosts.

The subgenus Schizotrypanum (under the genus *Trypanosoma*) encompasses all the parasites that are restricted to hosts of the Order Chiroptera (bats), except *T. cruzi*, which is a generalist parasite with a wide host range among mammals. Species in this subgenus in the current report include *T. cruzi*, *T. cruzi marinkellei*, *T. erneyi* (former *Trypanosoma* sp. 1946) and *T. dionisii* (Cavazzana Jr. et al., 2010) (Franzén et al., 2012) (Lima et al., 2012). Based on several protein and genetic markers, *T. cruzi* has been classified intraspecifically into six discrete typing units (DTUs): TcI, TcIIa, TcIIb, TcIIc, TcIId and TcIIe (Brisse et al., 2000) (Brisse et al., 2001), which were later renamed as DTUs I, IV, II, III, V and VI, respectively (Zingales et al., 2009). Recently, a seventh DTU, Tcbat, was proposed (Marcili et al., 2009) (Lima et al., 2015). *T. rangeli*, a trypanosome of wild animals, and *T. conorhini*, a rat trypanosome, generally cluster

together, and as a sister clade to the Schizotrypanum in molecular phylogenies (Stevens et al., 1999) (Da Silva et al., 2004) (Maia da Silva et al., 2009). *T. lewisi*, a cosmopolitan parasite of rats, is in turn, a sister clade of *T. rangeli* and *T. conorhini* (Marcili et al., 2009) (Cavazzana Jr. et al., 2010) (Lima et al., 2012) (Lima et al., 2013). *T. serpentis*, a more recently characterized snake trypanosome, branches off earlier in the trypanosomatid lineage, and is closer to the early diverging *T. brucei* (Viola et al., 2009) (Caballero et al., 2015).

The Order Kinetoplastida derives its name from its members, who show the presence of a unique organelle, called the "kinetoplast". The kinetoplast is a large, single mitochondrion starting near the base of the cell's flagellum, and runs through the length of the trypanosomatid cell. It houses the mitochondrial DNA of the cell, called the kinetoplast DNA (kDNA), which is a massive network of thousands of catenated circular DNA molecules of two types: the larger maxicircles (about a few dozen copies per network) and the smaller minicircles (upwards of 5000 copies per network) (Shapiro and Englund, 1995) (Lukes et al., 2002) (Shlomai, 2004) (Liu et al., 2005). Depending upon the species, the maxicircles range in sizes between ~16 and 40 kb; are homogeneous in a network, and are the functional homologs of eukaryotic mitochondrial DNA that carry genes encoding multiple subunits of the mitochondrial oxidation chain and the ribosomal RNAs. The kDNA, however, does not encode any tRNAs and all the tRNAs required for translation are imported from the cytosol (Maslov et al., 1984) (Alfonzo and Söll, 2009).

Several maxicircle transcripts are "cryptic", meaning that they have to be modified post-transcriptionally before they become a translatable, mature mRNAs. After transcription, these transcripts are edited in a process that involves the addition and deletion of one or more uridylate residues at precise sites within the coding region of the transcripts. This RNA editing mechanism creates start and stop codons, corrects frameshifts and creates entire open reading frames from

primary transcripts that are sometimes less than half their mature size before the editing (Estévez and Simpson, 1999) (Simpson et al., 2003). The precise sites of the editing events are "guided" and dictated by small, 3' oligouridylated RNA molecules, aptly named the "guide RNAs" (gRNAs), which are predominantly encoded by the minicircles, and less frequently, by the maxicircles themselves (Blum et al., 1990) (Blum and Simpson, 1990) (Sturm and Simpson, 1990b).

Presently, the "enzyme-cascade model" is the most widely accepted for the U-insertion / deletion mechanism (Blum et al., 1990) (Alfonzo et al., 1997). The core "editosome" complex consists of a number of proteins, several of which are very well characterized and studied, that work together to mediate the initial interaction between the gRNAs and the pre-edited mRNAs (Aphasizhev et al., 2003) (Simpson et al., 2004b) (Stuart et al., 2005) (Aphasizhev and Aphasizheva, 2008) (Osato et al., 2009) (Li et al., 2009). Initially, an "anchor duplex" is formed by the interaction between a short stretch of nucleotides in the 5' end of the gRNA and the 3' end of the mRNA, followed by the progressive addition or removal of U residues at precise sites of the pre-edited transcripts, thereby improving the complementarity between the interacting molecules. Thus, editing proceeds stepwise, in the 3' to 5' direction with respect to the mRNA, wherein upstream anchor sequences can be formed only by the downstream editing. Every gRNA only directs editing of a small stretch of the mRNA, termed the "editing block", and often several gRNAs are required to edit the entire length of a pre-edited mRNA (Blum et al., 1990) (Maslov and Simpson, 1992) (Corell et al., 1993) (Simpson et al., 1998). Some transcripts are only partially edited, either at their 5' ends or internally; while several others are extensively edited across the length of the transcripts, a process called "pan-editing".

The trypanosomatid maxcircles generally contain 18 protein encoding genes, alongside 2 ribosomal RNA genes, all syntenically clustered together in about an ~16kb "conserved" or "consensus" region. Outside this syntenic region is a stretch of repetitive sequences (of variable length in different species), called the "divergent region" (DR) or the "variable region" (VR), that makes up the rest of the length of the circular molecule. Although the exact functions of this region are unknown, it has been hypothesized to contain sequences that allow binding of topoisomerase II or other dimeric or multimeric proteins; and to contain the maxicircle replication origin and the promoter regions for the 12S rRNA gene (Muhich et al., 1985) (de Vries et al., 1988) (Gorbat et al., 1990) (Sloof et al., 1992) (Myler et al., 1993) (Vasil'eva et al., 2004) (Westenberger et al., 2006). Among the 20 maxicircle genes, the genes ND8, ND9, COIII, CYb, MURF4 (A6), G3, COII, MURF2, ND4, RPS12 and ND5 are transcribed from the same strand as the ribosomal genes, 12S and 9S, in the same direction and polarity. However, the genes ND9, MURF5, MURF1, ND1, COI, G4 and G5 (ND3) are transcribed from the opposite strand, in the opposite direction and polarity (Cruz et al., 1984) (Simpson et al., 1987) (Maslov et al., 1999). Discounting the two ribosomal RNA genes, the transcripts of five other genes are not known to undergo editing: MURF1 (ND2), ND1, COI, ND4 and ND5; while three genes are known to undergo partial editing: CYb and MURF2 undergo 5' editing, and COII undergoes internal editing; with nine genes undergoing pan (extensive) editing: ND8, ND9, ND7, COIII, A6, G3, G4, ND3 and RPS12, in the maxicircle genomes of *T. brucei* and *T. cruzi*. The editing of MURF5 is not characterized and it is presumed to undergo 5' editing (Westenberger et al., 2006) (Maslov, 2010) (Ruvalcaba-Trejo and Sturm, 2011).

Among the members of the *Trypanosoma* genus, whole length maxicircle sequence of *T. brucei brucei* is available on NCBI (GenBank: M94286.1) (Eperon et al., 1983) (Benne et al., 1983)

(Hensgens et al., 1984) (Feagin et al., 1985) (Benne et al., 1986) (Feagin et al., 1987) (Shaw et al., 1988) (Feagin et al., 1988a) (de Vries et al., 1988) (Bhat et al., 1990) (Koslowsky et al., 1990) (Sloof et al., 1992) (Maslov et al., 1992) (Souza et al., 1992) (Souza et al., 1993). Also, the complete maxcircle sequences of *T. cruzi* CL Brener (GenBank: DQ343645.1) and *T. cruzi* Esmeraldo (GenBank: DQ343646.1); and partial maxicircle sequence of *T. cruzi* Sylvio X10 (GenBank: FJ203996.1) are available in NCBI (Westenberger et al., 2006). More recently, whole genome sequencing has made the maxicircle sequences of *T. cruzi marinkellei* strain B7 (GenBank: KC427240.1) (Franzén et al., 2012) and *T. rangeli* strain SC-58 (GenBank: KJ803830.1) (Stoco et al., 2014) available. There is a general paucity of maxicircle sequences and the RNA editing information in the *Trypanosoma* genus.

In this report, we have sequenced and characterized the maxicircle genomes of six new strains of *T. cruzi* (*T. cruzi* CL, *T. cruzi* Tulahuen, *T. cruzi* M6241, *T. cruzi* CAN III, *T. cruzi* G and *T. cruzi* JR), *T. erneyi*, *T. dionisii*$^{+}$, *T. conorhini*, *T. lewisi* and *T. serpentis*; and compared them to the previously published maxicircle sequences of *T. brucei*, *T. cruzi* CL Brener, *T. cruz*i Esmeraldo, *T. cruzi* Sylvio X10, *T. cruzi marinkellei* and *T. rangeli*. We have also inferred the apparent RNA editing patterns, sequences of the fully-edited transcripts and the resultant proteins of all the genes in all the organisms. Phylogenetic analyses on the maxicircle and nuclear genes reveal expected relationships between the various organisms. Finally, quantification of the RNA editing patterns of these genes using a binary value-based strategy clarifies the evolution of RNA editing among this group.

Abbreviations for organisms herein

Tbru = T. brucei; *TcCAN = T. cruzi* CAN III*; TcCLB = T. cruzi* CL Brener*; TcEsm = T. cruzi* Esmeraldo*; TcJR = T. cruzi* JR cl. 4*; TcM62 = T. cruzi* M6241*; TcMar = T. cruzi marinkellei; Tcon = T. conorhini; TcCL = T. cruzi* CL*; TcG = T. cruzi* G*; TcSyl = T. cruzi* Sylvio X10*; TcTula = T. cruzi* Tulahuen*, Tdio = T. dionisii*; *Tern = T. erneyi* (*Trypanosoma* sp. 1946)*; Tlew = T. lewisi; Tran = T. rangeli; Tser = T. serpentis.*

Gene Abbreviations

A6 = ATPase 6; CO = Cytochrome Oxidase; CYb = (Apo) Cytochrome b; DHFR = Dihydrofolate reductase; G = G-rich regions; GAPDH = Glyceraldehyde 3-Phosphate Dehydrogenase; GSH1 = Gamma-Glutamylcysteine Synthetase; MURF = Mitochondrial Unidentified Reading Frame; ND = NADH Dehydrogenase; RPS12 = Ribosomal Protein Subunit 12.

Other abbreviations

AA(s) = amino-acid(s); Avg = average; bp(s) = base-pair(s); gDNA = genomic DNA; indels = insertion-deletion mutations; kb = kilo bases; min = minutes; MSA = Multiple Sequence Alignment; No. = Number; nt(s) = nucleotide(s); PCR = polymerase chain reaction; s = seconds; T = Thymidylate residue; T_m = Temperature of melting; U = uridylate residue

Materials and Methods:

Trypanosomatid strains, culturing and DNA extraction

The following strains of *T. cruzi*: *TcG* (TCC30), *TcEsm, TcCL* (TCC33), *TcMar* (TCC344), and *Tern* (TCC1946), *Tran* AM80 (TCC086), *Tcon* (TCC025E) were grown in LIT medium (Camargo, 1964) supplemented with 10% FBS. *Tdio* (TCC211) and *Tlew* (TCC034) were grown in Grace's insect medium (GIBCO, Life Technologies, USA) supplemented with 10% FBS. Genomic DNA of *Tser* (TCC1052) and live cultures of all the trypanosomatids above were obtained from the Trypanosomatid Culture Collection of the University of Sao Paulo (Brazil), TCC-USP. Cultures were grown at 25^{o}C and culture volumes were expanded every 48 hours. Cultures were maintained in logarithmic phase and harvested by centrifuging them at 3000 rpm for 8 mins at 4^{o}C. The harvested pellets were washed thrice with phosphate buffered saline at pH 7.4, and total gDNA (with kDNA) was isolated from the washed pellets using ZR Genomic DNA - Tissue MiniPrep kit (Zymo Research, USA), as described by the manufacturer. Complete and draft genome sequences of *TcJR* (cl. 4), *TcCAN* (cl. 1), *TcM62* (cl. 6) and *TcTula* (cl. 2) are freely available from The Genome Institute at Washington University School of Medicine (St. Louis, USA) – project ID 59941. Similarly, genomes of *TcCLB*, *TcEsm* and *TcSyl* are available from The Wellcome Trust (TriTrypDB) (Aslett et al., 2010).

Shotgun sequencing of trypanosomatid genomes

Total DNA from all the organisms obtained from TCC-USP were subjected to shotgun sequencing using the Roche 454 FLX pyrosequencing platform, as per the manufacturer's protocols, and the reads were assembled using Newbler software (Roche) [version 2.3 for *TcG*, *TcJR* and *TcCL*; version 2.8 for *Tdio*, *Tran*, *Tcon* and *Tlew*; version 2.9 for *TcMar*, *Tern* and

Tser] as described previously (Alves et al., 2011) (Alves et al., 2013a) (Alves et al., 2013b). Putative maxicircle sequences were identified and selected from the genome assemblies using the published maxicircle sequence of *TcCLB* (GenBank) as a query in BLASTN version 2.2.25 (Altschul et al., 1990), with a maximum E-value threshold of $1x10^{-6}$ (*-evalue 1e-6*).

Maxicircle contig closures and gap filling

The 2 maxicircle fragments (contigs) of *TcJR*, and all of the 4 contigs of *TcTula* could be assembled by finding overlaps between adjacent contig sequences, using Sequencher version 5.0 sequence analysis software (Gene Codes Corporation, USA). For the other contigs that could not be assembled by this method, PCRs were set up as described in **Chapter 3** (Please see **Supplementary Table 4.1** for PCR primers and their T_m). The maxicircle contigs of *Tern* have not been closed due to the non-availability of cultures and total DNA at this point.

Analyzing the DNA sequences

The DNA sequences were analyzed as essentially described in **Chapter 3**.

Analyzing the RNA editing

The pre-edited sequences of the edited genes were aligned and compared to the previously published, fully-edited transcripts of *Tbru* and *T. cruzi* in GeneDoc. By parsimonious addition of Us and deletion of maxicircle encoded Ts, we were able to build complete open reading frames that showed high identities to the published transcripts and protein sequences, as described in **Chapter 3**. The resulting gene and protein alignments have been shown in the **Supporting Information (SI) for Chapter 4** section following this chapter, along with their identity matrices.

RNA folding and analyses

The two ribosomal RNA genes (12S and 9S) in each of the species' maxicircles were folded as described in **Chapter 3**.

Phylogenetic analyses

Gene and protein sequences were aligned individually using MUSCLE (MEGA) and manually curated and concatenated. The following three types of alignment-datasets were generated for analysis using the maximum likelihood (ML) method: **1) a nuclear dataset** consisting of a concatenated nucleotide alignment of DHFR, GAPDH and GSH1 genes; **2) a maxicircle dataset** consisting of a concatenated nucleotide alignment of the following non-edited maxicircle genes: 12S, 9S, MURF1, ND1, COI, ND4 and ND5. Missing gene(s) in one species were not compensated for in the others; **3) an RNA editing quantification dataset** consisting of a binary (1/0) matrix quantifying the editing events at each locus in each of the putative fully edited codign sequences of all the species, from the following edited genes: A6, COIII, G3, G4, ND3 (conserved motif), ND7, ND8, ND9 and RPS12 from the pan-edited genes, CYb and MURF2 from the 5' edited genes and COII which is internally edited. MURF5 was excluded from any of the datasets due to its uncharacterized editing pattern. Again, any missing RNA editing quantification event in one species was not compensated for in others. In the binary value strategy, the presence of any editing event (either insertion or deletion of one or more U residues) at a specific locus in the inferred, fully edited, mature transcript is scored as "1" and its absence in the same locus in another species is scored as "0". A detail of this strategy has been explained in **Supplementary Figure 3.1 from the previous chapter**.

For both the nucleotide concatenated datasets 1 and 2 above, any ambiguous sites in the concatenated final alignment was removed using Gblocks (Talavera and Castresana, 2007) with all default parameters, except for the stringency parameter of "Allowed Gap positions", where "With half" option was used so that all the positions with a gap in less than 50% of the aligned sequences would still be used in the final alignment. For **the nuclear dataset**, Tamura's 3-parameter model (Tamura, 1992) was used as the model of nucleotide substitution, along with a discrete gamma distribution for among-site rate variance (Yang, 1996), while still allowing for some sites to be evolutionarily invariable (T92 + G + I). For **the maxicircle dataset**, the general time reversible model (Tavaré, 1986) was used as the substitution model, with a discrete gamma distribution for among-site rate variance, while allowing for some proportion of evolutionarily invariable sites (GTR + G + I). Maximum likelihood trees were constructed for the nuclear and maxicircle datasets, with 1000 bootstrap replicates, using MEGA 6.0. Finally, for the binary **RNA editing quantification dataset**, the concatenated matrix of editing events was imported into GeneDoc, visually curated, and then imported into raxmlGUI (Silvestro and Michalak, 2012) (Stamatakis, 2006). ML analysis was carried out using the "ML + rapid bootstrap" option with 1000 bootstrap replicates, and "BINGAMMA" option for using among site rate heterogeneity in the binary dataset. Tree visualization and formatting was done on MEGA 6.0. All alignment-datasets used in this report are available upon request.

Please note that for *TcG*, *TcJR*, *TcM62*, *TcCAN*, *TcCL*, *TcTula*, *Tdio*, *Tern*, *Tcon*, *Tlew* and *Tser*, the nuclear and the maxicircle datasets were either obtained from the draft genomes in our Tree of Life project, or from The Genome Institute at Washington University School of Medicine (St. Louis, USA). For *TcSyl* and *TcCLB*, the maxicircle sequences were obtained from NCBI (GenBank), while their nuclear genes were obtained from TriTrypDB. For *TcEsm*, maxicircle

sequence was obtained from NCBI, while the nuclear gene sequences were obtained from the draft genomes in our Tree of Life project. For *TcMar* and *Tran*, the maxicircle sequences were obtained from NCBI (strains B7 and SC-58, respectively), while their nuclear genes for the nuclear dataset was obtained from the draft genomes in our Tree of Life project (TCC344 and AM80, respectively). For *Tbru*, both the maxicircle and nuclear genes were obtained from TriTrypDB. The binary editing quantification for all the genes in all the organisms were based on the maxicircle sequences.

Results

Maxicircle genome characteristics

The partial maxicircle genomes of six strains of *T. cruzi* (*TcCL*, *TcTula*, *TcM62*, *TcCAN*, *TcG* and *TcJR*), *Tcon*, *Tlew*, *Tser*, *Tdio*$^{+}$ and *Tern* were assembled, analyzed and compared with the previously published maxicircle genomes of *Tbru*, *TcCLB*, *TcSyl* and *TcEsm*. The published maxicircle sequences of *TcMar* and *Tran* were added to the existing group for better representation and analysis of RNA editing among all the *Trypanosoma* clades. Please see **Tables 4.1A and 4.1B** for statistics on the maxicircle genomes. The maxicircle sequences that we analyzed range between ~13 kb for *Tdio* to over 25 kb for *Tran*, not including the highly repetitive divergent region which broke the contigs of many of these maxicircle genomes. The maxicircle of *Tern* is incomplete with several breaks in the circular genome sequence due to a lack of material to complete the sequence. The maxicircle of *Tdio* is putatively circularized, unless the designed primer spanned a repetitive sequence in the divergent region. The GC content of the genomes ranged between 23-32%. The maxicircle of *Tser* showed the highest GC content of any trypanosomatid in our analysis (32%) and most of the *T. cruzi* strains showed a modal GC content of 25%.

Table 4.1A. Maxicircle genome statistics (Non *T. cruzi Trypanosoma* members)[+]

Organism	*Tbru*	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio^*	*Tern***
Genome size [est. bps]	23016	19974	16350	25288 @	17135	13349	10014 @
Number of contigs	na	1	1	na	1	2	4
GC content (%) +	23	32	26	24	26	26	23
Coding sequence [bps] # >	14329	14634	14647	14757 @	14804	11732	6697
Coding sequence (%) # ~ +	62	73	90	58 $	86	88	72 $
Avg intergenic distance [bps] ! +	29	36	36	35 $	36	30	44 $
5' edited genes*	3	3	3	3	3	3	2
Internally edited genes	1	1	1	1	1	1	1
Pan-edited genes	9	9	9	9	9	7	4
Non-edited genes	7	7	7	7	7	7	3
Total number of genes #	20	20	20	20	20	18	10

Table 4.1B. Maxicircle genome statistics (*T. cruzi* strains)[+]

Organism	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl* **	*TcJR*	*TcEsm*	*TcMar*
Genome size [est. bps]	20631	17575	21510	20217	17388	18469	15185	20172	22296	20050
Number of contigs	na	1	4	1	1	1	na	2	na	na
GC content (%) +	25	25	25	26	25	25	26	25	24	23
Coding sequence [bps] # >	14705	14706	14695	14683	14705	14692	14495	14660	14570	14663
Coding sequence (%) # ~ +	71	84	68	73	85	80	95	73	65	73
Avg intergenic distance [bps] ! +	32	32	32	32	32	33	35	35	27	33
5' edited genes*	3	3	3	3	3	3	3	3	3	3
Internally edited genes	1	1	1	1	1	1	1	1	1	1
Pan-edited genes	9	9	9	9	9	9	9	9	9	9
Non-edited genes	7	7	7	7	7	7	7	7	7	7
Total number of genes #	20	20	20	20	20	20	20	20	20	20

Abbreviations for the names of the organisms are as per in the text.
+ All values rounded to nearest integer.
^ 2 (of the 20) genes are absent in Tdio; and maxicircle is putatively fully circularized. See text for details.
 # includes rRNA sequences, but not guide RNA genes.
> sum of the lengths of pre-edited genes encoded by the maxicircle; excludes the intergenic regions.
~ (coding sequence / genome size)*100; includes the divergent region in the total genome size.
! excludes the divergent region; starts from 12S and ends at the last codon of ND5, including the intergenic regions.

*** Includes MURF5 gene that we believe putatively 5' edited to create a start codon.**
**** The genomes of TcSyl and Tern are incomplete. Refer to the text.**
@ For Tern, genome size estimate includes the 665 Ns used to join the contigs; and for Tran, it includes the 8 Ns used in the ND1 gene.
$ For Tern and Tran, these characteristics do not include the Ns used.

The 20 known maxicircle genes (including the two ribosomal RNA genes) were all found to be perfectly syntenic among the maxicircles of all the trypanosomatids in this study (**Fig 4.1**), with the exception of *Tdio* and *Tern*. The maxicircle of *Tdio* is apparently missing two of the nine pan-edited genes (G3 and G4), and exhibits deletions in the MURF1, ND1 and ND4 genes. This explains its smaller total maxicircle size. The maxicircle sequence of *Tern* is incomplete and hence only 10 of the 20 known genes could be identified on its fragmented genome. Other than an apparent deletion of its MURF2 gene, it is unknown at this point whether any other deletions exist in the maxicircle of *Tern*. Please refer to **Tables 4.2A, 4.2B and 4.2C** for the putative co-ordinates of all the genes on the maxcircles of all the organisms, the pre-edited sizes of each individual gene and their putative start and stop codons. The sum of the pre-edited coding sequences of all the maxicircle genes (excluding the intergenic regions) were approximately ~14.5 kb, thereby occupying ~58-95% of their circular genomes (**Tables 4.1A and 4.1B**).

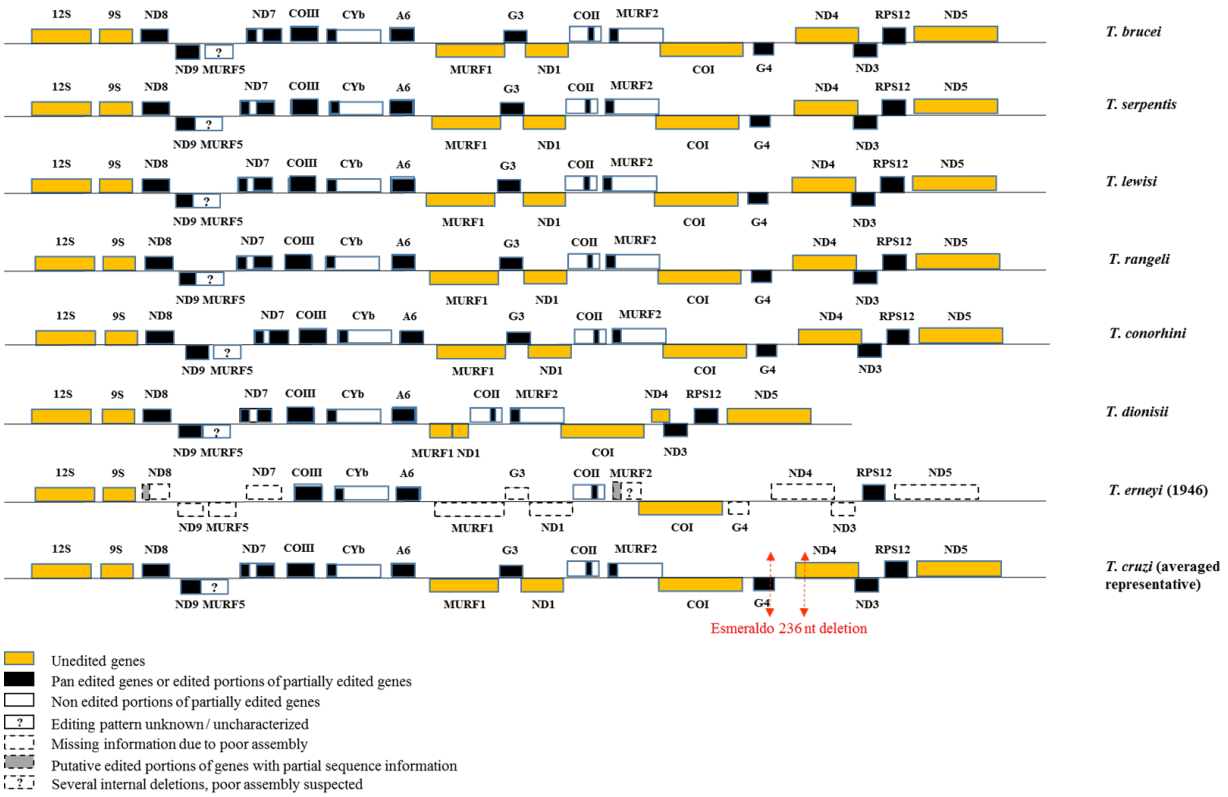


Fig 4.1. The syntenic 20-gene region of the trypanosomatid maxicircle genomes: The maxicircle genomes of all the organisms in this report are shown along with the inferred editing patterns of the genes. The figure is not drawn to scale. Genes above the line are transcribed in the same direction as the ribosomal RNA genes and the ones underneath are transcribed in the opposite direction.

As described in 'Materials and Methods', the start and stop codons and the editing patterns of each gene were established by aligning the homologous, pre-edited maxicircle sequences of all the organisms with the previously published, fully-edited transcripts and coding sequences of *Tbru*, *TcCLB*, *TcSyl*, and *TcEsm*. AUG is the most common start codon, although both UUG and UUA (encoding leucine) represented about 15% of the translation start signals. Less frequently, AUA, GUA, GUG, AAU, ACU, AGA and AAG are also found at the site of a putative start codon in some maxicircles. UAA and UAG are the universal stop codons, but UGA codon encodes Tryptophan, as reported previously (Osawa et al., 1992) (Jukes and Osawa, 1993). The

pre-edited sizes of all the genes were highly similar among these organisms, emphasizing the similarities among the editing patterns of the genes in the maxicircles of these trypanosomatids. The details of the editing of each gene is described in the following section.

Table 4.2A. Coding gene lengths with co-ordinates and start - stop codons for non-*T. cruzi* *Trypanosoma* members (Genes ordered as per their location on the maxicircle)

GENES		*Tbru*	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern* +
12S *	**CC**	1364-2512	1578-2738	672-1840	2963-4117	456-1636	328-1487	2210-3380
	CL	1149	1161	1169	1155	1181	1160	1171
9S *	**CC**	2549-3154	2781-3386	1890-2497	4179-4784	1714-2321	1541-2152	3418-4022
	CL	606	606	608	606	608	612	605
ND8	**CC**	3204-3469	3435-3711	2535-2818	4824-5111	2340-2633	2191-2467	4056-?
	CL	266	277	284	288	294	277	?
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUA-?
ND9 #	**CC**	3770-3506	4087-3767	3240-2881	5543-5187	3095-2702	2897-2523	?
	CL	265	321	360	357	394	375	?
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	?
MURF5 *#	**CC**	4012-3788	4363-4051	3431-3196	5803-5504	3350-3120	3136-2897	?
	CL	225	313	236	300	231	240	?
	Start-Stop	AUG-UAA	UUA-?	GAU-?	AGA-?	AAU-UAA	ACU-UAA	?
ND7	**CC**	4066-4767	4464-5199	3556-4327	5882-6692	3428-4233	3189-3871	?
	CL	702	736	772	811	806	683	?
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAA	AUG-UAG	AUG-UAG	?
COIII	**CC**	4813-5206	5278-5685	4407-4821	6758-7204	4291-4727	3932-4347	4212-4636
	CL	394	408	415	447	437	416	425
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
CYb	**CC**	5275-6354	5779-6858	4909-5988	7281-8360	4791-5870	4414-5493	4691-5770
	CL	1080	1080	1080	1080	1080	1080	1080
	Start-Stop	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AUG-UAA
ATPase6	**CC**	6397-6679	6898-7187	6026-6308	8404-8736	5913-6227	5532-5853	5805-6106
	CL	283	290	283	333	315	322	302
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF1 #	**CC**	8060-6735	8585-7260	7695-6371	10130-8805	7618-6293	6358-5922	?
	CL	1326	1326	1325	1326	1326	437 (5' Deletion)	?
	Start-Stop	UUG-UAA	UUG-UAA	UUG-?	UUG-UAA	UUG-UAA	?	?
G3	**CC**	8060-8167	8586-8711	7699-7828	10133-10255	7620-7746	Gene	?
	CL	108	126	130	123	127	Absent	?
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	-	?
ND1 #	**CC**	9107-8163	9652-8708	8783-7839	11134-10258	8685-7741	6675-6359	?
	CL	945	945	945	877	945	317 (3'	?

							Deletion)	
	Start-Stop	UUA-UAA	UUG-UAA	AUG-UAA	UUA-?	UUG-UAA	AUG-?	?
COII	**CC**	9111-9739	9653-10281	8782-9410	11142-11770	8694-9322	6687-7315	6498-7126
	CL	629	629	629	629	629	629	629
	Start-Stop	AUG-UAG	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF2	**CC**	9763-10818	10306-11361	9436-10489	11806-12862	9346-10397	7341-8381	7149-7798
	CL	1056	1056	1054	1057	1052	1041	650 (Deletions)
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-?	AUG-?
COI #	**CC**	12455-10809	12998-11352	12129-10480	14501-12853	12037-10388	10021-8372	9438-7789
	CL	1647	1647	1650	1649	1650	1650	1650
	Start-Stop	AUG-UAA	UUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	UUG-UAA
G4 *#	**CC**	12671-12488	13268-13067	12365-12177	14747-14561	12277-12092	Gene	?
	CL	184	202	189	187	186	Absent	?
	Start-Stop	AUG-UAG	GUG-UAG	AUG-UAA	AUG-UAA	GUG-UAA	-	?
ND4	**CC**	12780-14093	13392-14705	12489-13802	14872-16185	12401-13714	10077-10358	?
	CL	1314	1314	1314	1314	1314	282 (5' Deletion)	?
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	?	?
ND3 *#	**CC**	14285-14069	14922-14681	14023-13778	16430-16161	13964-13690	10582-10334	?
	CL	217	242	246	270	275	249	?
	Start-Stop	UUG-UAA	UUG-UAA	UUG-UAA	UUG-UAA	UUG-UAA	UUA-UAA	?
RPS12	**CC**	14310-14469	14947-15128	14047-14231	16453-16627	13992-14172	10605-10796	9636-9820
	CL	160	182	185	175	181	192	185
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG
ND5	**CC**	14490-16262	15153-16925	14260-16032	16659-18431	14207-15979	10826-12595	?
	CL	1773	1773	1773	1773	1773	1770	?
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAG	AAG-?	?

Table 4.2B. Coding gene lengths with co-ordinates and start - stop codons for *T. cruzi* DTUs VI, III and IV (Genes ordered as per their location on the maxicircle)

GENES		*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*
12S *	**CC**	4707-5867	656-1816	3234-4393	2978-4139	379-1543
	CL	1161	1161	1160	1162	1165
9S *	**CC**	5906-6514	1855-2462	4432-5038	4178-4786	1582-2191
	CL	609	608	607	609	610
ND8	**CC**	6559-6837	2507-2785	5083-5361	4831-5110	2236-2513
	CL	279	279	279	280	278
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG
ND9 #	**CC**	7238-6901	3186-2849	5762-5425	5511-5174	2917-2577
	CL	338	338	338	338	341
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF5 *#	**CC**	7496-7201	3444-3149	6020-5725	5769-5474	3175-2880
	CL	296	296	296	296	296
	Start-Stop	AUA-?	AUA-?	AUA-?	AUA-?	AUA-?
ND7	**CC**	7563-8317	3511-4265	6087-6841	5836-6592	3242-3999
	CL	755	755	755	757	758
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG
COIII	**CC**	8384-8806	4332-4755	6908-7330	6659-7081	4066-4485
	CL	423	424	423	423	420
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
CYb	**CC**	8881-9960	4830-5909	7405-8484	7156-8235	4560-5639
	CL	1080	1080	1080	1080	1080
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAA	AUG-UAA
ATPase6	**CC**	9998-10312	5947-6261	8522-8836	8273-8587	5677-5991
	CL	315	315	315	315	315
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF1 #	**CC**	11706-10381	7655-6330	10229-8905	9980-8656	7385-6060
	CL	1326	1326	1325	1325	1326
	Start-Stop	UUG-UAA	UUG-UAA	UUG-?	UUG-?	UUG-UAA
G3	**CC**	11708-11826	7657-7775	10231-10349	9982-10101	7387-7506
	CL	119	119	119	120	120
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
ND1 #	**CC**	12766-11822	8715-7771	11289-10345	11036-10097	8446-7502
	CL	945	945	945	940	945
	Start-Stop	UUG-UAA	UUG-UAA	UUG-UAA	UUG-?	UUG-UAA
COII	**CC**	12777-13405	8726-9354	11300-11928	11047-11675	8457-9085
	CL	629	629	629	629	629
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
MURF2	**CC**	13431-14486	9380-10435	11954-13003	11701-12750	9112-10163

	CL	1056	1056	1050	1050	1052
	Start-Stop	AUG-?	AUG-?	AUG-?	AUG-?	AUG-UAG
COI #	**CC**	16126-14477	12075-10426	14643-12994	14390-12741	11803-10154
	CL	1650	1650	1650	1650	1650
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
G4 *#	**CC**	16370-16180	12320-12129	14887-14697	14633-14444	12046-11857
	CL	191	192	191	190	190
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA
ND4	**CC**	16488-17801	12438-13751	15005-16318	14751-16053	12167-13480
	CL	1314	1314	1314	1303	1314
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-?	AUG-UAA
ND3 *#	**CC**	18041-17777	13991-13727	16558-16294	16293-16029	13721-13456
	CL	265	265	265	265	266
	Start-Stop	UUA-UAA	UUA-UAA	UUA-UAA	UUA-UAA	UUG-UAA
RPS12	**CC**	18063-18246	14013-14196	16580-16763	16315-16497	13742-13921
	CL	184	184	184	183	180
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG
ND5	**CC**	18274-20043	14224-15993	16791-18560	16525-18292	13949-15718
	CL	1770	1770	1770	1768	1770
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-?	AUG-UAG

Table 4.2C. Coding gene lengths with co-ordinates and start - stop codons for *T. cruzi* DTUs I, II and *TcMar* (Genes ordered as per their location on the maxicircle)

GENES		*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*	NCS
12S *	**CC**	741-1896	1-1159	922-2078	4751-5910	4601-5770	
	CL	1156	1159	1157	1160	1170	
9S *	**CC**	1940-2551	1203-1812	2122-2731	5947-6554	5811-6414	
	CL	612	610	610	608	604	
ND8	**CC**	2598-2874	1856-2133	2777-3053	6607-6877	6456-6733	
	CL	277	278	277	271	278	(1/17)
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	
ND9 #	**CC**	3287-2939	2547-2198	3464-3118	7295-6942	7138-6793	
	CL	349	350	347	354	346	!
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
MURF5 *#	**CC**	3539-3248	2801-2541	3718-3458	7539-7256	7382-7140	
	CL	292	261	261	284	243	!
	Start-Stop	AUA-?	AUA-UAG	AUA-UAG	AUA-?	UUG-UAA	
ND7	**CC**	3615-4363	2877-3623	3794-4541	7612-8367	7462-8219	
	CL	749	747	748	756	758	!
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	
COIII	**CC**	4437-4858	3697-4116	4614-5033	8435-8859	8283-8707	
	CL	422	420	420	425	425	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
CYb	**CC**	4937-6016	4189-5269	5112-6191	8917-9996	8781-9860	
	CL	1080	1081	1080	1080	1080	0
	Start-Stop	AUG-UAA	AUG-UAG	AUG-UAA	AUG-UAA	AUG-UAA	
ATPase6	**CC**	6053-6362	5308-5618	6236-6545	10038-10351	9902-10218	
	CL	310	311	310	314	317	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
MURF1 #	**CC**	7765-6440	7021-5696	7947-6622	11764-10431	11611-10286	
	CL	1326	1326	1326	1334	1326	!
	Start-Stop	UUG-UAA	UUG-UAA	UUG-UAA	UUG-?	UUG-UAA	
G3	**CC**	7767-7887	7023-7143	7949-8069	11766-11886	11613-11724	
	CL	121	121	121	121	112	!
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
ND1 #	**CC**	8826-7882	8082-7138	9008-8064	12826-11882	12663-11719	
	CL	945	945	945	945	945	!
	Start-Stop	UUG-UAA	UUG-UAA	UUG-UAA	UUG-UAA	UUG-UAA	
COII	**CC**	8837-9465	8093-8721	9019-9647	12838-13466	12671-13299	
	CL	629	629	629	629	629	0
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
MURF2	**CC**	9492-10544	8748-9800	9674-10726	13500-14544	13323-14370	

	CL	1053	1053	1053	1045	1048	0
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-?	AUG-?	
COI #	**CC**	12184-10535	11440-9791	12366-10717	16184-14535	16010-14361	
	CL	1650	1650	1650	1650	1650	(2/17)
	Start-Stop	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	AUG-UAA	
G4 *#	**CC**	12431-12239	11688-11495	12616-12421	16408-16240	16256-16063	
	CL	193	194	196	169 (5' Deletion)	194	!
	Start-Stop	AUA-UAA	AUA-UAA	AUA-UAA	?-UAA	AUG-UAA	
ND4	**CC**	12545-13858	11802-13116	12730-14043	16409-17622	16370-17683	
	CL	1314	1315	1314	1214 (5' Deletion)	1314	!
	Start-Stop	AUG-UAA	AUG-?	AUG-UAA	?	AUG-UAA	
ND3 *#	**CC**	14095-13834	13355-13092	14281-14019	17857-17598	17927-17659	
	CL	262	264	263	260	269	!
	Start-Stop	UUA-UAA	UUA-UAA	UUA-UAA	UUA-UAA	UUA-UAA	
RPS12	**CC**	14116-14297	13377-13557	14302-14484	17879-18058	17949-18131	
	CL	182	181	183	180	183	0
	Start-Stop	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	AUG-UAG	
ND5	**CC**	14326-16095	13586-15185	14513-16282	18085-19855	18156-19927	
	CL	1770	1600^ (Incomplete)	1770	1771	1772	!
	Start-Stop	AUG-UAG	AUG-?	AUG-UAG	AUG-?	AUG-?	

CC = Co-ding Co-ordinates for the pre-edited gene in maxicircle; from start to stop codon.
CL = Unedited (or pre-edited) Coding Length of the gene; from start to stop codon.
*** Genes with approximate coding co-ordinates.**
 # Genes transcribed on the strand opposite to the ribosomal RNA genes.
? Codon or co-ordinates uncharacterized or uncertain due to lack of sequence information or frameshift.
^ = truncated gene due to incomplete genome.
NCS = Frequency that the gene starts with a non-canonical start codon (not AUG).
! = NCS cannot be calculated due to missing codon information for some organisms.
+ = The co-ordinates of Tern includes the Ns used to join the contigs, i.e., the 665 Ns are counted.

We aligned the syntenic 20-gene regions (starting at 12S and ending at ND5, with the intergenic regions) of all the trypanosomatid maxicircles in this report and compared their nucleotide sequence identities. The overall identities generally followed the expected patterns, wherein, the maxicircles of the organisms whose nuclear genes are more closely related, show higher identities to each other (**Table 4.3**). The *T. cruzi* strains show the highest syntenic region identities to each other. Particularly, members of the DTUs TcVI and TcIII show a remarkable

100% identity to each other, closely followed by TcIV (*TcCAN*) that shows a 99% identity to TcVI and TcIII. Similarly, members of TcI show identities between 98-100% to each other. TcII (*TcEsm*, 75-91%) and *TcMar* (75-89%) are the *T. cruzi* strains, whose maxcicircles are the least identical to the maxicircles of all other *T. cruzi* members. Among the non-*T. cruzi* trypanosomes, the maxicircle sequence of *Tser* is the least identical to all other maxicircles, with a modal identity of 75%; followed closely by the maxicircle of *Tbru*, which shows a modal identity of 78% to the maxicircles of all the other trypanosomatids. The maxicircles of *Tern* (incomplete maxicircle) and *Tdio* are the most identical to the maxicircles of the *T. cruzi* strains, with modal identities of 86% and 84% respectively.

Table 4.3. Percent identity matrix of the syntenic 20-gene region of maxicircle genomes

	Tbru	*Tser*	*Tlew*	*Tran**	*Tcon*	*Tdio**	*Tern**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl**	*TcJR*	*TcEsm**	*TcMar*
Tbru	100																
Tser	75	100															
Tlew	78	75	100														
*Tran**	77	75	80	100													
Tcon	78	76	81	83	100												
*Tdio**	79	76	81	82	82	100											
*Tern**	81	79	84	84	84	86	100										
TcCLB	78	76	81	81	82	84	86	100									
TcCL	78	75	81	81	82	84	86	100	100								
TcTula	78	75	81	81	82	84	86	100	100	100							
TcM62	79	75	81	81	82	84	86	100	100	100	100						
TcCAN	78	76	81	81	82	84	86	99	99	99	99	100					
TcG	78	75	81	82	82	84	86	93	93	93	93	93	100				
*TcSyl**	78	75	81	81	82	84	86	93	93	93	93	93	98	100			
TcJR	78	76	81	82	82	84	86	93	93	93	93	93	99	99	100		
*TcEsm**	78	75	81	82	82	84	86	91	91	91	91	91	91	91	91	100	
TcMar	79	75	81	82	82	85	87	89	89	89	89	89	88	88	88	89	100

The matrix above shows the nucleotide percent identities of the syntenic 20-gene region of maxicircles (starting at 12S and ending at ND5, including the intergenic regions) between all the organisms studied.
*** indicates that the identities of these maxicircles were obtained after elimination of the absent genes or the partially deleted (and missing) regions in all the other maxicircles.**

To get a perspective on the identities of the nuclear genes, we computed the nucleotide sequence identities of three nuclear protein-encoding genes: GAPDH, DHFR and GSH1, among all the trypanosomatids in this report. The GAPDH gene of *Tbru* is the least identical to the others (64-69%) while the GAPDH of the members of TcI are the most identical to each other (between 99-100%). Similarly, the DHFR gene of *Tbru* is between 66-68% identical to all the other DHFR sequences, while members of TcI are all 100% identical. This observation extends to the GSH1 gene as well, wherein the GSH1 of *Tbru* is the least identical to that of the others (64-69%) and the GSH1 of TcI are between 99-100% identical. For all the three nuclear genes analyzed, the genes of *Tdio* (among the non-*T. cruzi* trypanosomatids) are the most identical to the *T. cruzi* strains. Please refer to the **Supplementary Tables 4.2, 4.3 and 4.4** for the three genes' identity matrices. Although the maxicircle identity ranges of all these trypanosomatids were comparable to the identities noticed for their nuclear counterparts, it must be noted that the nuclear genes used for comparison are discrete and do not involve any intergenic regions. The identities of each of the individual maxicircle genes is described in the following section.

The RNA editing patterns are remarkably similar among all the members of the *Trypanosoma* genus in this report

Below, we summarize all the maxicircle genes and their RNA editing in three groups: 1) genes that are not edited; 2) genes that are partially (5' or internally) edited; and 3) genes that are pan-edited.

Genes that are not known to undergo editing:

The ribosomal genes (12S and 9S), MURF1, ND1, COI, ND4 and ND5 were previously reported to be unedited in *Tbru*. See below for description of each of these genes in all of the organisms.

12S and 9S rRNA genes: The 12S gene ranges between 1149-1181 nts (**4.SI-S1a**) and the 9S ranges between 604-612 nts in length (**4.SI-S2a**). Their co-ordinates were derived based on homology comparison of these genes with the previously reported co-ordinates of *TcCLB* genes (Ruvalcaba-Trejo and Sturm, 2011). While the 12S gene is between 79-100% identical between various species (**4.SI-S1b**), the 9S gene is between 81-100% identical (**4.SI-S2b)**. The 12S and 9S sequences of all the organisms fold into the canonical rRNA secondary structures as previously described for *L. tar* (de la Cruz et al., 1985b) (de la Cruz et al., 1985a) (**Supplementary Figures 4.1A through 4.1R and 4.2A through 4.2R** for 12S and 9S, respectively).

MURF1 (ND2): Very limited information is available for the MURF1 gene, which is transcribed on the strand opposite to the ribosomal genes. Earlier, this gene was thought to be 5' edited in the *T. cruzi* group (Westenberger et al., 2006). Subsequently it was accepted to be unedited and to encode ND2 (Kannan and Burger, 2008). The gene is currently believed to be unedited and to

use a non-canonical 'UUG' start codon that encodes Leucine. Alignment of the homologous sequences of the trypanosomatids in this report to the published MURF1 sequences of several *T. cruzi* strains (Ruvalcaba-Trejo and Sturm, 2011) resulted in the alignment as shown in **4.SI-S3a**. All the sequences appear to be unedited with a final coding sequence length of approximately 1326 nt. However, some sequences are shorter or longer due to the presence of indels. For instance, the MURF1 sequences of *TcM62, TcTula* and *Tlew* show an indel (lack of a T residue), each, at positions 29, 1184 and 1255 respectively of their coding sequence, that changed the reading frame of the translated protein from that point. Similarly, the 10 indels of *TcEsm* were discussed previously (Ruvalcaba-Trejo and Sturm, 2011). The indels cause the translated proteins to be truncated, with several stop codons in the altered reading-frame, as shown in **4.SI-S3b**. Whether the indels in each of these organisms are corrected by RNA editing to remove the frameshift, is unknown. *Tdio* shows an extensive 5' deletion of its MURF1 gene, resulting in the presence of only the last 437 nt at the 3' end of the gene, which also shows several indels. The gene could not be identified in *Tern* due to the lack of sequence information for this gene. UUG is the universal start codon and UAA is the universal stop codon, for all the proteins in-frame. Though the coding sequences of the pre-edited transcripts are between 72-100% identical (**4.SI-S3c**), the proteins show a much lower identity range of 13-100%, due to the altered reading frame caused by the indels (**4.SI-S3d**).

ND1: The ND1 gene is also transcribed in the direction opposite to that of the ribosomal genes. Early reports comparing the maxcircle sequences of *Tbru* and *Ltar* identified this gene as an unedited open reading frame (Simpson et al., 1987). More recently, this gene has been established to be unedited and its sequence has been reported for *TcCLB*, *TcEsm* and *TcSyl* (Ruvalcaba-Trejo and Sturm, 2011). Comparison of the homologous sequences of all the

trypanosomatids in our report to these previously reported sequences show these ND1 transcripts to be unedited with a final coding sequence length of 945 nts (**4.SI-S4a**). However, the coding sequence of *TcM62* shows a deletion of five nts (after the 115th nt) and the coding sequence of *Tran* shows an indel (a missing T residue) at position 487 of the transcript, both of which change the translation frame. The coding sequence of *Tran* is missing 78 nts due to a lack of sequence information and we have assumed that this does not change the reading frame (indicated by Ns in the alignment). The coding sequence of *Tdio* shows only the presence of the first 317 nt on its 5' end with the rest of its gene being deleted. Additionally, this truncated ND1 of *Tdio* also shows the presence of a pre-mature stop codon after its 39th AA residue. ND1 was not characterized in *Tern* due to lack of sequence information for this gene. The protein alignment is shown in **4.SI-S4b**. Although the pre-edited coding sequences are between 67-100% identical (**4.SI-S4c**), the change of frame caused by the indels causes the resultant protein identities to drop and they are only between 11-100% identical (**4.SI-S4d**). UUG is the most common start codon (alongside UUA and AUG) and UAA is the universal stop codon for all the proteins that are in-frame.

COI: This gene, which is transcribed from the strand opposite to the ribosomal genes, has never been reported to be edited and is well conserved among all the trypanosomatids studied to date. We compared homologous regions of the maxicircles of all the trypanosomatids in this report to the previously published COI sequences of *TcCLB*, *TcSyl* and *TcEsm* (Ruvalcaba-Trejo and Sturm, 2011). Our alignment indicates that the gene is unedited among all the organisms and the coding sequence measures 1650 nt in length (**4.SI-S5a**). Coding sequences of *Tbru* and *Tser* are 1647 nt in length and lack an AA each, without a change in the frame; while the cpoding sequence of *Tran* is 1649 nt and lacks a T residue at position 1427, which is present in all others. Preliminary analyses indicate that this T residue is also present in the COI of *T. rangeli* AM80,

when we mapped the reads back to the COI gene sequence of the published *Tran* (SC-58). Thus, this indel may be due an error in assembly and has therefore been ignored in our analysis. The alignment of the translated proteins is shown in **4.SI-S5b**. AUG is the predominant start codon (though *Tser* and *Tern* use UUG) and UAA is the universal stop codon. The coding sequences are between 77-100% identical (**4.SI-S5c**), whereas the proteins are between 82-100% identical (**4.SI-S5d**). The COI proteins of the *T. cruzi* group are 98-100% identical. Thus, COI is the only gene in this report that is not edited, has no indels disrupting its reading frame, and shows a highly conserved protein.

ND4: This gene is also transcribed from the same strand as the ribosomal genes and was not reported to be edited in *Tbru* (Simpson et al., 1987). The 5' deletion of the 99 nt in the 5' end of ND4 in *TcEsm* and the indel in *TcSyl* coding sequences have been described previously (Ruvalcaba-Trejo and Sturm, 2011) (Westenberger et al., 2006). Although *TcEsm* was reported to have the rest of its ND4 protein conserved with the other trypanosomatids, this conserved protein can only be produced in an alternate reading frame. Our alignments indicate that there are indels in *TcEsm*, *TcSyl* and *TcM62*, which alter their translation reading frames to produce truncated proteins (**4.SI-S6a** and **4.SI-S6b**). Additionally, there is an extensive 5' deletion of ND4 in *Tdio*, leading to the presence of only 282 nt at the 3' end of the coding sequence, which also shows several indels, leading to frameshift and protein truncation. However, the ND4 sequences in all the other organisms that do not show any indels or frameshifts appear to be unedited and show a final coding length of 1314 nt. The gene was not identified in *Tern* due to lack of sequence information. AUG is the universal start codon (except in *TcEsm* and *Tdio* that lack a start codon) and UAA is the universal stop codon for all the proteins in frame. Although the coding sequences show identities in the range of 72-100% (**4.SI-S6c**), the truncated and out-

of-frame proteins show extremely poor identities starting at just about 5% and ranging up to 100% (**4.SI-S6d**).

ND5: This gene is also transcribed from the same strand as the ribosomal genes and was reportedly unedited in *Tbru* and *TcCLBren* (Westenberger et al., 2006). The same report discussed the indels in *TcEsm*. The *TcSyl* maxicircle was reportedly incomplete (Ruvalcaba-Trejo and Sturm, 2011) and therefore lacks the whole length of this gene. We used these sequences as references to align the homologous sequences of the organisms from this report. *TcM62* has 2 indels (lack of a T residue, each) at coding positions 988 and 1743. *TcMar* has an indel at positon 1305 and *Tdio* has several indels through the length of the coding sequence (see **4.SI-S7a**). The gene could not be characterized in *Tern* due to lack of sequence information. All these indels change the translation frame and introduce several stop codons through the length of the translated proteins (**4.SI-S7b**). AUG is the universal start codon, although *Tdio* apparently uses AAG (encoding Lysine), based on the homology of sequences. Of all the coding sequences that do not show any indels, members of the *T. cruzi* family and *Tcon* use UAG as their stop codon while all others (non-*T. cruzi* members) use UAA as their stop codon. The coding sequences show identities ranging between 70-100% (**4.SI-S7c**), whereas the proteins show a much lower identity range of 30-100% due to the several indels and frameshifts (**4.SI-S7d**).

Genes that undergo partial (5' or internal) editing:

Only 3 genes: CYb, COII and MURF2 were reported to undergo partial editing of their transcripts in *Tbru*. We have included MURF5 gene in this group as we believe that it must undergo some editing. See below:

CYb: This gene is also transcribed from the same strand as the ribosomal genes and was reported to be edited on its 5' end in *Tbru* (Feagin et al., 1988b). The fully edited, partial CYb RNA sequence of *Tbru* and cDNA sequence of *Leishmania tarentolae* (GenBank: L07542.1) (Cruz et al., 1984) (Shaw et al., 1988) were used as references to predict the RNA editing in all the other trypanosomatids in this report. The transcript is 5' edited in all the organisms with the final length of the coding sequence being 1113 nt (*Ltar* has 1116 nt due to an additional Phenylalanine residue it encodes right after the start codon (**4.SI-S8a**). The additional T residue encoded in the usually unedited portion of the CYb gene of *TcSyl* discussed earlier (Ruvalcaba-Trejo and Sturm, 2011) has been manually deleted to maintain the frame and is shown as '*' in the nucleotide alignment. The translated proteins are remarkably well-conserved and their alignment is shown in **4.SI-S8b**. AUG is the universal start codon and stop codon is either UAA or UAG. The fully-edited coding sequences are between 80-100% (**4.SI-S8c**) identical and the proteins are 91-100% identical (**4.SI-S8d**). See **Tables 4.4A, 4.4B and 4.4C** for the estimated number of Us inserted and deleted and the predicted editing patterns for all the edited genes.

COII: This gene is also transcribed from the same strand as the ribosomal genes and was found to be internally edited in *Ltar* (GenBank: L07544.1) (Cruz et al., 1984) (Shaw et al., 1989) and *T. cruzi* (GenBank: L22643.1) (Kim et al., 1994). Comparing the homologous sequences of the other organisms with these published references indicate that the COII transcripts of all the

organisms are internally edited in an identical fashion by the insertion of four uridine residues at the exact same locus in all of them (**4.SI-S9a**), thereby generating a fully-edited, coding length of 633 nt. The *in silico* translated protein alignment is shown in **4.SI-S9b**. The edited coding sequences are between 78-100% identical (**4.SI-S9c**) and the proteins are between 84-100% identical (**4.SI-S9d**). AUG is the universal start codon and UAA is the predominant stop codon (Only *Tbru* and *Tlew* use UAG as their stop codon).

MURF2: This gene is also transcribed from the same strand as the ribosomal genes and was reported to be 5' edited in both *L. tar* (GenBank: L07545.1) and *Tbru* (GenBank: M20253.1) (Shaw et al., 1988), and these fully-edited sequences were used as the editing reference. Our alignments indicate that all the MURF2 genes in all the organisms in this report are 5' edited, with the fully-edited, in-frame coding sequence measuring 1074 nt in length. However, most of the organisms' MURF2 gene shows several indels. *TcCL* shows the same 4 nt indel described for *TcCLB* (Ruvalcaba-Trejo and Sturm, 2011). The same report also highlighted the 13 indels in the MURF2 coding sequence of *TcEsm*. *TcTula* and *TcM62* show 2 single-nucleotide indels each, at coding positions 753 and 1049, and 442 and 937, respectively. *TcMar* also shows a 4 nt deletion after position 317 and two single-nucleotide indels at coding positions 923 and 933. *Tdio* encodes an additional T residue in one of the unedited portions of the transcript (positon 103). It also shows a 3 nt deletion and a 15 nt deletion further downstream in the unedited portion of the gene. *Tern* shows the start codon and a few subsequent codons but showed deletions of the coding sequence and also several indels. It is unclear whether poor assembly and lack of sequence information in *Tern* is the reason for this observation. Please refer to **4.SI-S10a** for the nucleotide alignment of the gene. All these indels alter the reading frame of the translated proteins and the alignment of the resultant, truncated proteins is shown in **4.SI-S10b**. AUG is the

universal start codon. Among the transcripts that maintained their frame, strains of *T. cruzi* use UAG, whereas the other *Trypanosoma* members (non *T. cruzi*) use UAA as their stop codons. Although the coding sequences are between 73-100% identical (**4.SI-S10c**), the proteins show much lower identities due to the frameshifts and are only between 15-100% identical (**4.SI-S10d**).

MURF5: Transcribed from the strand opposite to the one encoding the ribosomal genes, the MURF5 gene is one of those genes, whose expression, editing and even coding region remains unclear. It is currently believed to be non-edited and several reports have determined a conserved motif of its protein (Maslov, 2010) (Ruvalcaba-Trejo and Sturm, 2011). We identified this gene based on homology comparisons with the above-mentioned motif and the putative coding sequences vary in lengths between 225-313 nt for the organisms in this report (excluding *Tern*, where the gene could not be identified due to lack of sequence information). Although this conserved motif was found in all the organisms, a functional start codon (as for *Tbru*) could not be identified in these sequences (the co-ordinates of the gene are, therefore, an estimate). An in-frame stop codon was found for a few sequences; however, most of the others continued to extend into the upstream gene of ND9 (**4.SI-S11a**), likely due to several frame-shifts and indels throughout the gene, that were reported and discussed elsewhere (Ruvalcaba-Trejo and Sturm, 2011). This casts a shadow of doubt as to whether the gene is really unedited. We believe that the gene needs to be edited in its 5' end to generate a functional start codon. As for a stop codon, polyadenylation may generate an in-frame stop codon as postulated previously (Maslov, 2010), especially for the transcripts that seem to extend into the upstream ND9 gene. The translated protein alignment is shown in **4.SI-S11b**. The inferred coding sequences are between 56-100% identical (**4.SI-S11c**) and the proteins are between 39-100% identical (**4.SI-S11d**).

Table 4.4A. Inferred RNA editing patterns in genes of non-*T. cruzi Trypanosoma* members

GENES		*Tbru*	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*
ND8 (G1)	**U+[U-]**	213[41]	215[54]	214[60]	204[54]	202[58]	208[47]	?
	FCTL	438	438	438	438	438	438	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND9 (G2)	**U+[U-]**	338[18]	321[57]	308[83]	307[79]	297[106]	302[92]	?
	FCTL	585	585	585	585	585	585	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	?
ND7 (MURF3)	**U+[U-]**	540[81]	517[92]	509[120]	501[151]	498[143]	500[91]	?
	FCTL	1161	1161	1161	1161	1161	1092	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	?
COIII	**U+[U-]**	513[40]	497[38]	498[46]	499[79]	501[71]	505[54]	501[59]
	FCTL	867	867	867	867	867	867	867
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
CYb	**U+[U-]**	33[0]	33[0]	33[0]	33[0]	33[0]	33[0]	33[0]
	FCTL	1113	1113	1113	1113	1113	1113	1113
	PEP	5' edited	5' edited	5' edited	5' edited	5' edited	5' edited	5' edited
ATPase6 (MURF4)	**U+[U-]**	384[22]	378[23]	384[22]	366[54]	375[45]	378[55]	378[35]
	FCTL	645	645	645	645	645	645	645
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
G3	**U+[U-]**	130[13]	120[21]	122[27]	124[22]	120[22]	Gene	?
	FCTL	225	225	225	225	225	Absent	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	-	?
COII	**U+[U-]**	4[0]	4[0]	4[0]	4[0]	4[0]	4[0]	4[0]
	FCTL	633	633	633	633	633	633	633
	PEP	internal	internal	internal	internal	internal	internal	internal
MURF2	**U+[U-]**	22[4]	24[6]	23[3]	24[7]	24[2]	22[6]	25[6]
	FCTL	1074	1074	1074	1074	1074	1057	669
	PEP	5'	5'	5'	5'	5'	5'	5'
G4	**U+[U-]**	282[25]	277[38]	284[32]	283[29]	281[26]	Gene	?
	FCTL	441	441	441	441	441	Absent	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	-	?
ND3	**U+[U-]**	201[13]	184[33]	202[34]	198[63]	195[65]	193[55]	?
	FCTL	405	393	414	405	405	387	?
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	?
RPS12 (G6)	**U+[U-]**	120[28]	110[40]	111[44]	109[32]	110[39]	110[50]	110[43]
	FCTL	252	252	252	252	252	252	252
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
TOTAL	**U+**	**2780**	**2465**	**2478**	**2448**	**2438**	**2047**	**1051+**
	U-	**285**	**348**	**411**	**516**	**519**	**403**	**143+**

Table 4.4B. Inferred RNA editing patterns in genes of *T. cruzi* DTUs VI, III and IV.

GENES		*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*
ND8 (G1)	**U+[U-]**	203[44]	203[44]	203[44]	202[44]	202[42]
	FCTL	438	438	438	438	438
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND9 (G2)	**U+[U-]**	306[59]	306[59]	305[58]	305[58]	305[61]
	FCTL	585	585	585	585	585
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND7 (MURF3)	**U+[U-]**	512[106]	512[106]	512[106]	512[108]	511[108]
	FCTL	1161	1161	1161	1161	1161
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COIII	**U+[U-]**	503[59]	503[60]	503[59]	503[59]	503[56]
	FCTL	867	867	867	867	867
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
CYb	**U+[U-]**	33[0]	33[0]	33[0]	33[0]	33[0]
	FCTL	1113	1113	1113	1113	1113
	PEP	5' edited	5' edited	5' edited	5' edited	5' edited
ATPase6 (MURF4)	**U+[U-]**	374[44]	374[44]	374[44]	374[44]	374[44]
	FCTL	645	645	645	645	645
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
G3	**U+[U-]**	123[17]	123[17]	123[17]	123[18]	123[18]
	FCTL	225	225	225	225	225
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COII	**U+[U-]**	4[0]	4[0]	4[0]	4[0]	4[0]
	FCTL	633	633	633	633	633
	PEP	internal	internal	internal	internal	internal
MURF2	**U+[U-]**	24[2]	24[2]	24[2]	24[2]	24[2]
	FCTL	1078	1078	1072	1072	1074
	PEP	5'	5'	5'	5'	5'
G4	**U+[U-]**	272[22]	272[23]	272[22]	273[22]	272[21]
	FCTL	441	441	441	441	441
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND3	**U+[U-]**	198[58]	198[58]	198[58]	198[58]	197[58]
	FCTL	405	405	405	405	405
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
RPS12 (G6)	**U+[U-]**	110[42]	110[42]	110[42]	110[41]	111[39]
	FCTL	252	252	252	252	252
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
TOTAL	**U+**	**2662**	**2662**	**2661**	**2661**	**2659**
	U-	**453**	**455**	**452**	**454**	**449**

Table 4.4C. Inferred RNA editing patterns in genes of *T. cruzi* DTUs I, II and *TcMar*.

GENES		*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
ND8 (G1)	**U+[U-]**	201[40]	201[41]	201[40]	206[39]	202[42]
	FCTL	438	438	438	438	438
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND9 (G2)	**U+[U-]**	303[67]	304[69]	304[66]	303[72]	310[71]
	FCTL	585	585	585	585	585
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND7 (MURF3)	**U+[U-]**	509[97]	509[95]	509[96]	514[109]	513[110]
	FCTL	1161	1161	1161	1161	1161
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COIII	**U+[U-]**	505[60]	505[58]	505[58]	503[61]	501[59]
	FCTL	867	867	867	867	867
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
CYb	**U+[U-]**	33[0]	33[1]	33[0]	33[0]	33[0]
	FCTL	1113	1113	1113	1113	1113
	PEP	5' edited	5' edited	5' edited	5' edited	5' edited
ATPase6 (MURF4)	**U+[U-]**	375[40]	376[42]	375[40]	374[43]	374[46]
	FCTL	645	645	645	645	645
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
G3	**U+[U-]**	121[17]	121[17]	121[17]	121[17]	127[14]
	FCTL	225	225	225	225	225
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
COII	**U+[U-]**	4[0]	4[0]	4[0]	4[0]	4[0]
	FCTL	633	633	633	633	633
	PEP	internal	internal	internal	internal	internal
MURF2	**U+[U-]**	23[2]	23[2]	23[2]	24[4]	24[4]
	FCTL	1074	1074	1074	1065	1068
	PEP	5'	5'	5'	5'	5'
G4	**U+[U-]**	272[24]	272[25]	272[27]	244[19]	268[21]
	FCTL	441	441	441	394	441
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
ND3	**U+[U-]**	198[55]	198[57]	198[56]	198[53]	193[57]
	FCTL	405	405	405	405	405
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
RPS12 (G6)	**U+[U-]**	111[41]	111[40]	111[42]	110[38]	110[41]
	FCTL	252	252	252	252	252
	PEP	pan-edited	pan-edited	pan-edited	pan-edited	pan-edited
TOTAL	**U+**	**2655**	**2657**	**2656**	**2634**	**2659**
	U-	**443**	**447**	**444**	**455**	**465**

U+ = Number of Us inserted in the coding region (i.e., start to stop codon).
U- = Number of Us deleted in the coding region (i.e., start to stop codon).
FCTL = Final Coding Transcript Length (start codon to stop codon).
PEP = Predicted Editing Pattern.
? = No information due to poor sequence.
+ = More total U+ and U- expected in Tern since the genome is incomplete.

Genes that undergo extensive (pan) editing:

Nine of the twenty maxicircle genes of *Tbru* reportedly undergo pan (extensive) editing: ND8, ND9, ND7, COIII, A6, G3, G4, ND3 and RPS12. Please refer below for the description of editing for each of these genes in all the organisms herein.

ND8: The ND8 gene is transcribed from the same strand as the ribosomal genes. It was found to be pan-edited in *Tbru* (Souza et al., 1992) and its completely edited transcript (GenBank: M63820.1) was used as a reference to determine the editing patterns in the other trypanosomatids in this study. The gene is pan-edited in all the organisms and the fully edited coding sequences measure 438 nt in length. A previous report had predicted the ND8 editing pattern in *TcCLB*, *TcSyl* and *TcEsm* based on the editing of *Tbru* (Ruvalcaba-Trejo and Sturm, 2011). However, following the editing patterns of *Tbru* precisely leads to a pre-mature stop codon being incorporated after the 73rd AA residue in the members of TcIII, TcIV and TcVI, thereby generating a truncated protein (4.SI-S12a and S12b). This is due to a G→A transition event in the genes of these organisms that converts the 'UGA' codon encoding Tryptophan into the stop codon 'UAA' (**4.SI-S12a**). Curiously, despite the premature stop codon at position 74 of the protein, the rest of the protein is remarkably conserved through its entire length (**4.SI-S12b**). This suggests the presence of a selective pressure to maintain the functionality of the ND8 protein, much similar to the observations for the CYb and ND4 genes in *TcSyl* (Ruvalcaba-Trejo and Sturm, 2011). Shifting the position of the U insertion event by just one nucleotide eliminates this premature stop codon (by converting the 'UAA' stop codon into 'AUA' encoding an Isoleucine residue) and maintains the frame of the translated proteins (**4.SI-S12c** and **4.SI-S12d**). Though this shifted U insertion has not been tested, it is not impossible that such minor alteration in editing could occur to circumvent the problem. AUG and UAG are the universal start and stop

codons, respectively, except in *Tern*, where AUA is the start codon and the stop codon could not be identified due to missing sequence information. The fully edited coding sequences are between 94-100% identical (**4.SI-S12e**) and the proteins are between 97-100% identical (**4.SI-S12f**), not considering the truncated *Tern* sequence in both the cases.

ND9: Similar to ND8, the ND9 gene was also first studied in *Tbru* and was found to be pan edited (Souza et al., 1993). The gene is transcribed from the strand opposite to that of the ribosomal genes. The fully edited, mature ND9 transcript (GenBank: L05586.1) was used as the editing reference to predict the putative editing patterns of ND9 in all the other organisms. The gene is pan-edited in all the organisms and the fully edited coding sequence measures 585 nt in length (**4.SI-S13a**). The gene could not be identified in the maxicircle of *Tern* due to lack of sequence information. Alignment of the *in silico* translated proteins is shown in **4.SI-S13b**. AUG is the universal start codon and the stop codon is UAA for all the organisms, except *Tlew*, that uses UAG. The fully-edited coding sequences are between 91-100% identical (**4.SI-S13c**) and the proteins are 87-100% identical (**4.SI-S13d**) between all the organisms (excluding *Tern*).

ND7: This gene is transcribed from the same strand as the ribosomal genes and was first reported to be pan-edited in two distinct domains, with a 59 nt unedited stretch towards the beginning of the gene, in *Tbru*. This completely edited RNA sequence of *Tbru* (Koslowsky et al., 1990) was used as a reference to predict the editing of the ND7 transcripts of all the other trypanosomatids in this report. It is unclear whether this gene is present in the maxicircle of *Tern*, owing to poor sequence information. In all the others, the gene is pan edited in an almost identical fashion (with the 59 nt unedited region) and the fully edited coding sequences measure 1161 nt, except for *Tdio* that is only 1092 nt, due four small deleted stretches within the gene (See **4.SI-S14a**). Curiously, these deletions do not alter the reading frame of the *Tdio* transcript. However, it is

unclear whether these deletions remain neutral or are detrimental to the protein's functionality. Regardless, the protein is remarkably well-conserved among all the trypanosomatids (**4.SI-S14b**). The fully edited coding sequences are between 92-100% identical (**4.SI-S14c**) and the proteins are 93-100% identical (**4.SI-S14d**). AUG is the unambiguous start codon and UAG is the predominant stop codon, except *Tran*, that uses UAA as its stop codon.

COIII: This gene is transcribed from the same strand as the ribosomal genes and was found to be pan-edited in both *Tbru* (Feagin et al., 1988a) and *T. cruzi* (GenBank: EF058194.1). It is interesting to note that the fully-edited transcript sequence of *T. cruzi* varies considerably from that of *Tbru* and the former was used as the editing-reference for all the strains of *T. cruzi.* Editing was inferred for all the other trypanosmatids by locally matching and aligning their nucleotide sequences with either of the two references; albeit maintaining the same number of editing events. The fully edited coding sequences are 867 nt in length and pan-edited (**4.SI-S15a**), thereby creating a protein that is 288 amino acids in length (**4.SI- S15b**). AUG and UAA are the universal start and stop codons respectively. The fully edited coding sequences are between 88-100% identical (**4.SI- S15c**) while the proteins are between 86-100% identical (**4.SI-S15d**). What deserves a special mention here is that the proteins of most stains of *T. cruzi* are all 100% identical.

ATPase6 (MURF4): This gene is also transcribed from the same strand as the ribosomal genes and was identified to be pan-edited in *Tbru* (Bhat et al., 1990). The RNA sequence of the fully-edited transcript was, therefore, used as the reference to identify the editing patterns in the A6 transcripts of all the other organisms herein. All the transcripts are pan-edited with a final, fully-edited coding sequence length of 645 nt (**4.SI-S16a**). AUG and UAA are the universal start and

stop codons, respectively. The mature coding sequences are between 92-100% identical (**4.SI-S16c**) and the translated proteins (**4.SI-S16b**) are between 91-100% identical (**4.SI-S16d**).

G3: This small, pan-edited gene is transcribed from the same strand as the ribosomal RNA genes and the published, fully-edited transcript sequence of G3 (CR3) of *Tbru* (GenBank: U12395.1) was used as a reference to determine the editing patterns of this transcript in all the other trypanosomatids in this report. G3 is pan-edited in all of the organisms, with a final, mature coding sequence length of 225 nt (**4.SI-S17a**). This gene is completely absent in *Tdio*; and its presence in *Tern* is unclear due to poor sequence information. The translated proteins are as shown in **4.SI-S17b**. AUG and UAA are the universal start and stop codons, respectively. The fully-edited coding sequences are between 88-100% identical (**4.SI-S17c**) and the proteins are between 86-100% identical (**4.SI-S17d**).

G4: This gene is also transcribed from the strand opposite to the ribosomal genes and was reported to be pan-edited in *Tbru* (Corell et al., 1994). This fully-edited sequence was used as a reference to elucidate the editing patterns in all other trypanosomatids in this report. Our alignments indicate that G4 is pan-edited in all the organisms, with a final coding sequence length of 441 nt (**4.SI-S18a**). The 5' deletion of G4 in *TcEsm* was reported earlier (Ruvalcaba-Trejo and Sturm, 2011). This gene is absent in *Tdio* and is not detectable in *Tern* due to poor sequence information. The translated protein alignment is shown in **4.SI-S18b**. Although AUG is the predominant start codon, AUA and GUG, encoding Isoleucine and Valine, respectively, are also seen. Similarly, UAA is the predominant stop codon, with UAG being used less frequently. The fully-edited coding sequences are between 92-100% identical (**4.SI-S18c**) and the proteins are 88-100% identical (**4.SI-S18d**).

ND3 (G5): The G5 gene is also transcribed from the strand opposite to the ribosomal genes and is believed to putatively encode ND3. There are some uncertainties about the protein it encodes and its editing; in that, there are reports about this protein being mis-edited or alternatively edited (Read et al., 1994) (Maslov et al., 1994b). Our alignment indicates that this transcript is pan-edited in all the trypanosomatids in this report, with the majority of the fully-edited coding sequences measuring 405 nts in length; although indels cause a reduction or increase in the lengths of these coding sequences (**4.SI-S19a**). Coding sequence of *Tser* shows a deletion with a final coding sequence length of 393 nts, while that of *Tlew* shows an insertion with the final coding sequence length of 414 nts. Coding sequence of *Tdio* shows both an insertion as well as a deletion with a final coding sequence length of 387 nts. It is unclear whether the gene is present in *Tern* due to poor sequence information. The translated proteins are believed to start with a non-canonical but conserved UUA or UUG codon (encoding Leucine); although a Methionine residue exists in frame at the 42nd position of the translated proteins of all these trypanosomatids; except in *Tdio*, that shows a deletion at that locus (**4.SI-S19b**). UAA is the universal stop codon. Identities for the transcripts and the proteins have been calculated for the conserved stretch beginning with Leucine. The conserved stretch ranges between 75-100% in identities (**4.SI-S19c**) and the proteins are between 81-100% identical (**4.SI-S19d**).

RPS12 (G6): G6 is also transcribed from the same strand as the ribosomal genes and was reported to be pan-edited and to encode RPS12 in *L. tar* (Maslov et al., 1992). The same report also predicted the fully-edited sequence of *Tbru* RPS12 and its protein. This putative fully-edited RPS12 sequence of *Tbru* was used as a reference to predict the editing patterns of the gene in all the organisms in this report. Our alignments (**4.SI-S20a** and **4.SI-S20b**) indicate that the transcript is pan-edited in all the organisms, with a mature coding sequence length of 252 nts.

AUG and UAG are the universal start and stop codons, respectively. The fully edited coding sequences are between 92-100% identical and the proteins are between 88-100% identical among all the organisms in this report (**4.SI-S20c** and **4.SI-S20d**).

Phylogenetic relationships and evolution of RNA editing:

ML analyses were performed, as detailed in the Materials and Methods section, on a **nuclear dataset** consisting of catenated nucleotide sequences of three protein encoding nuclear genes: GAPDH, DHFR and GSH1. For the non- *T. cruzi* protozoans, the relationships obtained from our nuclear dataset (**Fig 4.2A**) corroborate several previously reported phylogenies (Viola et al., 2009) (Lima et al., 2012) (Acosta et al., 2013) (Lima et al., 2013) (Caballero et al., 2015); with a minor change in the positions of *Tern* and *Tdio* with respect to *T. cruzi*. Our results show *Tdio* to be closer to the *T. cruzi* strains compared to *Tern*, as was seen in the phylogeny for PRAC gene (Caballero et al., 2015). As for the *T. cruzi* strains, *TcM62* (TcIII) is the closest to the TcI members, followed by *TcCAN* (TcIV) to TcI and TcIII. Members of TcVI cluster with TcII (*TcEsm*). These results recapitulate the relationships established previously (Machado and Ayala, 2001) (Brisse et al., 2003) (Westenberger et al., 2005) (Flores-López and Machado, 2011) (Lewis et al., 2011) (Tomasini and Diosque, 2015) between several *T. cruzi* DTUs based on several nuclear loci. Also, *TcMar* forms the outgroup to all the other *T. cruzi* strains, as was reported previously (Machado and Ayala, 2001) (Cavazzana Jr. et al., 2010).

The ML analysis performed on the **maxicircle dataset**, consisting of a catenated subset of unedited maxicircle genes generated an almost identical phylogenetic tree to that of the nuclear genes (**Fig 4.2B**). Peculiar differences are observed in the relationships between the various *T. cruzi* strains. The maxicircle sequences of TcI members appear to be closer to the members of TcIII, TcVI and TcVI, while that of TcII appears to be the furthest from I, III, IV and VI. The maxicircle of *TcMar*, again appears to be the closest to all the other *T. cruzi* DTUs. The relationships between the non-*T. cruzi* members, based on the maxicircle sequences, remain unchanged from the nuclear-genes based tree.

Finally, we also generated a **binary editing dataset** that defines the patterns of editing and quantitates the RNA editing among all the organisms in this report (See Materials and Methods for details). We expected that a ML analysis on this dataset should group the organisms based on the similarities in their editing patterns, and in effect, help predict the evolution of RNA editing among the different organisms. Not surprisingly, closely related members showed more similar patterns of RNA editing (**Fig 4.2C**), thereby reflecting the relationships seen in the nuclear genes and maxicircle subset trees. In particular, the RNA editing binary tree showed a striking resemblance to the tree obtained using the maxicircle dataset (**Fig 4.2B**), except for the relationship between the members of TcVI and TcIII, in addition to *Tern* and *Tdio*. The editing tree indicates that the RNA editing in the members of TcVI is progressively and increasingly similar to the editing in TcIII and TcIV. Also, the editing in TcIII, TcIV and TcVI is closer to the editing in TcI than to TcII or *TcMar*. Diverging from the nuclear genes-based relationships, the RNA editing in *Tern* appears to be closer to the *T. cruzi* strains, compared to the editing patterns in *Tdio*. However, the maxicircle of *Tern* is incomplete, and therefore, not all the genes and their RNA editing have been characterized and included in this dataset (see Materials and Methods). The altered topology of *Tern* with respect to the *T. cruzi* clade in this tree may be attributable to its incomplete genome and RNA editing information. Overall, RNA editing has evolved to be similar in closely related organisms and also in parallel with the nuclear genes, in all the organisms, except within the *T. cruzi* clade.

In conclusion, all the trees unambiguously indicate that *TcMar* is the closest to all the other *T. cruzi* strains; and the different strains of *T. cruzi* show unique variations in the internal relationships, depending upon the dataset sampled (nuclear or maxicircle) for the analyses. The

non-*T. cruzi* members maintain their relationships with respect to each other, thereby reflecting their evolutionary similarity.

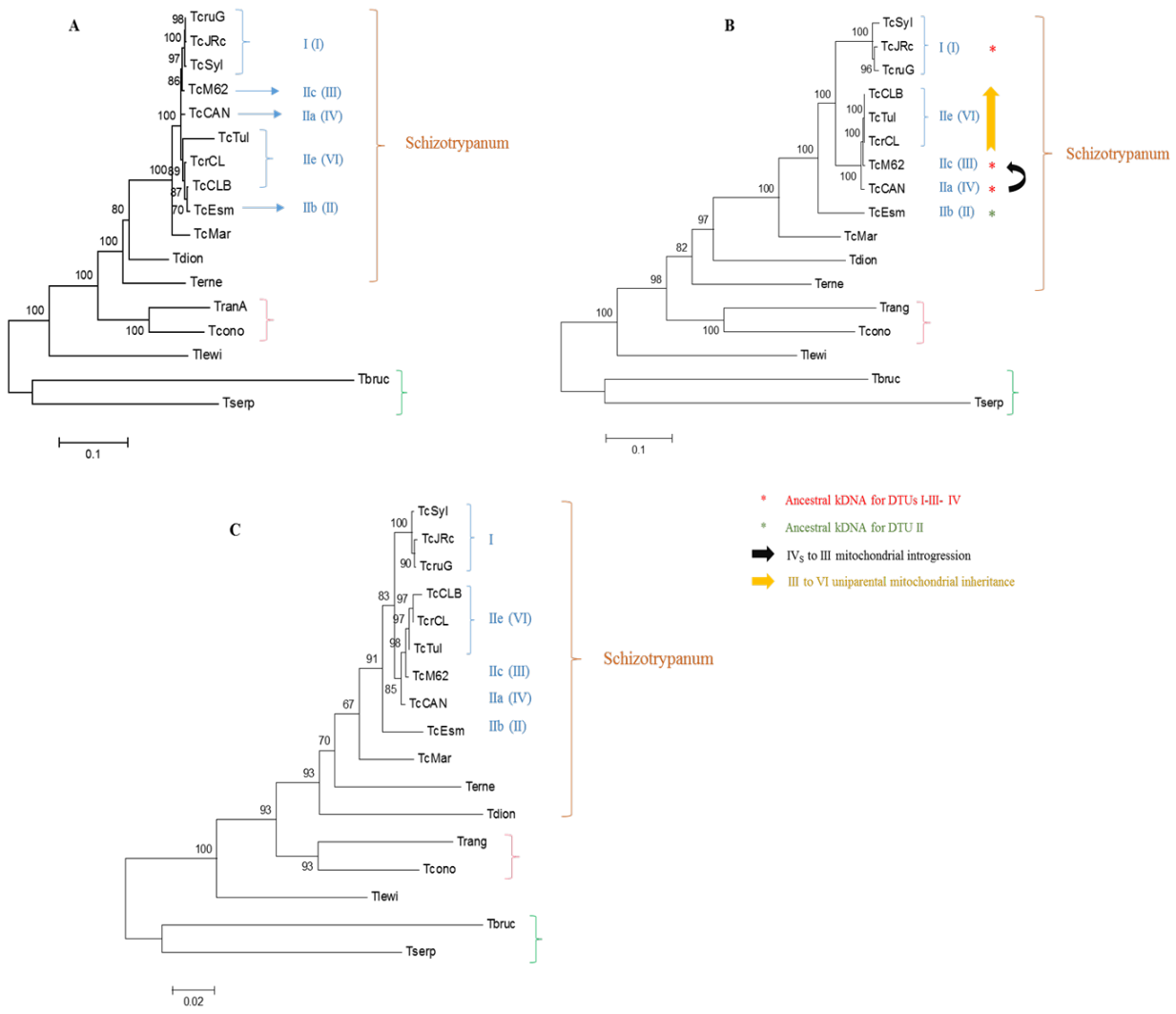


Fig 4.2. Maximum likelihood phylogenies depicting the relationships between the trypanosomatids in this report. (A) Phylogenetic tree based on the nuclear dataset (4667 characters). **(B)** Phylogenetic tree based on the maxicircle dataset (8738 characters). **(C)** Phylogenetic tree based on the RNA editing quantification (binary) dataset (1649 characters). Bootstrap values from 1000 pseudoreplicates are shown on the nodes. The scale bar represents the number of substitutions per site.

[Tbruc = *T. brucei;* TcCAN = *T. cruzi* CANIII; TcCLB = *T. cruzi* CL Brener; TcEsm = *T. cruzi* Esmeraldo; TcJRc = *T. cruzi* JR cl. 4; TcM62 = *T. cruzi* M6241; TcMar = *T. cruzi marinkellei* (B7 for maxicircle and binary dataset; TCC 344 for the nuclear dataset); Tcono = *T. conorhini*; TcrCL = *T. cruzi* CL; TcruG = *T. cruzi* G; TcSyl = *T. cruzi* Sylvio X10; TcTul = *T. cruzi* Tulahuen, Tdion = *T. dionisii;* Terne = *T. erneyi* (sp. 1946); Tlewi = *T. lewisi*; TranA = *T. rangeli* AM80 (for nuclear dataset); Trang = *T. rangeli* SC-58 (for maxicircle and binary datasets); Tserp = *T. serpentis*.]

Discussion

Herein, we report the syntenic maxicircle genomes of several strains of *T. cruzi* and non-*T. cruzi* members of the genus *Trypanosoma* and compare them with the published maxicircle genomes of *TcCLB*, *TcEsm*, *TcSyl* and *Tbru*.

The maxicircle genomes of all the organisms in this report showed the presence of the 18 protein encoding genes, along with the two ribosomal RNA genes, with the exception of the maxicircle of *Tdio*. The *Tdio* maxicircle showed specific deletions of portions of several genes (see the results above for gene-by-gene analysis). Similarly, the maxicircle of *Tern* is incomplete and therefore, 10 of its 20 genes have not been characterized in this report. The polarities of all the genes in all the organisms were as reported earlier for *Tbru*, *TcCLB*, *TcEsm* and *TcSyl*. The mitochondrial genomes of all these trypanosomatids showed a low GC content of 23-32%, which is not unexpected for these organisms (Shioiri and Takahata, 2001).

T. cruzi strains have been classified in to six DTUs. Replication among the *T. cruzi* strains is mostly clonal (Tibayrenc and Ayala, 2002) (Tibayrenc, 2003), although several reports have indicated the possibility of the occurrence of a rare genetic exchange via hybridization between the different strains (Machado and Ayala, 2001) (Brisse et al., 2003) (Sturm et al., 2003) (Gaunt et al., 2003). Reports examined different genomic loci of several DTUs and reported that the DTUs I, II, III and IV appeared homozygous, while the DTUs V and VI appeared heterozygous (Machado and Ayala, 2001) (Gaunt et al., 2003) (Westenberger et al., 2005). In their analyses, Westenberger et al (Westenberger et al., 2005), reported that the DTUs I and III were the closest to each other, with the DTUs III and IV being the next closest. DTUs I and II were the farthest from each other. They also reported that despite the homozygosity of III and IV, their sequences

appeared to share some SNPs with DTUs I and II, thereby hypothesizing a possible ancient hybridization event between DTUs I and II to form the hybrid progeny DTUs III and IV through subsequent homologous recombination and clonal evolution. Similarly, DTUs V and VI were proposed to be the hybrid progeny from a more recent single "backcross" hybridization event between II and III. However, de Freitas et al (de Freitas et al., 2006) contradicted Westenberger et al's results and proposed that the DTUs I, II and III were the ancestral *T. cruzi* DTUs and that at least two independent hybridization events have occurred between the DTUs II and III to create viable hybrid progenies. Subsequently, Lewis et al (Lewis et al., 2011) agreed on the possibility of an ancient hybridization event between DTUs I and II to form the DTUs III and IV, as per Westenberger et al's report. They also corroborated de Freitas et al's hypothesis that the more recent hybrid DTUs V and VI have been formed after two independent hybridization events between DTUs II and III. At around the same time Flores-Lopez and Machado (Flores-López and Machado, 2011) argued that the DTU relationships in Westenberger et al's analyses were skewed due to the non-inclusion of a suitable outgroup. They reported that the DTUs III and IV appeared to cluster with I, with a high support, but not with II. Therefore, they discounted a possible hybridization event between DTUs I and II to form the progeny III and IV. However, they agreed with the postulated single hybridization event between II with III to generate the hybrid DTUs V and VI. Finally, in an attempt to put an end to the DTU controversy, Tomasini and Diosque (Tomasini and Diosque, 2015) reanalyzed all the available sequence data and reviewed the literature to define the relationships as it stands today. They have concluded that the DTUs I and III are sister clades, with I-III-IV forming a monophyletic group. They say that the ancestral *T. cruzi* DTU that separated from *TcMar* divided into two groups: a) DTU II and b) the DTUs I-III-IV. Subsequently, IV differentiated geographically into IV_N and IV_S, and I-III

diverged into separate clades, clonally. Finally, DTUs V and VI have resulted more recently by two independent hybridization events between the DTUs II and III. In keeping with all the hypotheses and results so far, the phylogenetic trees generated from our ML analysis on the three nuclear genes (GAPDH, DHFR and GSH1) [**Fig 4.2A**] shows *TcM62* (III) to be the closest to *TcG*, *TcSyl* and *TcJR* (members of DTU I); and *TcCAN* (IV), *TcM62* (III) and DTU I forming a monophyletic group. Members of the DTU VI (*TcCLB*, *TcTul* and *TcCL*) clustered together with *TcEsm* (DTU II) in our analysis, reminiscent of the results from the concatenation of 13 MLST fragments by Diosque et al (Diosque et al., 2014). These results hint towards the plausibility of the three sampled nuclear genes (in our dataset) being the DTU II haplotype in DTU VI. A representative of DTU V is unavailable for comparison in the current analyses.

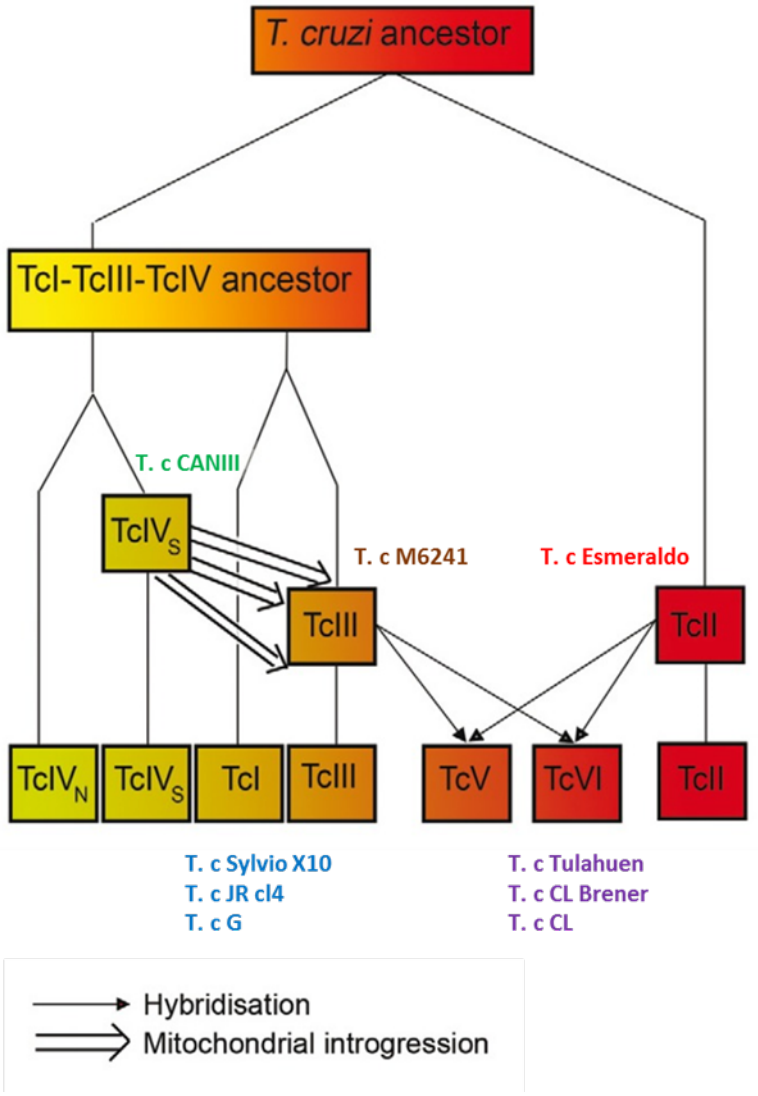


Figure 4.3: The revisited model for *Trypanosoma cruzi* evolution: Figure adapted and republished from *Memórias do Instituto Oswaldo Cruz*, 110.3, May 2015, Nicolás Tomasini, and Patricio Diosque, Evolution of *Trypanosoma Cruzi*: Clarifying Hybridisations, Mitochondrial Introgressions and Phylogenetic Relationships between Major Lineages, Pages 403–413, Open Access article from *PMC* under the terms of the Creative Commons Attribution Non-Commercial License.

As for the maxicircle inheritance during hybridization events, it is important to note that the maxicircles are inherited uniparentally, although the minicircles are inherited from both parental strains (Gibson et al., 1997) (Gaunt et al., 2003). Early analyses based on the maxicircle genes of

CYb and portions of COII and ND1 had regrouped the six DTUs into 3 major clades: clade A, comprising TcI, clade B, encompassing the DTUs TcIII, TcIV, TcV and TcVI, and clade C, comprising TcII (Machado and Ayala, 2001) (Brisse et al., 2003). This led to the hypothesis (Westenberger et al., 2006) (Sturm and Campbell, 2010) that TcI should have been the parent maxicircle donor to the hybrid progeny DTUs of TcIII and TcIV, during the proposed initial hybridization event between TcI and TcII. Subsequently, TcIII should have been the parent maxicircle donor to the hybrid DTUs TcV and TcVI, during the apparent backcrossing hybridization event between TcIII and TcII. More recently, however, since TcI-III-IV have been shown to be monophyletic, and V and VI are proposed to have resulted from two independent hybridization events between II and III, a new scenario of maxicircular introgression events have been proposed (Tomasini and Diosque, 2015). It is now believed that IV_S transferred its kDNA to III, after the I-III split, through an uncharacterized mechanism of mitochondrial introgression. Thereafter, multiple events of such IV_S to III mitochondrial introgressions happened, and the IV_S kDNA 'surfed' through the expanding III population, establishing itself as the current III kDNA. Finally, when DTUs V and VI appeared through two independent II-III hybridization events, the kDNA of III (the introgressed IV_S kDNA) was inherited by the new hybrid DTUs both the times. Indeed, phylogenetic analysis on our maxicircle genes dataset (**Fig 4.2B**) demonstrates the closeness of the maxicircle sequences of DTUs III, IV and VI. Again, since I-III-IV are supposedly one of the original ancestral DTUs, the maxicircle analyses indicate that the DTU I members are monophyletic with that of III-IV-VI group. The kDNA of DTU II does not appear to have been inherited or taken part in any introgression, and therefore, II forms an outgroup to all the other DTUs.

The overall resemblance of the binary RNA editing evolution tree (**Fig 4.2C**) to the maxicircle tree (**Fig 4.2B**) in the *T. cruzi* group, rather than to the nuclear tree, may be due to the fact that RNA editing, by itself, is specific to the maxicircle sequence; and the maxicircle sequences of III, IV, V and VI are more similar to I than to II, as described above. Thus, the RNA editing patterns intuitively correlate with the maxicircle similarities and inheritance patterns.

T. cruzi marinkellei, a bat-associated subspecies of *T. cruzi*, has consistently been recognized as the closest to the other *T. cruzi* DTUs. The Schizotrypanum subgenus of *Trypanosoma* forms a monophyletic group and comprises all the species that are bat-restricted, except *T. cruzi* that infects all mammals (Marcili et al., 2009) (Cavazzana Jr. et al., 2010). The blood and culture forms of all the parasites of the Schizotrypanum are morphologically indistinguishable from *T. cruzi*, thereby earning them the distinction of '*T. cruzi*-like' parasites. With the identification of a new *Trypanosoma* member, *T. erneyi* (former Trypanosoma sp. 1946), that groups as a sister-clade to *TcMar*, the subgenus Schizotrypanum includes the following species in this report: *T. cruzi*, *TcMar*, *Tern* and *Tdio* (Lima et al., 2012). *Tran*, that infects several mammals, including bats, and *Tcon*, that normally infects rats, are not only phylogenetically closer to each other, but also are consistently reported to group as a sister clade to the Schizotrypanum (Stevens et al., 1999) (Da Silva et al., 2004) (Maia da Silva et al., 2009) (Cavazzana Jr. et al., 2010) (Lima et al., 2013) (Caballero et al., 2015). The species of *T. cruzi*, *TcMar*, *Tern*, *Tdio*, *Tran* and *Tcon* from this current report have all been broadly classified into the "*T. cruzi* clade" (Hamilton et al., 2012) (Caballero et al., 2015). *Tlew*, a cosmopolitan trypanosome of domestic rats that can infect immunocompromised humans and non-human primates, forms an outgroup to the *T. cruzi* clade (Caballero et al., 2015) (Truc et al., 2013). *Tser*, the snake trypanosome, is more basal in the trypanosome family and clusters closer to the *T. brucei* family as has been previously observed

(Viola et al., 2009) (Caballero et al., 2015). The phylogenetic relationships of the non-*T.cruzi* trypanosomes are conserved and remain unchanged at both the nuclear and the mitochondrial level. In fact, their RNA editing also appears to have evolved in parallel to the nuclear and mitochondrial genes (**Figs 4.2A, 4.2B and 4.2C**).

Not surprisingly, these very relationships are reflected in the maxicircle as well as nuclear gene identities analyzed within the *T. cruzi* populations, as well as between other members of the Schizotrypanum and the broader *T. cruzi* clade (**Table 4.3 and Supplementary Tables 2, 3 and 4**).

The complicated process of U insertion-deletion RNA editing of the mitochondrial transcripts of the trypanosomatids is believed to have evolved in an ancestor to the kinetoplastids-bodonids (Simpson and Maslov, 1999) (Simpson et al., 2000). The recent discovery of apparently edited mitochondrial transcripts from a Neobodonid member confirms that RNA editing is an ancient phenomenon existing across the Order Kinetoplastida (Deschamps et al., 2011). Several rooted phylogenies (Maslov et al., 1994a) (Landweber and Gilbert, 1994) have indicated that species diverging early in the kinetoplastid lineage (including the bodonids, cryptobiids and trypanosomatids like *T. brucei*, the *Trypanosoma* genus and even the genus *Herpetomonas*) have more extensive RNA editing, or more pan edited genes, compared to the species that diverged later (like the Leishmaninae family or the symbiont harboring trypanosomatids). It was also reported that pan-editing is more ancient than 5' editing (Simpson and Maslov, 1994a). The fact that recently diverged species show lower numbers of pan edited genes and higher numbers of 5' or unedited genes led to the hypothesis that the more extensively edited genes in the past have been somehow replaced by genes requiring far lesser amounts of editing by a putative

retroposition mechanism, where the cellular reverse transcriptase-like activities could reverse transcribe partially or fully-edited transcripts into cDNA, which could replace the existing pan edited gene in the maxicircle by the homologous recombination mechanism, thereby leading to genes that require only 5' editing or no editing at all, respectively (Landweber, 1992) (Maslov et al., 1994a) (Landweber and Gilbert, 1993) (Simpson and Maslov, 1999) (Jan Arts and Benne, 1996) (Simpson and Maslov, 1994a) (Gabriel and Boeke, 1991) (González et al., 1997). This was also believed to be evolutionarily favorable, since, elimination of editing removes the genetic load imposed on the cell and also helps the cells in case of minicircle mis-segregation during replication (thereby leading to a loss of gRNAs) and transkinetoplastidy-like situations (Maslov et al., 1994a) (Simpson et al., 2000). This entire premise appears to be true, with all the trypanosomatids in this report undergoing editing in an identical pattern, both in terms of the number of genes edited as well as the extents of editing of each individual gene, much unlike among the endosymbiont bearing and insect trypanosomatids (**Chapter 3**).

One of the main functions of RNA editing is to post-transcriptionally correct frameshifts and create well-defined open reading frames for translation. It is only after this extensive process of editing, that the coding sequences of most genes show higher identities between each other and a fair amount of conservation at the level of translated proteins (Matrices in **Supporting Information** section). On the other hand, the unedited genes are normally the only maxicircle genes that show high levels of identities in both their coding sequences and proteins, without undergoing any editing. Thus, it is intuitive that the unedited genes should apparently face a higher evolutionary pressure to maintain their sequences and reading frames, simply due to the absence of editing of their transcripts. Surprisingly, however, several unedited genes in maxcircles of the *Trypanosoma* members show a high degree of insertion-deletion mutations

(indels) that range from a single nucleotide to a stretch of a few nucleotides that disrupt their translation reading frames (See **Results** section). Similar instances of indels in the non-edited genes or the unedited portions of partially edited genes have been reported in the past for some *T. cruzi* strains (Ruvalcaba-Trejo and Sturm, 2011). It is unknown at this point whether these indels are corrected by the mechanism of RNA editing. If they are shown to be edited, these may then represent the first ever case of editing of the apparently 'non-edited' genes. It is also tempting, however, to propose that these indels may be observed in the maxicircles of some of these trypanosomatids, due to the lack of a selective pressure to maintain their gene sequences, especially when these organisms are being cultured for a long time in a laboratory environment. Further, it may also represent a candidate of genes, whose functional loss may be tolerated by the cell. Such indels may, in fact, be deleterious to the organisms in the wild.

Curiously, most of the genes harboring such indels encode the various subunits of Complex I of the mitochondrial oxidative chain [ND1, ND2 (MURF1), ND4 and ND5], except the 5' edited MURF2 gene, whose function is unknown. Additionally, note that the genes showing major deletions in *Tdio*, are again Complex I subunit genes [ND1, ND2 (MURF1) and ND4]. This is reminiscent of the controversy surrounding the trypanosomatid Complex I; its role in mitochondrial bioenergetics and its necessity in the organism's life cycle (Verner et al., 2011) (Carranza et al., 2009) (Opperdoes and Michels, 2008).

A wealth of information is available about the extensive RNA editing needed in the procyclic (insect stage) form of *T. brucei* and about its very limited RNA editing (only a few genes are edited) in the bloodstream form. Such "compartmentalized" editing activities and mitochondrial gene regulation, depending upon its host, is not seen in all the trypanosomatids. For instance, *L.*

tarentolae and *T. cruzi* are both known to maintain their RNA editing activities at the same level throughout their life cycle and differentiation stages; i.e. all the mitochondrial genes are edited at all life stages. Other than these and a few other trypanosomatids that are extensively studied and well characterized, there is a general paucity of information on the life cycles and bioenergetics of most of the trypanosomatids discovered as yet. Therefore, biochemical studies on the RNA editing levels, coupled with details on the mitochondrial metabolism in their various hosts is the need of the hour, to help us understand and better appreciate the roles and significance of RNA editing.

Conclusion:

The data represented in this report elucidates the maxicircular genome organization of several members of the genus *Trypanosoma* (*T. cruzi* related and unrelated parasites). We have characterized the presence and absence of all the known maxicircle genes in all the organisms, along with the synteny of their genes. We have also inferred the editing patterns of all the genes of all the organisms and estimated the identities between the fully edited transcripts and proteins of all the organisms. Phylogenetic analyses on the concatenated datasets of maxicircle and nuclear genes reflect the expected the relationships between all the organisms. We quantified the RNA editing of all the edited genes using a binary matrix and a ML reconstruction on this RNA editing dataset revealed that the phenomenon of RNA editing has actually evolved in parallel to the nuclear and maxicircle genes.

Supplementary Tables for Chapter 4

Supplementary Table 4.1: Primers used for gap-closures and circularization of the maxicircle contigs (wherever circularization was possible).

Name	Primer sequence (5'-3')	T_m (^oC)
Tdio-gap F	CAGAAATGGATGCCAGCCGTTGCGG	69.5
Tdio-gap R	GTCTTATACTCACATAACTTTTCTGTACCACG	64.6
Tdio-circ F+	TCATTTTGTGTACCTTCGGGTACCG	62.8
Tdio-circ R+	TCTTACTTGGTACATATATAACAACTGTGCC	63.3
Tdio-del1 F*	CATATCCTACAAATCCACACAAAATAGC	61.7
Tdio-del1 R*	TGCCTACCTAATTGAAATAGTACC	59.4
Tdio-del2 F*	TCCAAATGCAATTGGTATAGACCCAATGC	64.6
Tdio-del2 R*	GGAGGGATTGGCAATTGTACAACTTAGG	66.1
TcG-circ F^	GACTGTAAATCTCATTAAATAAAGC	56.4
TcG-circ R^	GAAATTTTCATTACAAGGATTATCC	56.4

+ Putatively circularized the maxicircle.
*** Tdio-del1 primer set confirmed the deletion of genes G4 and 5' end of ND4; while Tdio-del2 primer set confirmed the deletion between MURF1 to ND1 genes. Refer to the text for details.**
^ Primer set did not circularize the maxicircle but helped extend the sequence at its 5' end.

Supplementary Table 4.2. Percent identity matrix for GAPDH gene for the trypanosomatids.

	Tbru	*Tser**	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser*	78	**100**															
Tlew	82	81	**100**														
Tran	79	80	84	**100**													
Tcon	80	80	86	90	**100**												
Tdio	81	82	87	86	89	**100**											
Tern	81	80	88	86	89	92	**100**										
TcCLB	81	80	86	87	88	93	92	**100**									
TcCL	81	80	87	87	89	93	92	100	**100**								
TcTula	81	81	87	87	88	93	93	99	99	**100**							
TcM62	81	80	86	87	88	93	92	99	99	99	**100**						
TcCAN	81	80	87	87	89	93	92	99	99	99	99	**100**					
TcG	81	80	86	87	88	93	92	99	99	99	99	99	**100**				
TcSyl	81	80	86	87	88	93	92	99	99	99	99	99	100	**100**			
TcJR	81	80	86	87	88	93	92	99	99	99	99	99	100	100	**100**		
TcEsm	81	81	87	87	88	93	92	99	99	100	99	99	99	99	99	**100**	
TcMar	81	81	86	87	88	93	92	98	98	98	98	98	97	97	97	98	**100**

The matrix above shows the nucleotide percent identities of the GAPDH gene between all the organisms studied.
*** indicates that the identities of *Tser* GAPDH gene were obtained after elimination of the absent gene portions in all the other organisms' GAPDH genes. (*Tser* GAPDH gene sequence misses its 3' 259 nts)**

Supplementary Table 4.3. Percent identity matrix for DHFR gene for the trypanosomatids.

	Tbru	*Tser**	*Tlew**	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser*	68	**100**															
Tlew*	66	72	**100**														
Tran	66	75	74	**100**													
Tcon	66	74	75	88	**100**												
Tdio	67	76	75	83	82	**100**											
Tern	67	75	76	83	83	89	**100**										
TcCLB	67	76	75	82	83	90	88	**100**									
TcCL	67	76	75	82	83	90	88	100	**100**								
TcTula	68	76	75	82	83	91	88	99	99	**100**							
TcM62	68	76	75	82	83	91	88	99	99	100	**100**						
TcCAN	67	76	75	82	83	91	88	99	99	99	99	**100**					
TcG	67	76	75	82	82	90	88	99	99	99	99	99	**100**				
TcSyl	67	76	75	82	82	90	88	99	98	99	99	99	100	**100**			
TcJR	67	76	75	82	82	90	88	99	99	99	99	99	100	100	**100**		
TcEsm	67	76	75	82	83	91	88	100	99	99	99	99	99	99	99	**100**	
TcMar	68	76	74	81	81	89	87	95	95	96	96	95	95	95	95	95	**100**

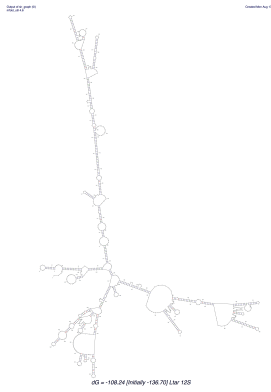
The matrix above shows the nucleotide percent identities of the DHFR gene between all the organisms studied.
*** indicates that the identities of the gene in *Tser* and *Tlew* were obtained after elimination of the absent gene portions in all the other organisms' DHFR genes. (*Tser* DHFR gene sequence misses its 5' 386 nts and *Tlew* DHFR sequence misses its 3' 86 nts)**

Supplementary Table 4.4. Percent identity matrix for GSH1 gene for the trypanosomatids.

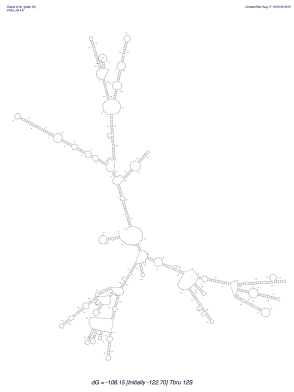
	Tbru	*Tser**	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser*	68	**100**															
Tlew	67	70	**100**														
Tran	66	70	76	**100**													
Tcon	67	72	74	85	**100**												
Tdio	68	72	76	81	80	**100**											
Tern	68	70	74	78	78	85	**100**										
TcCLB	68	71	76	80	80	86	82	**100**									
TcCL	68	71	75	80	80	86	83	99	**100**								
TcTula	64	66	70	74	74	80	76	91	90	**100**							
TcM62	69	71	75	80	80	86	83	98	99	89	**100**						
TcCAN	68	72	76	80	80	86	83	98	99	90	99	**100**					
TcG	69	72	76	80	80	86	82	98	99	90	99	99	**100**				
TcSyl	69	72	76	80	80	86	83	98	99	90	99	99	100	**100**			
TcJR	69	72	76	80	80	86	82	98	99	89	99	99	100	99	**100**		
TcEsm	68	71	76	80	80	86	82	100	99	91	98	98	98	98	98	**100**	
TcMar	68	71	75	80	80	86	82	95	95	88	96	96	96	96	96	96	**100**

The matrix above shows the nucleotide percent identities of the GSH1 gene between all the organisms studied.
*** indicates that the identities of the gene in *Tser* were obtained after elimination of the absent gene portion in all the other organisms' GSH1 genes. (*Tser* GSH1 gene sequence misses its 3'~300 nts)**

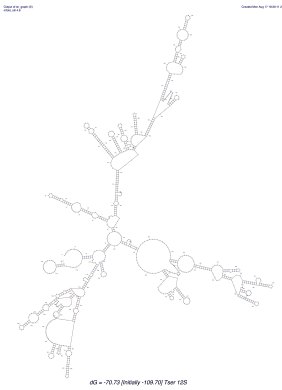
Supplementary Figures for Chapter 4



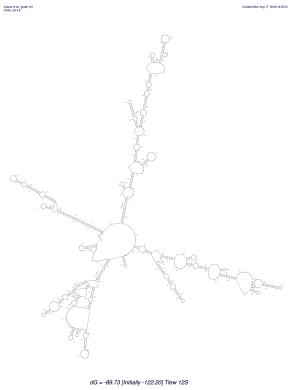
4.1A



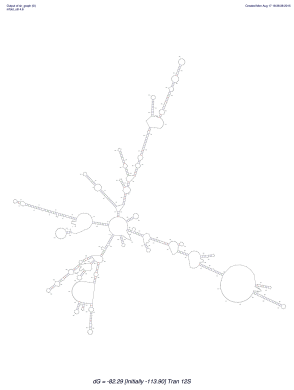
4.1B



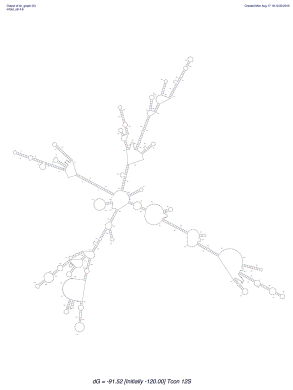
4.1C



4.1D

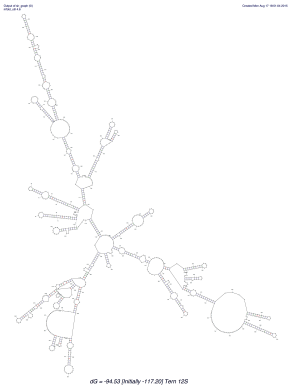


4.1E



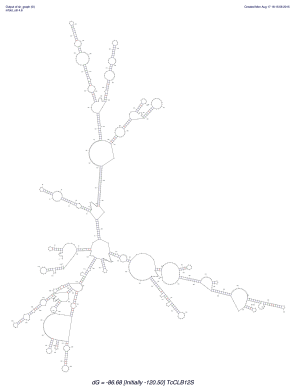
4.1F

4.1G

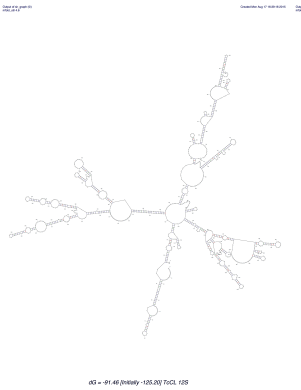


4.1H

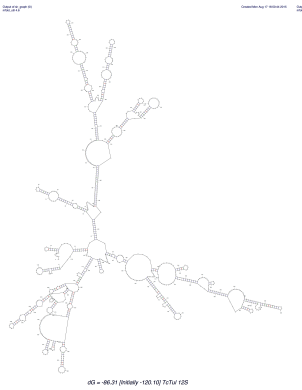
Supplementary Figure 4.1A. *Leishmania tarentolae* 12S folding;
Supplementary Figure 4.1B. *Tbru* 12S folding; Supplementary Figure 4.1C. *Tser* 12S folding;
Supplementary Figure 4.1D. *Tlew* 12S folding; Supplementary Figure 4.1E. *Tran* 12S folding;
Supplementary Figure 4.1F. *Tcon* 12S folding; Supplementary Figure 4.1G. *Tdio* 12S folding;
Supplementary Figure 4.1H. *Tern* 12S folding;



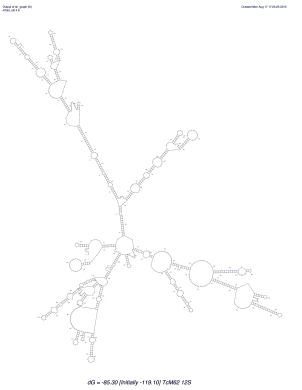
4.1I



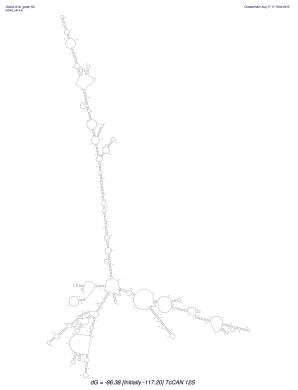
4.1J



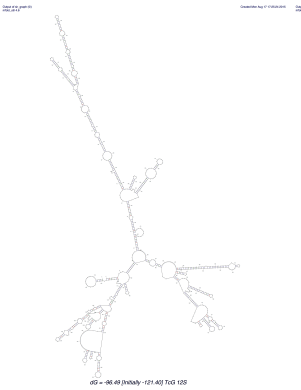
4.1K



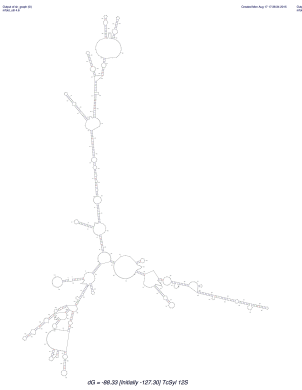
4.1L



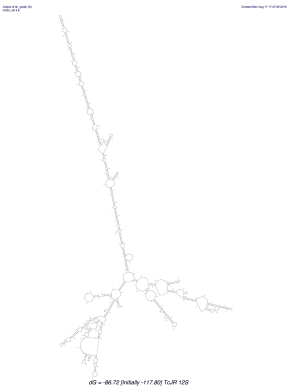
4.1M



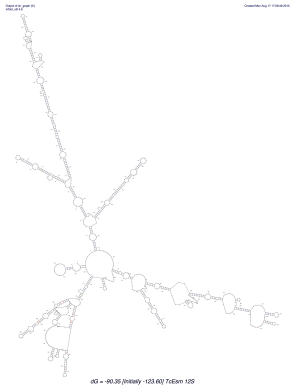
4.1N



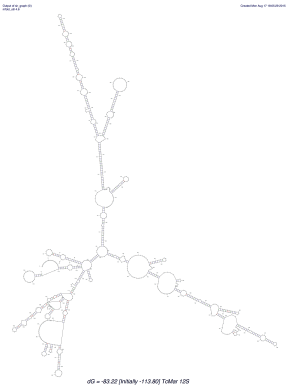
4.1O



4.1P

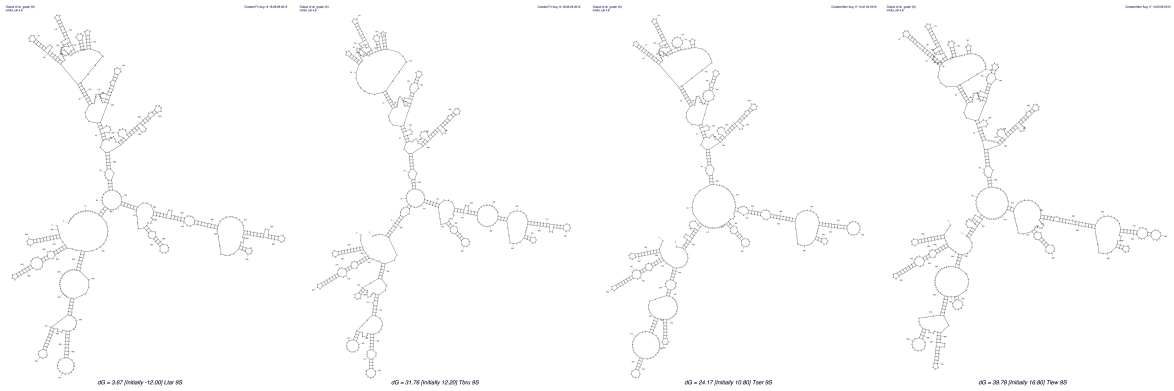


4.1Q

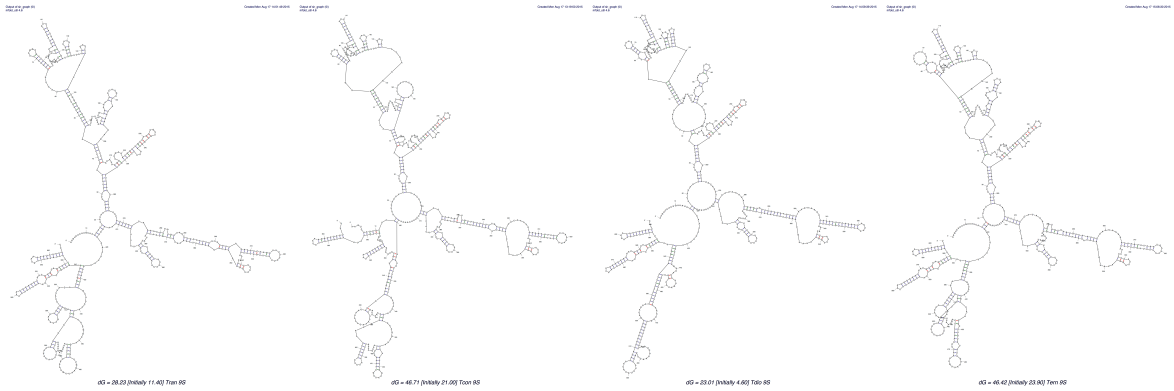


4.1R

Supplementary Figure 4.1I. *TcCLB* 12S folding; Supplementary Figure 4.1J. *TcCL* 12S folding; Supplementary Figure 4.1K. *TcTula* 12S folding; Supplementary Figure 4.1L. *TcM62* 12S folding; Supplementary Figure 4.1M. *TcCAN* 12S folding; Supplementary Figure 4.1N. *TcG* 12S folding; Supplementary Figure 4.1O. *TcSyl* 12S folding; Supplementary Figure 4.1P. *TcJR* 12S folding' Supplementary Figure 4.1Q. *TcEsm* 12S folding; Supplementary Figure 4.1R. *TcMar* 12S folding.

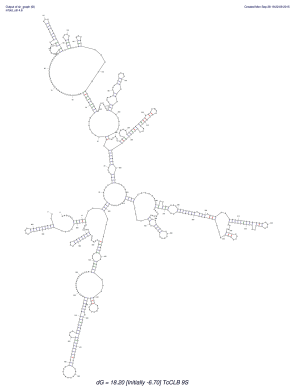


4.2A **4.2B** **4.2C** **4.2D**

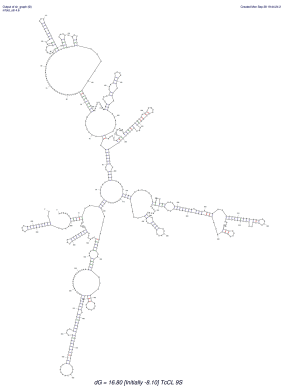


4.2E **4.2F** **4.2G** **4.2H**

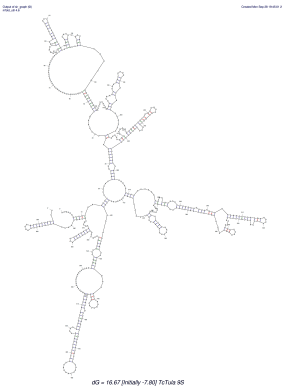
Supplementary Figure 4.2A. *L. tar* 9S folding; Supplementary Figure 4.2B. *Tbru* 9S folding; Supplementary Figure 4.2C. *Tser* 9S folding; Supplementary Figure 4.2D. *Tlew* 9S folding; Supplementary Figure 4.2E. *Tran* 9S folding; Supplementary Figure 4.2F. *Tcon* 9S folding; Supplementary Figure 4.2G. *Tdio* 9S folding; Supplementary Figure 4.2H. *Tern* 9S folding;



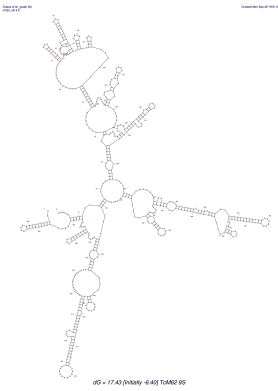
4.2I



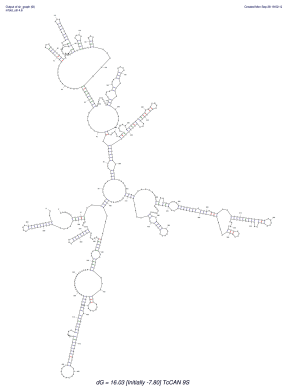
4.2J



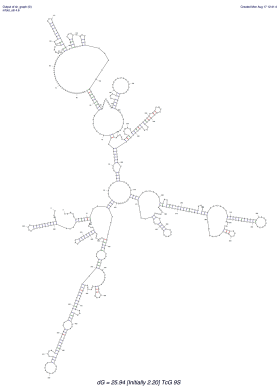
4.2K



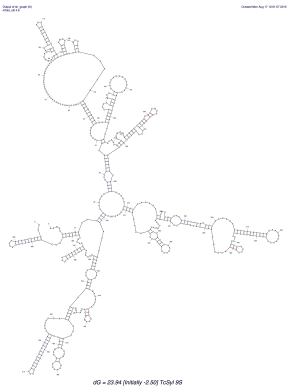
4.2L



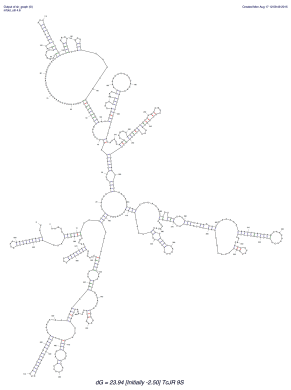
4.2M



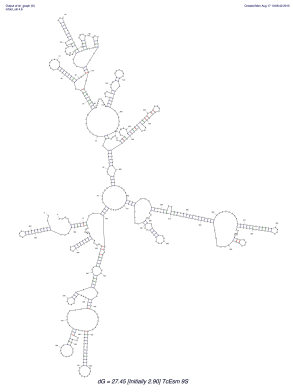
4.2N



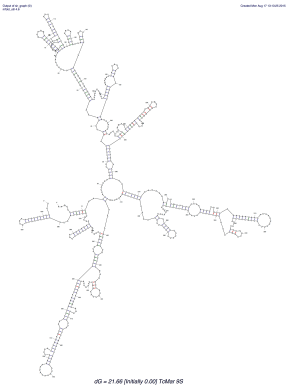
4.2O



4.2P



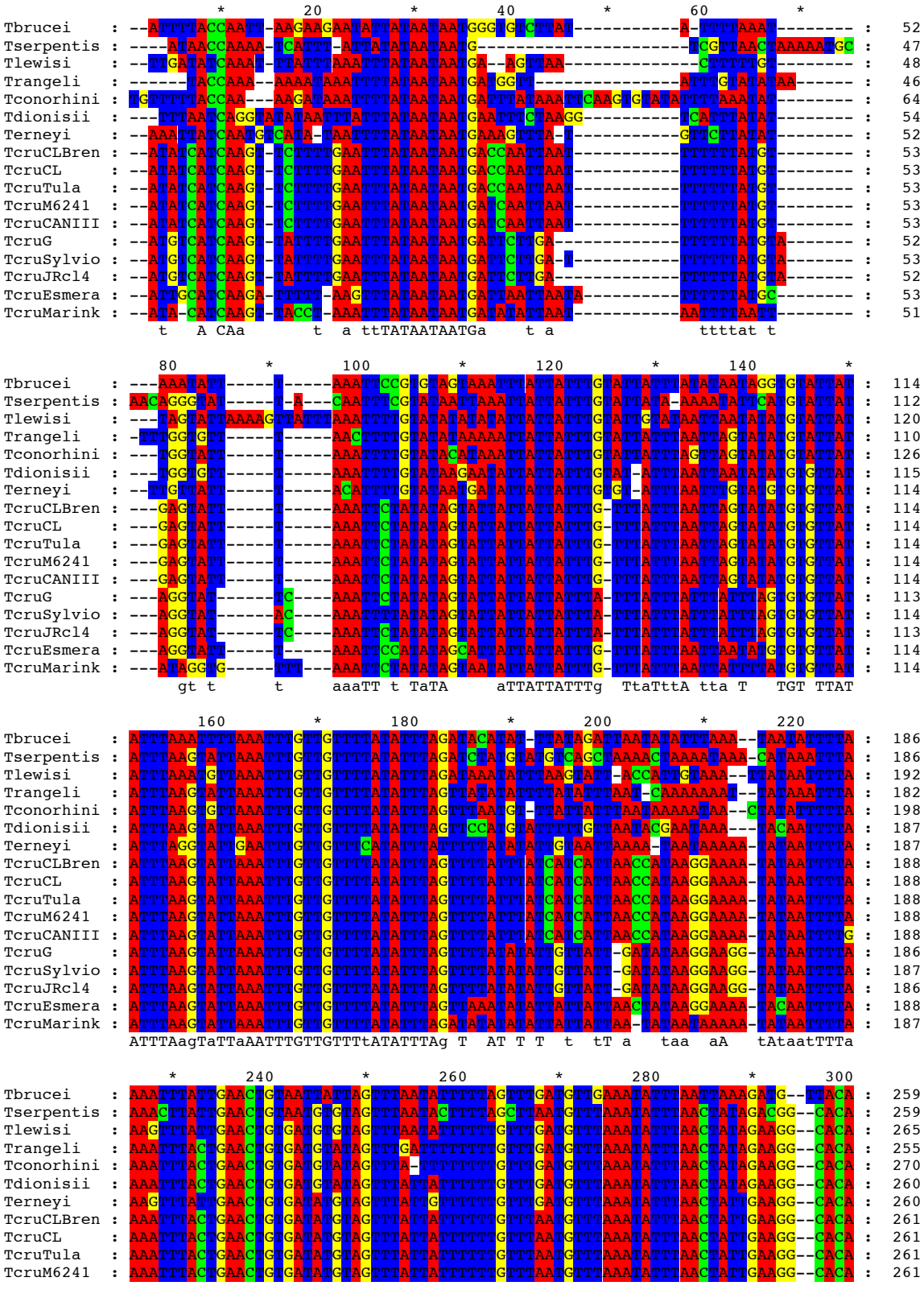
4.2Q

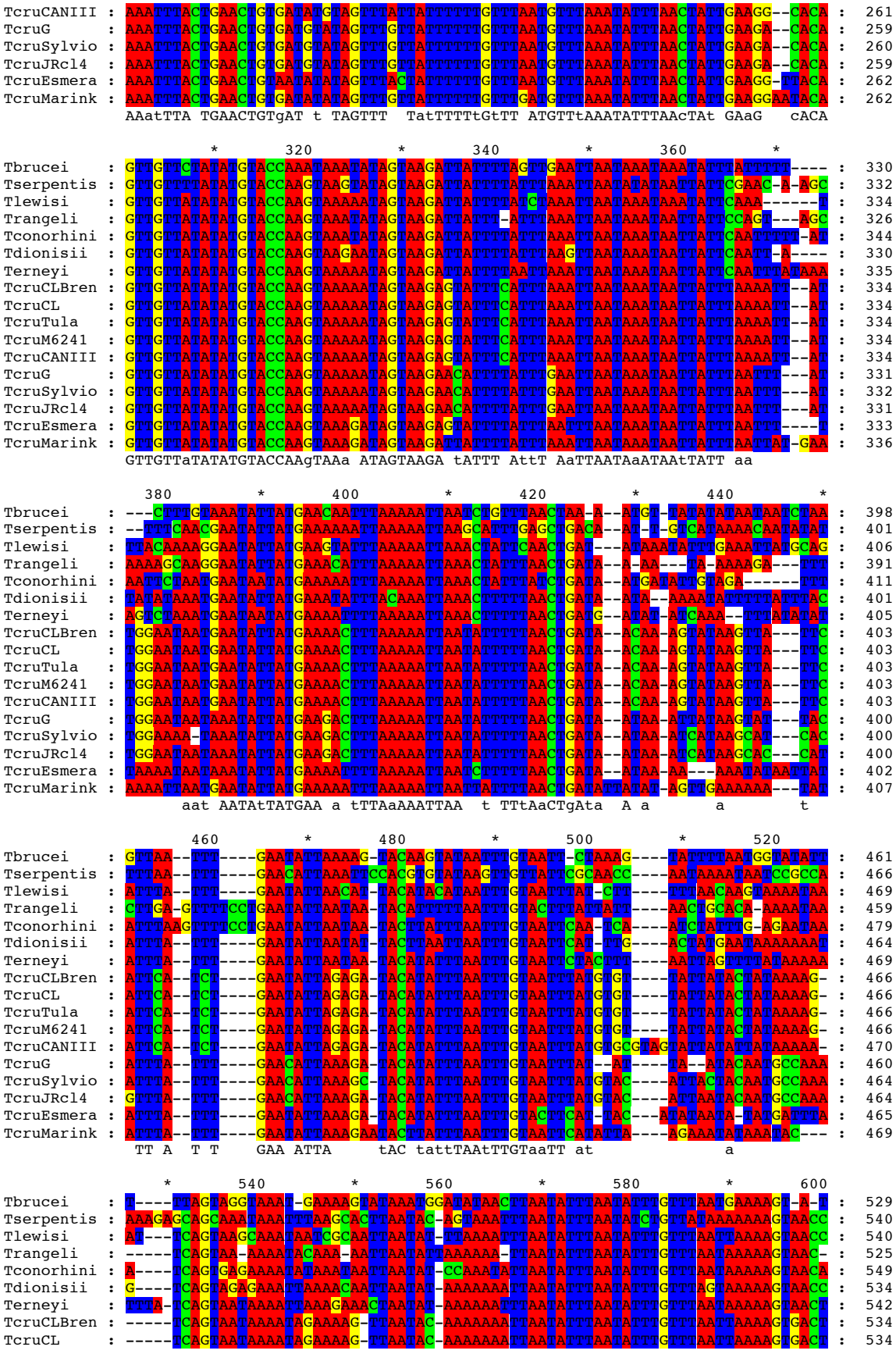


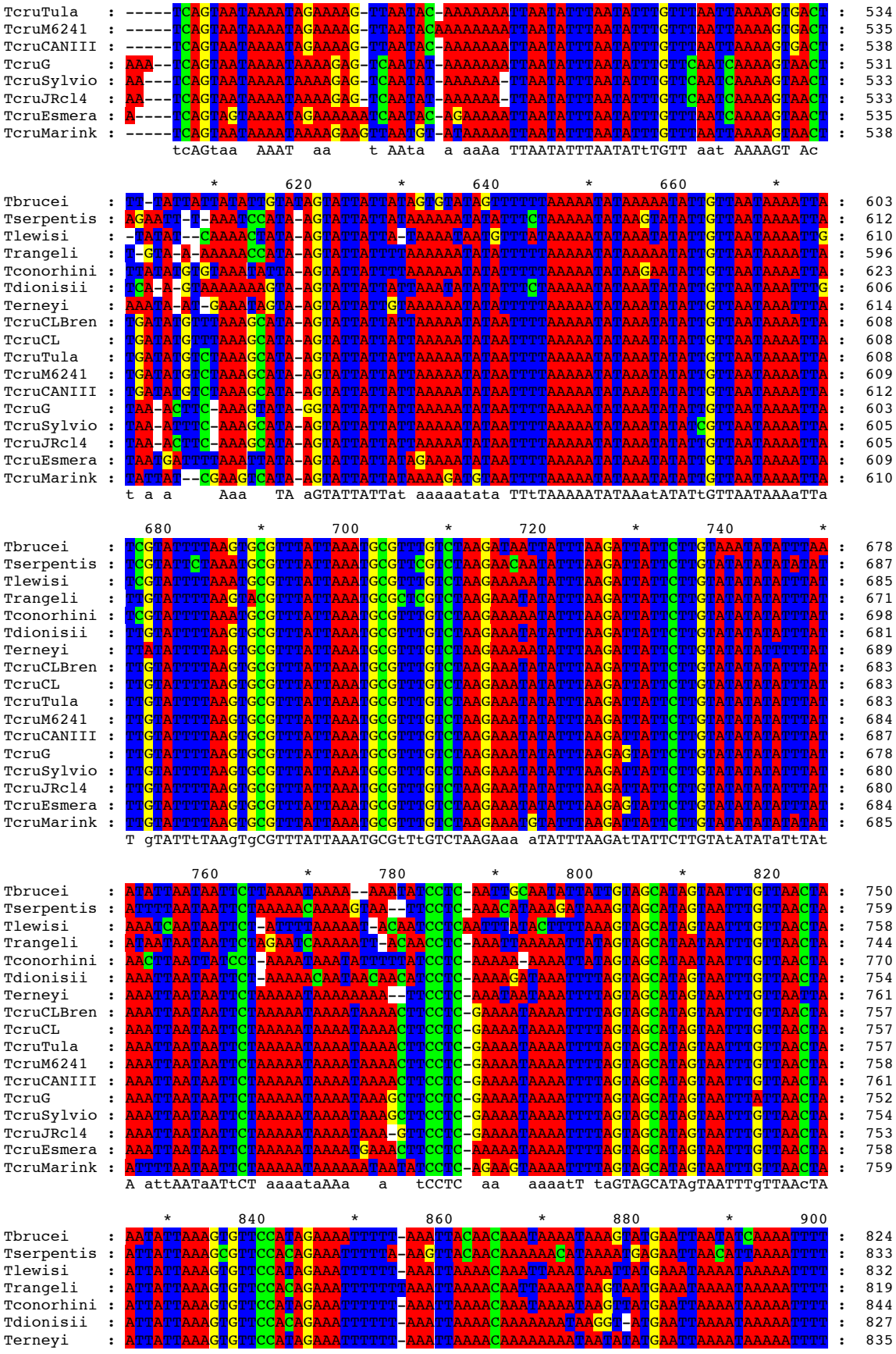
4.2R

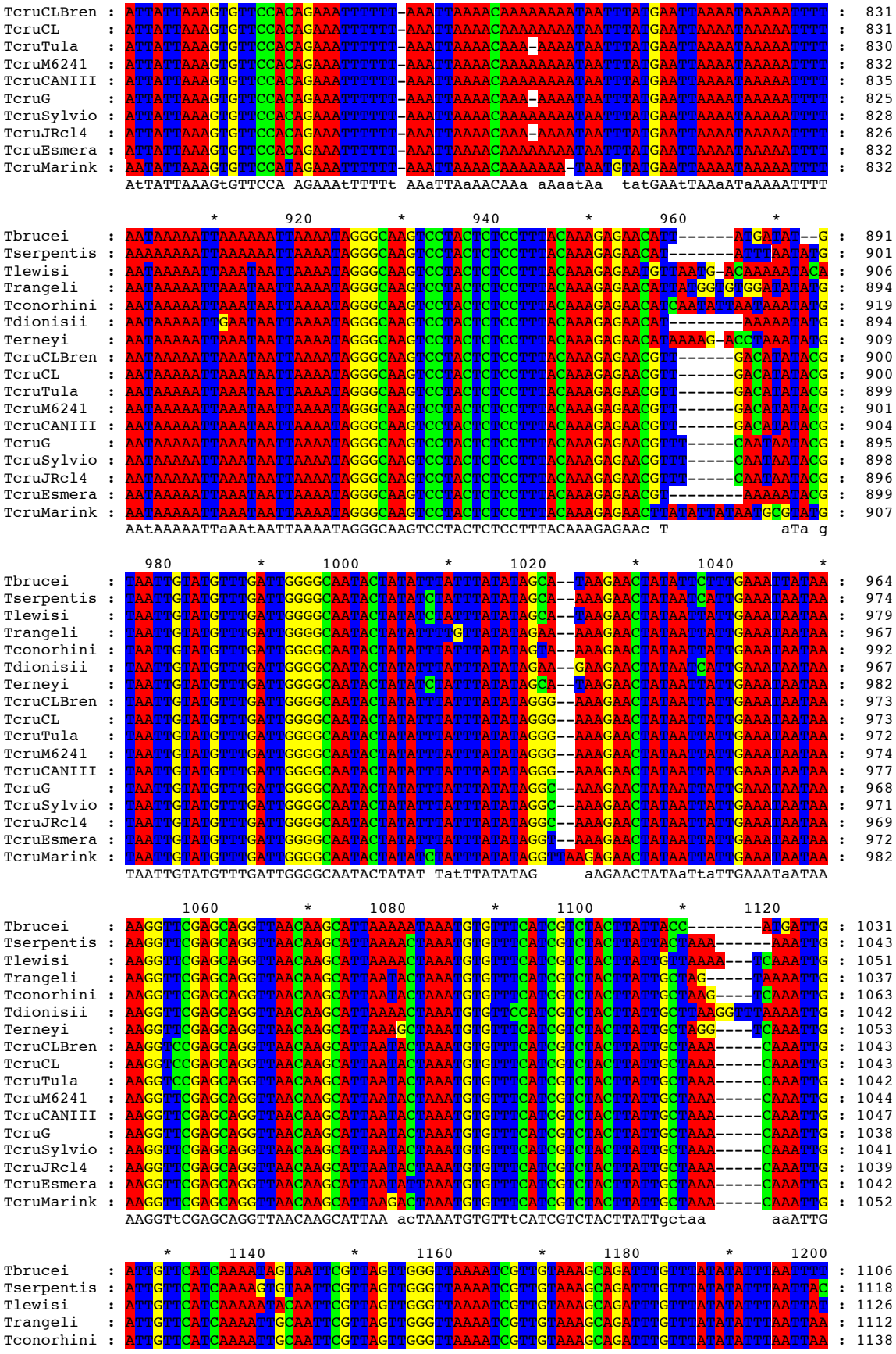
Supplementary Figure 4.2I. *TcCLB* 9S folding; Supplementary Figure 4.2J. *TcCL* 9S folding; Supplementary Figure 4.2K. *TcTula* 9S folding; Supplementary Figure 4.2L. *TcM62* 9S folding; Supplementary Figure 4.2M. *TcCAN* 9S folding; Supplementary Figure 4.2N. *TcG* 9S folding; Supplementary Figure 4.2O. *TcSyl* 9S folding; Supplementary Figure 4.2P. *TcJR* 9S folding; Supplementary Figure 4.2Q. *TcEsm* 9S folding; Supplementary Figure 4.2R. *TcMar* 9S foldi

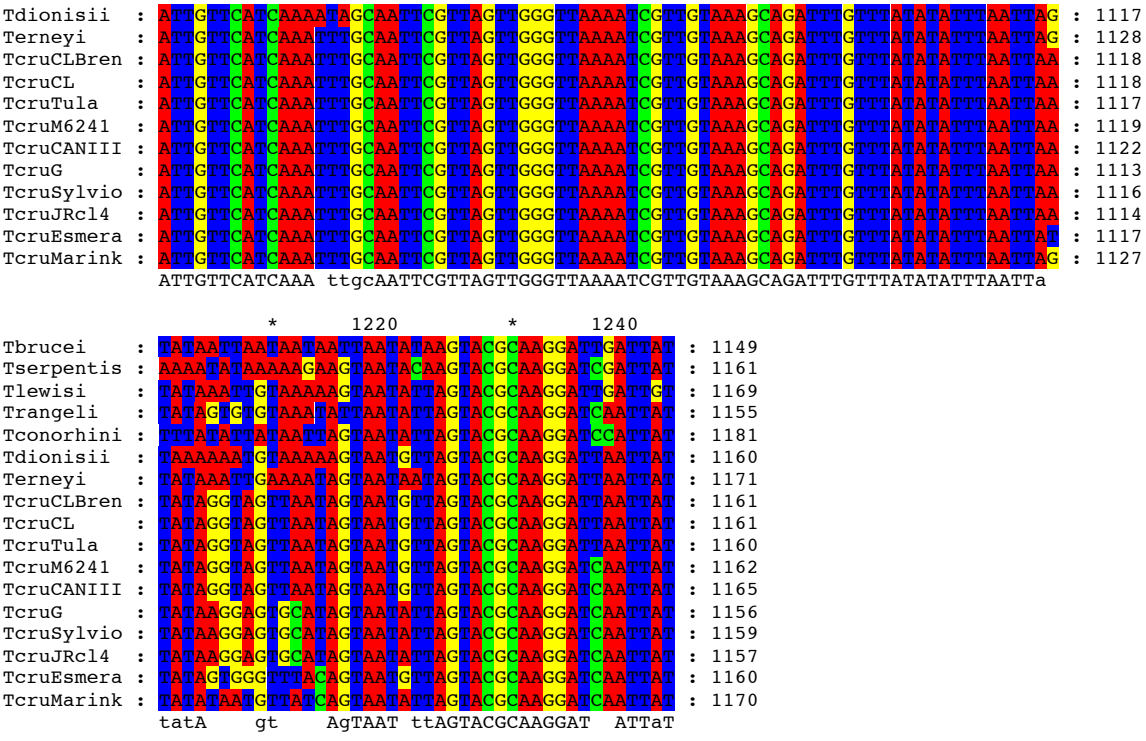
Supporting Information (SI) for Chapter 4






TcruTula : -----TCAGTAATAAAATAGAAAAG-TTAATAC-AAAAAAATTAATATTTAATATTTGTTTAATTAAAAGTGACT : 534
TcruM6241 : -----TCAGTAATAAAATAGAAAAG-TTAATAC-AAAAAAATTAATATTTAATATTTGTTTAATTAAAAGTGACT : 535
TcruCANIII : -----TCAGTAATAAAATAGAAAAG-TTAATAC-AAAAAAATTAATATTTAATATTTGTTTAATTAAAAGTGACT : 538
TcruG : AAA--TCAGTAATAAAATAAAAGAG-TCAATAT-AAAAAAATTAATATTTAATATTTGTTCAATCAAAAGTAACT : 531
TcruSylvio : AA---TCAGTAATAAAATAAAAGAG-TCAATAT-AAAAAA-TTAATATTTAATATTTGTTCAATCAAAAGTAACT : 533
TcruJRcl4 : AA---TCAGTAATAAAATAAAAGAG-TCAATAT-AAAAAA-TTAATATTTAATATTTGTTCAATCAAAAGTAACT : 533
TcruEsmera : A----TCAGTAGTAAAATAGAAAAAAATCAATAC-AGAAAAATTAATATTTAATATTTGTTTAATCAAAAGTAACT : 535
TcruMarink : -----TCAGTAATAAAATAAAAGAAGTTAATGT-ATAAAAATTAATATTTAATATTTGTTTAATTAAAAGTAACT : 538
tcAGtaa AAAT aa t AAta a aaAa TTAATATTTAATATTtGTTT aat AAAAGT Ac
620 640 660
Tbrucei : TT-TATTATTATATTGTATAGTATTATTATTAGTGTATAGTTTTTTAAAAATATAAAAATATTGTTAATAAAATTA : 603
Tserpentis : AGAATT-T-AAATCCAATA-AGTATTATTATTAAAAAATATATTTCTAAAAATATAAGTATATTGTTAATAAAATTA : 612
Tlewisi : -TATAT--CAAAACTATA-AGTATTATTA-TAAAATAAATGTTTATAAAAATATAAATATATTGTTAATAAAATTG : 610
Trangeli : T-GTA-A-AAAAACCAATA-AGTATTATTTTAAAAAATATATTTTTAAAAATATAAAAATATTGTTAATAAAATTA : 596
Tconorhini : TTATATGTGTAAATATTA-AGTATTATTTTAAAAAATATATTTTAAAAATATAAGAATATTGTTAATAAAATTA : 623
Tdionisii : TCA-A-GTAAAAAAGTA-AGTATTATTATTAAATATATATTTCTAAAAATATAAATATATTGTTAATAAATTTG : 606
Terneyi : AAATA-AT-GAAATAGTA-AGTATTATTGTAAAAAATATATTTTTAAAAATATAAATATATTGTTAATAAATTTA : 614
TcruCLBren : TGATATGTTTAAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 608
TcruCL : TGATATGTTTAAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 608
TcruTula : TGATATGTCTAAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 608
TcruM6241 : TGATATGTCTAAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 609
TcruCANIII : TGATATGTCTAAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 612
TcruG : TAA-ACTTC-AAAGTATA-GGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 603
TcruSylvio : TAA-ATTTC-AAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATCGTTAATAAAATTA : 605
TcruJRcl4 : TAA-ACTTC-AAAGCATA-AGTATTATTATTAAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 605
TcruEsmera : TAATGATTTTAAATTATA-AGTATTATTATAGAAAATATAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 609
TcruMarink : TATTAT--CGAAGTCATA-AGTATTATTATAAAAGATGTAATTTTAAAAATATAAATATATTGTTAATAAAATTA : 610
t a a Aaa TA aGTATTATTat aaaaatata TTtTAAAAATATAAatATATtGTTAATAAAaTTa
680 700 720 740
Tbrucei : TCGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGATAATATATTTAAGATTATTCTTGTAAATATATTTAA : 678
Tserpentis : TCGTATTCTAAATGCGTTTATTAAATGCGTTCGTCTAAGAACAATATTTAAGATTATTCTTGTATATATATATAT : 687
Tlewisi : TCGTATTTTAAATGCGTTTATTAAATGCGTTTGTCTAAGAAAAATATTTAAGATTATTCTTGTATATATATTTAT : 685
Trangeli : TTGTATTTTAAGTACGTTTATTAAATGCGCTCGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 671
Tconorhini : TCGTATTTTAAATGCGTTTATTAAATGCGTTTGTCTAAGAAAAATATTTAAGATTATTCTTGTATATATATTTAT : 698
Tdionisii : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 681
Terneyi : TTATATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAAAATATTTAAGATTATTCTTGTATATATTTTTAT : 689
TcruCLBren : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 683
TcruCL : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 683
TcruTula : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 683
TcruM6241 : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 684
TcruCANIII : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 687
TcruG : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGAGTATTCTTGTATATATATTTAT : 678
TcruSylvio : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 680
TcruJRcl4 : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGATTATTCTTGTATATATATTTAT : 680
TcruEsmera : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATATATTTAAGAGTATTCTTGTATATATATTTAT : 684
TcruMarink : TTGTATTTTAAGTGCGTTTATTAAATGCGTTTGTCTAAGAAATGATTTAAGATTATTCTTGTATATATATATAT : 685
T gTATTtTAAgTgCGTTTATTAAATGCGtTtGTCTAAGAaa aTATTTAAGAtTATTCTTGTAtATATaTtTAt
760 780 800 820
Tbrucei : ATATTAATAATTCTTAAAATAAAA--AAATATCCTC-AATTGCAATATTATTGTAGCATAGTAATTTGTTAACTTA : 750
Tserpentis : ATTTTAATAATTCTAAAAACAAAAGTAA--TTCCTC-AAACATAAAGATAAAGTAGCATAGTAATTTGTTAACTTA : 759
Tlewisi : AAATCAATAATTCT-ATTTTAAAAAT-ACAATCCTCAATTTATACTTTTAAAGTAGCATAGTAATTTGTTAACTTA : 758
Trangeli : ATAATAATAATTCTAGAATCAAAAATT-ACAACCTC-AAATTAAAAATTATAGTAGCATAATAATTTGTTAACTTA : 744
Tconorhini : AACTTAATTATCCT-AAAATAAATATTTTATCCTC-AAAAA-AAAATTATAGTAGCATAATAATTTGTTAACTTA : 770
Tdionisii : AAATTAATAATTCT-AAAAACAATAACAACATCCTC-AAAAGATAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 754
Terneyi : AAATTAATAATTCTAAAAATAAAAAAA--TTCCTC-AAATAATAAATTTTAGTAGCATAGTAATTTGTTAATTA : 761
TcruCLBren : AAATTAATAATTCTAAAAATAAAATAAAACTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 757
TcruCL : AAATTAATAATTCTAAAAATAAAATAAAACTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 757
TcruTula : AAATTAATAATTCTAAAAATAAAATAAAACTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 757
TcruM6241 : AAATTAATAATTCTAAAAATAAAATAAAACTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 758
TcruCANIII : AAATTAATAATTCTAAAAATAAAATAAAACTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 761
TcruG : AAATTAATAATTCTAAAAATAAAATAAAGCTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTATTAACTTA : 752
TcruSylvio : AAATTAATAATTCTAAAAATAAAATAAAGCTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 754
TcruJRcl4 : AAATTAATAATTCTAAAAATAAAATAAA-GTTCCTC-GAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 753
TcruEsmera : AAATTAATAATTCTAAAAATAAAATGAAACTTCCTC-AAAAATAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 758
TcruMarink : ATTTTAATAATTCTAAAAATAAAAAATAATATCCTC-AGAAGTAAAATTTTAGTAGCATAGTAATTTGTTAACTTA : 759
A attAATaATtCT aaaataAAa a tCCTC aa aaaatT taGTAGCATAgTAATTTgTTAAcTA
840 860 880 900
Tbrucei : AATATTAAAGTGTTCCATAGAAAATTTTT-AAATTACAACAAATAAAATAAAGTAATGAATTAATATCAAAATTTT : 824
Tserpentis : ATTATTAAAGCGTTCCACAGAAATTTTTA-AAGTTACAACAAAAAACATAAAATGAGAATTAACATTAAAATTTT : 833
Tlewisi : ATTATTAAAGTGTTCCATAGAAATTTTTT-AAATTAAAACAAATTAAATAAATTATGAAATAAAATAAAAATTTT : 832
Trangeli : ATTATTAAAGTGTTCCACAGAAATTTTTTTAAATTAAAACAATTAAAATAAGTAATGAAATAAAATAAAAATTTT : 819
Tconorhini : ATTATTAAAGTGTTCCATAGAAATTTTTT-AAATTAAAACAAATAAAATAAGTTATGAATTAAAATAAAAATTTT : 844
Tdionisii : ATTATTAAAGTGTTCCACAGAAATTTTTT-AAATTAAAACAAAAAAATAAGGT-ATGAATTAAAATAAAAATTTT : 827
Terneyi : ATTATTAAAGTGTTCCATAGAAATTTTTT-AAATTAAAACAAAAAAATAATATATGAATTAAAATAAAAATTTT : 835



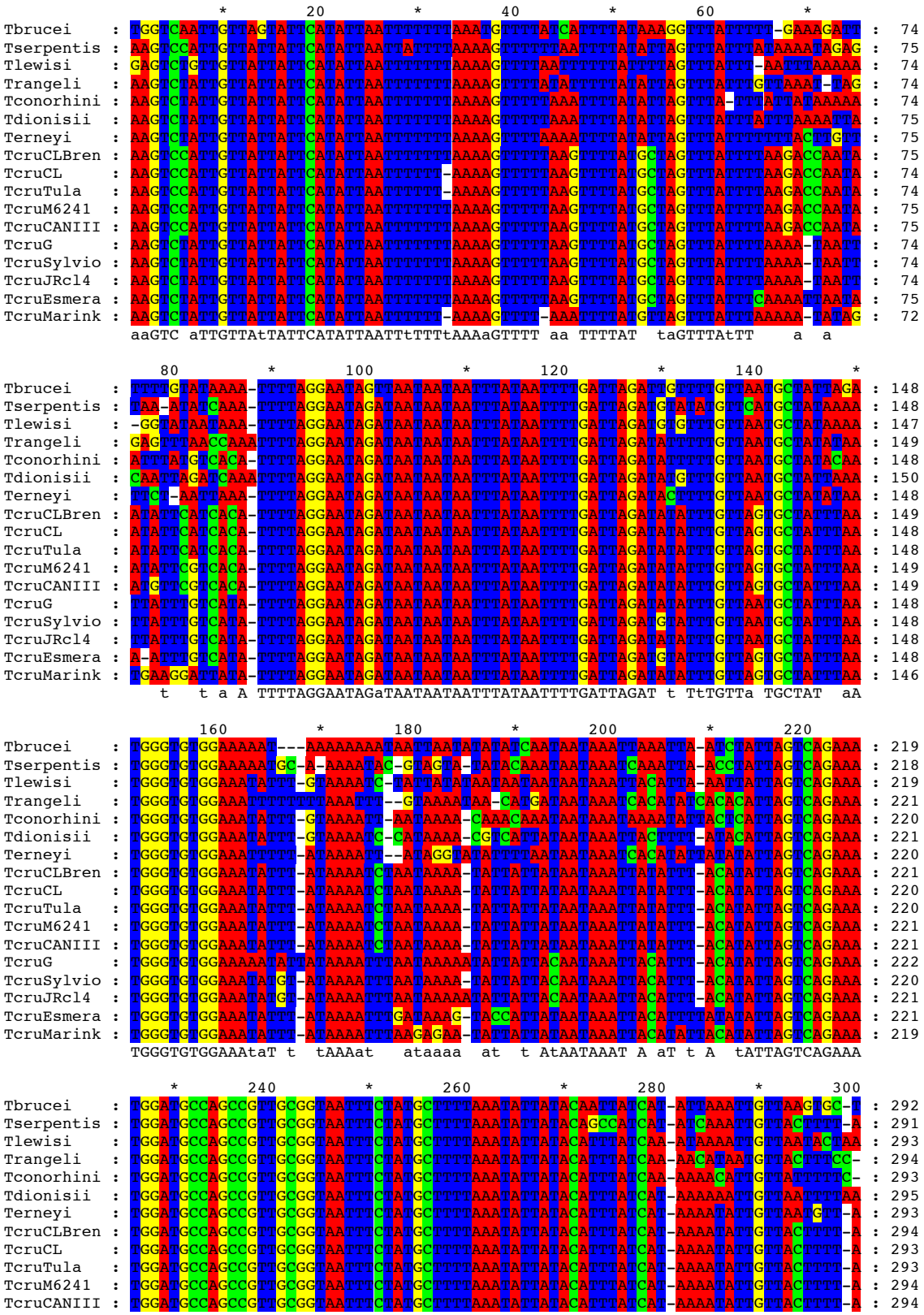


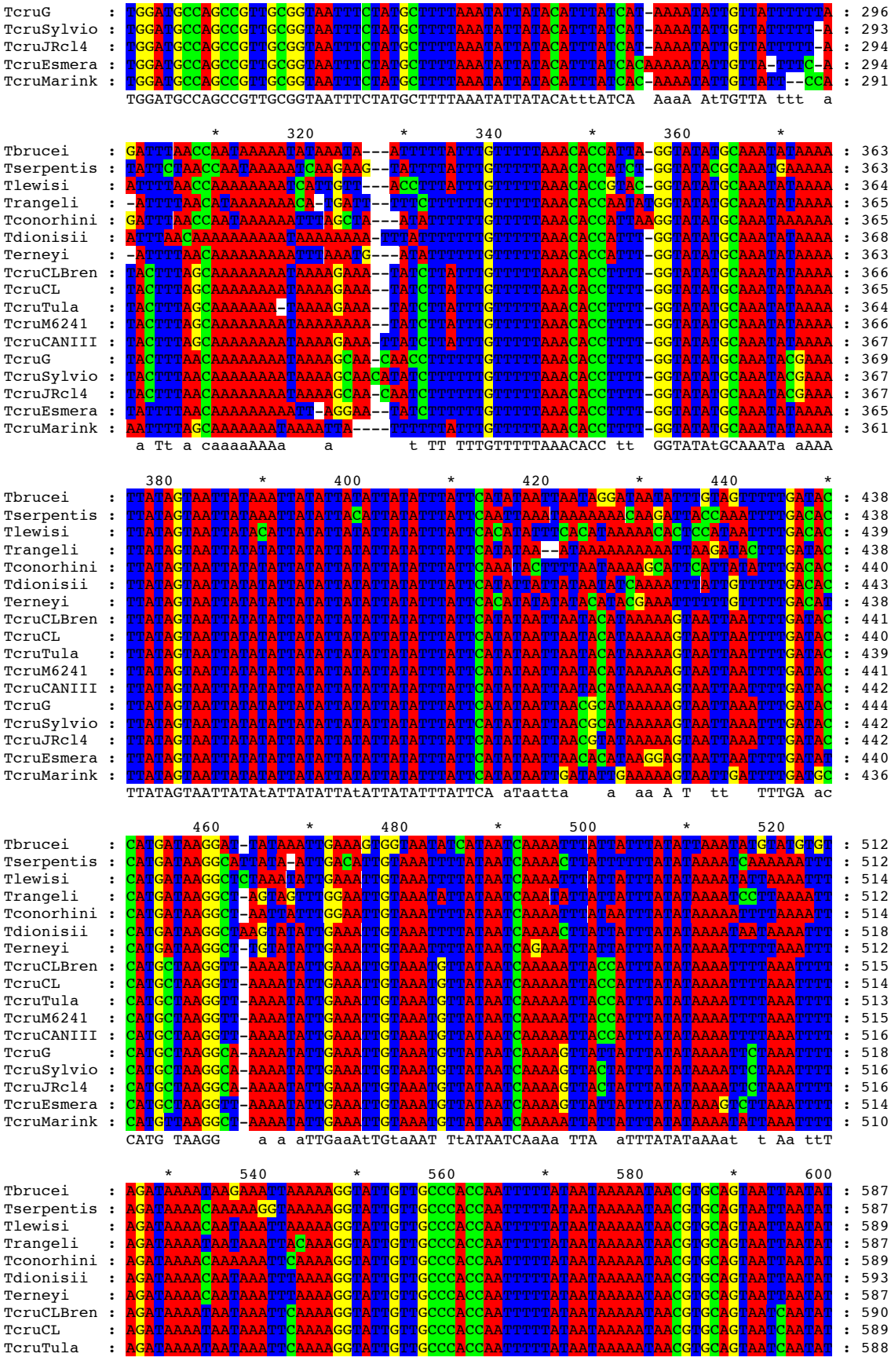
4.SI-S1a: GeneDoc 12S gene alignment: Reds in the alignments are As, yellows are Gs, greens are Cs and blues are Ts. All colored nucleotides are encoded by the maxicircle. Gaps in the alignments are indicated by '-'. Consensus in each column is denoted as stretches of plain alphabets under *Terneyi* (lower case alphabets mean lower confidence).

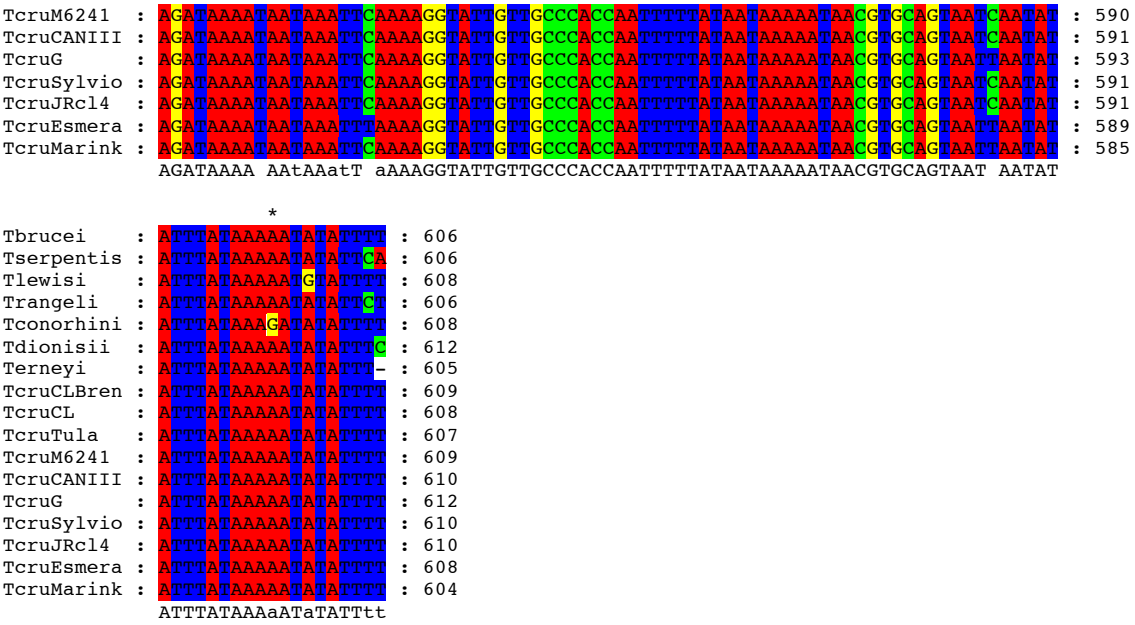
Abbreviations for the names of the organisms in the alignments are as follows: Tbrucei = *T. brucei*; TcruCLBren = *T. cruzi* CL Brener; TcruCL = *T. cruzi* CL; TcruTula = *T. cruzi* Tulahuen; TcruM6241 = *T. cruzi* M6241; TcruCANIII = *T. cruzi* CANIII; TcruG = *T. cruzi* G; TcruSylvio = *T. cruzi* Sylvio X10; TcruJRcl4 = T. cruzi JRcl4; TcruEsmera = *T. cruzi* Esmeraldo; TcruMarink = *T. cruzi marinkellei*; Tconorhini = *T. conorhini*; Tlewisi = *T. lewisi*; Tserpentis = *T. serpentis*; Trangeli = *T. rangeli*; Tdionisii = *T. dionisii*; Terneyi = *T. erneyi* (former *Trypanosoma* sp. 1946).

4.SI-S1b. 12S nucleotide percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	79	**100**															
Tlew	81	81	**100**														
Tran	82	82	86	**100**													
Tcon	82	82	86	90	**100**												
Tdio	81	82	86	87	87	**100**											
Tern	82	82	86	87	88	88	**100**										
TcCLB	81	81	86	87	87	87	88	**100**									
TcCL	81	81	86	87	87	87	88	100	**100**								
TcTula	81	81	86	87	87	87	88	100	100	**100**							
TcM62	81	81	86	87	87	87	88	100	100	100	**100**						
TcCAN	81	81	86	87	87	87	88	99	99	99	100	**100**					
TcG	81	81	85	86	86	87	87	93	93	93	93	93	**100**				
TcSyl	80	81	85	86	86	87	87	92	92	93	93	93	99	**100**			
TcJR	81	81	85	86	87	86	87	93	93	93	93	93	99	99	**100**		
TcEsm	83	81	85	87	87	87	87	93	93	93	93	93	93	92	93	**100**	
TcMar	82	82	84	86	87	87	87	89	89	89	90	90	89	89	90	90	**100**



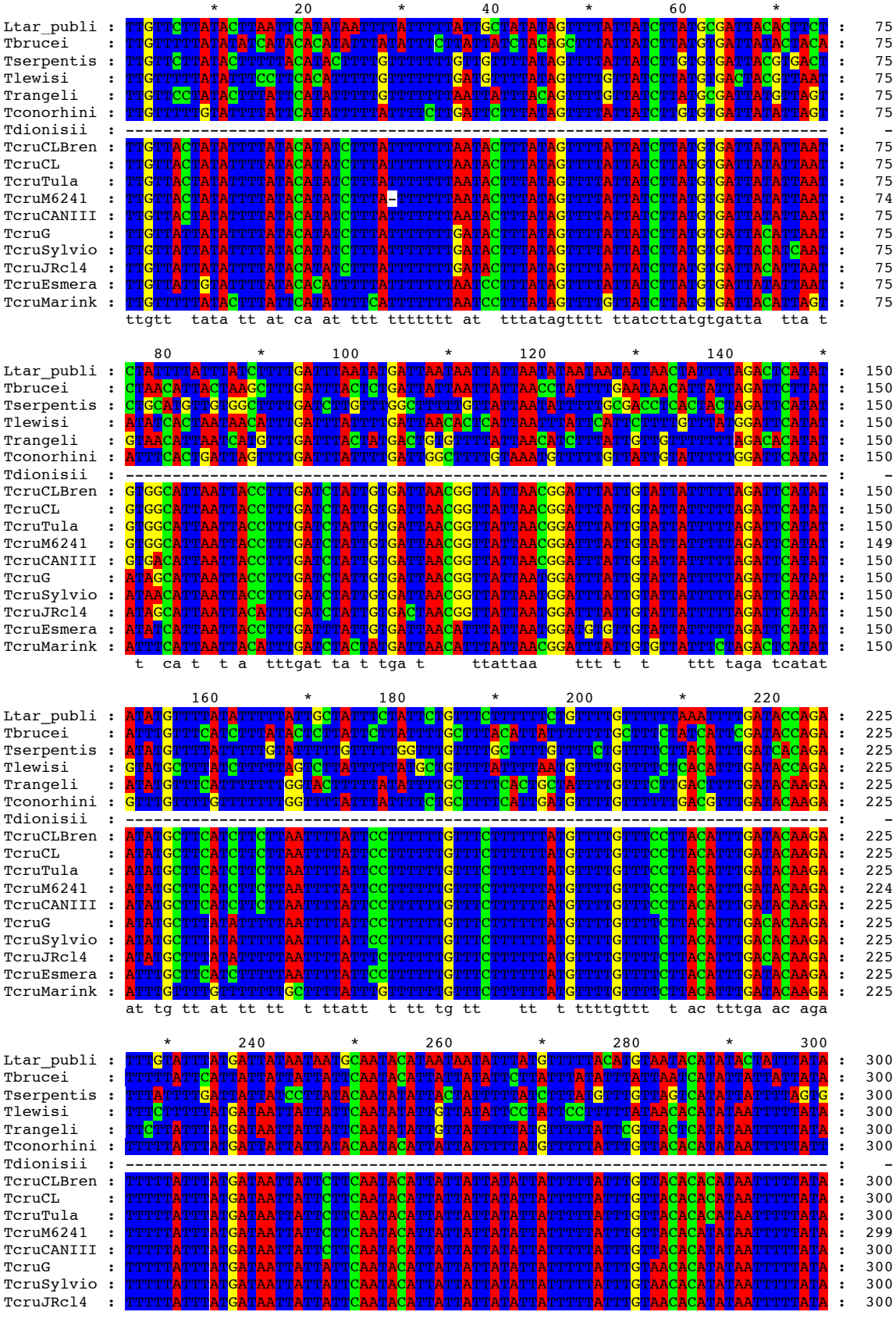


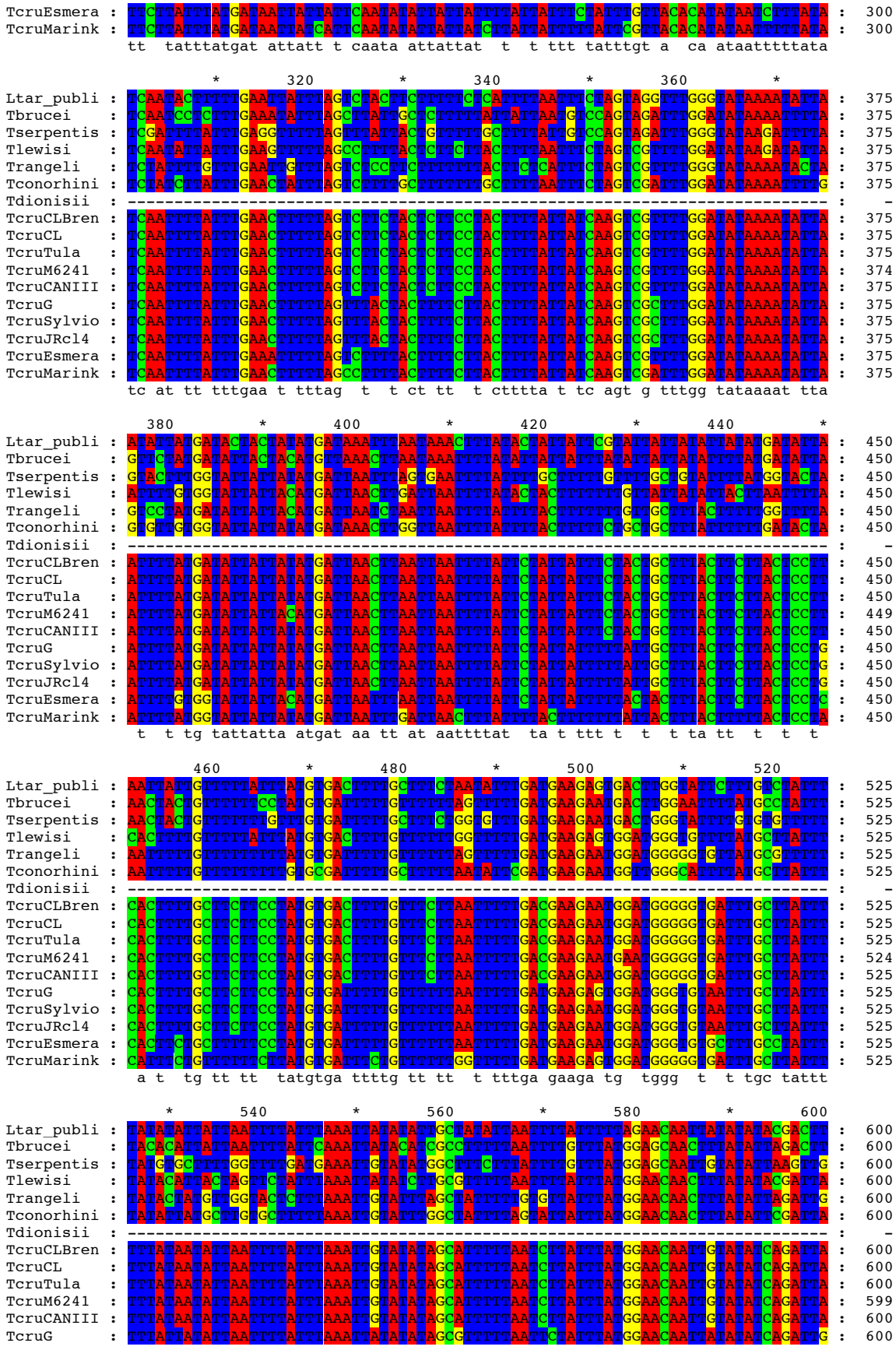


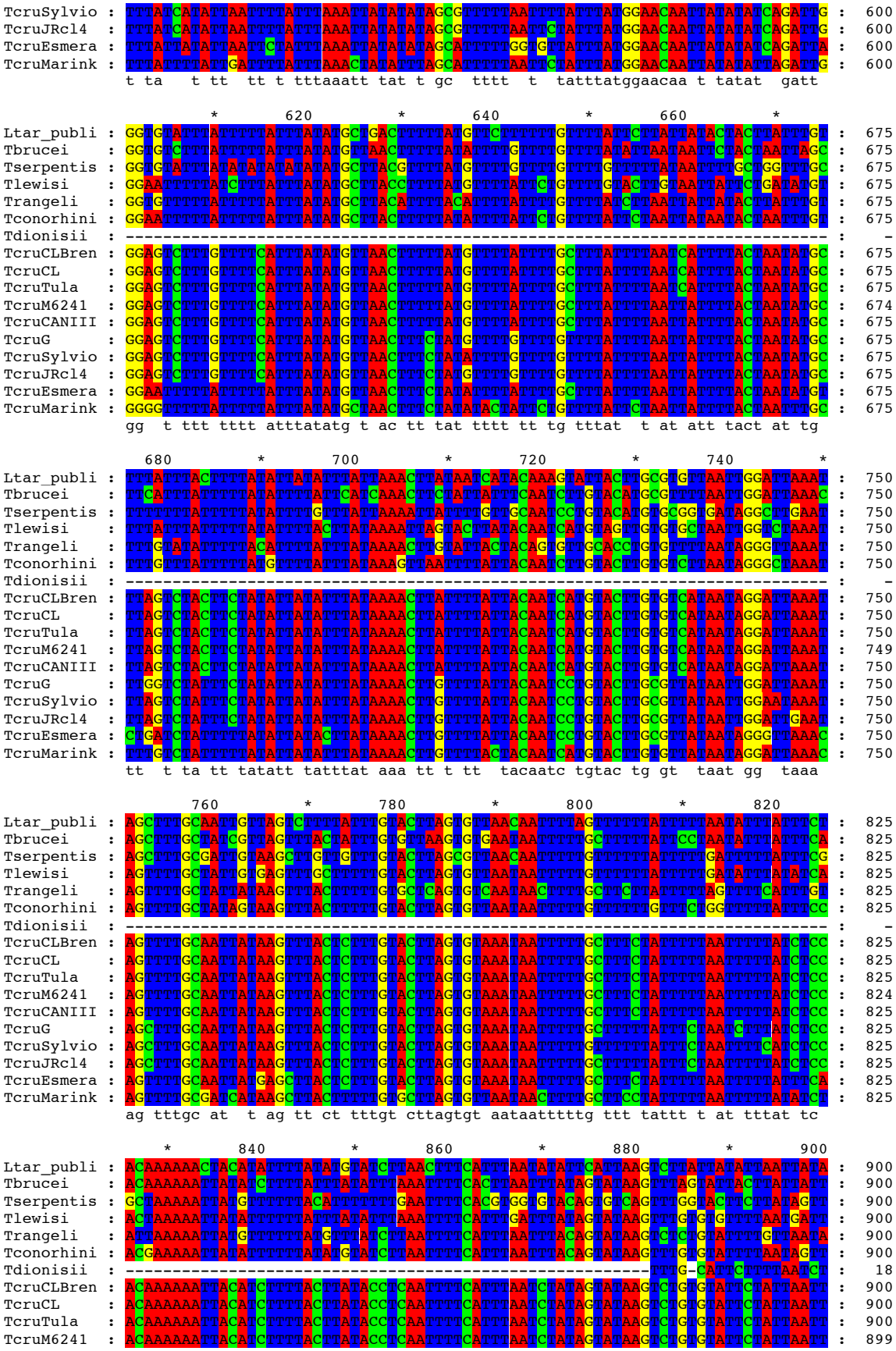
4.SI-S2a: GeneDoc 9S gene alignment: Abbreviations for the names of the organisms, color code, gaps and consensus are as mentioned in S1.

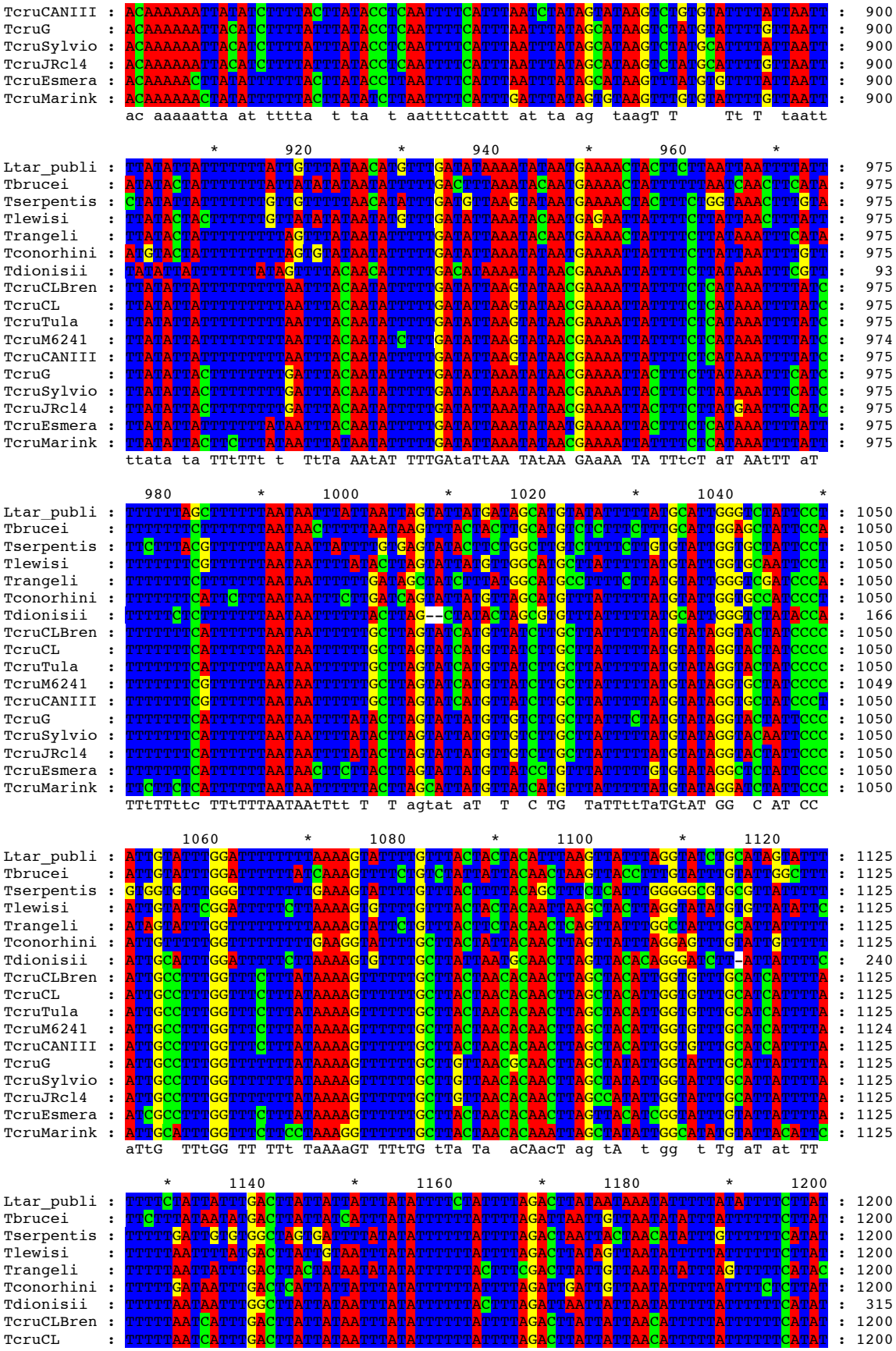
4.SI-S2b. 9S nucleotide percent identity matrix.

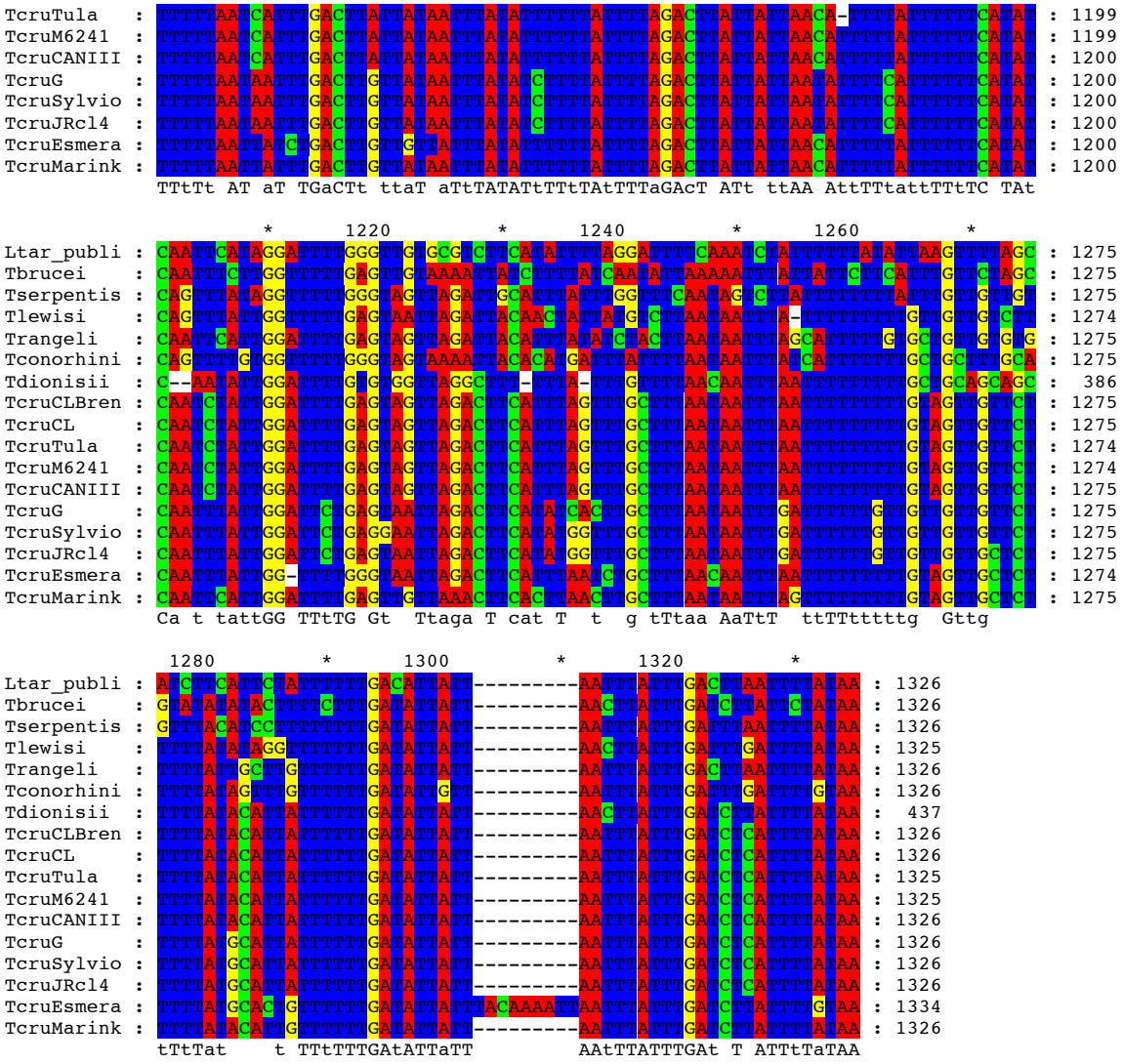
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	82	**100**															
Tlew	83	84	**100**														
Tran	81	84	85	**100**													
Tcon	83	83	87	87	**100**												
Tdio	83	84	87	87	87	**100**											
Tern	84	85	86	87	88	89	**100**										
TcCLB	83	84	86	86	86	88	87	**100**									
TcCL	83	84	86	86	86	88	87	100	**100**								
TcTula	83	84	86	86	86	88	87	100	100	**100**							
TcM62	83	83	86	86	86	88	87	100	100	100	**100**						
TcCAN	83	83	86	86	87	88	87	100	100	100	100	**100**					
TcG	83	84	85	86	87	88	87	95	95	95	95	95	**100**				
TcSyl	82	84	85	86	87	88	87	95	95	95	95	95	99	**100**			
TcJR	82	84	85	86	87	88	87	95	95	95	95	95	99	99	**100**		
TcEsm	82	84	86	87	87	88	88	94	94	94	94	94	94	95	94	**100**	
TcMar	83	84	87	88	86	88	88	93	93	93	93	92	92	92	92	92	**100**



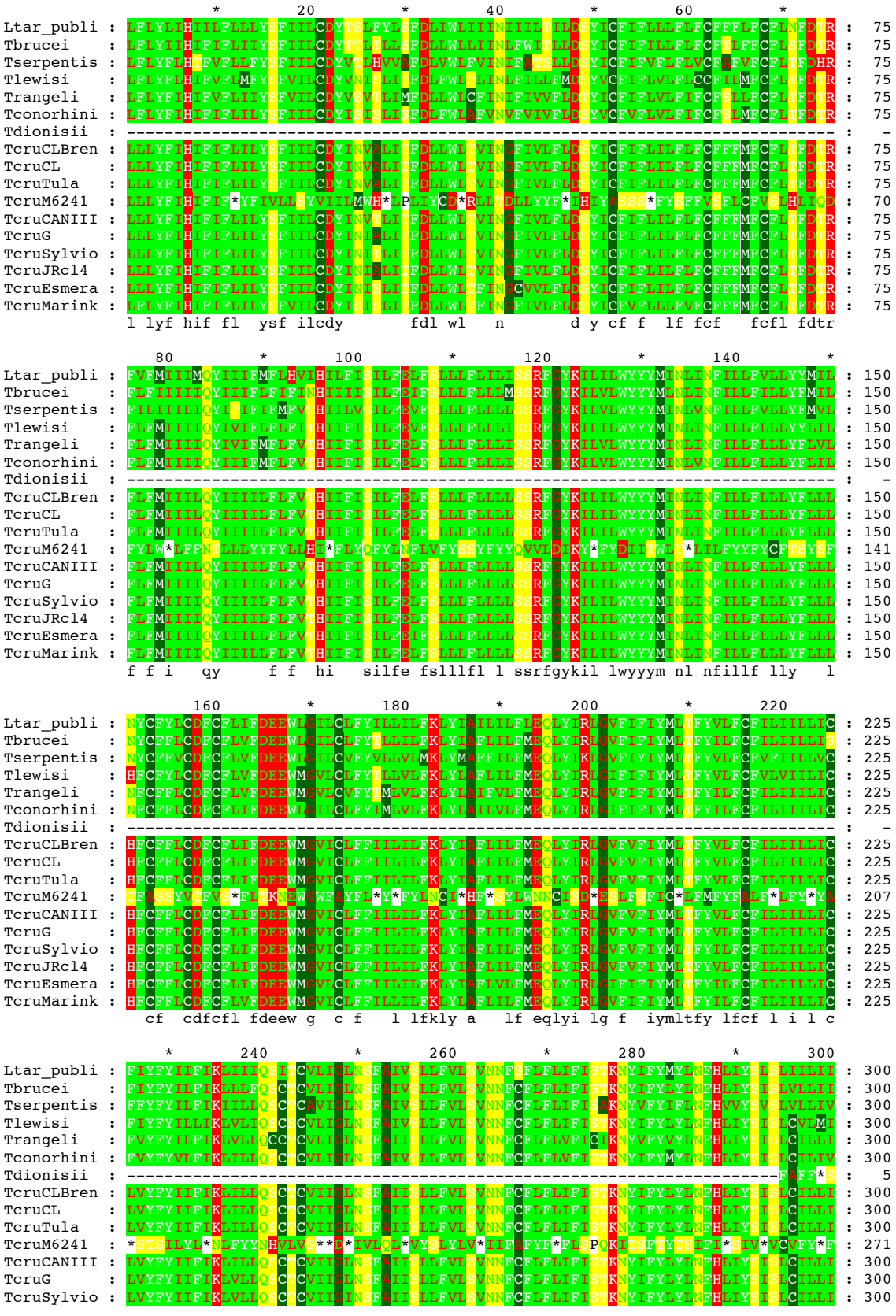


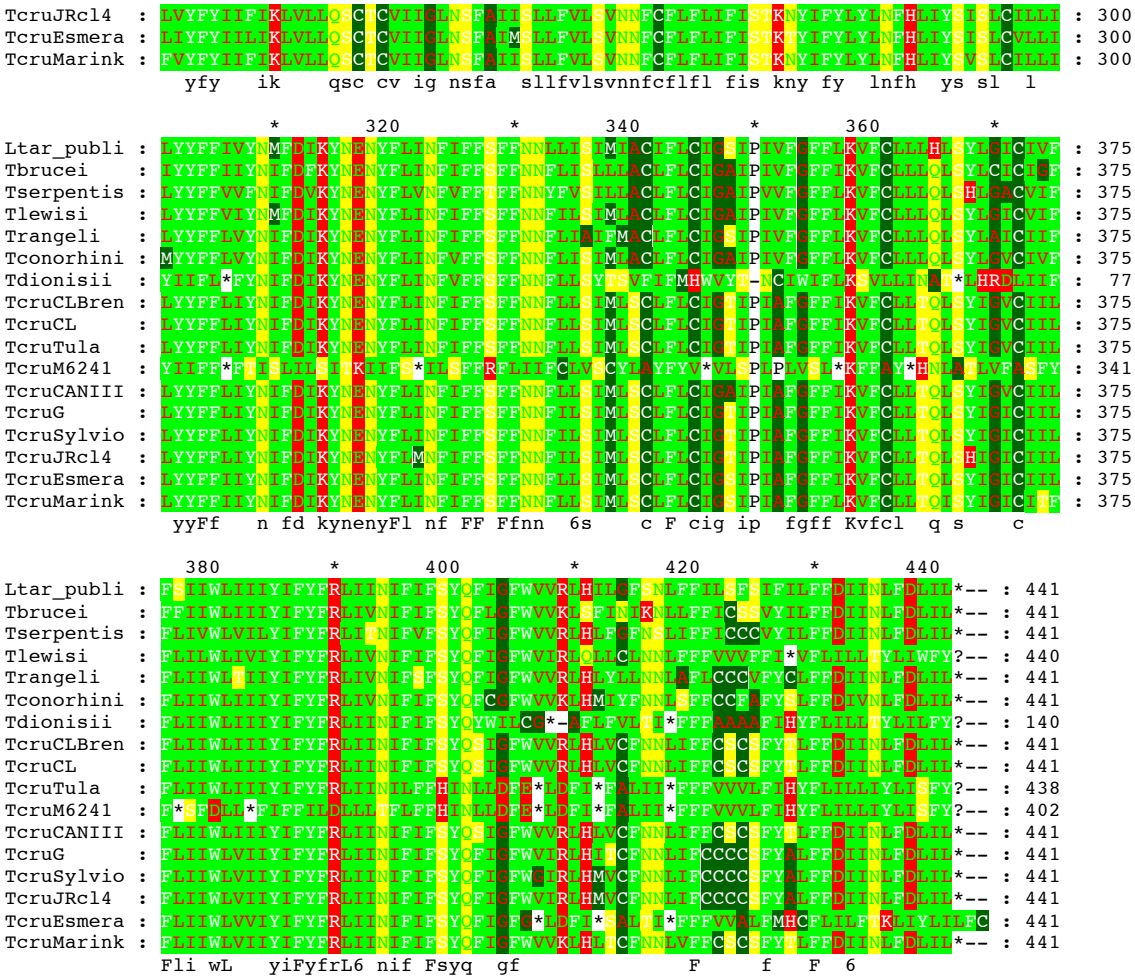

TcruSylvio : TTTATCATATTAATTTTATTTAAATTATATATAGCGTTTTTAATTTTATTTATGGAACAATTATATATCAGATTG : 600
TcruJRcl4 : TTTATCATATTAATTTTATTTAAATTATATATAGCGTTTTTAATTCTATTTATGGAACAATTATATATCAGATTG : 600
TcruEsmera : TTTATTATATTAATTCTATTTAAATTATATATAGCATTTTTGGTGTATTTATGGAACAATTATATATCAGATTA : 600
TcruMarink : TTTATTTTATTGATTTTATTTAAACTATATTTAGCATTTTTAATTCTATTTATGGAACAATTATATATTAGATTG : 600
t ta t tt tt t tttaaatt tat t gc tttt t tatttatggaacaa t tatat gatt
620 640 660
Ltar_publi : GGTGCTATTTATTTTATTTATATGCTGACTTTTTATGTTCTTTTTGTTTTATTCTTATTATACTACTTATTGC : 675
Tbrucei : GGTGCTCTTTATTTTATTTATATGTTAACTTTTTATATTTTGTTTTGTTTTATATTAATAATTCTACTAATTAGC : 675
Tserpentis : GGTGCTATTTATATATATATATGCTTACGTTTTATGTTTTGTTTTGTTTTGTTTTTATAATTTGCTGGTTTGC : 675
Tlewisi : GGAATTTTTATCTTTATTTTATATGCTTACCTTTTATGTTTTATTCTGTTTTGTACTTGTAATTATTCTGATATGT : 675
Trangeli : GGTGTTTTTATTTTATTTATATGCTTACATTTTACATTTTATTTTGTTTTATCTTAATTATTATACTTATTGT : 675
Tconorhini : GGAATTTTTATTTTATTTATATGCTTACTTTTTATATTTTATTCTGTTTTATTCTAATTATAATACTAATTTGT : 675
Tdionisii : ---------- : -
TcruCLBren : GGAGCTCTTGGTTTCATTTATATGTTAACTTTTTATGTTTTATTTTGCTTTATTTTAATCATTTTACTAATATGC : 675
TcruCL : GGAGCTCTTGGTTTCATTTATATGTTAACTTTTTATGTTTTATTTTGCTTTATTTTAATCATTTTACTAATATGC : 675
TcruTula : GGAGCTCTTGGTTTCATTTATATGTTAACTTTTTATGTTTTATTTTGCTTTATTTTAATCATTTTACTAATATGC : 675
TcruM6241 : GGAGCTCTTGGTTTCATTTATATGTTAACTTTTTATGTTTTATTTTGCTTTATTTTAATTATTTTACTAATATGC : 674
TcruCANIII : GGAGCTCTTGGTTTCATTTATATGTTAACTTTTTATGTTTTATTTTGCTTTATTTTAATTATTTTACTAATATGC : 675
TcruG : GGAGCTCTTGGTTTCATTTATATGTTAACTTTCTATGTTTTGTTTTGTTTTATTTTAATTATTTTACTAATATGC : 675
TcruSylvio : GGAGCTCTTGGTTTCATTTATATGTTAACTTTCTATGTTTTGTTTTGTTTTATTTTAATTATTTTACTAATATGC : 675
TcruJRcl4 : GGAGCTCTTGGTTTCATTTATATGTTAACTTTCTATGTTTTGTTTTGTTTTATTTTAATTATTTTACTAATATGC : 675
TcruEsmera : GGAATTTTTATTTTATTTATATGTTAACTTTCTATATTTTATTTTGCTTTATTTTAATTATTTTACTAATATGT : 675
TcruMarink : GGGGTTTTTATTTTATTTATATGCTAACTTTCTATATACTATTCTGTTTTATTCTAATTATTTTACTAATTTGC : 675
gg t ttt tttt atttatatg t ac tt tat tttt tt tg tttat t at att tact at tg
680 700 720 740
Ltar_publi : TTTATTTACTTTTATATTATATTTATTAAACTTATAATCATACAAAGTATTACTTGCGTGTTAATTGGATTAAAT : 750
Tbrucei : TTCATTTATTTTTATATTTATTCATCAAACTTCTATTATTTCAATCTTGTACATGCGTTTAATTGGATTAAAC : 750
Tserpentis : TTTTTTTATTTTTATATTTGTTTATTAAAATTATTTTGTTGCAATCCTGTACATGTGCGGTGATAGGCTTGAAT : 750
Tlewisi : TTTATTTATTTTTATATTTACTTATAAAATTAGTACTTATACAATCATGTAGTTGTGTGCTAATTGGTCTAAAT : 750
Trangeli : TTTGTATATTTTTACATTTATTTATAAAACTTGTATTACTACAGTGTTGCACCTGTGTTTTAATAGGGTTAAAT : 750
Tconorhini : TTTGTTTATTTTTATGTTTTATTTATAAAGTTAATTTTATTACAATCTTGTACTTGTGTCTTAATAGGGCTAAAT : 750
Tdionisii : ---------- : -
TcruCLBren : TTAGTCTACTTCTATATTATATTTATTAAAACTTATTTTATTACAATCATGTACTTGTGTCATAATAGGATTAAAT : 750
TcruCL : TTAGTCTACTTCTATATTATATTTATTAAAACTTATTTTATTACAATCATGTACTTGTGTCATAATAGGATTAAAT : 750
TcruTula : TTAGTCTACTTCTATATTATATTTATTAAAACTTATTTTATTACAATCATGTACTTGTGTCATAATAGGATTAAAT : 750
TcruM6241 : TTAGTCTACTTCTATATTATATTTATTAAAACTTATTTTATTACAATCATGTACTTGTGTCATAATAGGATTAAAT : 749
TcruCANIII : TTAGTCTACTTCTATATTATATTTATTAAAACTTATTTTATTACAATCATGTACTTGTGTCATAATAGGATTAAAT : 750
TcruG : TTGGTCTATTTCTATATTATATTTATTAAAACTTGTTTTATTACAATCCTGTACTTGCGTTATAATTGGATTAAAT : 750
TcruSylvio : TTAGTCTATTTCTATATTATATTTATTAAAACTTGTTTTATTACAATCCTGTACTTGCGTTATAATTGGAATAAAT : 750
TcruJRcl4 : TTAGTCTATTTCTATATTATATTTATTAAAACTTGTTTTATTACAATCCTGTACTTGCGTTATAATTGGATTGAAT : 750
TcruEsmera : CTGATCTATTTTTATATTATACTTATTAAAACTTGTTTTATTACAATCCTGTACTTGCGTTATAATAGGGTTAAAC : 750
TcruMarink : TTTGTCTATTTTTATATTATATTTATTAAAACTTGTTTTACTACAATCATGTACTTGTGTTATAATAGGATTAAAC : 750
tt t ta tt tatatt tatttat aaa tt t tt tacaatc tgtac tg gt taat gg taaa
760 780 800 820
Ltar_publi : AGCTTTGCAAATGTTAGCTTTTATTTGTACTTAGTGTTAACAATTTTAGTTTTTATTTTTAATATTTATTCT : 825
Tbrucei : AGCTTTGCTATCGTTAGTTTACTATTTGTGTTAAGTGGAATAATTTTTGCTTTTTATTCTAATATTTATTCA : 825
Tserpentis : AGCTTTGCGATGTAAGCTTGTTGTTTGTACTTAGCGTTAACAATTTTTGTTTTTTATTTTGATTTTATTTCG : 825
Tlewisi : AGTTTTGCTATTGTGAGTTTGCTTTTTGTACTTAGTGTTAATAATTTTTGTTTTTTATTTTGATATTTATATCA : 825
Trangeli : AGTTTTGCTATTATAAGTTTACTTTTTGTGCTCAGTGTCAATAACTTTTGCTTCTTATTTTTAGTTTTCATTTGT : 825
Tconorhini : AGTTTTGCTATAGTAAGTTTACTTTTTGTACTTAGTGTTAATAATTTTTGTTTTTTGTTTCTGGTTTTTATTCCC : 825
Tdionisii : ---------- : -
TcruCLBren : AGTTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATCTCC : 825
TcruCL : AGTTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATCTCC : 825
TcruTula : AGTTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATCTCC : 825
TcruM6241 : AGTTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATCTCC : 824
TcruCANIII : AGTTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATCTCC : 825
TcruG : AGCTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTTTATTTCTAATCTTTATCTCC : 825
TcruSylvio : AGCTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGTTTTTTATTTCTAATTTTCATCTCC : 825
TcruJRcl4 : AGCTTTGCAATTATAAGTTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTTTATTTCTAATTTTTATCTCC : 825
TcruEsmera : AGTTTTGCAATTATGAGCTTACTCTTTGTACTTAGTGTAAATAATTTTTGCTTTCTATTTTTAATTTTTATTCA : 825
TcruMarink : AGTTTTGCGATCATAAGCTTACTTTTTGTGCTTAGTGTTAATAACTTTTGCTTCCTATTTTTAATTTTTATATCT : 825
ag tttgc at t ag tt ct tttgt cttagtgt aataatttttg ttt tattt t at tttat tc
840 860 880 900
Ltar_publi : ACAAAAAACTACATATTTTATATGTATCTTTAACTTTCATTTAATATATTCATTAAGTCTTATTATATTAATTATA : 900
Tbrucei : ACAAAAAATTATATCTTTTATTTATATTTAAATTTTCACTTAATTTATAGTATAAGTTTAGTATTACTTATTATT : 900
Tserpentis : GCTAAAAATTATGTTTTTTACATTTTTTGAATTTTCACGTGGTGTACAGTGTCAGTTTGGTACTTCTTATAGTT : 900
Tlewisi : ACTAAAAATTATATTTTTTATTTATATTTAAATTTTCATTTGATTTATAGTATAAGTTTGTGTGTTTTAATGATT : 900
Trangeli : ATTAAAAATTATGTTTTTTATGTTTATCTTAATTTCATTTAATTTACAGTATAAGTCTCTGTATTTTGTTAATA : 900
Tconorhini : ACGAAAAATTATATTTTTTATAGTATCTTAATTTTCATTTAATTTACAGTATAAGTTTGTGTATTTTAATAGTT : 900
Tdionisii : ---------- : 18
TcruCLBren : ACAAAAAATTACATCTTTTACTTATACCTCAATTTTCATTTAATCTATAGTATAAGTCTCGTGTATTCTATTAATT : 900
TcruCL : ACAAAAAATTACATCTTTTACTTATACCTCAATTTTCATTTAATCTATAGTATAAGTCTCGTGTATTCTATTAATT : 900
TcruTula : ACAAAAAATTACATCTTTTACTTATACCTCAATTTTCATTTAATCTATAGTATAAGTCTCGTGTATTCTATTAATT : 900
TcruM6241 : ACAAAAAATTACATCTTTTACTTATACCTCAATTTTCATTTAATCTATAGTATAAGTCTCGTGTATTCTATTAATT : 899


TcruCANIII : ACAAAAAATTATATCTTTTACTTATACCTCAATTTTCATTTAATCTATAGTATAAGTCTGTGTATTTTATTAATT : 900
TcruG : ACAAAAAATTACATCTTTTATTTATACCTCAATTTTCATTTAATTTATAGCATAAGTCTATGTATTTTGTTAATT : 900
TcruSylvio : ACAAAAAATTACATCTTTTATTTATACCTCAATTTTCATTTAATTTATAGCATAAGTCTATGCATTTTATTAATT : 900
TcruJRcl4 : ACAAAAAATTACATCTTTTATTTATACCTCAATTTTCATTTAATTTATAGCATAAGTCTATGCATTTTGTTAATT : 900
TcruEsmera : ACAAAAACTTATATTTTTTACTTATACCTTAATTTTCATTTAATTTATAGCATAAGTTTATGTGTTTTATTAATT : 900
TcruMarink : ACAAAAAACTATATTTTTTACTTATACTTAATTTTCATTTGATTTATAGTGTAAGTTTGTGTATTTTGTTAATT : 900
ac aaaaatta at tttta t ta t aattttcattt at ta ag taagT T Tt T taatt
920 940 960
Ltar_publi : TTATATTATTTTTTTATTGTTTTATAACATGTTTGATATAAAATATAATGAAAACTACTTCTTAATTAATTTTATT : 975
Tbrucei : ATATACTATTTTTTTATTATATAATAATTTTTGACTTTAAATACAATGAAAACTATTTTTAATCAACTTCATA : 975
Tserpentis : CTATATTATTTTTTTGTTGTTTTTAACATATTTGATGTTAAGTATAATGAAAACTACTTTCTGGTAAACTTTGTA : 975
Tlewisi : TTATACTACTTTTTTGTTATATATAATATGTTTGATATTAAATACAATGAGAATTATTTTCTTATTAACTTTATT : 975
Trangeli : TTATACTATTTTTTTTTAGTTTATAATATTTTTGATATTAAATACAATGAAAACTATTTTCTTATAAATTTCATA : 975
Tconorhini : ATGTACTATTTTTTTTTAGTGTATAATATTTTTGATATTAAATATAATGAAAATTATTTTCTTATTAATTTTGTT : 975
Tdionisii : TATATTATTTTTTTATAGTTTTACAACATTTTTGACATAAAATATAACGAAAATTATTTTCTTATAAATTTCGTT : 93
TcruCLBren : TTATATTATTTTTTTTTTAATTTACAATATTTTTGATATTAAGTATAACGAAAATTATTTTCTCATAAATTTTATC : 975
TcruCL : TTATATTATTTTTTTTTTAATTTACAATATTTTTGATATTAAGTATAACGAAAATTATTTTCTCATAAATTTTATC : 975
TcruTula : TTATATTATTTTTTTTTTAATTTACAATATTTTTGATATTAAGTATAACGAAAATTATTTTCTCATAAATTTTATC : 975
TcruM6241 : TTATATTATTTTTTTTTTAATTTACAATATTTTTGATATTAAGTATAACGAAAATTATTTTCTCATAAATTTTATC : 974
TcruCANIII : TTATATTATTTTTTTTTTAATTTACAATATTTTTGATATTAAGTATAACGAAAATTATTTTCTCATAAATTTTATC : 975
TcruG : TTATATTACTTTTTTTTGATTTACAATATTTTTGATATTAAATATAACGAAAATTACTTTCTTATAAATTTCATC : 975
TcruSylvio : TTATATTACTTTTTTTTGATTTACAATATTTTTGATATTAAATATAACGAAAATTACTTTCTTATAAATTTCATC : 975
TcruJRcl4 : TTATATTACTTTTTTTTGATTTACAATATTTTTGATATTAAATATAACGAAAATTACTTTCTTATAAATTTCATC : 975
TcruEsmera : TTATATTATTTTTTTTATAATTTACAATATTTTTGATATTAAATATAATGAAAATTACTTTCTCATAAATTTTATT : 975
TcruMarink : TTATATTACTTCTTTTATAATTTATAATATTTTTGATATTAAATATAACGAAAATTATTTTCTCATAAATTTTATT : 975
ttata ta TTtTTt t TtTa AAtAT TTTGAtaTtAA TAtAA GAaAA TA TTtcT aT AAtTT aT
980 1000 1020 1040
Ltar_publi : TTTTTTAGCTTTTTTAATAATTTATTAATTAGTATTATGATAGCATGTATATTTTTATGCATTGGGACTATTCCT : 1050
Tbrucei : TTTTTTTCTTTTTTTAATAACTTTTTAATAAGTTTACTACTTGCATGTCTCTTTCTTTGCATTGGAGCTATTCCA : 1050
Tserpentis : TTCTTTACGTTTTTTAATAATTATTTTGTGAGTATACTTCTGGCTTGTCTTTTCTTGTGTATTGGTGCTATTCCT : 1050
Tlewisi : TTTTTTTCGTTTTTTAATAATTTTATACTTAGTATTATGTTGGCATGCTTATTTTTATGTATTGGTGCAATTCCT : 1050
Trangeli : TTTTTTTCTTTTTTTAATAATTTTTGATAGCTATCTTTATGGCATGCCTTTTCTTATGTATTGGTGCGATCCCA : 1050
Tconorhini : TTTTTTTCATTCTTTAATAATTTCTTGATCAGTATTATGTTAGCATGTTTATTTTTATGTATTGGTGCCATCCCT : 1050
Tdionisii : TTTTTCTCTTTTTTTAATAATTTTTTACTTAG---CTATACTAGCGTGTTTATTTTTATGCATTGGCTCTATACCA : 166
TcruCLBren : TTTTTTTCATTTTTTAATAATTTTTTGCTTAGTATCATGTTATCTTGCTTATTTTTATGTATAGGTACTATCCCC : 1050
TcruCL : TTTTTTTCATTTTTTAATAATTTTTTGCTTAGTATCATGTTATCTTGCTTATTTTTATGTATAGGTACTATCCCC : 1050
TcruTula : TTTTTTTCATTTTTTAATAATTTTTTGCTTAGTATCATGTTATCTTGCTTATTTTTATGTATAGGTACTATCCCC : 1050
TcruM6241 : TTTTTTTCGTTTTTTAATAATTTTTTGCTTAGTATCATGTTATCTTGCTTATTTTTATGTATAGGTGCTATCCCC : 1049
TcruCANIII : TTTTTTTCGTTTTTTAATAATTTTTTGCTTAGTATCATGTTATCTTGCTTATTTTTATGTATAGGTGCTATCCCT : 1050
TcruG : TTTTTTTCATTTTTTAATAATTTTATACTTAGTATTATGTTGTCTTGCTTATTTCTATGTATAGGTACTATTCCC : 1050
TcruSylvio : TTTTTTTCATTTTTTAATAATTTTATACTTAGTATTATGTTGTCTTGCTTATTTTTATGTATAGGTACAATTCCC : 1050
TcruJRcl4 : TTTTTTTCATTTTTTAATAATTTTTATACTTAGTATTATGTTGTCTTGCTTATTTTTATGTATAGGTACTATTCCC : 1050
TcruEsmera : TTTTTTTCATTTTTTAATAACTTCTTACTTAGTATTATGTTATCCTGTTTATTTTTGTGTATAGGCTCTATTCCC : 1050
TcruMarink : TTCTTCTCATTTTTTAATAATTTTTTACTTAGCATTATGTTATCATGTTTATTTTTATGTATAGGATCTATTCCC : 1050
TTtTTttc TTtTTTAATAAtTtt T T agtat aT T C TG TaTTttTaTGtAT GG C AT CC
1060 1080 1100 1120
Ltar_publi : ATTGTATTTGGATTTTTTTTAAAAGTATTTTGTTTACTACTACATTTAAGTTATTTAGGTATCTGCATAGTATTT : 1125
Tbrucei : ATTGTATTTGGATTTTTTATCAAAGTTTTCTGTCTATTATTACAACTAAGTTACCTTTGTATTTGTATTGGCTTT : 1125
Tserpentis : GTGGTGTTTGGGTTTTTTTTGAAAGTATTTTGTTTACTTTTACAGCTTTCTCATTTGGGGGCGTGGCGTTATTTTT : 1125
Tlewisi : ATTGTATTCGGATTTTCTTTAAAAGTGTTTTGTTTACTACTACAATTAAGCTACTTAGGTATATGTGTTATATTC : 1125
Trangeli : ATAGTATTTGGTTTTTTTTTAAAAGTATTCTGTTTACTTCTACAACTCAGTTATTTGGCTATTGCATTATTTTT : 1125
Tconorhini : ATTGTTTTTGGTTTTTTTTTGAAGGTATTTTGCTTACTATTACAACTTAGTTATTTAGGAGTTTGTATTGTTTTT : 1125
Tdionisii : ATTGCATTTGGATTTTTCTTAAAAGTGTTTTGCTTATTAATGCAACTTAGTTACACAGGGATCTT--ATTATTTTC : 240
TcruCLBren : ATTGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGCTACATTGGTGTTTGCATCATTTTA : 1125
TcruCL : ATTGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGCTACATTGGTGTTTGCATCATTTTA : 1125
TcruTula : ATTGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGCTACATTGGTGTTTGCATCATTTTA : 1125
TcruM6241 : ATTGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGCTACATTGGTGTTTGCATCATTTTA : 1124
TcruCANIII : ATTGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGCTACATTGGTGTTTGCATCATTTTA : 1125
TcruG : ATTGCCTTTGGTTTTTTTATAAAAGTTTTTTGCTTGTTAACGCAACTTAGCTATATTGGTGATTTGCATTATTTTA : 1125
TcruSylvio : ATTGCCTTTGGTTTTTTTATAAAAGTTTTTTGCTTGTTAACACAACTTAGCTATATTGGTGATTTGCATTATTTTA : 1125
TcruJRcl4 : ATTGCCTTTGGTTTTTTTATAAAAGTTTTTTGCTTGTTAACACAACTTAGCCATATTGGTGATTTGCATTATTTTA : 1125
TcruEsmera : ATCGCCTTTGGTTTCTTTATAAAAGTTTTTTGCTTACTAACACAACTTAGTTACATCGGTGATTTGTATTATTTTA : 1125
TcruMarink : ATTGCATTTGGTTTCTTCCTAAAAGGTTTTTTGCTTACTAACACAAATTAGCTATATTGGCATATGTATTACATTC : 1125
aTtG TTtGG TT TTt TaAAaGT TTtTG tTa Ta aCAacT ag tA t gg t Tg aT at TT
1140 1160 1180 1200
Ltar_publi : TTTTCTATTATTTGACTTATTATTATTTATATTTTCTATTTTAGACTTATAATAAATATTTTTATATTTTCTTAT : 1200
Tbrucei : TTCTTTATAATATGACTTATTATCATTTATATTTTTTATTTTAGATTAATTGTTAATATATTTATTTTTCTTAT : 1200
Tserpentis : TTTTTGATTGTGGTGGCTAGTGATTTTATATATTTTTTATTTTAGACTAATTACTAACATATTTGTTTTTCATAT : 1200
Tlewisi : TTTTTAATTTTATGACTTATTGTAATTTATATTTTTTATTTTAGACTTATAGTTAATATTTTTATTTTTCTTAT : 1200
Trangeli : TTTTTAATTATTTGACTTACTATAATATATATTTTTTACTTTCGACTTATTGTTAATATATTTAGTTTTCATAAC : 1200
Tconorhini : TTTTTGATAATTTGACTCATTATTATTTATATTTTTTATTTTAGATTGATTGTTAATATTTTTATTTTCTCTTAT : 1200
Tdionisii : TTTTTAATAATTTGGCTTATTATAATTTATATTTTTTACTTTAGATTAATTATTAATATTTTTATTTTTCATAT : 315
TcruCLBren : TTTTTAATCATTTGACTTATTATAATTTATATTTTTTATTTTAGACTTATTATTAACATTTTTATTTTTCATAT : 1200
TcruCL : TTTTTAATCATTTGACTTATTATAATTTATATTTTTTATTTTAGACTTATTATTAACATTTTTATTTTTCATAT : 1200



4.SI-S3a: GeneDoc MURF1 / ND2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. *L.tarentolae* MURF1 sequence has been used as a reference (Ltar_publi).





4.SI-S3b: GeneDoc MURF1 / ND2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. This alignment accounts for the frameshifts caused by the indels [shown in the corresponding nucleotide sequence alignment above (S3a)] and shows the resultant truncated proteins in the altered reading frame. '*' indicates stop codon in the alignment. Several stop codons are seen in the altered frame of TcruTula, TcruM6241, TcruEsmera, Tlewisi and Tdionisii. The reference has also been translated.

4.SI-S3c. MURF1 nucleotide percent identity matrix.

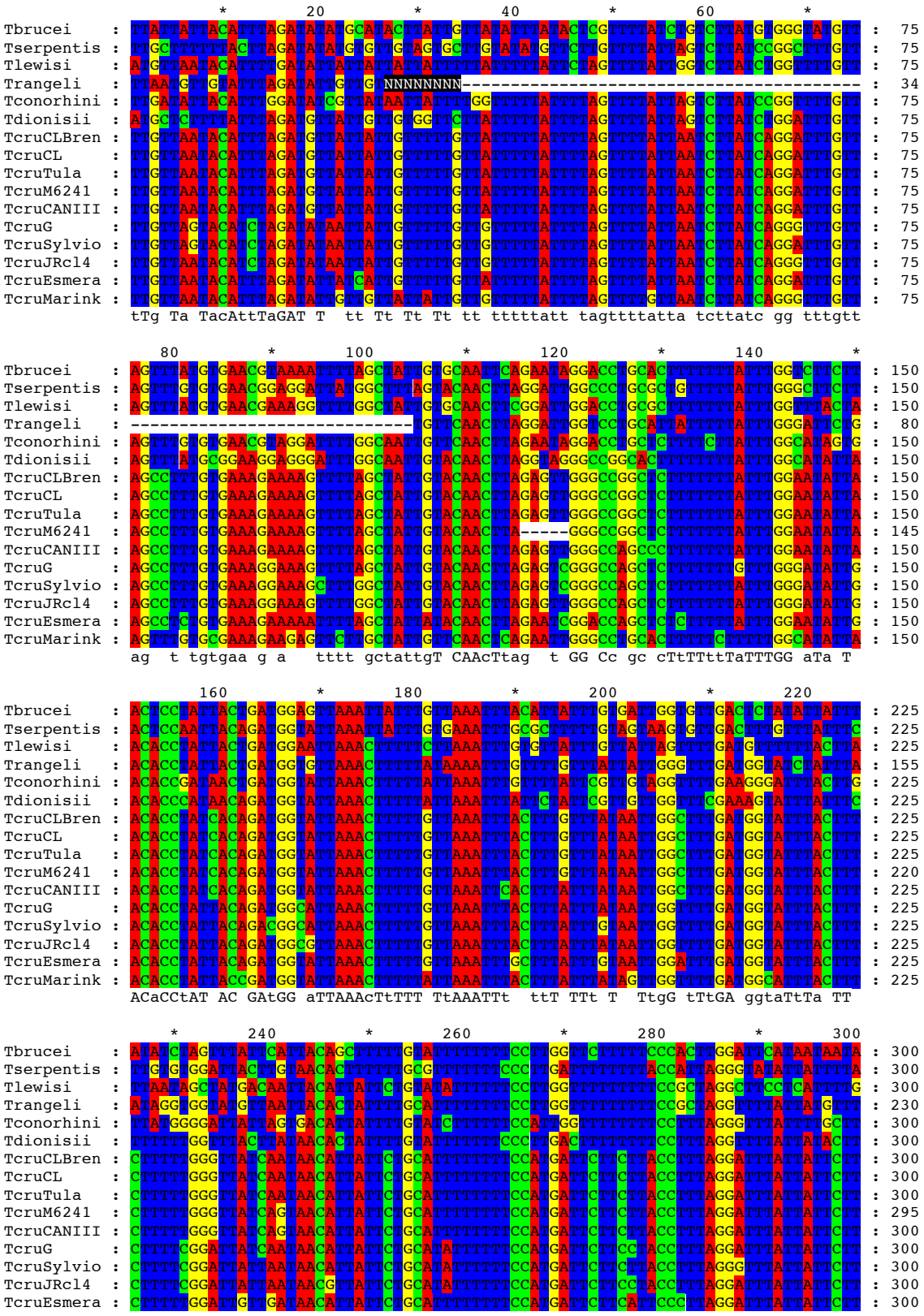
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	72	**100**														
Tlew	77	73	**100**													
Tran	75	73	78	**100**												
Tcon	75	74	81	81	**100**											
Tdio*	73	69	76	73	75	**100**										
TcCLB	75	72	80	79	79	80	**100**									
TcCL	75	82	80	79	79	80	100	**100**								
TcTula	75	72	80	79	79	80	100	100	**100**							
TcM62	75	72	80	79	79	80	99	99	99	**100**						
TcCAN	75	73	80	79	80	81	99	99	99	99	**100**					
TcG	76	72	81	78	79	79	93	93	93	93	93	**100**				
TcSyl	76	72	81	78	79	79	93	93	93	93	93	98	**100**			
TcJR	76	73	80	78	79	79	93	93	93	93	93	99	99	**100**		
TcEsm	77	72	81	79	80	79	91	91	91	91	91	90	90	90	**100**	
TcMar	75	72	81	80	81	80	88	88	88	87	88	88	87	87	89	**100**

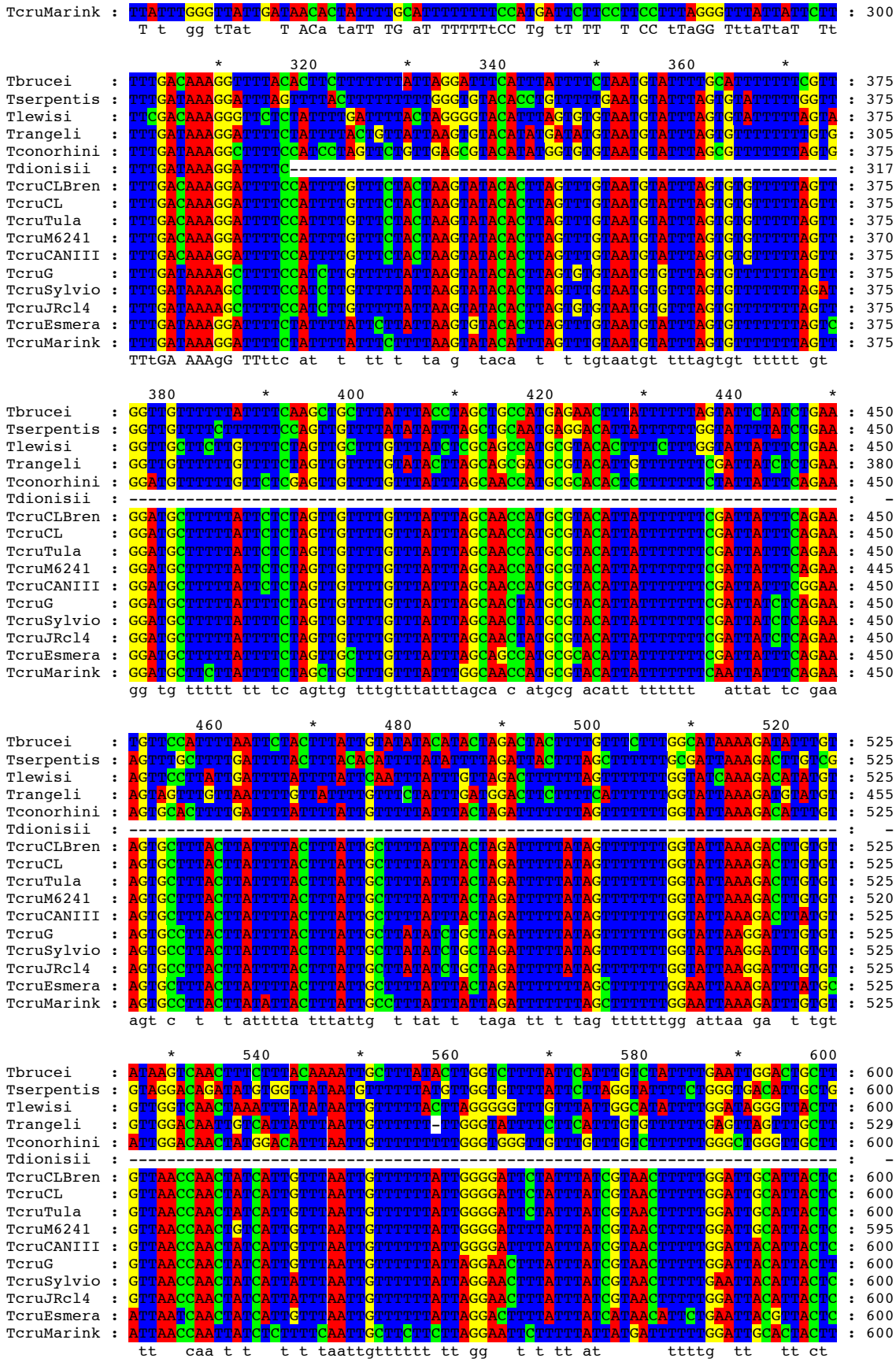
* For the available length of *Tdio* MURF1 transcript.

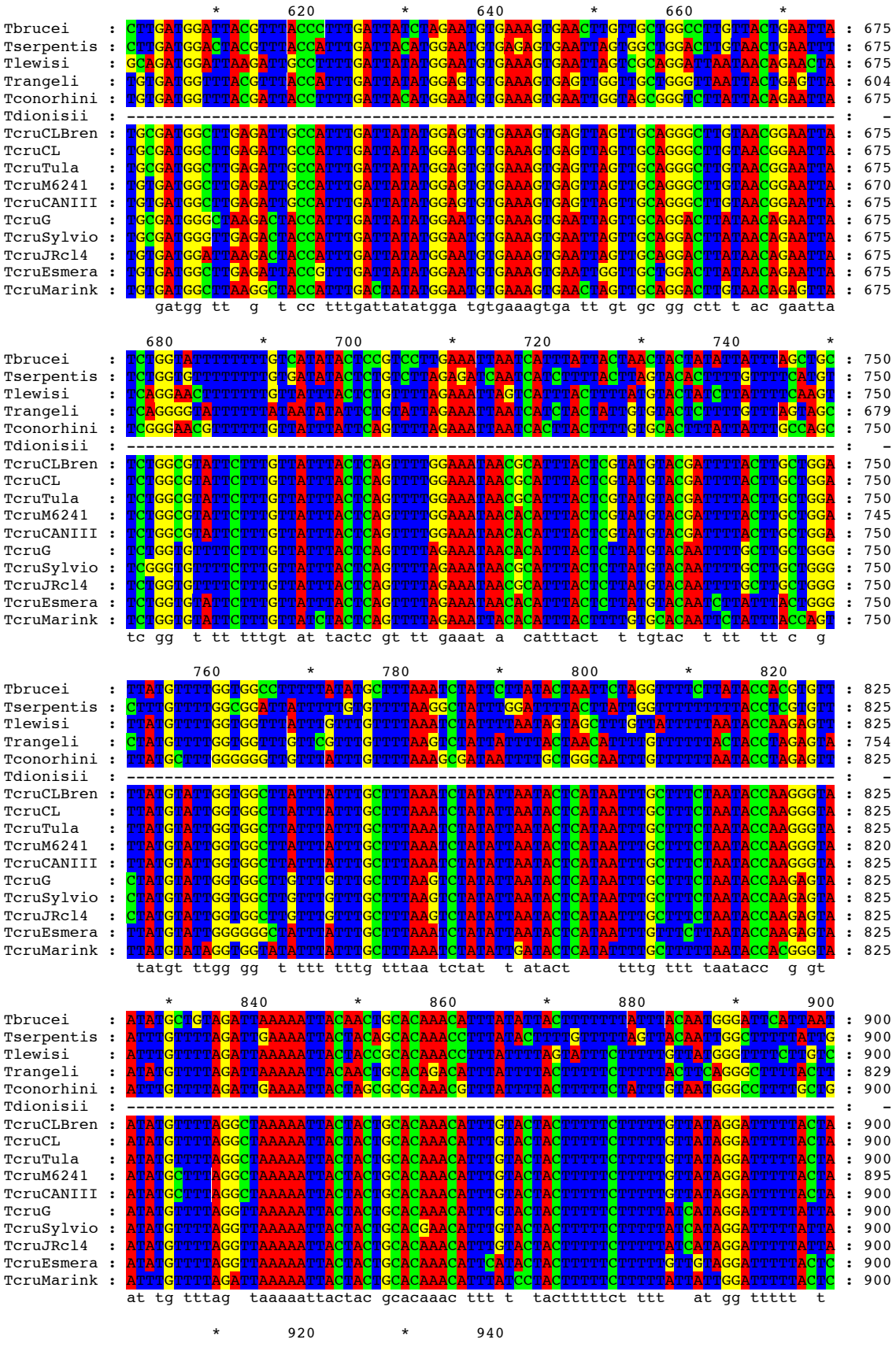
4.SI-S3d. MURF1 protein percent identity matrix.

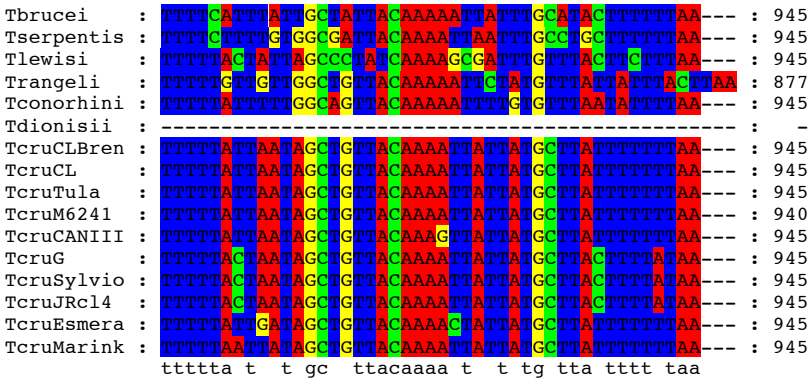
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	76	**100**														
Tlew	77	71	**100**													
Tran	80	75	80	**100**												
Tcon	82	74	80	86	**100**											
Tdio*	44	41	54	46	48	**100**										
TcCLB	80	73	80	83	83	48	**100**									
TcCL	80	73	80	83	83	48	100	**100**								
TcTula	74	67	80	76	77	55	92	92	**100**							
TcM62	14	13	18	14	14	30	14	14	24	**100**						
TcCAN	80	73	80	83	83	48	100	100	91	14	**100**					
TcG	80	73	81	83	83	45	97	97	90	14	97	**100**				
TcSyl	80	72	80	83	83	46	96	96	89	14	96	99	**100**			
TcJR	79	73	80	83	83	44	97	97	90	14	96	99	99	**100**		
TcEsm	76	67	82	77	79	54	88	88	90	19	88	89	89	88	**100**	
TcMar	81	73	81	85	85	48	93	93	85	14	93	93	93	93	87	**100**

* For the available length of *Tdio* MURF1 protein. This matrix shows the identities of the truncated MURF1 proteins.

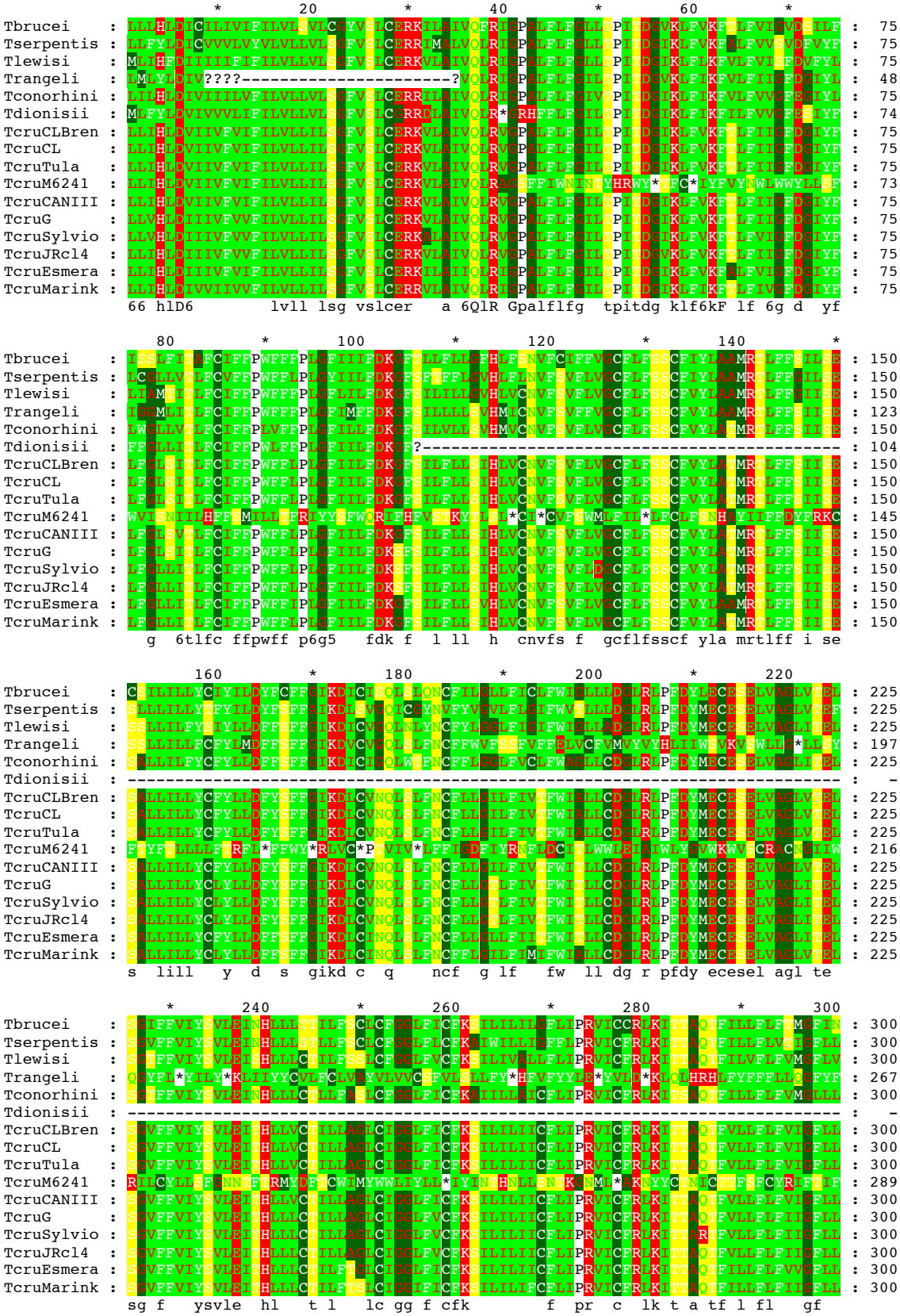

Tbrucei
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink
NNNNNNNN

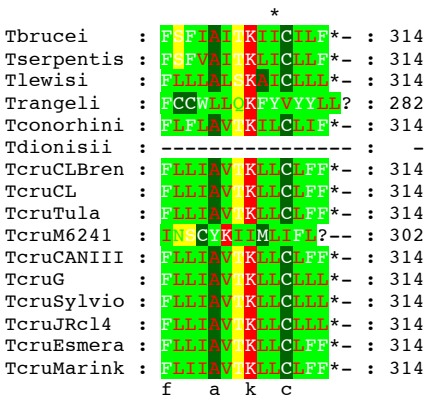

TcruMarink : TTATTTGGGTTATTGATAACACTATTTTGCATTTTTTTTCCATGATTCTTCCTTCCTTTAGGGTTTATTATTCTT : 300
T t gg tTat T ACa taTT TG aT TTTTTtCC Tg tT TT T CC tTaGG TttaTtaT Tt
320 340 360
Tbrucei : TTTGACAAAGGTTTTACACTTCTTTTTTATTAGGATTTCATTTATTTTCTAATGTATTTTGCATTTTTTTCGTT : 375
Tserpentis : TTTGATAAAGGATTTAGTTTTACTTTTTTTTTTGGGTGTACACCTGTTTTTGAATGTATTTTAGTGTATTTTTGGTT : 375
Tlewisi : TTCGACAAAGGGTTTCTCTATTTTGATTTTACTAGGGGTTACATTTAGTGTGTAATGTATTTAGTGTATTTTTACTA : 375
Trangeli : TTTGATAAAGGATTTTCTATTTTACTGTTATTAAGTGTACATATGATATGTAATGTATTTAGTGTTTTTTTTGTG : 305
Tconorhini : TTTGATAAAGGCTTTTCCATCCTAGTTCTGTTGAGCGTACATATGGTGTGTAATGTATTTAGCGTTTTTTTAGTG : 375
Tdionisii : TTTGATAAAGGATTTTC---------- : 317
TcruCLBren : TTTGACAAAGGATTTTCCATTTTGTTTCTACTAAGTATACACTTAGTTTGTAATGTATTTAGTGTGTTTTTAGTT : 375
TcruCL : TTTGACAAAGGATTTTCCATTTTGTTTCTACTAAGTATACACTTAGTTTGTAATGTATTTAGTGTGTTTTTAGTT : 375
TcruTula : TTTGACAAAGGATTTTCCATTTTGTTTCTACTAAGTATACACTTAGTTTGTAATGTATTTAGTGTGTTTTTAGTT : 375
TcruM6241 : TTTGACAAAGGATTTTCCATTTTGTTTCTACTAAGTATACACTTAGTTTGTAATGTATTTAGTGTGTTTTTAGTT : 370
TcruCANIII : TTTGACAAAGGATTTTCCATTTTGTTTCTACTAAGTATACACTTAGTTTGTAATGTATTTAGTGTGTTTTTAGTT : 375
TcruG : TTTGATAAAAGCTTTTCCATCTTGTTTTTATTAAGTATACACTTAGTTGTGTAATGTGTTTAGTGTTTTTTTAGTT : 375
TcruSylvio : TTTGATAAAAGCTTTTCCATCTTGTTTTTATTAAGTATACACTTAGTTGTGTAATGTGTTTAGTGTTTTTTTAGAT : 375
TcruJRcl4 : TTTGATAAAAGCTTTTCCATCTTGTTTTTATTAAGTATACACTTAGTTGTGTAATGTGTTTAGTGTTTTTTTAGTT : 375
TcruEsmera : TTTGATAAAGGATTTTCTATTTTATTCTTATTAAGTGTACACTTAGTTTGTAATGTATTTAGTGTTTTTTTAGTC : 375
TcruMarink : TTTGATAAAGGATTTTCTATTTTATTCTTTTAAGTATACATTTAGTTTGTAATGTATTTAGTGTTTTTTTAGTT : 375
TTtGA AAAgG TTttc at t tt t ta g taca t t tgtaatgt tttagtgt ttttt gt
380 400 420 440
Tbrucei : GGTTGTTTTTTATTTCAAGCTGCTTTATTTACCTAGCTGCCATGAGAACTTTATTTTTTAGTATTCTATCTGAA : 450
Tserpentis : GGTTGTTTTCTTTTTCCAGTTGTTTTATATTTAGCTGCAAAGAGGACATTATTTTTGGTATTTATCTGAA : 450
Tlewisi : GGTTGCTTCTTGTTTTCTAGTTGCTTTGTTTATCTCGCAGCCATGCGTACACTTTCTTTGGTATTATTTCTGAA : 450
Trangeli : GGTTGTTTTTTGTTTCTAGTTGTTTTGTATACTTAGCAGCGATGCGTACATTGTTTTTTTCGATTATCTCTGAA : 380
Tconorhini : GGATGTTTTTTGTTCTCGAGTTGTTTTGTTTATTTAGCAACCATGCGCACACTCTTTTTTTCTATTATTTCAGAA : 450
Tdionisii : ---------- : -
TcruCLBren : GGATGCTTTTTATTCTCTAGTTGTTTTGTTTATTTAGCAACCATGCGTACATTATTTTTTTCGATTATTTCAGAA : 450
TcruCL : GGATGCTTTTTATTCTCTAGTTGTTTTGTTTATTTAGCAACCATGCGTACATTATTTTTTTCGATTATTTCAGAA : 450
TcruTula : GGATGCTTTTTATTCTCTAGTTGTTTTGTTTATTTAGCAACCATGCGTACATTATTTTTTTCGATTATTTCAGAA : 450
TcruM6241 : GGATGCTTTTTATTCTCTAGTTGTTTTGTTTATTTAGCAACCATGCGTACATTATTTTTTTCGATTATTTCAGAA : 445
TcruCANIII : GGATGCTTTTTATTCTCTAGTTGTTTTGTTTATTTAGCAACCATGCGTACATTATTTTTTTCGATTATTTCGGAA : 450
TcruG : GGATGCTTTTTATTTCTAGTTGTTTTGTTTATTTAGCAACTATGCGTACATTATTTTTTTCGATTATCTCAGAA : 450
TcruSylvio : GGATGCTTTTTATTTCTAGTTGTTTTGTTTATTTAGCAACTATGCGTACATTATTTTTTTCGATTATCTCAGAA : 450
TcruJRcl4 : GGATGCTTTTTATTTCTAGTTGTTTTGTTTATTTAGCAACTATGCGTACATTATTTTTTTCGATTATCTCAGAA : 450
TcruEsmera : GGATGCTTTTTATTTCTAGTTGCTTTGTTTATTTAGCAGCCATGCGCACATTATTTTTTTCGATTATTTCAGAA : 450
TcruMarink : GGATGCTTCTTATTTTCTAGCTGCTTTGTTTATTTGGCAACCATGCGTACATTATTTTTTTCAATTATTTCAGAA : 450
gg tg ttttt tt tc agttg tttgtttatttagca c atgcg acatt tttttt attat tc gaa
460 480 500 520
Tbrucei : TGTTCCATTTTAATTCTACTTTATTGCTATATACATACTAGACTACTTTGTTTCTTTGGCATAAAAGATATTTGT : 525
Tserpentis : AGTTTGCTTTTGATTTTACTTTACACATTTTATATTTAGATTACTTTAGCTTTTTTGCGATTAAAGACTTGTCGG : 525
Tlewisi : AGTTCCTTATTGATTTTATTTTATTCAATTTATTTGTTAGACTTTTTTAGTTTTTTGGTATCAAAGACATAGGT : 525
Trangeli : AGTAGTTTGTTAATTTTGTTATTTTGTTTCTAATTGAAGGACTTCTTTTCATTTTTTGGTATTAAAGATGTATGT : 455
Tconorhini : AGTGCACTTTGATTTTATTTTATTGTTTTATTTACTAGATTTTTTTTAGTTTTTTTGGTATTAAAGACATTTGT : 525
Tdionisii : ---------- : -
TcruCLBren : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTATAGTTTTTTTGGTATTAAAGACTTGTGT : 525
TcruCL : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTATAGTTTTTTTGGTATTAAAGACTTGTGT : 525
TcruTula : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTATAGTTTTTTTGGTATTAAAGACTTGTGT : 525
TcruM6241 : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTATAGTTTTTTTGGTATTAAAGACTTGTGT : 520
TcruCANIII : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTATAGTTTTTTTGGTATTAAAGACTTATGT : 525
TcruG : AGTGCCTTACTTATTTTACTTTATTGCTTTATTCTGCTAGATTTTTATAGTTTTTTTGGTATTAAGGATTTGTGT : 525
TcruSylvio : AGTGCCTTACTTATTTTACTTTATTGCTTTATTCTGCTAGATTTTTATAGTTTTTTTGGTATTAAGGATTTGTGT : 525
TcruJRcl4 : AGTGCCTTACTTATTTTACTTTATTGCTTTATTCTGCTAGATTTTTATAGTTTTTTTGGTATTAAGGATTTGTGT : 525
TcruEsmera : AGTGCTTTACTTATTTTACTTTATTGCTTTTATTTACTAGATTTTTTTAGCTTTTTTGGAATTAAAGATTTAGGC : 525
TcruMarink : AGTGCCTTACTTATATTACTTTATTGCCTTTATTTATTAGATTTTTTTAGCTTTTTTGGAATTAAAGATTTGTGT : 525
agt c t t atttta tttattg t tat t taga tt t tag ttttttgg attaa ga t tgt
540 560 580 600
Tbrucei : ATAAGTCAACTTTCTTTACAAAATTGCTTTATACTTGGCTTTTATTCATTGCTATTTGAATTGGACTGCTT : 600
Tserpentis : GTAGGACAGATATGTGGTTATAAATGTTTTTTATGTTGGTGTTTAATCTTAGGTATTTTCTGGGGACATTGCTG : 600
Tlewisi : GTTGGTCAACTAAATTTATATAATTGTTTTTACTTAGGGGGTTTGTTTATTGGCATATTTTGGAATGGGTTACTT : 600
Trangeli : GTTGGACAATTGTCATTATTTAATTGTTTTTT-TTGGGTATTTGCTTCATTTGTGTTTTTTGAGTTAGTTGCTT : 529
Tconorhini : ATTGGACAACTATGGACATTTAATTGTTTTTTTTGGGTGGGTTGTTTGTTGCTTTTTGGGCTGGGTTGCTT : 600
Tdionisii : ---------- : -
TcruCLBren : GTTAACCAACTATCATTGCTTTAATTGTTTTTTATTGGGGATTCTATTTATTCGTAACTTTTTGGATTGCATTACTC : 600
TcruCL : GTTAACCAACTATCATTGCTTTAATTGTTTTTTATTGGGGATTCTATTTATTCGTAACTTTTTGGATTGCATTACTC : 600
TcruTula : GTTAACCAACTATCATTGCTTTAATTGTTTTTTATTGGGGATTCTATTTATTCGTAACTTTTTGGATTGCATTACTC : 600
TcruM6241 : GTTAACCAACTGTCATTGCTTTAATTGTTTTTTATTGGGGATTTTATTTATTCGTAACTTTTTGGATTGCATTACTC : 595
TcruCANIII : GTTAACCAACTATCATTGCTTTAATTGTTTTTTATTGGGGATTTTATTTATTCGTAACTTTTTGGATTACATTACTC : 600
TcruG : GTTAACCAACTATCATTGCTTTAATTGTTTTTTATTAGGAACTTTATTTATTCGTAACTTTTTGGATTACATTACTT : 600
TcruSylvio : GTTAACCAACTATCATTATTTAATTGTTTTTTATTAGGAACTTTATTTATTCGTAACTTTTTGGATTACATTACTC : 600
TcruJRcl4 : GTTAACCAACTATCATTATTTAATTGTTTTTTATTAGGAACTTTATTTATTCGTAACTTTTTGGATTACATTACTC : 600
TcruEsmera : ATTAATCAACTATCATTGCTTTAATTGTTTTTTATTAGGACTTTTATTTATTCATAACATTCTGAATTACGTTACTC : 600
TcruMarink : ATTAACCAATTATCTCTTTTCAATTGCTTCTTCTTTAGGAATTCTTTTTATTATGATTTTTTGGATTGCACTACTT : 600
tt caa t t t t taattgtttttt tt gg t t tt at ttttg tt tt ct


620
640
660
Tbrucei : 675
Tserpentis : 675
Tlewisi : 675
Trangeli : 604
Tconorhini : 675
Tdionisii : -
TcruCLBren : 675
TcruCL : 675
TcruTula : 675
TcruM6241 : 670
TcruCANIII : 675
TcruG : 675
TcruSylvio : 675
TcruJRcl4 : 675
TcruEsmera : 675
TcruMarink : 675
gatgg tt g t cc tttgattatatgga tgtgaaagtga tt gt gc gg ctt t ac gaatta
680
700
720
740
Tbrucei : 750
Tserpentis : 750
Tlewisi : 750
Trangeli : 679
Tconorhini : 750
Tdionisii : -
TcruCLBren : 750
TcruCL : 750
TcruTula : 750
TcruM6241 : 745
TcruCANIII : 750
TcruG : 750
TcruSylvio : 750
TcruJRcl4 : 750
TcruEsmera : 750
TcruMarink : 750
tc gg t tt tttgt at tactc gt tt gaaat a catttact t tgtac t tt tt c g
760
780
800
820
Tbrucei : 825
Tserpentis : 825
Tlewisi : 825
Trangeli : 754
Tconorhini : 825
Tdionisii : -
TcruCLBren : 825
TcruCL : 825
TcruTula : 825
TcruM6241 : 820
TcruCANIII : 825
TcruG : 825
TcruSylvio : 825
TcruJRcl4 : 825
TcruEsmera : 825
TcruMarink : 825
tatgt ttgg gg t ttt tttg tttaa tctat t atact tttg ttt taatacc g gt
840
860
880
900
Tbrucei : 900
Tserpentis : 900
Tlewisi : 900
Trangeli : 829
Tconorhini : 900
Tdionisii : -
TcruCLBren : 900
TcruCL : 900
TcruTula : 900
TcruM6241 : 895
TcruCANIII : 900
TcruG : 900
TcruSylvio : 900
TcruJRcl4 : 900
TcruEsmera : 900
TcruMarink : 900
at tg tttag taaaaattactac gcacaaac ttt t tacttttttct ttt at gg ttttt t
920
940



4.SI-S4a: GeneDoc ND1 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. "NNNN" represents lack of sequence information for *T. rangeli*.





4.SI-S4b: GeneDoc ND1 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. This alignment accounts for the frameshifts caused by the indels [shown in the corresponding nucleotide sequence alignment above (S4a)] and shows the resultant truncated proteins in the alternate frame. '*' indicates stop codon in the alignment. Several stop codons are seen in the altered frame of TcruM6241 and Trangeli.

4.SI-S4c: ND1 nucleotide percent identity matrix.

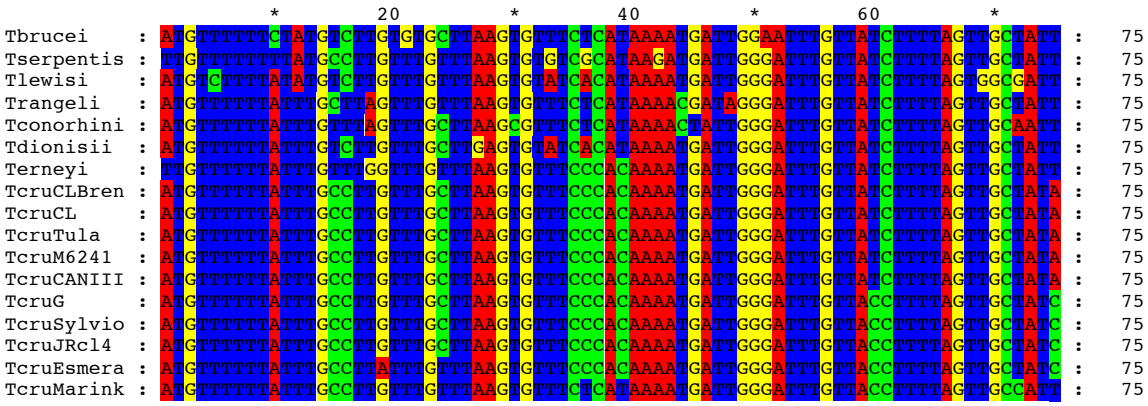
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	72	**100**														
Tlew	72	70	**100**													
Tran	72	70	75	**100**												
Tcon	71	71	77	78	**100**											
Tdio*	67	75	74	77	78	**100**										
TcCLB	72	69	77	76	77	79	**100**									
TcCL	72	69	77	76	77	79	100	**100**								
TcTula	72	69	77	76	77	79	100	100	**100**							
TcM62	72	70	77	76	77	79	99	99	99	**100**						
TcCAN	72	69	77	75	76	79	99	99	99	99	**100**					
TcG	71	70	77	75	76	76	92	92	92	92	92	**100**				
TcSyl	71	70	76	75	77	77	92	92	92	91	91	98	**100**			
TcJR	72	71	77	76	77	76	92	92	92	92	92	99	98	**100**		
TcEsm	73	70	77	77	78	77	91	91	91	91	91	90	90	90	**100**	
TcMar	73	71	77	76	78	79	87	87	87	87	86	86	85	86	87	**100**

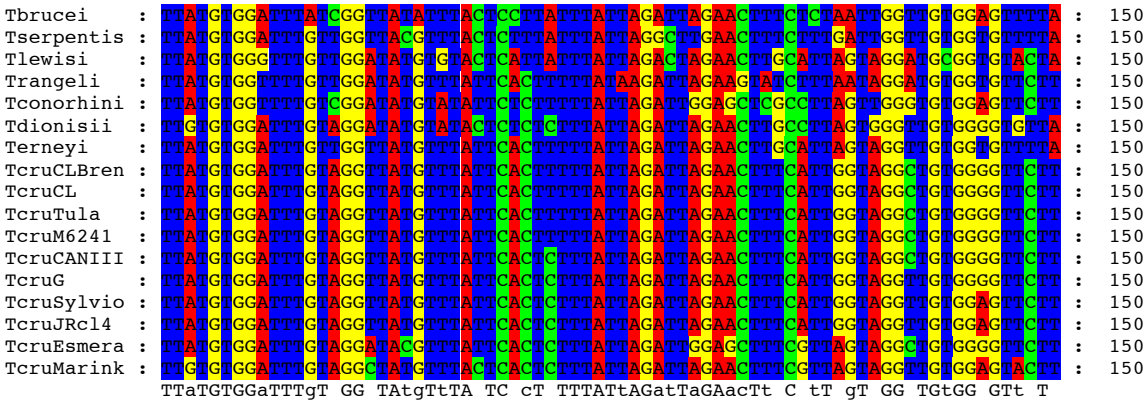
***** For the available length of *Tdio* ND1 transcript.

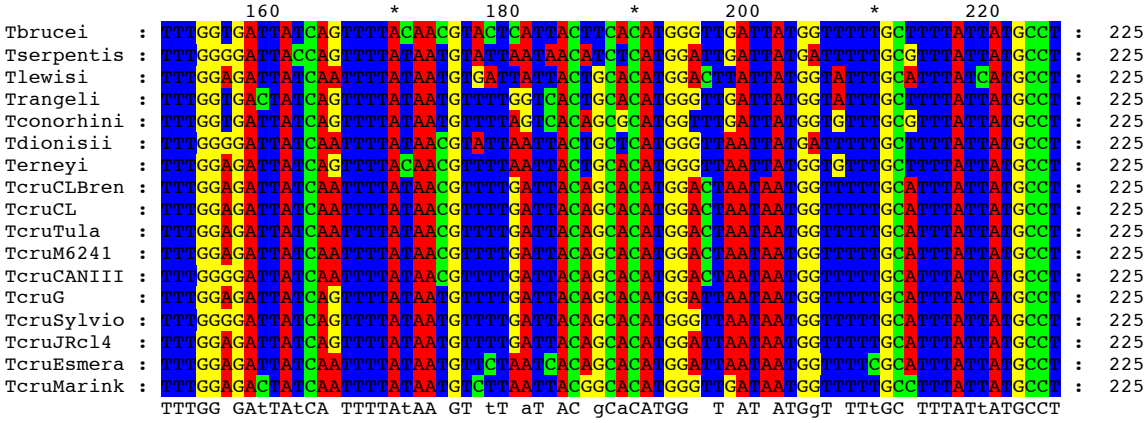
4.SI-S4d: ND1 protein percent identity matrix.

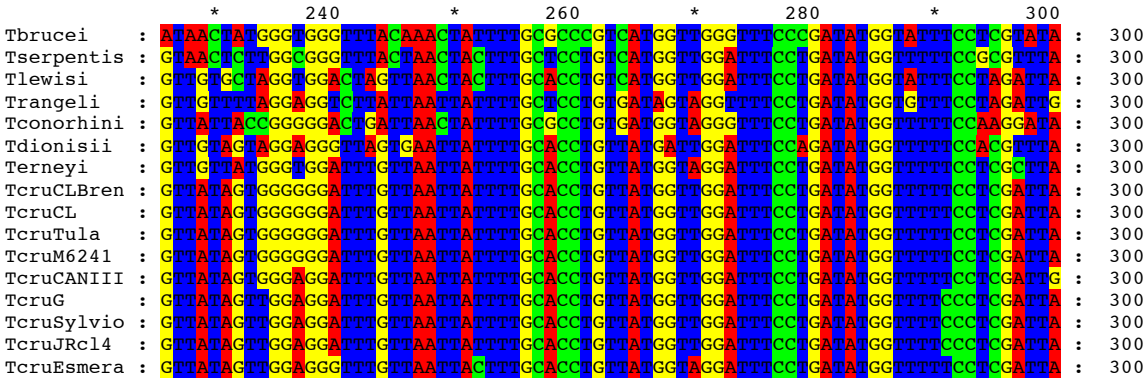
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	72	**100**														
Tlew	75	72	**100**													
Tran	45	43	49	**100**												
Tcon	72	72	79	51	**100**											
Tdio*	67	73	71	73	76	**100**										
TcCLB	75	72	79	51	80	78	**100**									
TcCL	75	72	79	51	80	78	100	**100**								
TcTula	75	72	79	51	80	78	100	100	**100**							
TcM62	17	16	18	11	16	37	20	20	20	**100**						
TcCAN	75	73	79	50	80	77	99	99	99	20	**100**					
TcG	75	73	80	51	80	75	96	96	96	19	96	**100**				
TcSyl	74	73	79	51	80	77	94	94	94	19	94	98	**100**			
TcJR	75	73	80	51	80	75	96	96	96	19	96	99	98	**100**		
TcEsm	77	75	81	51	83	78	93	93	93	19	93	92	92	92	**100**	
TcMar	75	74	79	52	83	79	92	92	92	18	91	91	90	91	91	**100**

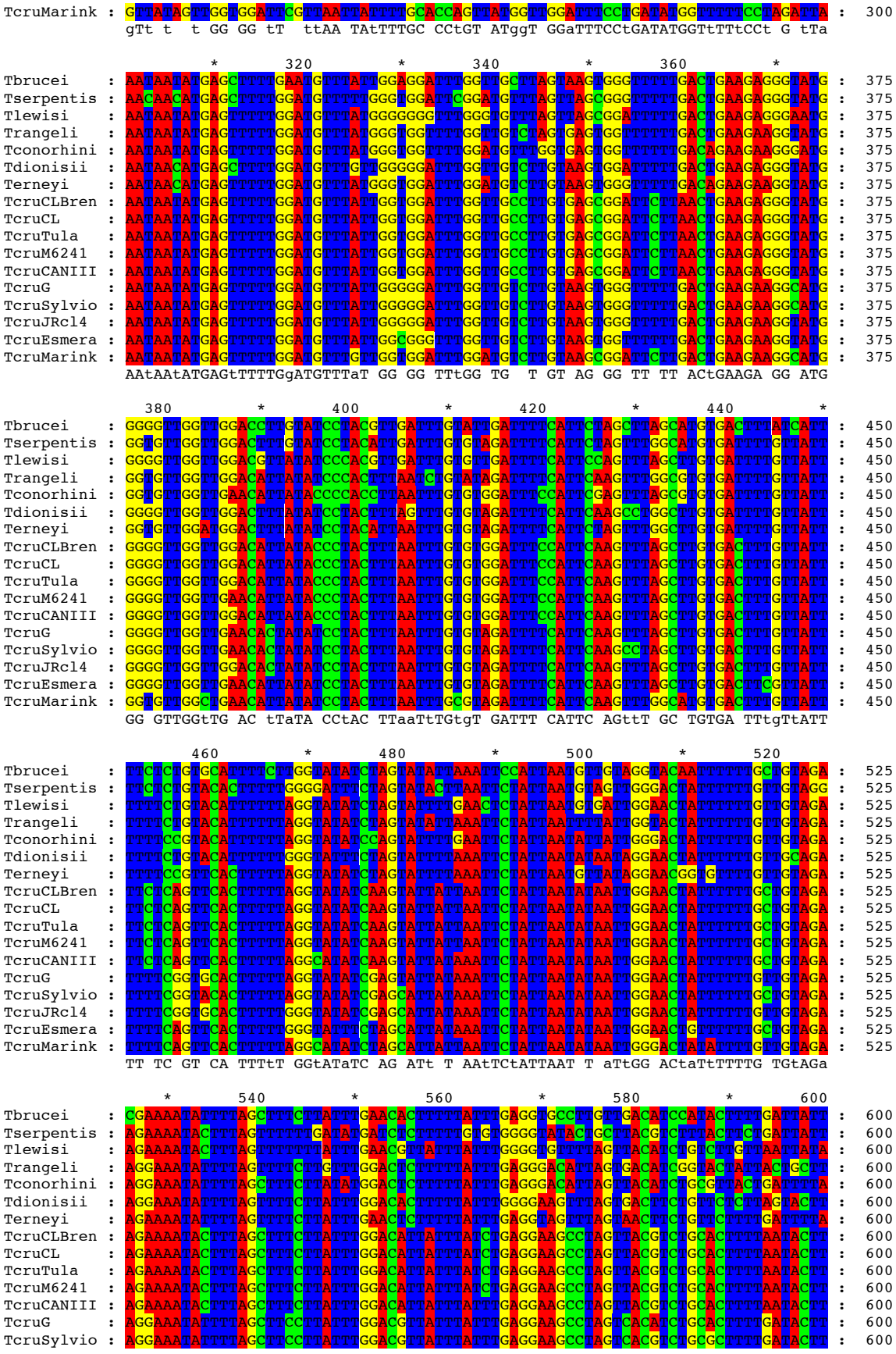
***** For the available length of *Tdio* ND1 transcript. This matrix shows the identities of the truncated ND1 proteins.

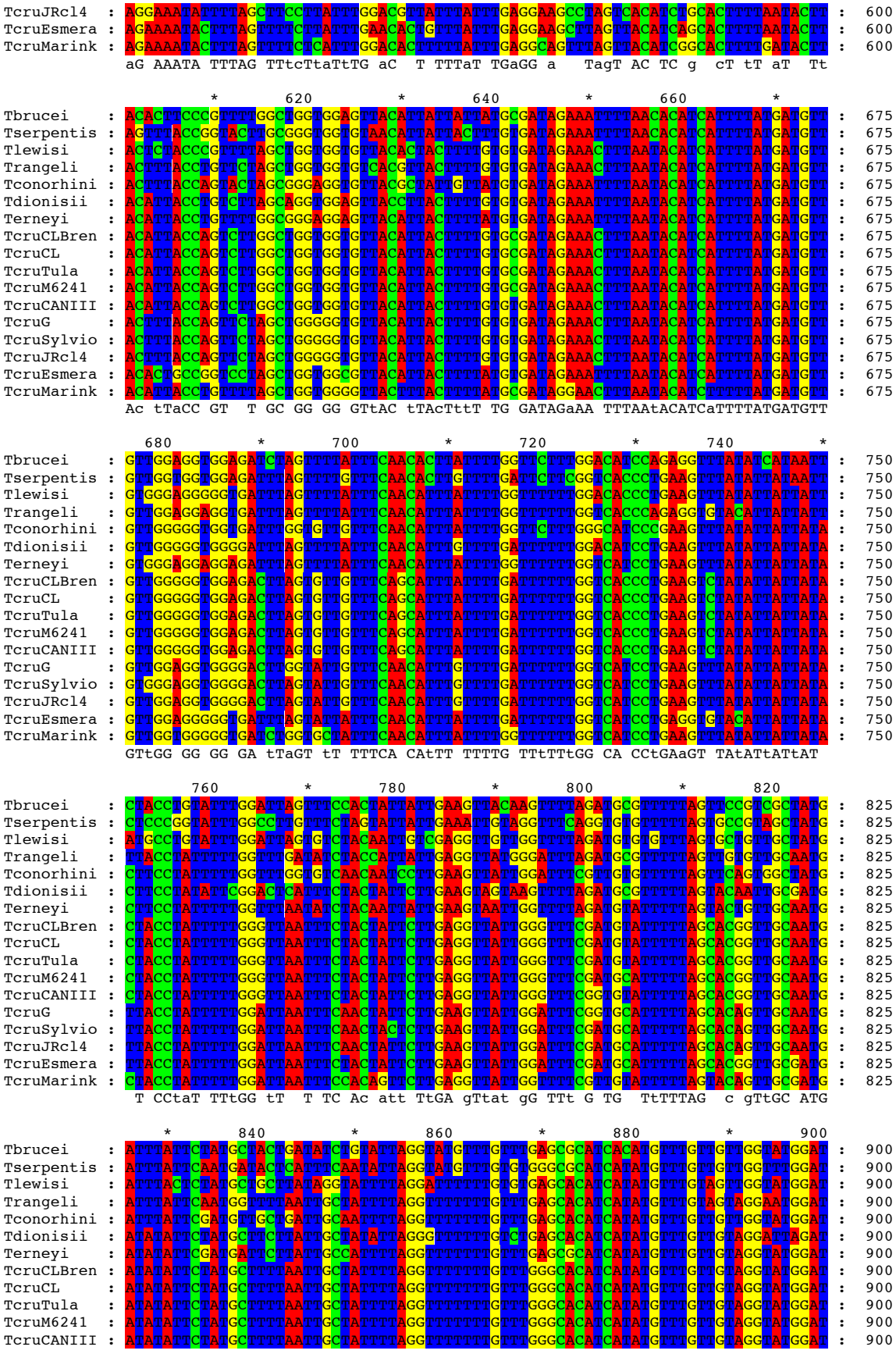

Tbrucei : 75
Tserpentis : 75
Tlewisi : 75
Trangeli : 75
Tconorhini : 75
Tdionisii : 75
Terneyi : 75
TcruCLBren : 75
TcruCL : 75
TcruTula : 75
TcruM6241 : 75
TcruCANIII : 75
TcruG : 75
TcruSylvio : 75
TcruJRcl4 : 75
TcruEsmera : 75
TcruMarink : 75

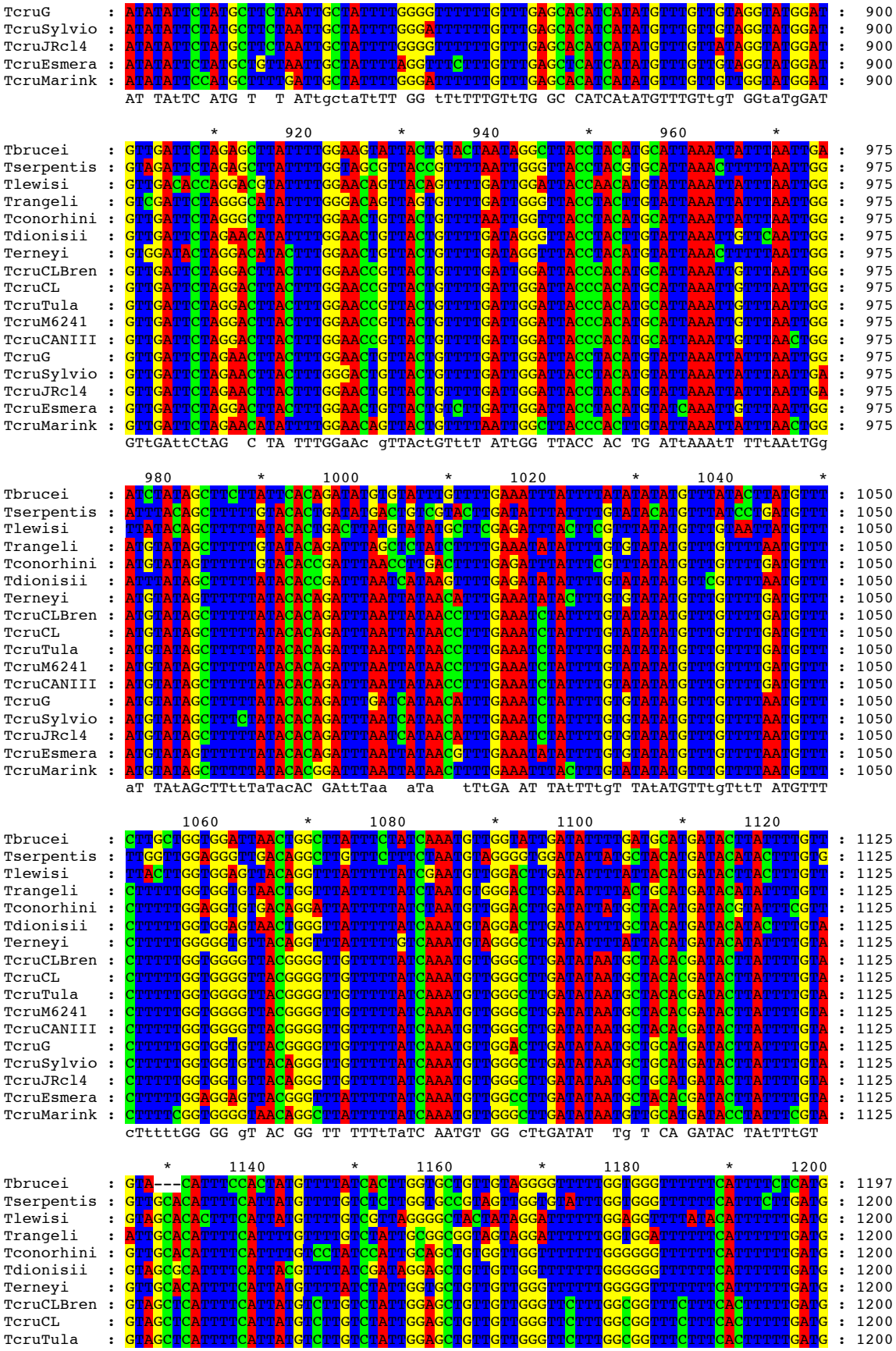

Tbrucei : 150
Tserpentis : 150
Tlewisi : 150
Trangeli : 150
Tconorhini : 150
Tdionisii : 150
Terneyi : 150
TcruCLBren : 150
TcruCL : 150
TcruTula : 150
TcruM6241 : 150
TcruCANIII : 150
TcruG : 150
TcruSylvio : 150
TcruJRcl4 : 150
TcruEsmera : 150
TcruMarink : 150

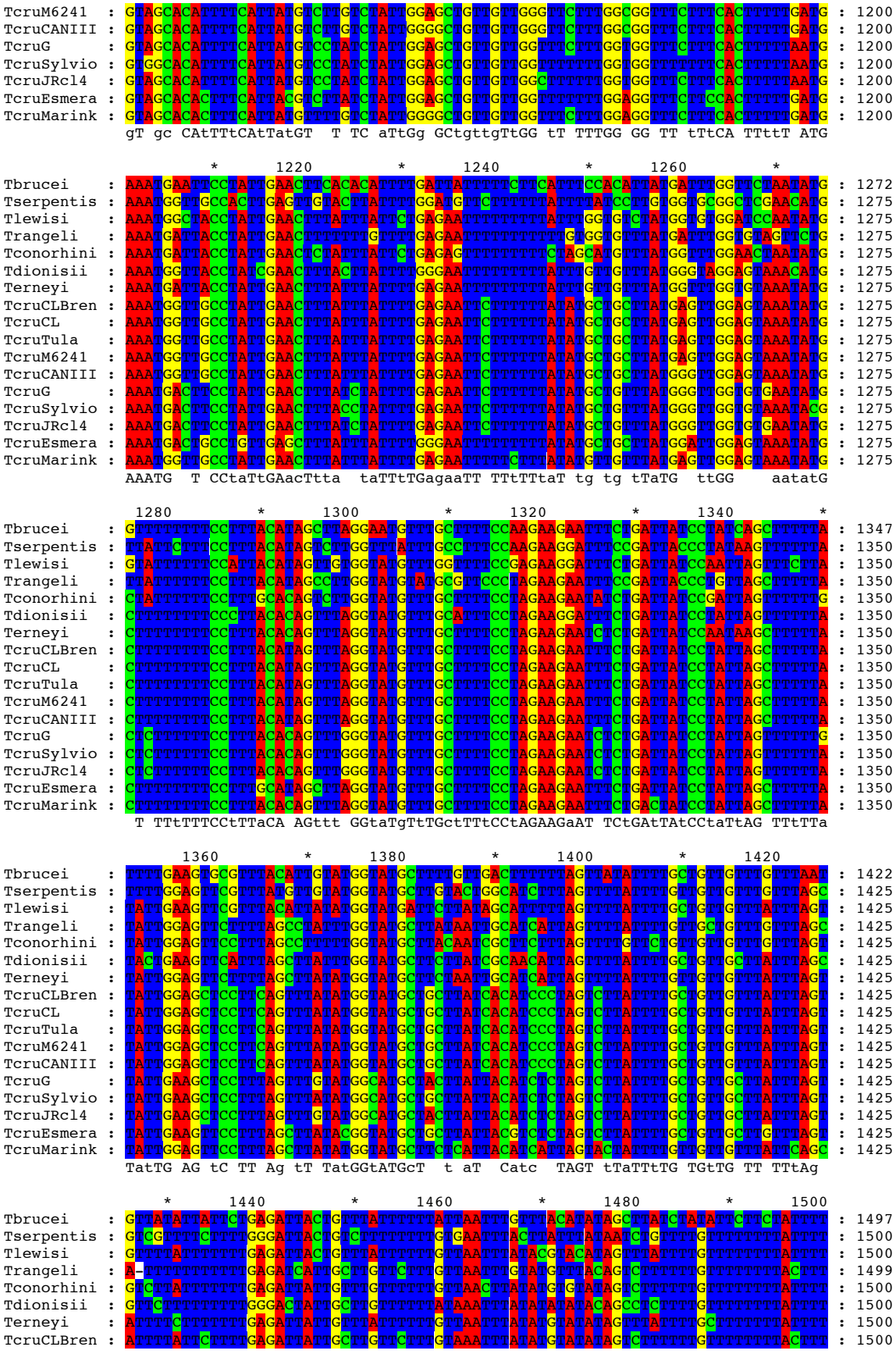

Tbrucei : 225
Tserpentis : 225
Tlewisi : 225
Trangeli : 225
Tconorhini : 225
Tdionisii : 225
Terneyi : 225
TcruCLBren : 225
TcruCL : 225
TcruTula : 225
TcruM6241 : 225
TcruCANIII : 225
TcruG : 225
TcruSylvio : 225
TcruJRcl4 : 225
TcruEsmera : 225
TcruMarink : 225

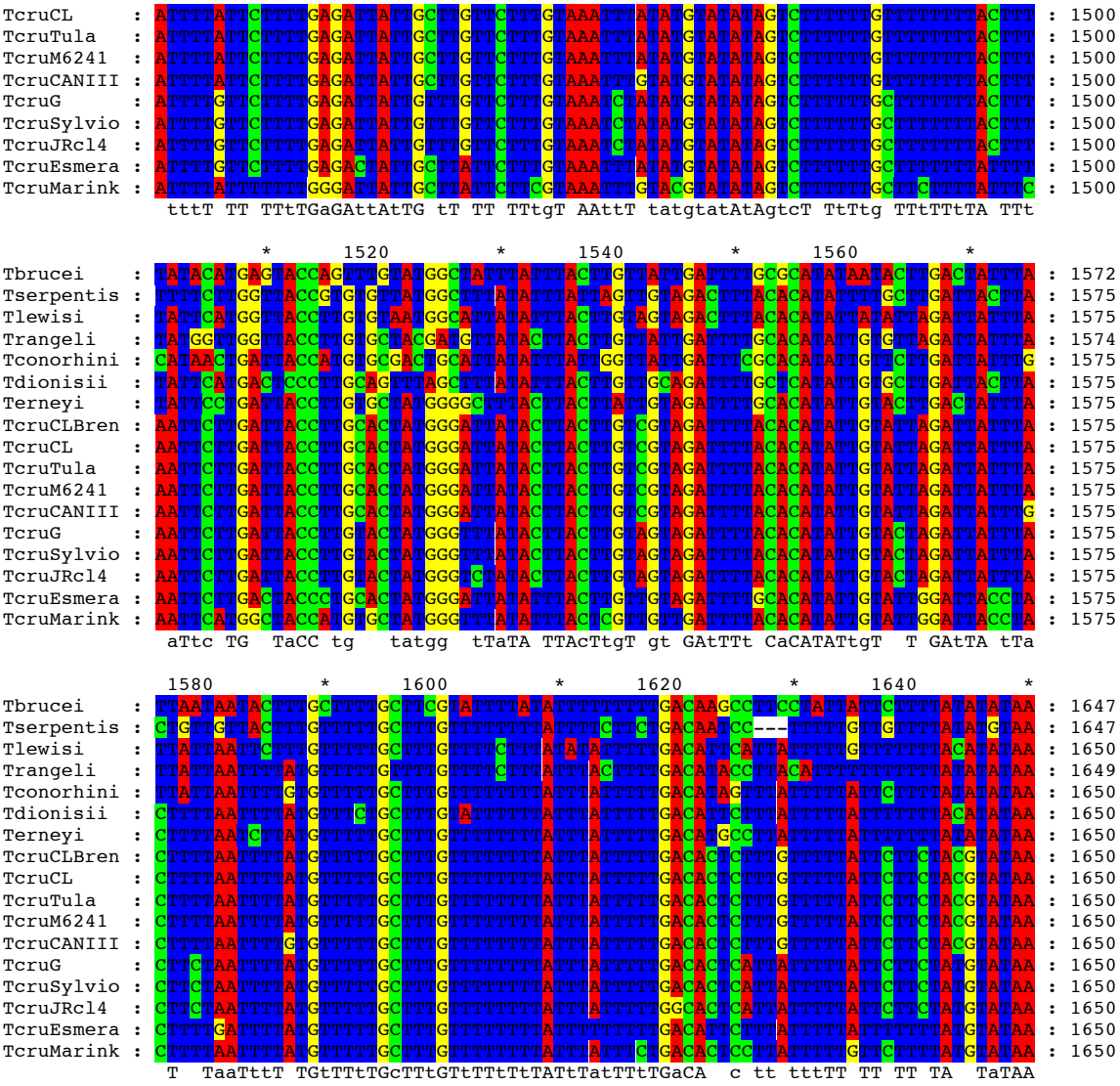

Tbrucei : 300
Tserpentis : 300
Tlewisi : 300
Trangeli : 300
Tconorhini : 300
Tdionisii : 300
Terneyi : 300
TcruCLBren : 300
TcruCL : 300
TcruTula : 300
TcruM6241 : 300
TcruCANIII : 300
TcruG : 300
TcruSylvio : 300
TcruJRcl4 : 300
TcruEsmera : 300



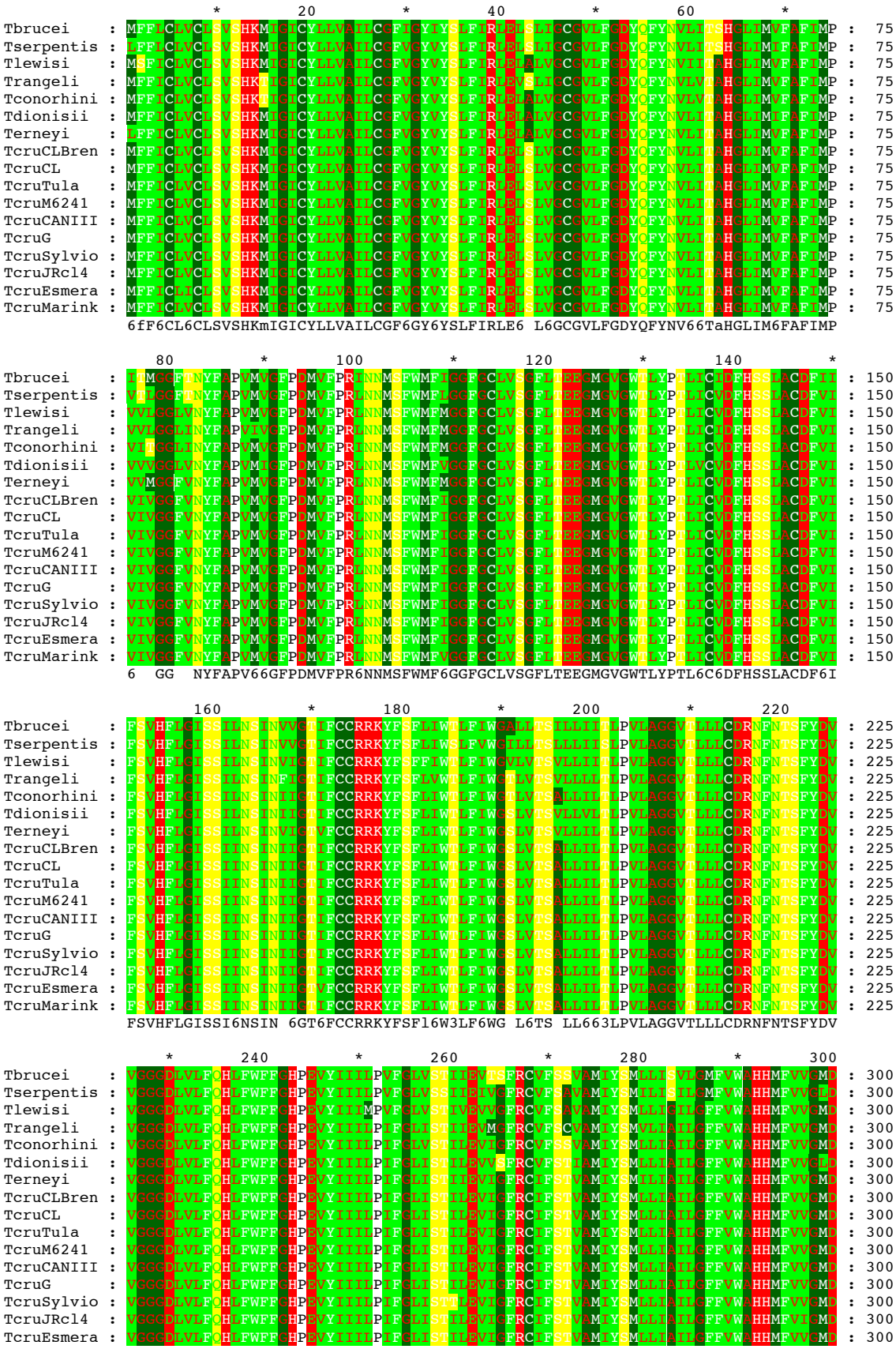

TcruJRcl4 : 600
TcruEsmera : 600
TcruMarink : 600
Tbrucei : 675
Tserpentis : 675
Tlewisi : 675
Trangeli : 675
Tconorhini : 675
Tdionisii : 675
Terneyi : 675
TcruCLBren : 675
TcruCL : 675
TcruTula : 675
TcruM6241 : 675
TcruCANIII : 675
TcruG : 675
TcruSylvio : 675
TcruJRcl4 : 675
TcruEsmera : 675
TcruMarink : 675
Tbrucei : 750
Tserpentis : 750
Tlewisi : 750
Trangeli : 750
Tconorhini : 750
Tdionisii : 750
Terneyi : 750
TcruCLBren : 750
TcruCL : 750
TcruTula : 750
TcruM6241 : 750
TcruCANIII : 750
TcruG : 750
TcruSylvio : 750
TcruJRcl4 : 750
TcruEsmera : 750
TcruMarink : 750
Tbrucei : 825
Tserpentis : 825
Tlewisi : 825
Trangeli : 825
Tconorhini : 825
Tdionisii : 825
Terneyi : 825
TcruCLBren : 825
TcruCL : 825
TcruTula : 825
TcruM6241 : 825
TcruCANIII : 825
TcruG : 825
TcruSylvio : 825
TcruJRcl4 : 825
TcruEsmera : 825
TcruMarink : 825
Tbrucei : 900
Tserpentis : 900
Tlewisi : 900
Trangeli : 900
Tconorhini : 900
Tdionisii : 900
Terneyi : 900
TcruCLBren : 900
TcruCL : 900
TcruTula : 900
TcruM6241 : 900
TcruCANIII : 900

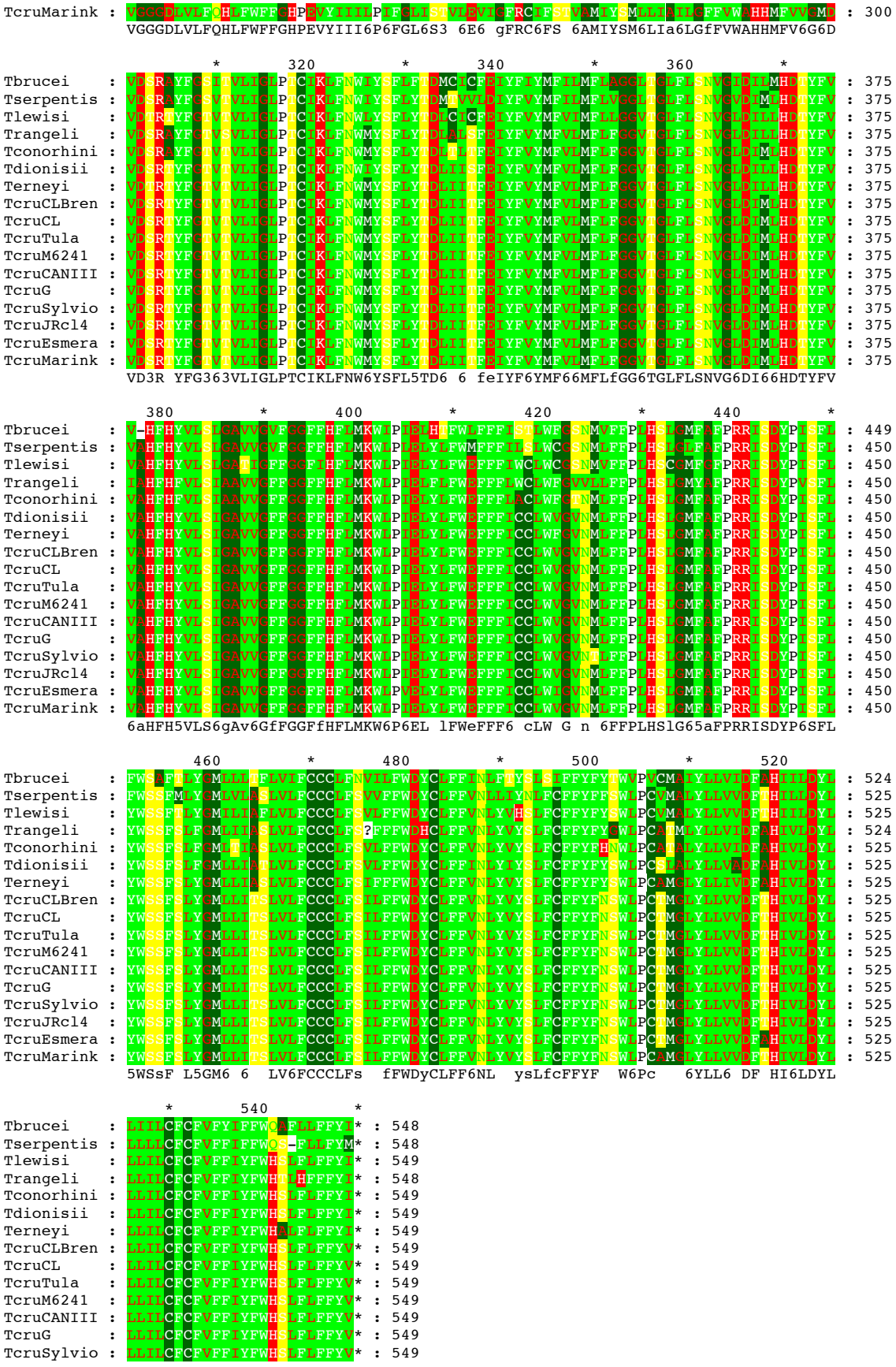

TcruG : 900
TcruSylvio : 900
TcruJRcl4 : 900
TcruEsmera : 900
TcruMarink : 900
AT TAtTC ATG T T ATtgctaTtTT GG tTtTTTGTtTG GC CATCAtATGTTTGTtgT GGtaTgGAT
Tbrucei : 975
Tserpentis : 975
Tlewisi : 975
Trangeli : 975
Tconorhini : 975
Tdionisii : 975
Terneyi : 975
TcruCLBren : 975
TcruCL : 975
TcruTula : 975
TcruM6241 : 975
TcruCANIII : 975
TcruG : 975
TcruSylvio : 975
TcruJRcl4 : 975
TcruEsmera : 975
TcruMarink : 975
GTtGAttCtAG C TA TTTGGaAc gTTActGTttT ATtGG TTACC AC TG ATtAAAtT TTtAAtTGg
Tbrucei : 1050
Tserpentis : 1050
Tlewisi : 1050
Trangeli : 1050
Tconorhini : 1050
Tdionisii : 1050
Terneyi : 1050
TcruCLBren : 1050
TcruCL : 1050
TcruTula : 1050
TcruM6241 : 1050
TcruCANIII : 1050
TcruG : 1050
TcruSylvio : 1050
TcruJRcl4 : 1050
TcruEsmera : 1050
TcruMarink : 1050
aT TAtAGcTTttTaTacAC GAttTaa aTa tTtGA AT TAtTTtgT TAtATGTTtgTttT ATGTTT
Tbrucei : 1125
Tserpentis : 1125
Tlewisi : 1125
Trangeli : 1125
Tconorhini : 1125
Tdionisii : 1125
Terneyi : 1125
TcruCLBren : 1125
TcruCL : 1125
TcruTula : 1125
TcruM6241 : 1125
TcruCANIII : 1125
TcruG : 1125
TcruSylvio : 1125
TcruJRcl4 : 1125
TcruEsmera : 1125
TcruMarink : 1125
cTttttGG GG gT AC GG TT TTTtTaTC AATGT GG cTtGATAT Tg T CA GATAC TAtTTtGT
Tbrucei : 1197
Tserpentis : 1200
Tlewisi : 1200
Trangeli : 1200
Tconorhini : 1200
Tdionisii : 1200
Terneyi : 1200
TcruCLBren : 1200
TcruCL : 1200
TcruTula : 1200


TcruM6241 : 1200
TcruCANIII : 1200
TcruG : 1200
TcruSylvio : 1200
TcruJRcl4 : 1200
TcruEsmera : 1200
TcruMarink : 1200
gT gc CAtTTtCAtTatGT T TC aTtGg GCtgttgTtGG tT TTTGG GG TT tTtCA TTttT ATG
1220 1240 1260
Tbrucei : 1272
Tserpentis : 1275
Tlewisi : 1275
Trangeli : 1275
Tconorhini : 1275
Tdionisii : 1275
Terneyi : 1275
TcruCLBren : 1275
TcruCL : 1275
TcruTula : 1275
TcruM6241 : 1275
TcruCANIII : 1275
TcruG : 1275
TcruSylvio : 1275
TcruJRcl4 : 1275
TcruEsmera : 1275
TcruMarink : 1275
AAATG T CCtaTtGAacTtta taTTtTGagaaTT TTtTTtaT tg tg tTaTG ttGG aatatG
1280 1300 1320 1340
Tbrucei : 1347
Tserpentis : 1350
Tlewisi : 1350
Trangeli : 1350
Tconorhini : 1350
Tdionisii : 1350
Terneyi : 1350
TcruCLBren : 1350
TcruCL : 1350
TcruTula : 1350
TcruM6241 : 1350
TcruCANIII : 1350
TcruG : 1350
TcruSylvio : 1350
TcruJRcl4 : 1350
TcruEsmera : 1350
TcruMarink : 1350
T TTtTTTCCtTTaCA AGttt GGtaTgTtTGctTTtCCtAGAAGaAT TCtGAtTAtCCtaTtAG TTtTTa
1360 1380 1400 1420
Tbrucei : 1422
Tserpentis : 1425
Tlewisi : 1425
Trangeli : 1425
Tconorhini : 1425
Tdionisii : 1425
Terneyi : 1425
TcruCLBren : 1425
TcruCL : 1425
TcruTula : 1425
TcruM6241 : 1425
TcruCANIII : 1425
TcruG : 1425
TcruSylvio : 1425
TcruJRcl4 : 1425
TcruEsmera : 1425
TcruMarink : 1425
TatTG AG tC TT Ag tT TatGGtATGcT t aT Catc TAGT tTaTTtTG TGtTG TT TTtAg
1440 1460 1480 1500
Tbrucei : 1497
Tserpentis : 1500
Tlewisi : 1500
Trangeli : 1499
Tconorhini : 1500
Tdionisii : 1500
Terneyi : 1500
TcruCLBren : 1500



4.SI-S5a: GeneDoc COI nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1.




TcruMarink : VGGGDLVLFQHLFWFFGHPEVYIIILPIFGLISTVDNVICFRCIFSTVAMIYSMLLIAIIGFFVWAHHMFVVCMD : 300
VGGGDLVLFQHLFWFFGHPEVYIII6P6FGL6S3 6E6 gFRC6FS 6AMIYSM6LIa6LGfFVWAHHMFV6G6D
320
340
360
Tbrucei : VDSRAYFGSIVLIGLPTCIKLFNWIYSFLFTDMCICFEIYFIYMFILMFLAGGLTGLFLSNVGIFIIMHDTYFV : 375
Tserpentis : VDSRAYFGSVVLIGLPTCIKLFNWIYSFLYTDMTVVTGIYFVYMPILMFLVGGLTGLFLSNVGVFIMLHDTYFV : 375
Tlewisi : VDARMYFGSVVLIGLPTCIKLFNWLYSFLYTDLCICFEIYFVYMPVIMFLLGGVTGLFLSNVGLLILLHDTYFV : 375
Trangeli : VDSRAYFGSVSVLIGLPTCIKLFNWMYSFLYTDLALSFEIYFVYMPVLMFLFGGVTGLFLSNVGLLILLHDTYFV : 375
Tconorhini : VDSRAYFGSVVVLIGLPTCIKLFNWMYSFLYTDLTLTFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
Tdionisii : VDSRMYFGSVVLIGLPTCIKLFNWIYSFLYTDLIISFEIYFVYMPVLMFLFGGVTGLFLSNVGLLILLHDTYFV : 375
Terneyi : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLILLHDTYFV : 375
TcruCLBren : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruCL : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruTula : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruM6241 : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruCANIII : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruG : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruSylvio : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruJRcl4 : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruEsmera : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
TcruMarink : VDSRMYFGSVVLIGLPTCIKLFNWMYSFLYTDLIITFEIYFVYMPVLMFLFGGVTGLFLSNVGLLIMLHDTYFV : 375
VD3R YFG363VLIGLPTCIKLFNW6YSFL5TD6 6 feIYF6YMF66MFLfGG6TGLFLSNVG6DI66HDTYFV
380
400
420
440
Tbrucei : V-HFHYVLSLGAVVGVFGGFFHFLMKWLPGDLHTFWLFFFISTLWFGSNMVFFPLHSLGMFAFPRRISDYPISFL : 449
Tserpentis : VAHFHYVLSLGAVVGVFGGFFHFLMKWLPGDLYLFWMFFFILSLWCGSNMLFFPLHSLGLFAFPRRISDYPISFL : 450
Tlewisi : VAHFHYVLSLGATIGFFGGFIHFLMKWLPGDLYLFWGFFFIWCLWCGSNMVFFPLHSCGMFGFPRRISDYPISFL : 450
Trangeli : LAHFHFVLSLAAVVGFFGGFFHFLMKWLPGDLFLFWGFFFLWCLWFGVVLLFFPLHSLGMYAFPRRISDYPVSFL : 450
Tconorhini : VAHFHFVLSLAAVVGFFGGFFHFLMKWLPGDLYLFWGFFFLACLWFGTNMLFFPLHSLGMFAFPRRISDYPISFL : 450
Tdionisii : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
Terneyi : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWFGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruCLBren : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruCL : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruTula : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruM6241 : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruCANIII : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruG : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruSylvio : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruJRcl4 : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruEsmera : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWIGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
TcruMarink : VAHFHYVLSLGAVVGFFGGFFHFLMKWLPGDLYLFWGFFFICCLWVGVNMLFFPLHSLGMFAFPRRISDYPISFL : 450
6aHFH5VLS6gAv6GfFGGFfHFLMKW6P6EL lFWeFFF6 cLW G n 6FFPLHSlG65aFPRRISDYP6SFL
460
480
500
520
Tbrucei : FWSAFTLYGMLLLTPLVIFCCCLFNVILFWIYCLFFINLFTYSLSIFFYFYTWVPVCMAIYLLVIPFAHIILDYL : 524
Tserpentis : FWSSFMLYGMLVLASLVIFCCCLFSVVFFWIYCLFFVNLLIYNLFCFFYFFSWLPCVMAIYLLVVCFTHILLDYL : 525
Tlewisi : YWSSFTLYGMLILAFLVLFCCCLFSVLFFWIYCLFFVNLYVHSLFCFFYFYSWLPCVMAIYLLVVIFTHIILDYL : 525
Trangeli : YWSSFSLFGMLLIASLVLFCCCLFS?FFFWDHCLFFVNLYVYSLFCFFYFYGWLPCAMMLYLLVIIFAHIVLDYL : 524
Tconorhini : YWSSFSLFGMLIIASLVLFCCCLFSVLFFWIYCLFFVNLYVYSLFCFFYFHNWLPCAMAIYLLVIIFAHIVLDYL : 525
Tdionisii : YWSSFSLFGMLLIATLVLFCCCLFSVLFFWIYCLFFINLYIYSLFCFFYFYSWLPCSLALYLLVAIFAHIVLDYL : 525
Terneyi : YWSSFSLYGMLLIASLVLFCCCLFSIFFFWIYCLFFVNLYVYSLFCFFYFYSWLPCAMGLYLLIVIFAHIVLDYL : 525
TcruCLBren : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruCL : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVCFTHIVLDYL : 525
TcruTula : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruM6241 : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruCANIII : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruG : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruSylvio : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruJRcl4 : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVIFTHIVLDYL : 525
TcruEsmera : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCTMGLYLLVVCFAHIVLDYL : 525
TcruMarink : YWSSFSLYGMLLITSLVLFCCCLFSILFFWIYCLFFVNLYVYSLFCFFYFNSWLPCAMGLYLLVVIFTHIVLDYL : 525
5WSsF L5GM6 6 LV6FCCCLFs fFWDyCLFF6NL ysLfcFFYF W6Pc 6YLL6 DF HI6LDYL
540
Tbrucei : LLIICFCFVFYIFFWQAFLLFFYI* : 548
Tserpentis : LLLLCFCFVFFIFFWQS-FLLFYM* : 548
Tlewisi : LLIICFCFVFFIYFWHSFLFFYI* : 549
Trangeli : LLIICFCFVFFIYFWHHFFFYI* : 548
Tconorhini : LLIICFCFVFFIYFWHSLFLFFYI* : 549
Tdionisii : LLIICFCFVFFIYFWHSFLFFYI* : 549
Terneyi : LLIICFCFVFFIYFWHHLFLFFYI* : 549
TcruCLBren : LLIICFCFVFFIYFWHSFLFFYV* : 549
TcruCL : LLIICFCFVFFIYFWHSLFLFFYV* : 549
TcruTula : LLIICFCFVFFIYFWHSLFLFFYV* : 549
TcruM6241 : LLIICFCFVFFIYFWHSLFLFFYV* : 549
TcruCANIII : LLIICFCFVFFIYFWHSLFLFFYV* : 549
TcruG : LLIICFCFVFFIYFWHSLFLFFYV* : 549
TcruSylvio : LLIICFCFVFFIYFWHSLFLFFYV* : 549

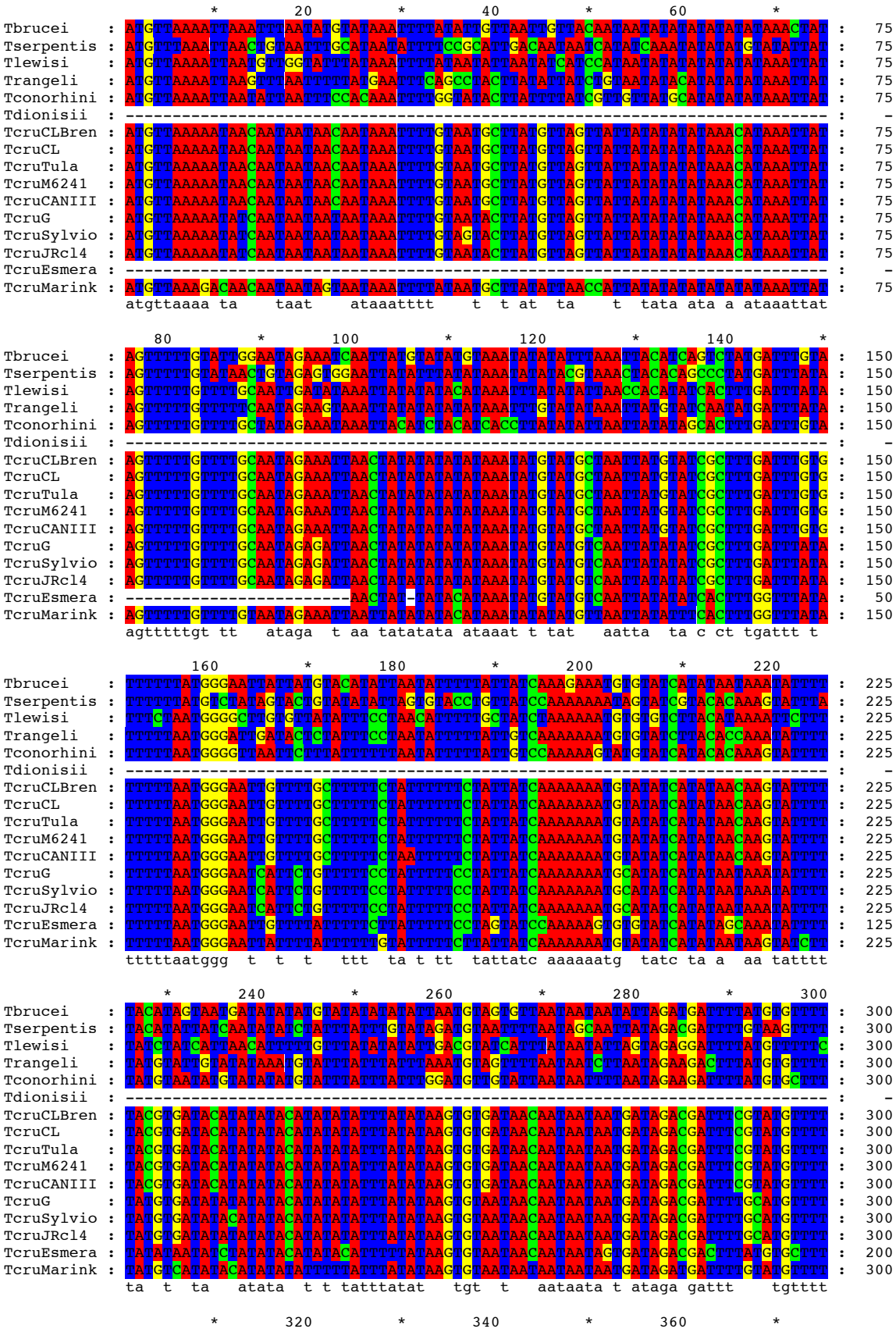
4.SI-S5b: GeneDoc COI protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '?' in Trangeli alignment indicates missing codon information. '*' indicates stop codon in the alignment.

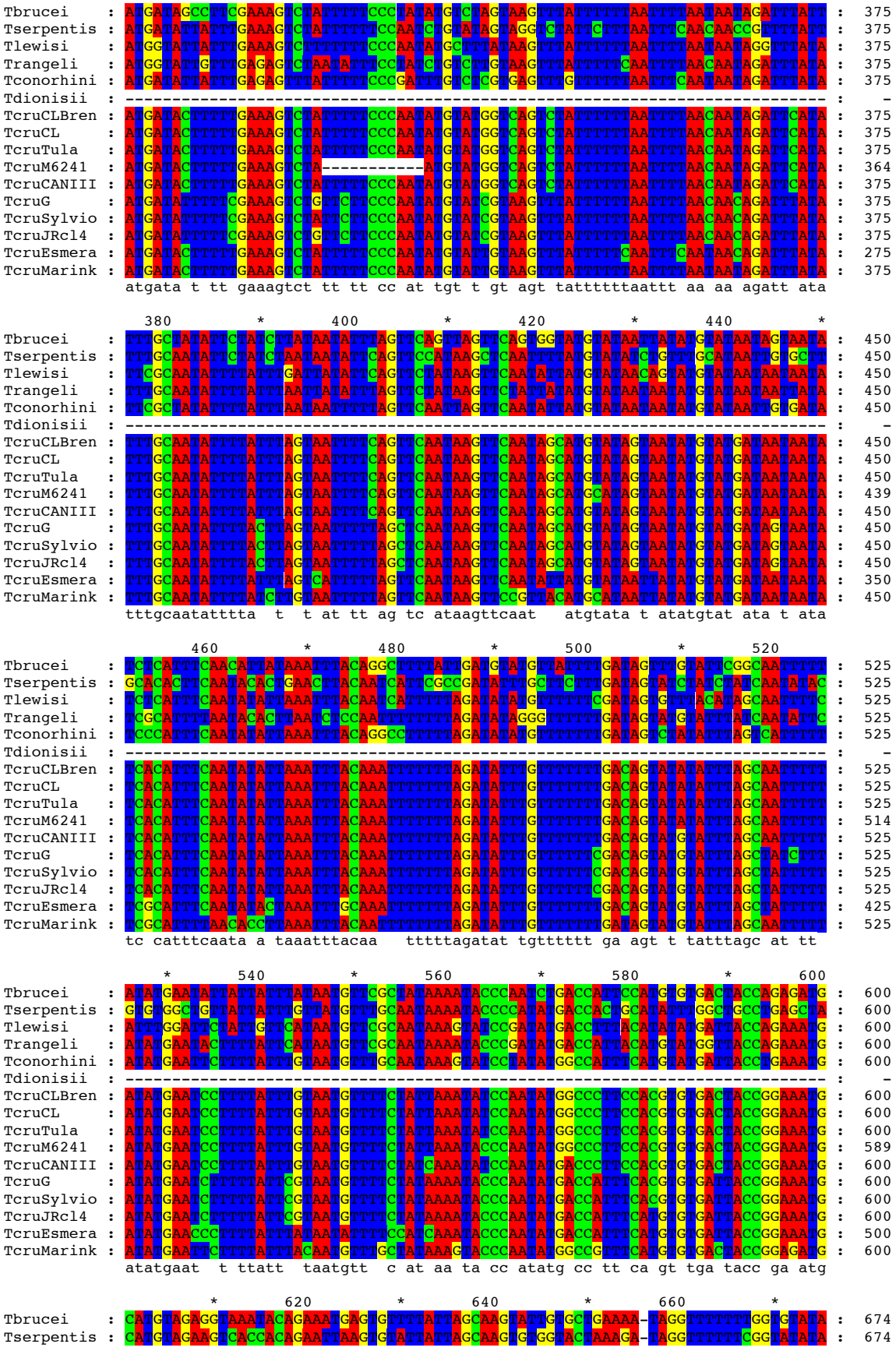
4.SI-S5c: COI nucleotide percent identity matrix.

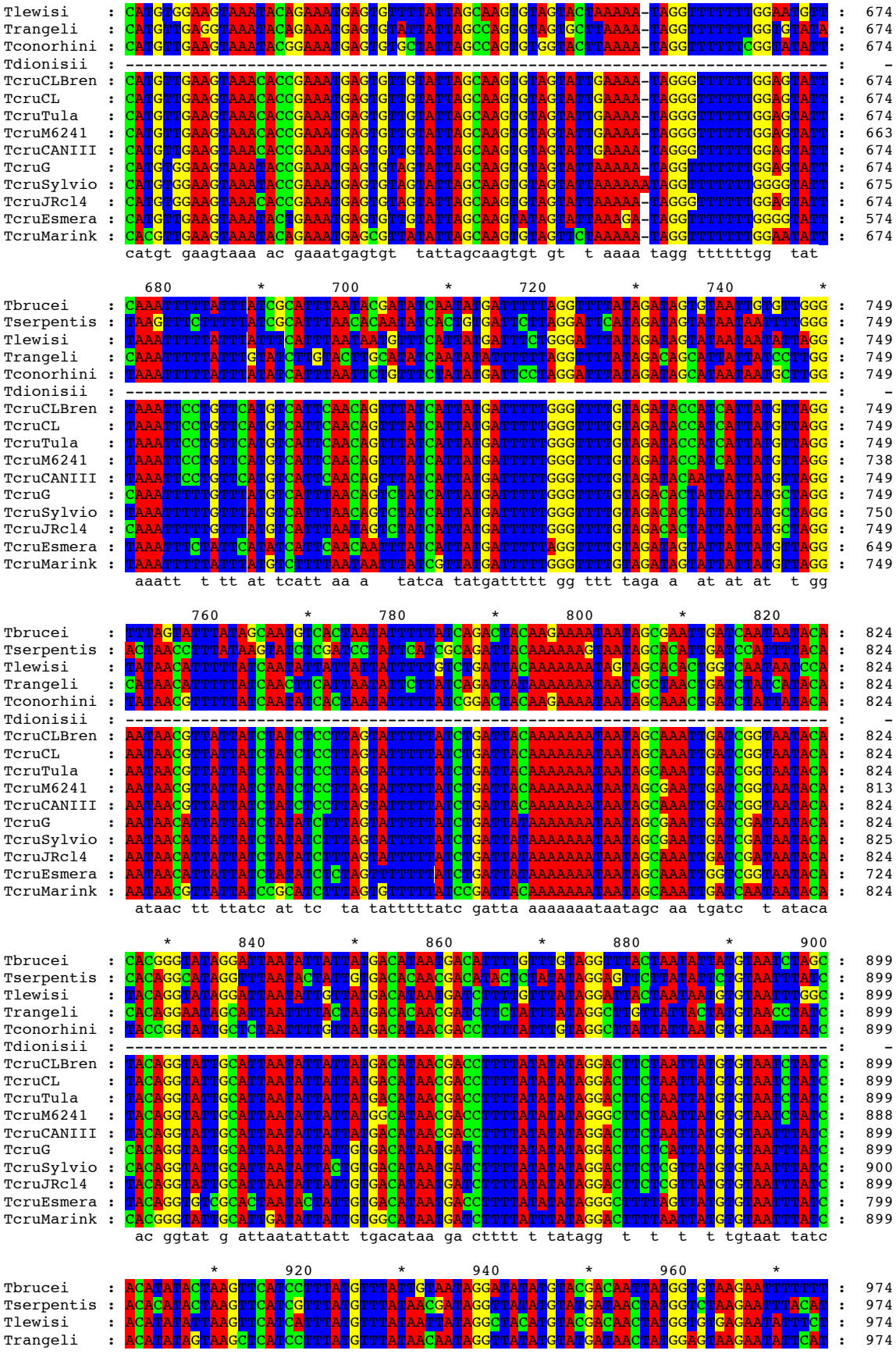
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	79	**100**															
Tlew	79	79	**100**														
Tran	78	77	82	**100**													
Tcon	78	78	81	85	**100**												
Tdio	79	79	82	82	81	**100**											
Tern	81	80	84	86	85	87	**100**										
TcCLB	78	78	83	83	83	85	87	**100**									
TcCL	78	78	83	83	83	85	87	100	**100**								
TcTula	78	78	83	83	83	85	87	100	100	**100**							
TcM62	78	78	83	83	83	85	87	100	100	100	**100**						
TcCAN	78	78	83	83	83	85	87	99	99	99	99	**100**					
TcG	79	78	82	84	84	86	88	93	93	93	93	93	**100**				
TcSyl	79	77	82	84	83	86	88	92	92	92	93	92	98	**100**			
TcJR	79	78	82	84	83	86	88	93	93	93	93	93	99	98	**100**		
TcEsm	78	77	82	84	83	86	87	91	91	91	91	91	91	91	91	**100**	
TcMar	78	78	83	84	83	85	87	89	89	89	89	89	89	88	88	89	**100**

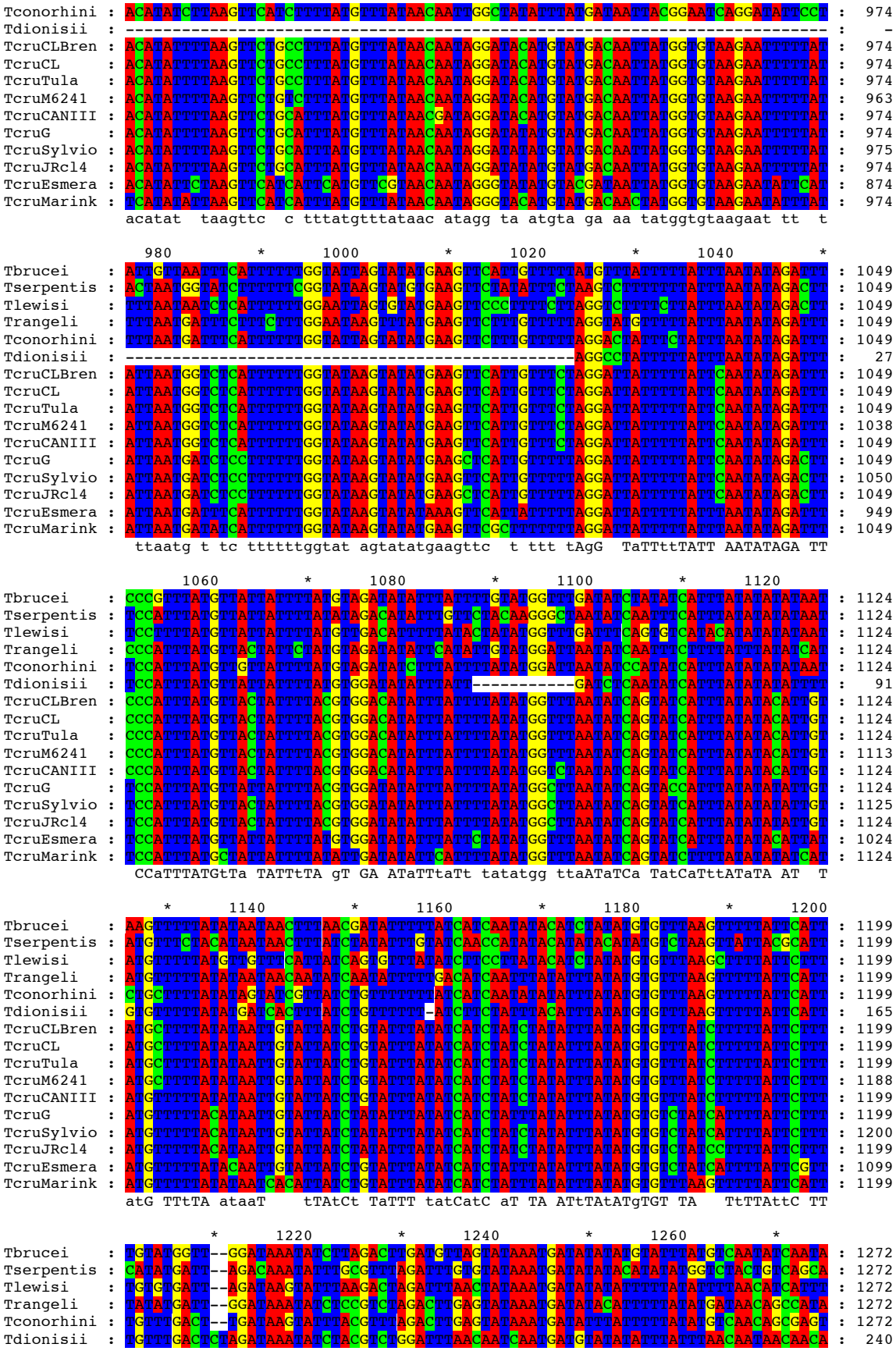
4.SI-S5d: COI protein percent identity matrix.

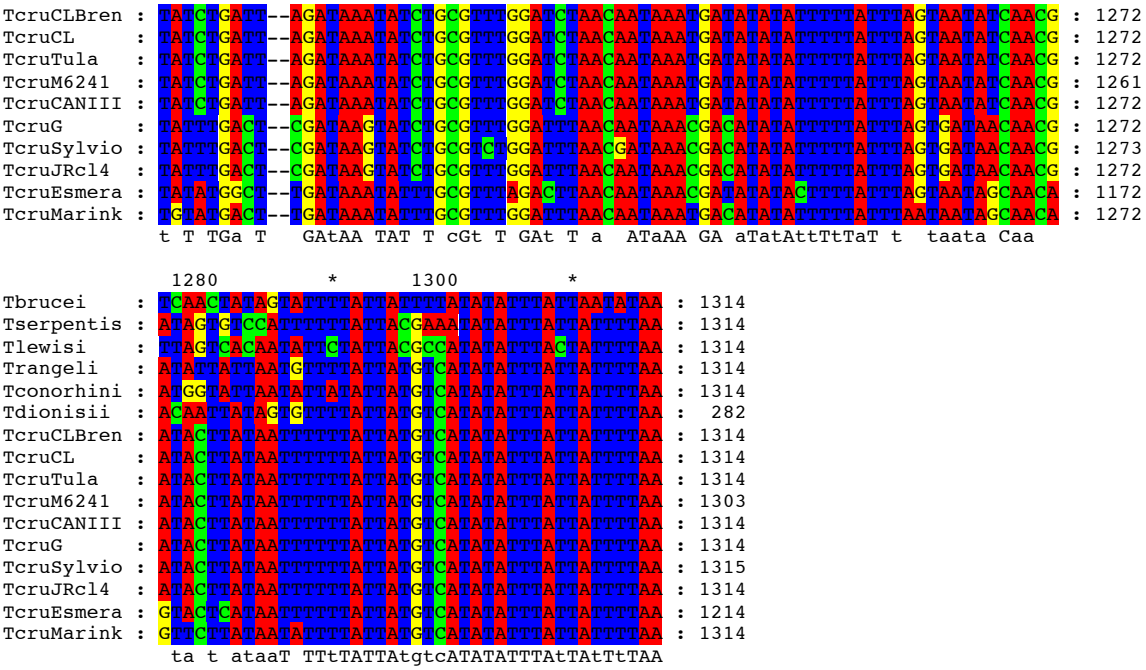
Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
100																
86	**100**															
84	87	**100**														
82	83	87	**100**													
84	86	90	93	**100**												
84	86	91	90	93	**100**											
84	87	92	91	93	94	**100**										
84	87	91	90	94	94	96	**100**									
84	87	91	90	94	94	96	100	**100**								
84	87	91	90	94	94	96	100	100	**100**							
84	87	91	90	94	94	96	100	100	100	**100**						
84	87	91	90	94	94	96	100	100	100	100	**100**					
84	87	91	90	94	94	96	100	100	100	100	100	**100**				
84	87	91	90	93	94	96	100	100	100	100	100	100	**100**			
84	87	91	90	94	94	96	100	100	100	100	100	100	99	**100**		
84	87	90	89	93	94	96	99	99	99	99	99	99	99	99	**100**	
84	87	91	90	94	94	96	99	99	99	99	99	99	99	99	98	**100**


Tbrucei
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink

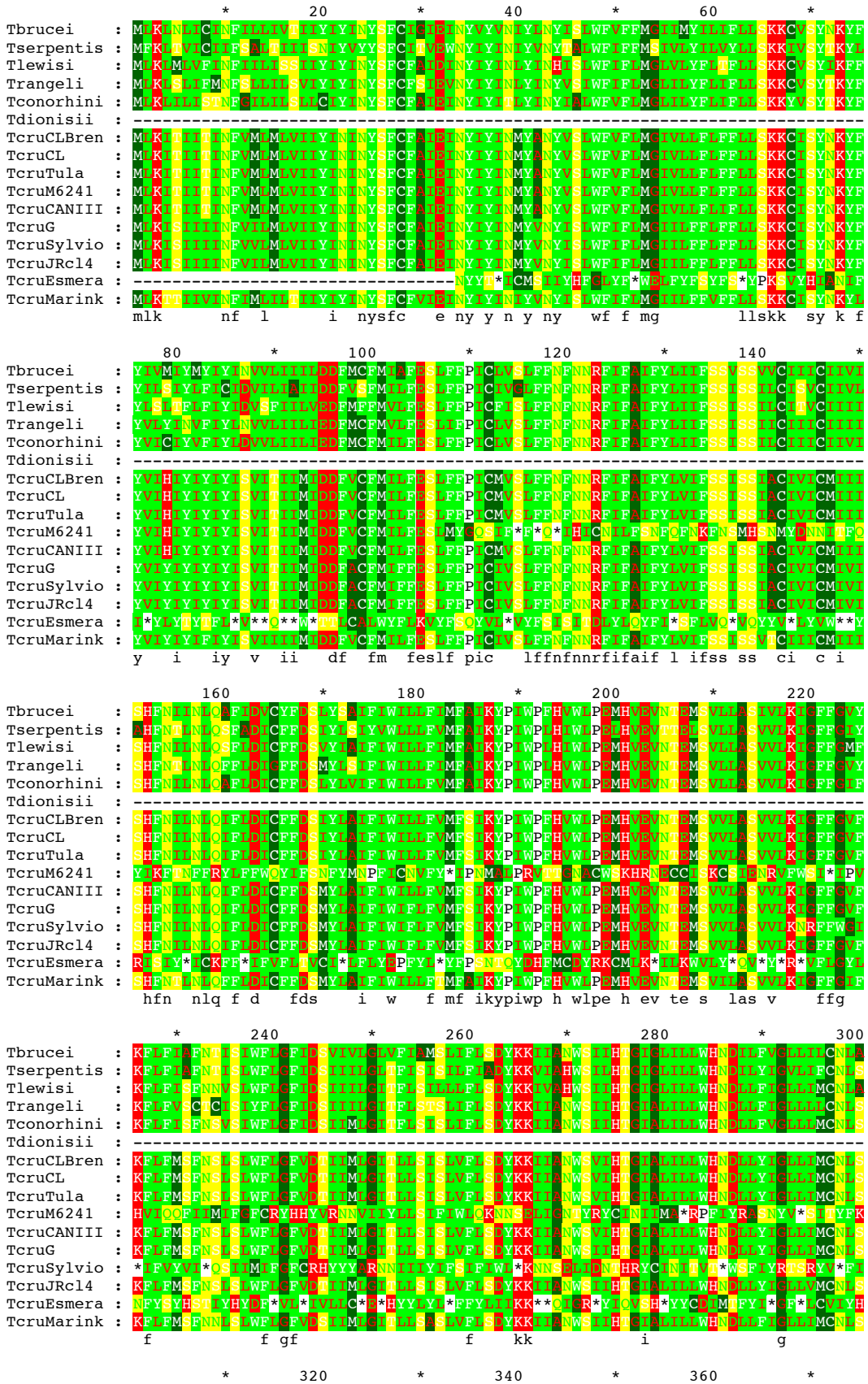

Tbrucei : 375
Tserpentis : 375
Tlewisi : 375
Trangeli : 375
Tconorhini : 375
Tdionisii : -
TcruCLBren : 375
TcruCL : 375
TcruTula : 375
TcruM6241 : 364
TcruCANIII : 375
TcruG : 375
TcruSylvio : 375
TcruJRcl4 : 375
TcruEsmera : 275
TcruMarink : 375
atgata t tt gaaagtct tt tt cc at tgt t gt agt tatttttaattt aa aa agatt ata
380 400 420 440
Tbrucei : 450
Tserpentis : 450
Tlewisi : 450
Trangeli : 450
Tconorhini : 450
Tdionisii : -
TcruCLBren : 450
TcruCL : 450
TcruTula : 450
TcruM6241 : 439
TcruCANIII : 450
TcruG : 450
TcruSylvio : 450
TcruJRcl4 : 450
TcruEsmera : 350
TcruMarink : 450
tttgcaatatttta t t at tt ag tc ataagttcaat atgtata t atatgtat ata t ata
460 480 500 520
Tbrucei : 525
Tserpentis : 525
Tlewisi : 525
Trangeli : 525
Tconorhini : 525
Tdionisii : -
TcruCLBren : 525
TcruCL : 525
TcruTula : 525
TcruM6241 : 514
TcruCANIII : 525
TcruG : 525
TcruSylvio : 525
TcruJRcl4 : 525
TcruEsmera : 425
TcruMarink : 525
tc catttcaata a taaatttacaa tttttagatat tgtttttt ga agt t tatttagc at tt
540 560 580 600
Tbrucei : 600
Tserpentis : 600
Tlewisi : 600
Trangeli : 600
Tconorhini : 600
Tdionisii : -
TcruCLBren : 600
TcruCL : 600
TcruTula : 600
TcruM6241 : 589
TcruCANIII : 600
TcruG : 600
TcruSylvio : 600
TcruJRcl4 : 600
TcruEsmera : 500
TcruMarink : 600
atatgaat t ttatt taatgtt c at aa ta cc atatg cc tt ca gt tga tacc ga atg
620 640 660
Tbrucei : 674
Tserpentis : 674

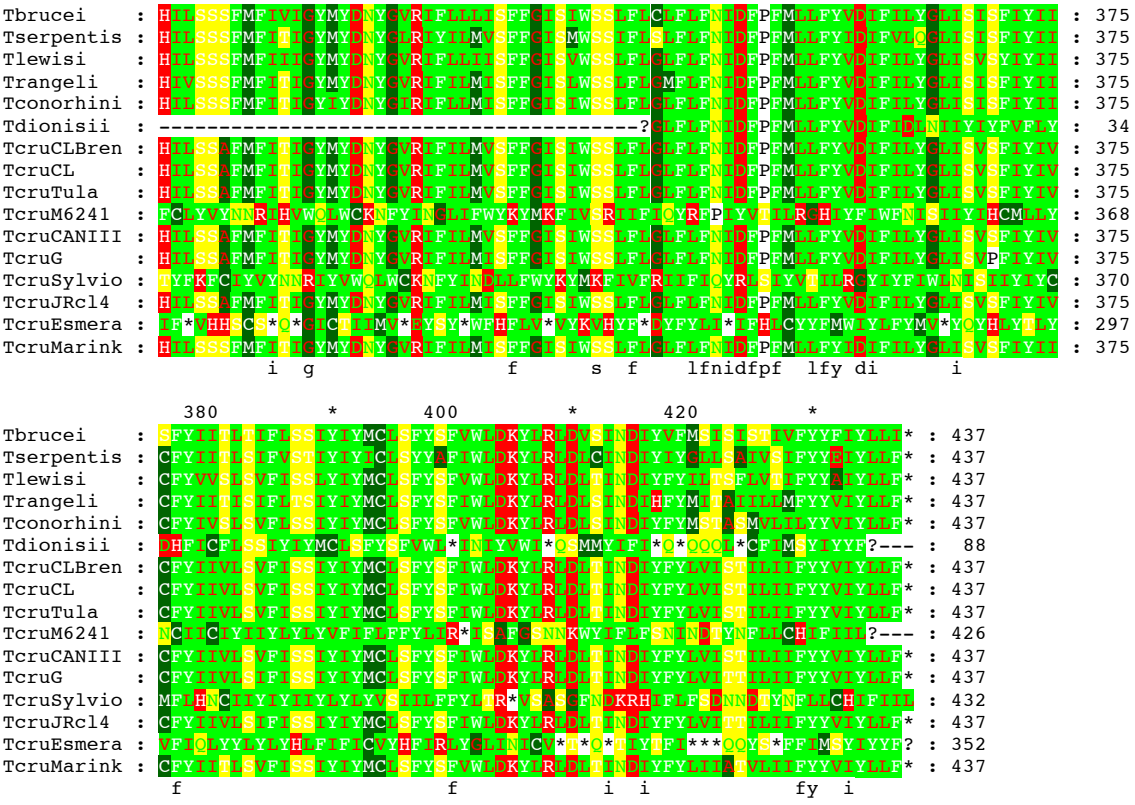

Tlewisi : 674
Trangeli : 674
Tconorhini : 674
Tdionisii : -
TcruCLBren : 674
TcruCL : 674
TcruTula : 674
TcruM6241 : 663
TcruCANIII : 674
TcruG : 674
TcruSylvio : 675
TcruJRcl4 : 674
TcruEsmera : 574
TcruMarink : 674
catgt gaagtaaa ac gaaatgagtgt tattagcaagtgt gt t aaaa tagg ttttttgg tat
680 * 700 * 720 * 740 *
Tbrucei : 749
Tserpentis : 749
Tlewisi : 749
Trangeli : 749
Tconorhini : 749
Tdionisii : -
TcruCLBren : 749
TcruCL : 749
TcruTula : 749
TcruM6241 : 738
TcruCANIII : 749
TcruG : 749
TcruSylvio : 750
TcruJRcl4 : 749
TcruEsmera : 649
TcruMarink : 749
aaatt t tt at tcatt aa a tatca tatgattttt gg ttt taga a at at at t gg
760 * 780 * 800 * 820
Tbrucei : 824
Tserpentis : 824
Tlewisi : 824
Trangeli : 824
Tconorhini : 824
Tdionisii : -
TcruCLBren : 824
TcruCL : 824
TcruTula : 824
TcruM6241 : 813
TcruCANIII : 824
TcruG : 824
TcruSylvio : 825
TcruJRcl4 : 824
TcruEsmera : 724
TcruMarink : 824
ataac tt ttatc at tc ta tatttttatc gatta aaaaaaataatagc aa tgatc t ataca
* 840 * 860 * 880 * 900
Tbrucei : 899
Tserpentis : 899
Tlewisi : 899
Trangeli : 899
Tconorhini : 899
Tdionisii : -
TcruCLBren : 899
TcruCL : 899
TcruTula : 899
TcruM6241 : 888
TcruCANIII : 899
TcruG : 899
TcruSylvio : 900
TcruJRcl4 : 899
TcruEsmera : 799
TcruMarink : 899
ac ggtat g attaatattatt tgacataa ga ctttt t tatagg t t t t tgtaat tatc
* 920 * 940 * 960 *
Tbrucei : 974
Tserpentis : 974
Tlewisi : 974
Trangeli : 974


Tconorhini : 974
Tdionisii : -
TcruCLBren : 974
TcruCL : 974
TcruTula : 974
TcruM6241 : 963
TcruCANIII : 974
TcruG : 974
TcruSylvio : 975
TcruJRcl4 : 974
TcruEsmera : 874
TcruMarink : 974
980
1000
1020
1040
Tbrucei : 1049
Tserpentis : 1049
Tlewisi : 1049
Trangeli : 1049
Tconorhini : 1049
Tdionisii : 27
TcruCLBren : 1049
TcruCL : 1049
TcruTula : 1049
TcruM6241 : 1038
TcruCANIII : 1049
TcruG : 1049
TcruSylvio : 1050
TcruJRcl4 : 1049
TcruEsmera : 949
TcruMarink : 1049
1060
1080
1100
1120
Tbrucei : 1124
Tserpentis : 1124
Tlewisi : 1124
Trangeli : 1124
Tconorhini : 1124
Tdionisii : 91
TcruCLBren : 1124
TcruCL : 1124
TcruTula : 1124
TcruM6241 : 1113
TcruCANIII : 1124
TcruG : 1124
TcruSylvio : 1125
TcruJRcl4 : 1124
TcruEsmera : 1024
TcruMarink : 1124
1140
1160
1180
1200
Tbrucei : 1199
Tserpentis : 1199
Tlewisi : 1199
Trangeli : 1199
Tconorhini : 1199
Tdionisii : 165
TcruCLBren : 1199
TcruCL : 1199
TcruTula : 1199
TcruM6241 : 1188
TcruCANIII : 1199
TcruG : 1199
TcruSylvio : 1200
TcruJRcl4 : 1199
TcruEsmera : 1099
TcruMarink : 1199
1220
1240
1260
Tbrucei : 1272
Tserpentis : 1272
Tlewisi : 1272
Trangeli : 1272
Tconorhini : 1272
Tdionisii : 240



4.SI-S6a: GeneDoc ND4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1.


Tbrucei
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink



4.SI-S6b: GeneDoc ND4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. This alignment accounts for the frameshifts caused by the indels [shown in the corresponding nucleotide sequence alignment above (S6a)] and shows the resultant truncated proteins in the alternate frame. '*' indicates stop codon in the alignment. Several stop codons are seen in the altered frame of TcruM6241 and TcruSylvio, TcruEsmera and Tdionisii.

4.SI-S6c: ND4 nucleotide percent identity matrix

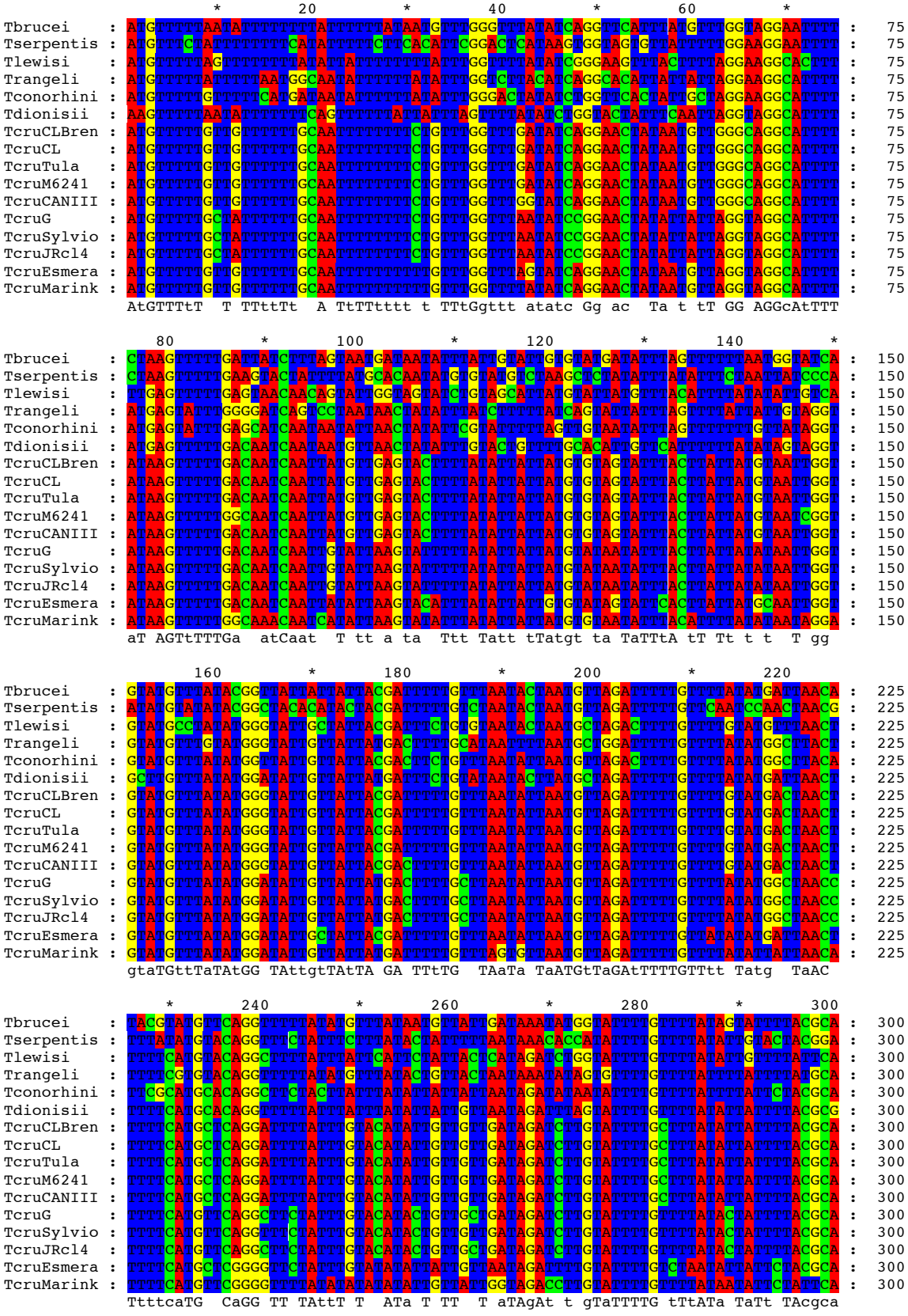
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm**	*TcMar*
Tbru	**100**															
Tser	73	**100**														
Tlew	78	72	**100**													
Tran	79	73	78	**100**												
Tcon	79	72	80	82	**100**											
Tdio*	82	75	78	80	82	**100**										
TcCLB	78	73	77	78	79	84	**100**									
TcCL	78	73	77	78	79	84	100	**100**								
TcTula	78	73	77	78	79	84	100	100	**100**							
TcM62	78	72	77	78	79	84	100	100	100	**100**						
TcCAN	78	73	78	78	79	85	99	99	99	99	**100**					
TcG	78	73	78	79	79	84	92	92	92	92	92	**100**				
TcSyl	78	73	78	79	79	84	92	92	92	92	92	99	**100**			
TcJR	78	73	78	79	79	84	93	93	93	92	93	99	99	**100**		
TcEsm*	78	74	78	79	80	84	88	88	88	88	89	89	89	89	**100**	
TcMar	80	73	80	80	81	87	88	88	88	87	88	88	88	88	88	**100**

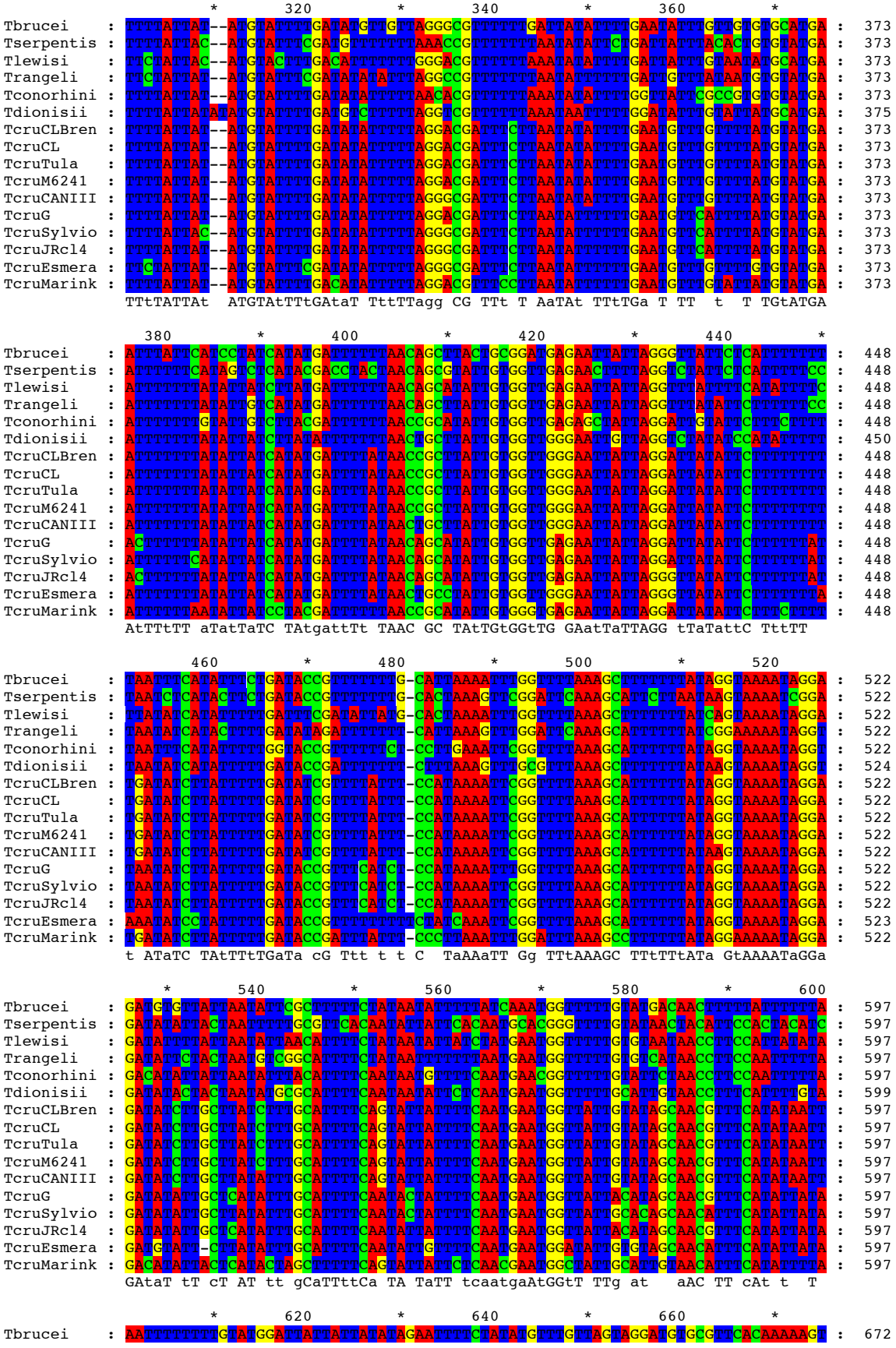
* For the available lengths of *TcEsm* and *Tdio* ND4 transcripts.

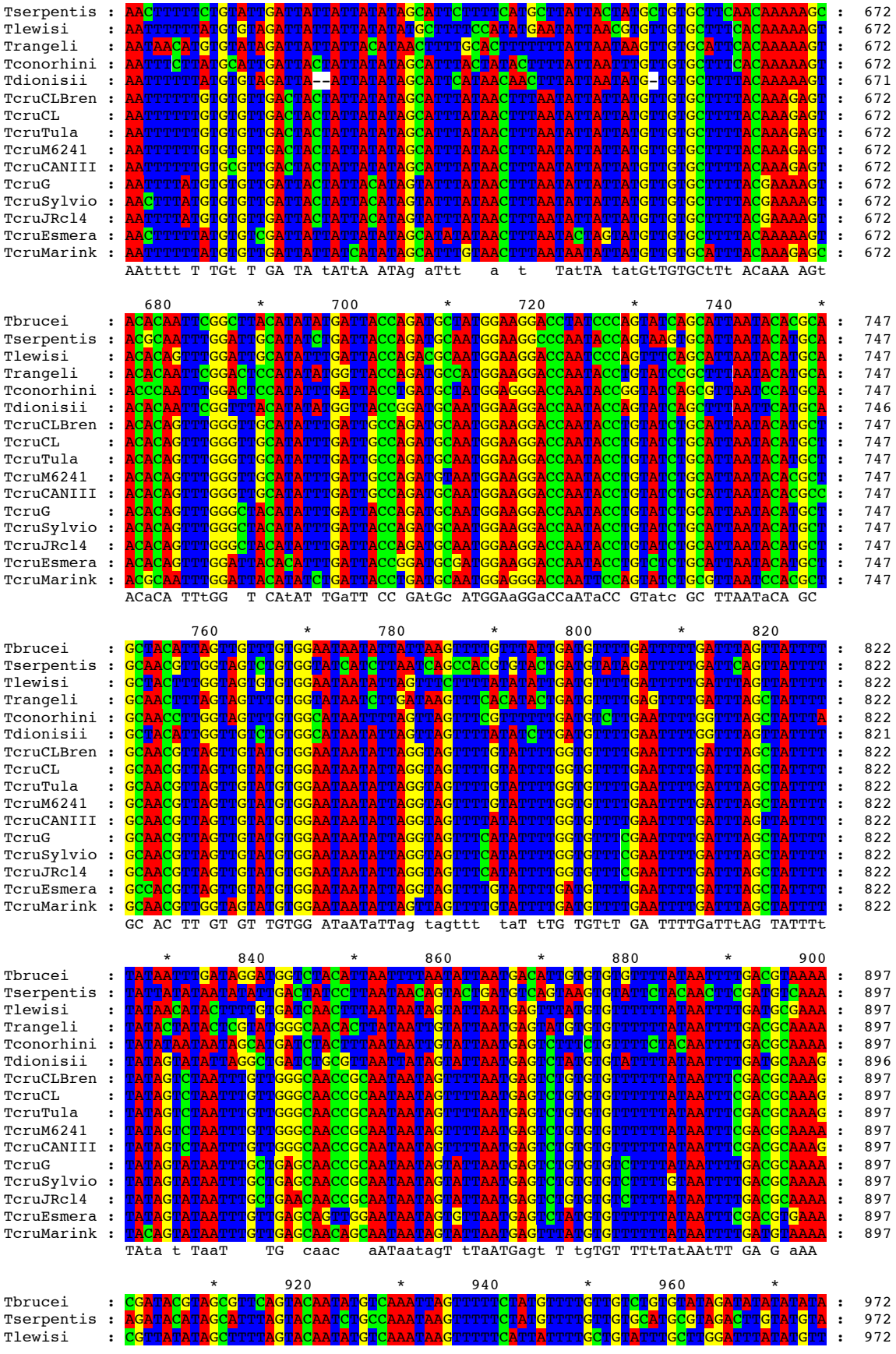
4.SI-S6d: ND4 protein percent identity matrix

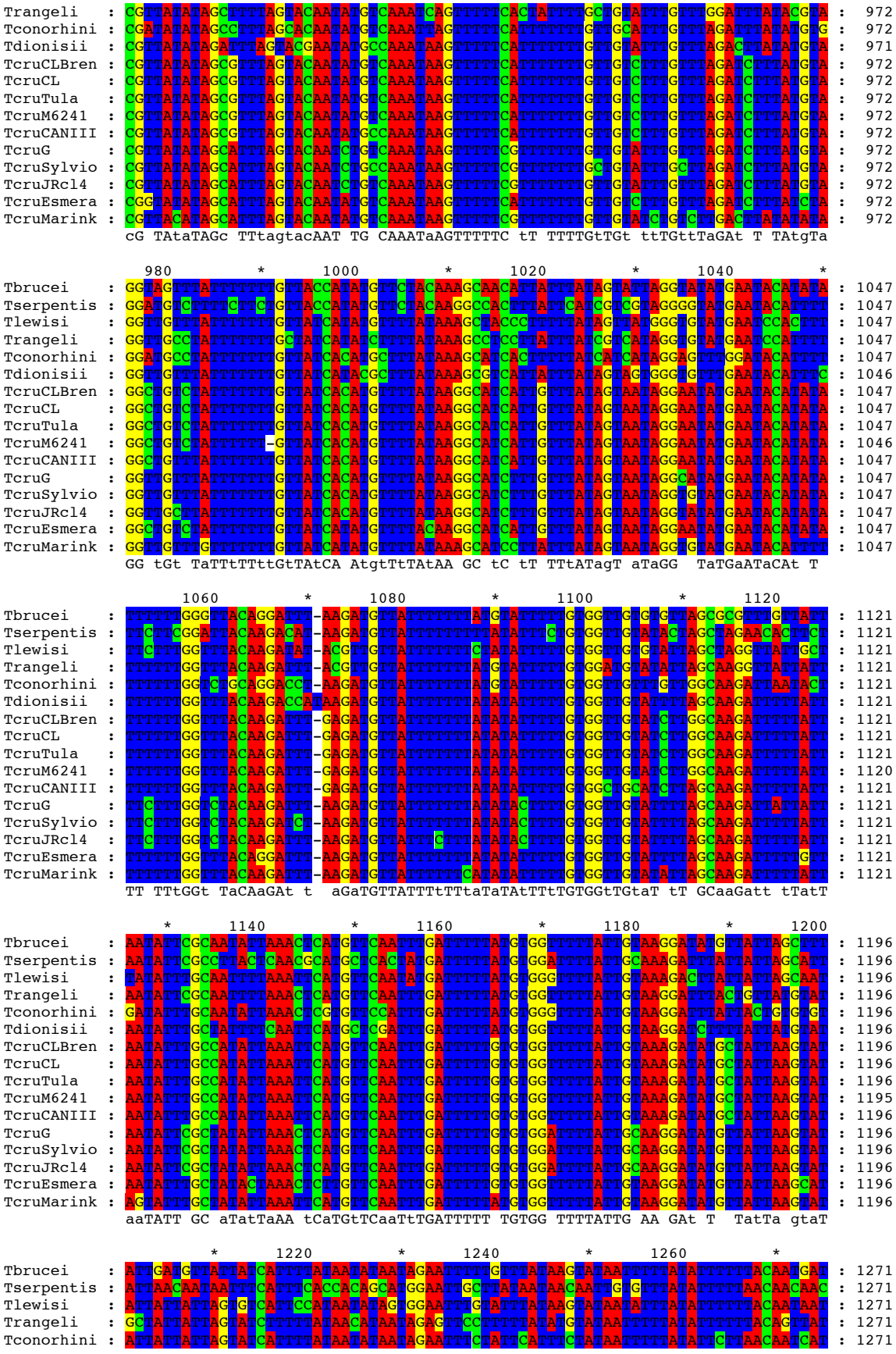
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm**	*TcMar*
Tbru	**100**															
Tser	72	**100**														
Tlew	76	72	**100**													
Tran	78	72	80	**100**												
Tcon	80	73	83	86	**100**											
Tdio*	32	31	30	33	33	**100**										
TcCLB	76	73	79	79	82	34	**100**									
TcCL	76	73	79	79	82	34	100	**100**								
TcTula	76	73	79	79	82	34	100	100	**100**							
TcM62	23	21	21	23	23	15	31	31	31	**100**						
TcCAN	76	72	79	80	82	34	100	100	100	31	**100**					
TcG	77	72	79	80	82	34	95	95	95	29	95	**100**				
TcSyl	42	37	40	42	43	5	50	50	50	27	50	53	**100**			
TcJR	77	72	79	80	82	34	95	95	95	29	95	100	53	**100**		
TcEsm*	9	8	10	10	10	9	9	9	9	6	9	10	10	10	**100**	
TcMar	78	74	81	82	83	32	90	90	90	27	90	91	47	91	10	**100**

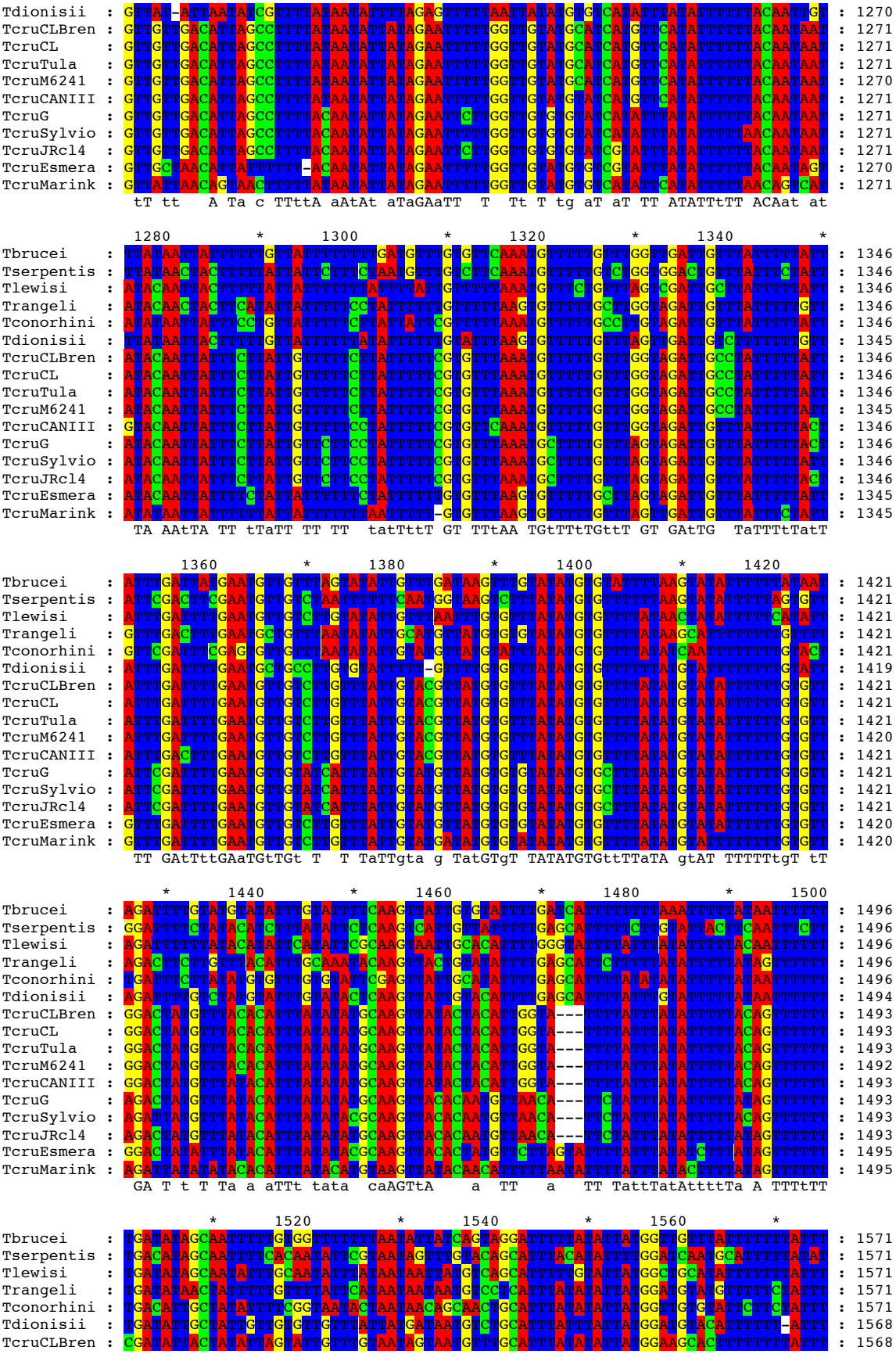
* For the available lengths of *TcEsm* and *Tdio* ND4 transcripts. This matrix shows the identities of the truncated ND4 proteins.

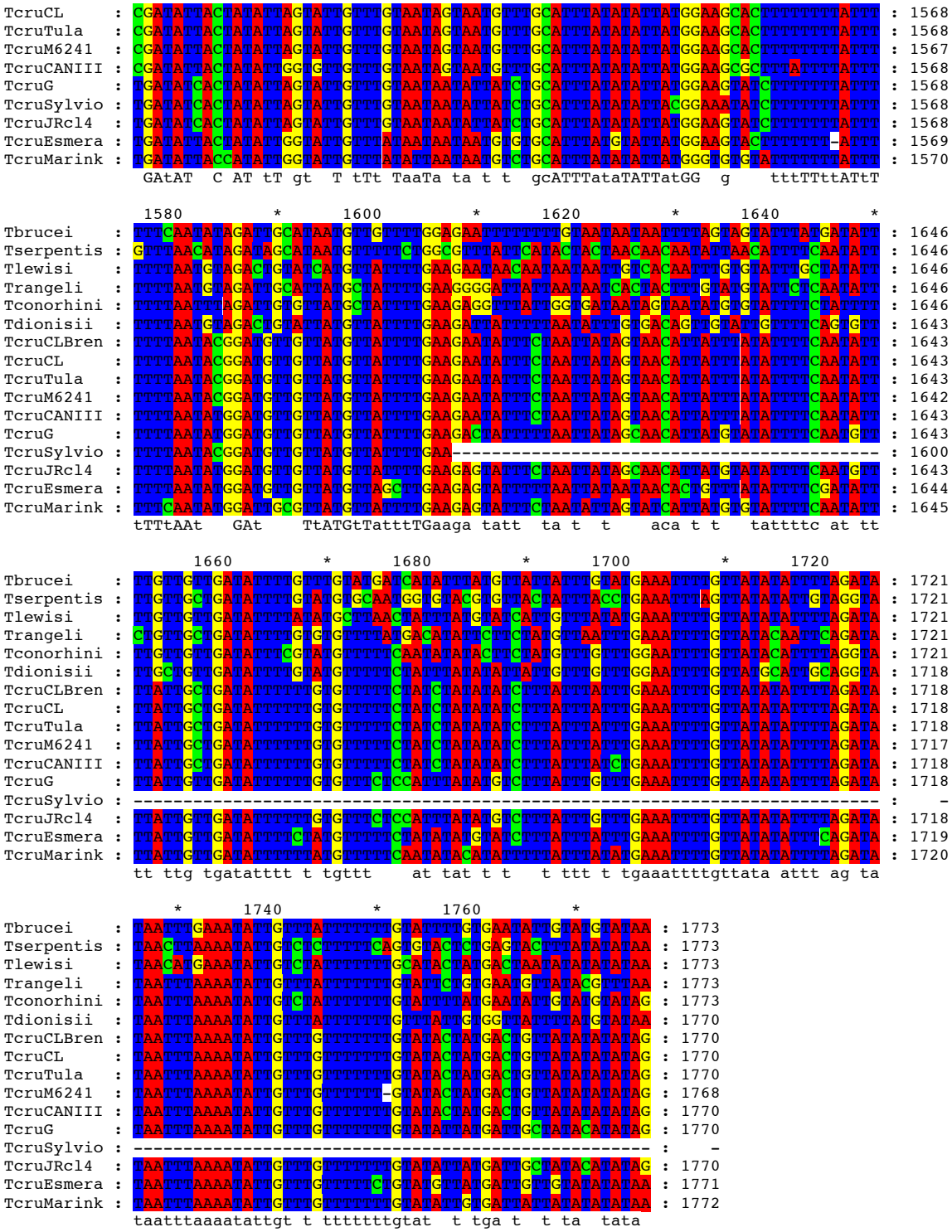

Tbrucei
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink



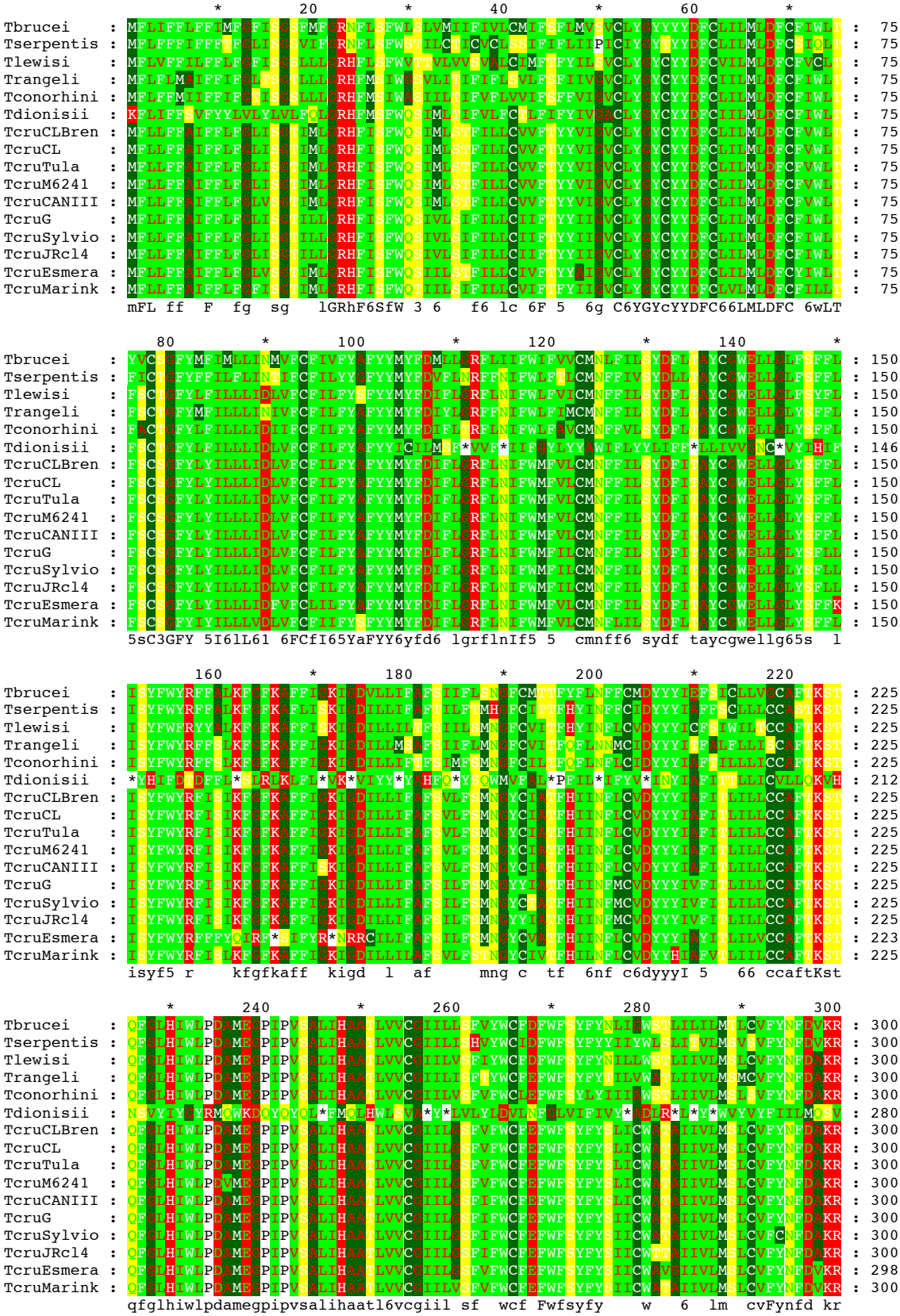

Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink
Tbrucei
680
700
720
740
760
780
800
820
840
860
880
900
920
940
960

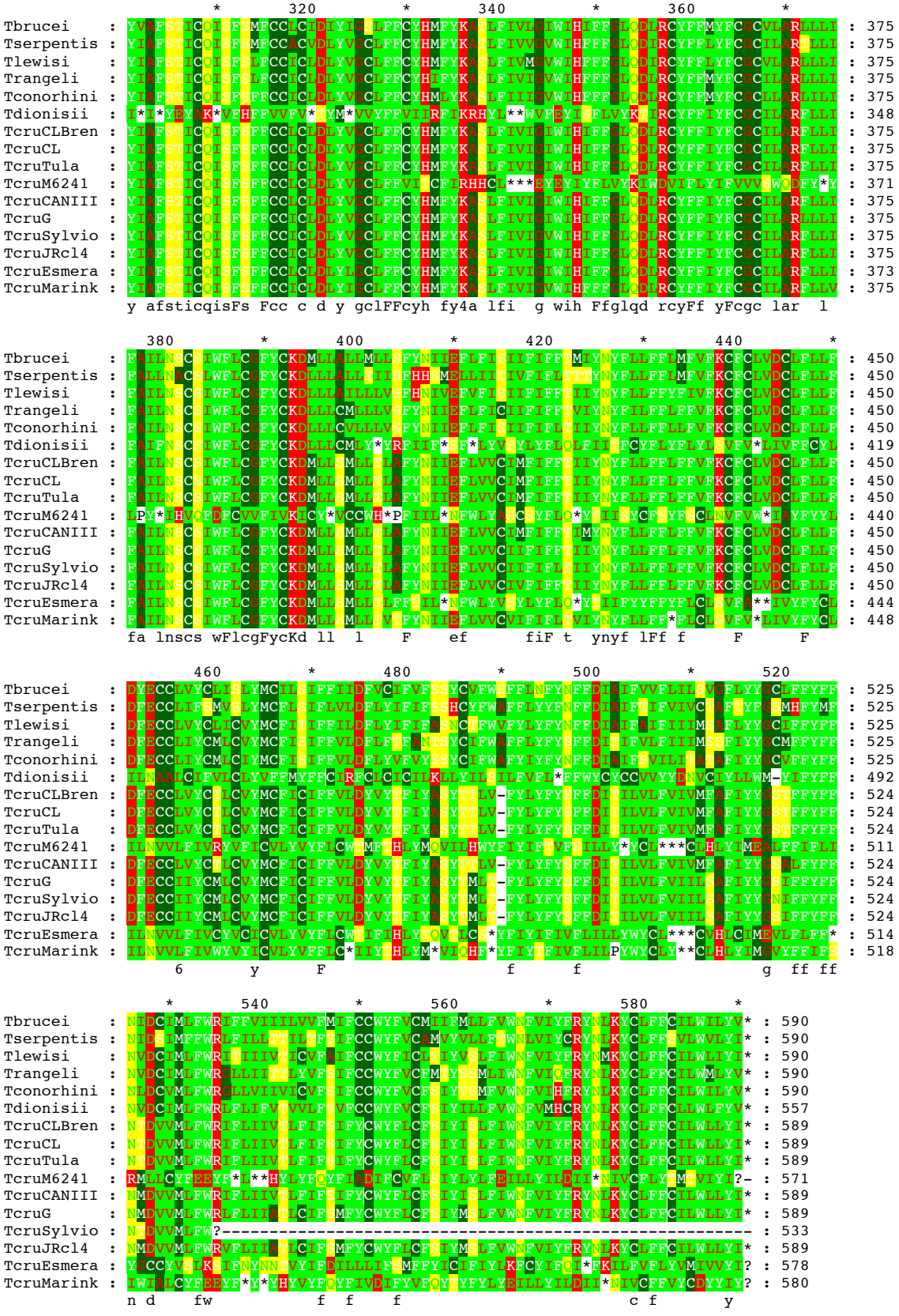

Trangeli : 972
Tconorhini : 972
Tdionisii : 971
TcruCLBren : 972
TcruCL : 972
TcruTula : 972
TcruM6241 : 972
TcruCANIII : 972
TcruG : 972
TcruSylvio : 972
TcruJRcl4 : 972
TcruEsmera : 972
TcruMarink : 972
cG TAtaTAGc TTtagtacAAT TG CAAATaAGTTTTC tT TTTTGtTGt ttTGttTaGAt T TAtgTa
980 * 1000 * 1020 * 1040 *
Tbrucei : 1047
Tserpentis : 1047
Tlewisi : 1047
Trangeli : 1047
Tconorhini : 1047
Tdionisii : 1046
TcruCLBren : 1047
TcruCL : 1047
TcruTula : 1047
TcruM6241 : 1046
TcruCANIII : 1047
TcruG : 1047
TcruSylvio : 1047
TcruJRcl4 : 1047
TcruEsmera : 1047
TcruMarink : 1047
GG tGt TaTTtTTttGtTAtCA AtgtTtTAtAA GC tC tT TTtATagT aTaGG TaTGaATaCAt T
1060 * 1080 * 1100 * 1120
Tbrucei : 1121
Tserpentis : 1121
Tlewisi : 1121
Trangeli : 1121
Tconorhini : 1121
Tdionisii : 1121
TcruCLBren : 1121
TcruCL : 1121
TcruTula : 1121
TcruM6241 : 1120
TcruCANIII : 1121
TcruG : 1121
TcruSylvio : 1121
TcruJRcl4 : 1121
TcruEsmera : 1121
TcruMarink : 1121
TT TTtGGt TaCAaGAt t aGaTGTTATTTtTTtaTaTAtTTtTGTGGtTGtaT tT GCaaGatt tTatT
* 1140 * 1160 * 1180 * 1200
Tbrucei : 1196
Tserpentis : 1196
Tlewisi : 1196
Trangeli : 1196
Tconorhini : 1196
Tdionisii : 1196
TcruCLBren : 1196
TcruCL : 1196
TcruTula : 1196
TcruM6241 : 1195
TcruCANIII : 1196
TcruG : 1196
TcruSylvio : 1196
TcruJRcl4 : 1196
TcruEsmera : 1196
TcruMarink : 1196
aaTATT GC aTatTaAA tCaTGtTCaaTtTGATTTTT TGTGG TTTTATTG AA GAt T TatTa gtaT
* 1220 * 1240 * 1260 *
Tbrucei : 1271
Tserpentis : 1271
Tlewisi : 1271
Trangeli : 1271
Tconorhini : 1271


Tdionisii : GTTAT-ATTAATATCGTTTTATAATATTTTAGAGTTTTTAATTATATGTGTCATATTTATATTTTTTACAATTGT : 1270
TcruCLBren : GTTGTTGACATTAGCCTTTTATAATATTATAGAATTTTTGGTTGTATGCATCATGTTCATATTTTTTACAATAAT : 1271
TcruCL : GTTGTTGACATTAGCCTTTTATAATATTATAGAATTTTTGGTTGTATGCATCATGTTCATATTTTTTACAATAAT : 1271
TcruTula : GTTGTTGACATTAGCCTTTTATAATATTATAGAATTTTTGGTTGTATGCATCATGTTCATATTTTTTACAATAAT : 1271
TcruM6241 : GTTGTTGACATTAGCCTTTTATAATATTATAGAATTTTTGGTTGTATGCATCATGTTCATATTTTTTACAATAAT : 1270
TcruCANIII : GTTGTTGACATTAGCCTTTTATAATATTATAGAATTTTTGGTTGTATGTATCATGTTCATATTTTTTACAATAAT : 1271
TcruG : GTTGTTGACATTAGCCTTTTACAATATTATAGAATTCTTGGTTGTGGTGTATCATATTTATATTTTTTACAATAAT : 1271
TcruSylvio : GTTGTTGACATTAGCCTTTTACAATATTATAGAATTTTTGGTTGTGGTGTATCATATTTATATTTTTAACAATAAT : 1271
TcruJRcl4 : GTTGTTGACATTAGCCTTTTACAATATTATAGAATTCTTGGTTGTGGTGTATCGTATTTATATTTTTTACAATAAT : 1271
TcruEsmera : GTTGCTAACATTATTTTT-ACAATATTATAGAATTTTGGTTGTATGTGTCGTATTTATATTTTTTACAATAGT : 1270
TcruMarink : GTTATTAACACTAACTTTTTATAATATTATAGAATTTTTGGTTGTATGTGTCATATTCATATTTTTAACAGTCAT : 1271
tT tt A Ta c TTttA aAtAt aTaGAaTT T Tt T tg aT aT TT ATATTtTT ACAat at
1280 * 1300 * 1320 * 1340 *
Tbrucei : TTATAATTATTTTTGTTATTTTTTTGATGTTTGTGTTCAAATGTTTTGTTTGGTTGATTGTTTATTTTTATT : 1346
Tserpentis : TTATAACTACTTTTTATTATTCTTTCTAATGTTTGTCTTCAAATGTTTTGTCTGGTGGACTGTTTATTTCTATT : 1346
Tlewisi : ATACAATTACTTTTTATTATTTTTTTATTTATTGTTTTTAAATGTTTCTGTTTAGTCGATTGCTTATTTTTATT : 1346
Trangeli : ATACAACTACTTCATATTATTTTTCCTATTTTTTGTTTTTAAGTGTTTTGCTTGGTAGATTGTTTATTTTTGTT : 1346
Tconorhini : ATATAATTATTTCCTGTTATTTTTCTTATTATTCGTTTTTAAATGTTTTGCCTTGGTAGATTGTTTATTTTTATT : 1346
Tdionisii : TTATAATTACTTTTTGTATTTTTTATTTTTTGTATTTAAGTGTTTTGTTTAGTTGATTTAGTTTTTTTGTT : 1345
TcruCLBren : ATACAATTATTTCTTATTGTTTTTCTTATTTTTCGTGTTTAAATGTTTTGTTTGGTAGATTGCCTATTTTTATT : 1346
TcruCL : ATACAATTATTTCTTATTGTTTTTCTTATTTTTCGTGTTTAAATGTTTTGTTTGGTAGATTGCCTATTTTTATT : 1346
TcruTula : ATACAATTATTTCTTATTGTTTTTCTTATTTTTCGTGTTTAAATGTTTTGTTTGGTAGATTGCCTATTTTTATT : 1346
TcruM6241 : ATACAATTATTTCTTATTGTTTTTCTTATTTTTCGTGTTTAAATGTTTTGTTTGGTAGATTGCCTATTTTTATT : 1345
TcruCANIII : GTACAATTATTTCTTATTGTTTTTCCTATTTTTCGTGTTTCAAATGTTTTGTTTGGTAGATTGTTTATTTTTACT : 1346
TcruG : ATACAATTATTTCTTATTGTTCTTCCTATTTTTCGTGTTTAAATGCTTTTGTTTAGTAGATTGTTTATTTTTACT : 1346
TcruSylvio : ATACAATTATTTCTTATTGTTCTTCCTATTTTTCGTGTTTAAATGCTTTTGTTTAGTAGATTGTTTATTTTTATT : 1346
TcruJRcl4 : ATACAATTATTTCTTATTGTTCTTCCTATTTTTCGTGTTTAAATGCTTTTGTTTAGTAGATTGTTTATTTTTACT : 1346
TcruEsmera : ATACAATTATTTTCTATTATTTTTTCTATTTTTTGTGTTTAAGTGTTTTGCTTAGTAGATTGTTTATTTTTATT : 1345
TcruMarink : ATATAATTATTTTTTATTATTTTTTTAATTTTT-GTGTTTAAGTGTTTTGTTTAGTTGTTTATTTCTATT : 1345
TA AAtTA TT tTaTT TT TT tatTttT GT TTtAA TGtTTtTGttT GT GAtTG TaTTTtTatT
1360 * 1380 * 1400 * 1420
Tbrucei : ATTTGATTATGAATGTTGTTTTAGTATATTGTTTGATAAGTTTGTATATGTGTATTTAAGTATATTTTTATAAT : 1421
Tserpentis : ATTCGACTTCGAATGTTGTCTAATTTTTCAATGGTAAGTCTTTATATGTGTTTTTAAGTATATTTTAGTGTT : 1421
Tlewisi : ATTTGATTTTGAATGTTGTCTTGTATATTGTTTAATTTGTGTTTATATGTGTTTTATAACTATATTTTCATATT : 1421
Trangeli : GTTTGACTTTGAATGCTGTTTAAATATTGCATGTTATGTGTGTATATGTGTTTTATAACCATTTTTTTGTTTT : 1421
Tconorhini : GTTCGATTTCGAGTGTTCTTTAAATATTGTATGTTATGTATTTATATGTGTTTTATATCAATTTTTTTGTACT : 1421
Tdionisii : ATTTGATTTTGAATGCTGCCTTTGTGTATTTTT-GTTTTGTGTTTATATGTGTTTTATATGTATTTTTTGTGT : 1419
TcruCLBren : ATTTGATTTTGAATGTTGTCTTGTTTATTGTACGTTATGTGTTTATATGTGTTTTATATGTATATTTTTTGTGT : 1421
TcruCL : ATTTGATTTTGAATGTTGTCTTGTTTATTGTACGTTATGTGTTTATATGTGTTTTATATGTATATTTTTTGTGT : 1421
TcruTula : ATTTGATTTTGAATGTTGTCTTGTTTATTGTACGTTATGTGTTTATATGTGTTTTATATGTATATTTTTTGTGT : 1421
TcruM6241 : ATTTGATTTTGAATGTTGTCTTGTTTATTGTACGTTATGTGTTTATATGTGTTTTATATGTATATTTTTTGTGT : 1420
TcruCANIII : ATTTGACTTTGAATGTTGTCTTGTTTATTGTACGTTATGTGTTTATATGTGTTTTATATGTATATTTTTTGTGT : 1421
TcruG : ATTCGATTTTGAATGTTGTATCATTTATTGTATGTTATGTGTGTATATGTGCTTTATATGTATATTTTTTGTGT : 1421
TcruSylvio : ATTCGATTTTGAATGTTGTATCATTTATTGTATGTTATGTGTGTATATGTGCTTTATATGTATATTTTTTGTGT : 1421
TcruJRcl4 : ATTCGATTTTGAATGTTGTATCATTTATTGTATGTTATGTGTGTATATGTGCTTTATATGTATATTTTTTGTGT : 1421
TcruEsmera : GTTTGATTTTGAATGTTGTCTTGTTTATTGTATGTTATGTGTGTATATGTGTTTTATATGTATATTTTTTGTGT : 1420
TcruMarink : GTTTGATTTTGAATGTTGTCTTGTTTATTGTATGATATGTGTATATATGTGTTTTATATGTATTTTTTTGTGT : 1420
TT GAtTttGAaTGtTGt T T TaTTgta g TatGTgT TATATGTGTttTTaTA gtAT TTTTTtgT tT
* 1440 * 1460 * 1480 * 1500
Tbrucei : AGATTTTGTATGTATATTTGTATTTCTCAAGTTATTGTGTATTTTGATCATTTTTTTAAATTTTTATAATTTTT : 1496
Tserpentis : GGATTTTCTATACATCTTTATATTCTCAAGTCATTGTTATTTTGAGCATTTTTCTTGTATTACTTCAATTTCTT : 1496
Tlewisi : AGATTTTTTATACATATTCATATTCGCAAGTAATTGCACATTTTGGGTATTTTATTTATATTTTTACAATTTTTT : 1496
Trangeli : AGACTTCTTGTTTACATTTGCAAATACAAGTTACTGTATATTTTGAGCATTCTTTTTATATTTTTATAGTTTTTT : 1496
Tconorhini : TGATTTCTTATATGTGTTTGTGTATTCGAGTTATTGCATATTTTGAGCATTTTATATATTTTTATAATTTTTT : 1496
Tdionisii : AGATTTTGTCTATGTATTTTGTATACTCAAGTTATTGTGTACATTTTGAGCATTTTTATTTTTATAATTTTTT : 1494
TcruCLBren : GGACTATGTTTACACATTTATATATGCAAGTTATACTACATTGGTA---TTTTATTTATATTTTTACAGTTTTTT : 1493
TcruCL : GGACTATGTTTACACATTTATATATGCAAGTTATACTACATTGGTA---TTTTATTTATATTTTTACAGTTTTTT : 1493
TcruTula : GGACTATGTTTACACATTTATATATGCAAGTTATACTACATTGGTA---TTTTATTTATATTTTTACAGTTTTTT : 1493
TcruM6241 : GGACTATGTTTACACATTTATATATGCAAGTTATACTACATTGGTA---TTTTATTTATATTTTTACAGTTTTTT : 1492
TcruCANIII : GGACTATGTTTATACATTTATATATGCAAGTTATACTACATTGGTA---TTTTATTTATATTTTTACAGTTTTTT : 1493
TcruG : AGACTATGTTTATACATTTATATATGCAAGTTACACAATGTTAACA---TTCTATTTATATTTTTATAGTTTTTT : 1493
TcruSylvio : AGATTATGTTTATACATTTATATACGCAAGTTACACAATGTTAACA---TTCTATTTATATTTTTACAGTTTTTT : 1493
TcruJRcl4 : AGACTATGTTTATACATTTATATATGCAAGTTACACAATGTTAACA---TTCTATTTATATTTTTATAGTTTTTT : 1493
TcruEsmera : GGACTATATTTATACACATTTATATACGCAAGTTACACTATGTTCTTAGTATTTTATTTATATCTTTTATAGTTTTTT : 1495
TcruMarink : AGATTATATATACACATTTATACATGTAAGTTATACAACATTTTTAATATTTTATTTATATCTTTTTTTTTTTT : 1495
GA T t T Ta a aTTt tata caAGTtA a TT a TT TattTatAttttTa A TTTtTT
* 1520 * 1540 * 1560 *
Tbrucei : TGATATAGCAATTTTTGTGGTTTTTTTAATATTATCAGTAGGATTTTTATATTATGGTTGTTTATTTTTTATTT : 1571
Tserpentis : TGACATAGCAATTTTCACAATATTCGTAATAGTTTGTACAGCATTTACATATTTTGGATCAATGCATTTTATATT : 1571
Tlewisi : TGATATAGCAATATTTGCAATATTTATAATAATTATTGCAGCATTTTTGTATTATGGCTGCATATTTTTTATTT : 1571
Trangeli : TGATATAACTATTTTTCTTTTATTCATAATAATAATGTCCTCATTTATATATTATGGATGCTATGTTTTCTATTT : 1571
Tconorhini : TGACATTGCTATATTTTCGGGTAAAACTAATAACAGCAACTGCATTTATATATTATGGTTGCTGTATTCTTCTATTT : 1571
Tdionisii : TGATATTGCTATTGTTGTGTTGTTTATTAGATAAATGTCTGCATTTATTTATTATGGAATGCACATTTTTT-ATTT : 1568
TcruCLBren : CGATATTACTATATTACATTGTTTGTAATAGTAATGTTTGCATTTATATATTATGGAAGCACTTTTTTTATTT : 1568



4.SI-S7a: GeneDoc ND5 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. *TcSyl* ND5 is incomplete due to the incomplete maxicircle.





4.SI-S7b: GeneDoc ND5 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. This alignment accounts for the frameshifts caused by the indels [shown in the corresponding nucleotide sequence alignment above (S7a)] and shows the resultant truncated proteins in the alternate frame. '*' indicates stop codon in the alignment. Several stop codons are seen in the altered frame of TcruM6241, TcruEsmera, TcruMarink and Tdionisii. '?' in TcruSylvio is due to incomplete gene information.

4.SI-S7c. ND5 nucleotide percent identity matrix.

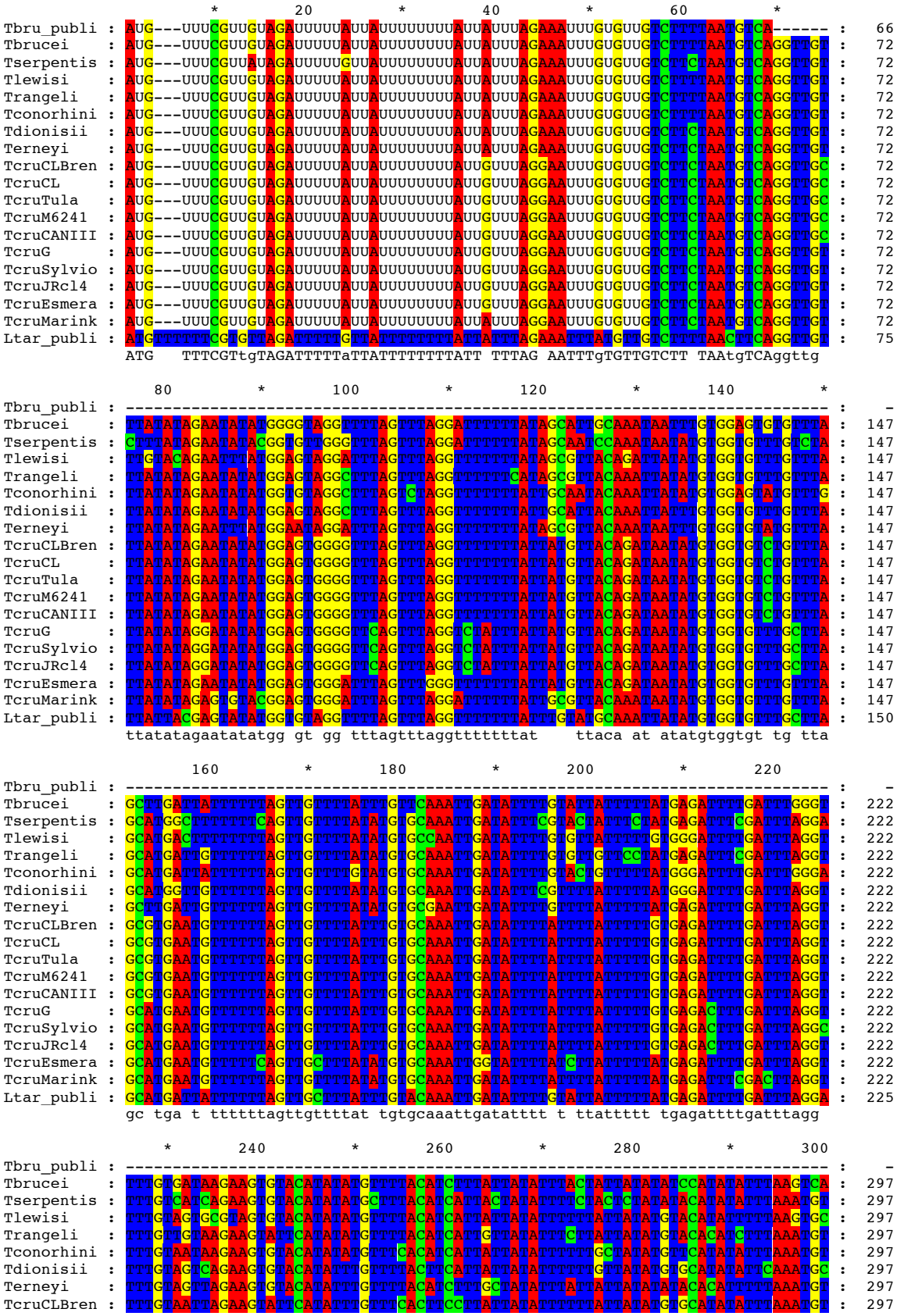
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl**	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	73	**100**														
Tlew	79	73	**100**													
Tran	77	70	79	**100**												
Tcon	78	71	77	80	**100**											
Tdio	78	70	80	78	78	**100**										
TcCLB	78	71	79	79	78	80	**100**									
TcCL	78	71	79	79	78	80	100	**100**								
TcTula	78	71	79	79	78	80	100	100	**100**							
TcM62	78	71	79	79	78	80	100	100	100	**100**						
TcCAN	78	71	79	78	78	80	98	98	98	98	**100**					
TcG	78	71	78	80	79	80	92	92	92	91	91	**100**				
TcSyl*	78	71	78	80	79	80	91	91	91	91	91	98	**100**			
TcJR	78	71	78	80	79	80	92	92	92	91	91	99	98	**100**		
TcEsm	78	71	78	79	78	81	91	91	91	91	90	89	89	89	**100**	
TcMar	79	71	80	79	79	81	88	88	88	88	87	87	86	87	87	**100**

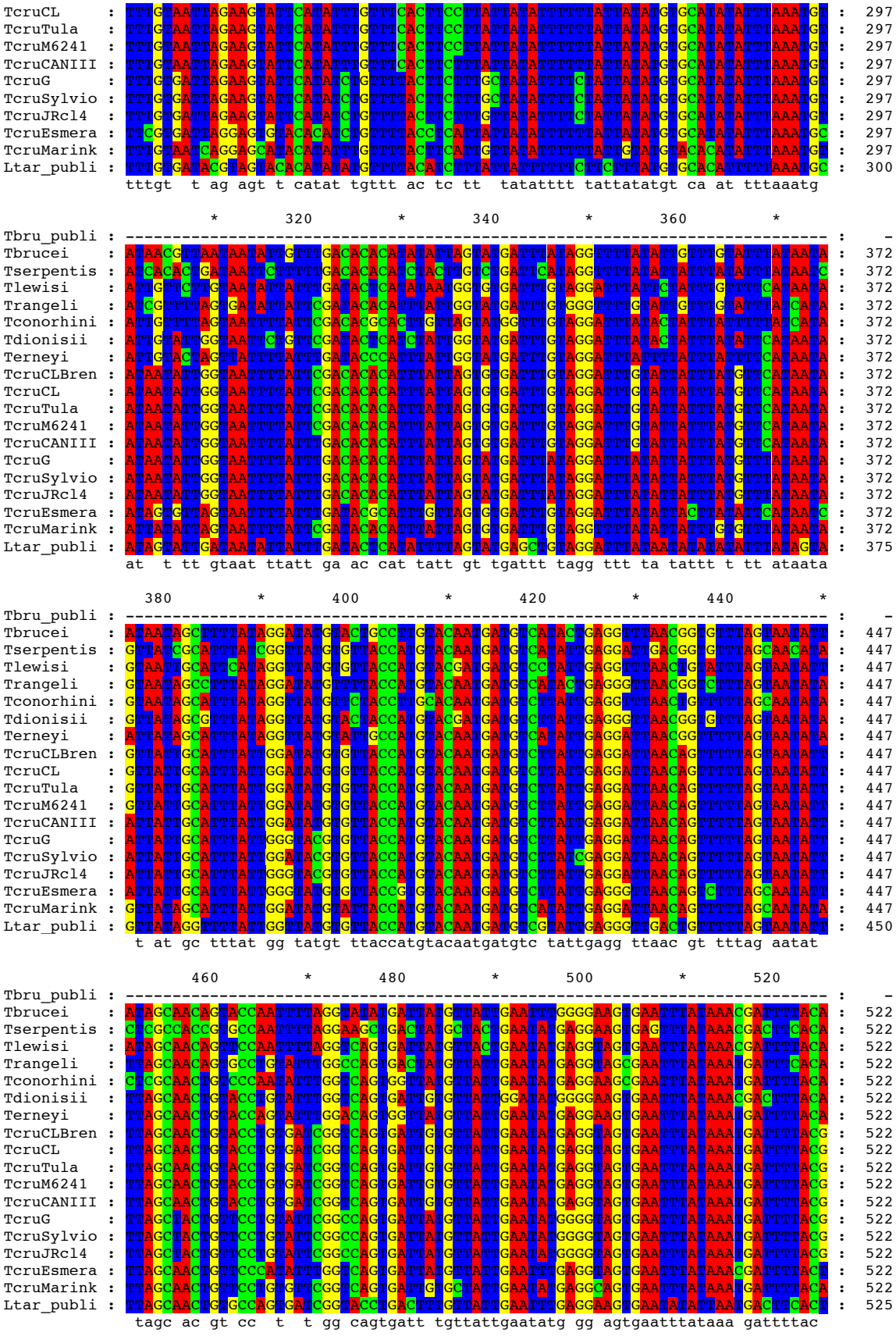
* For the available length of *TcSyl* ND5 transcript.

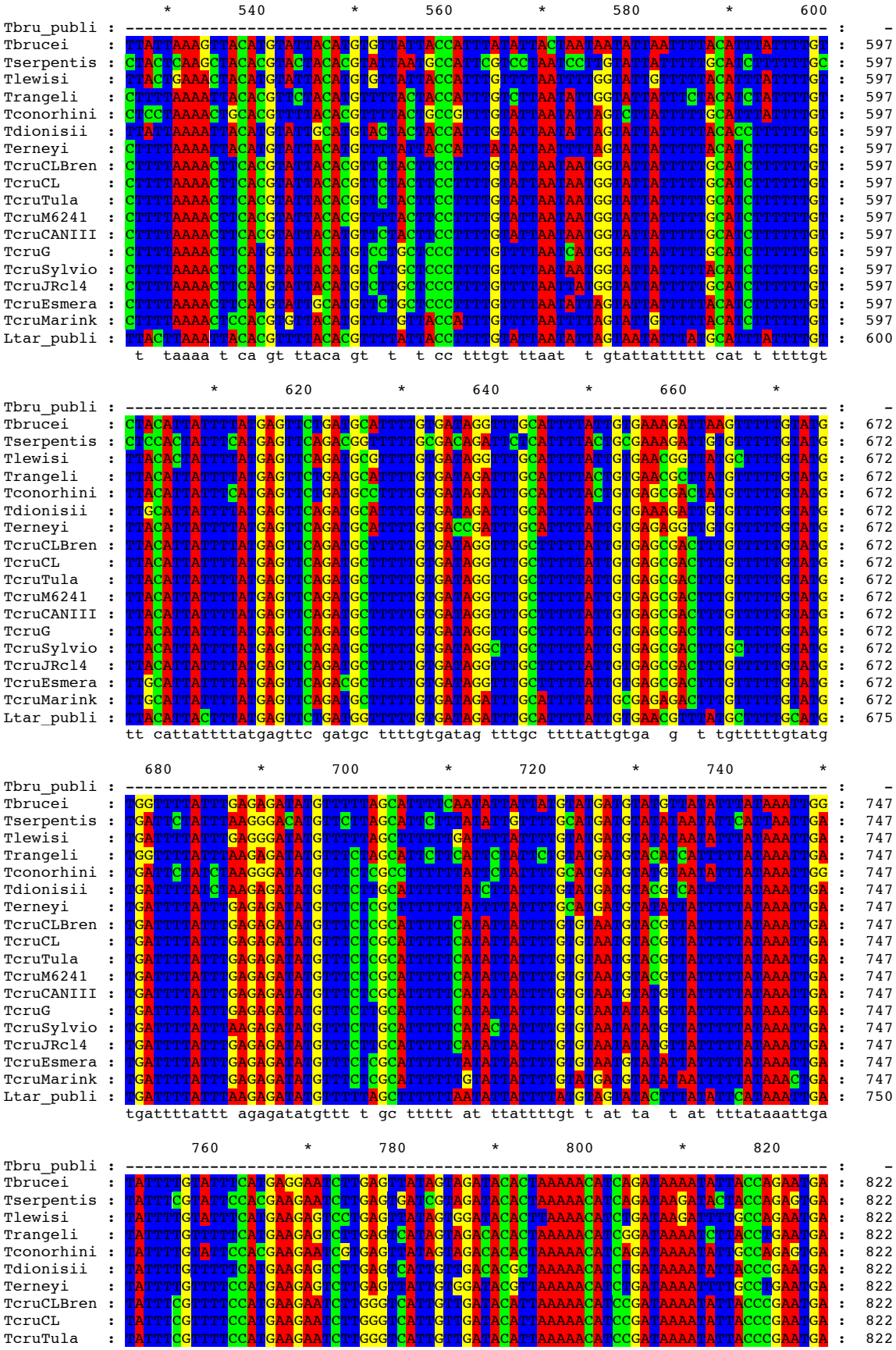
4.SI-S7d. ND5 protein percent identity matrix.

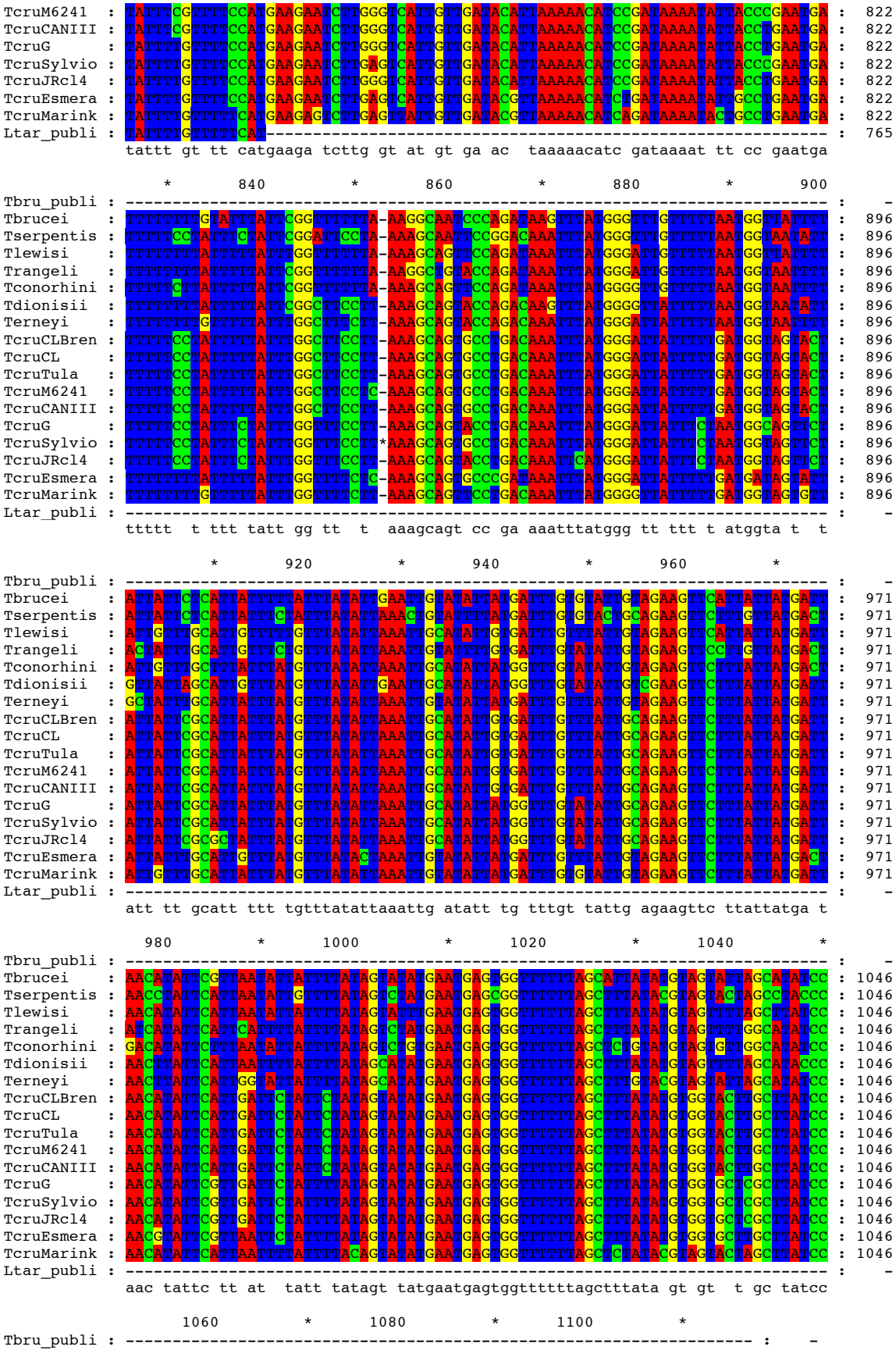
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl**	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	72	**100**														
Tlew	76	73	**100**													
Tran	75	72	78	**100**												
Tcon	76	73	79	84	**100**											
Tdio	35	33	36	37	37	**100**										
TcCLB	74	70	78	79	79	38	**100**									
TcCL	74	70	78	79	79	38	100	**100**								
TcTula	74	70	78	79	79	38	100	100	**100**							
TcM62	47	46	51	50	52	30	63	63	63	**100**						
TcCAN	73	69	78	78	78	38	99	99	99	62	**100**					
TcG	74	70	79	80	80	38	94	94	94	60	94	**100**				
TcSyl*	73	71	79	81	81	35	94	94	94	64	94	99	**100**			
TcJR	73	70	78	80	80	38	94	94	94	60	94	99	98	**100**		
TcEsm	56	53	58	57	58	34	69	69	69	66	69	67	72	67	**100**	
TcMar	60	57	63	63	64	32	72	72	72	64	71	71	77	71	74	**100**

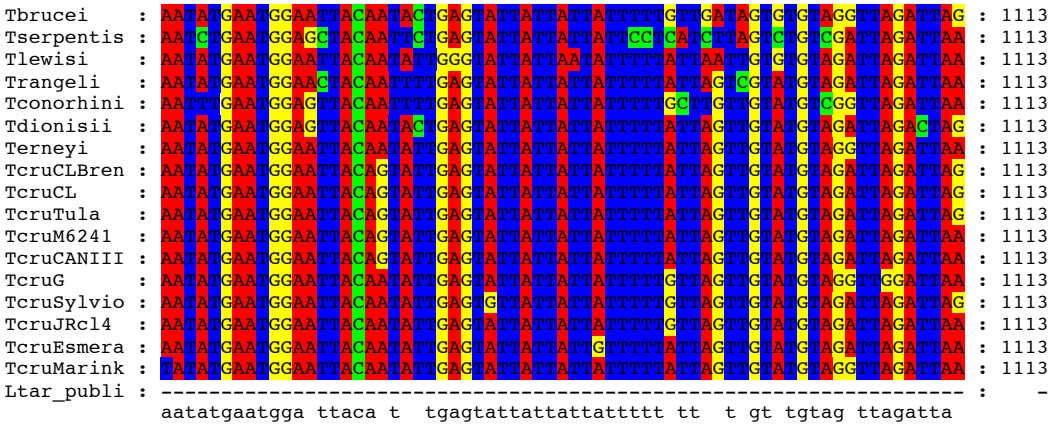
* For the available length of *TcSyl* ND5 transcript. This matrix shows the identities of the truncated ND5 proteins.



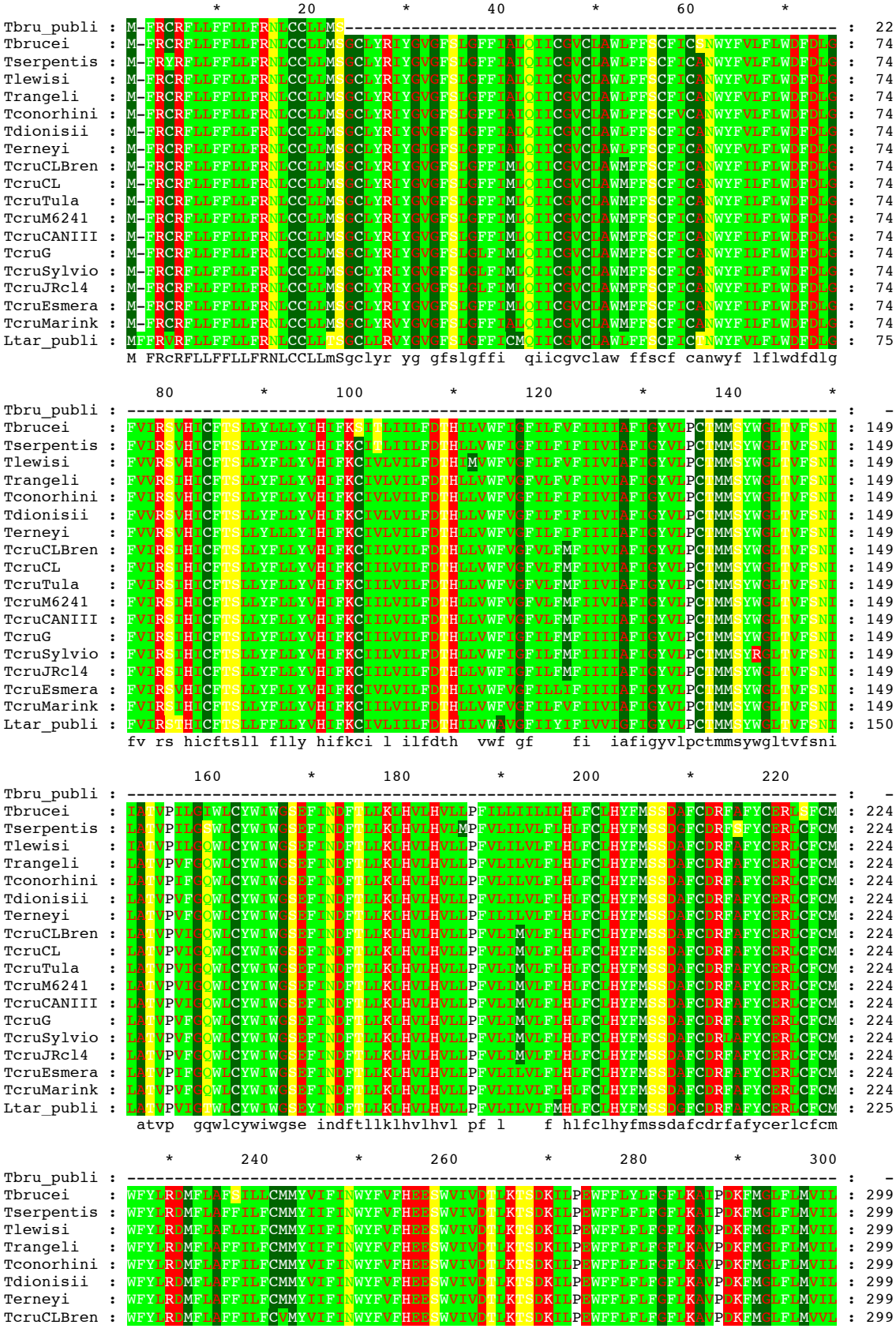

TcruCL : 297
TcruTula : 297
TcruM6241 : 297
TcruCANIII : 297
TcruG : 297
TcruSylvio : 297
TcruJRcl4 : 297
TcruEsmera : 297
TcruMarink : 297
Ltar_publi : 300
tttgt t ag agt t catat tgttt ac tc tt tatatttt tattatatgt ca at tttaaatg
320 340 360
Tbru_publi :
Tbrucei : 372
Tserpentis : 372
Tlewisi : 372
Trangeli : 372
Tconorhini : 372
Tdionisii : 372
Terneyi : 372
TcruCLBren : 372
TcruCL : 372
TcruTula : 372
TcruM6241 : 372
TcruCANIII : 372
TcruG : 372
TcruSylvio : 372
TcruJRcl4 : 372
TcruEsmera : 372
TcruMarink : 372
Ltar_publi : 375
at t tt gtaat ttatt ga ac cat tatt gt tgattt tagg ttt ta tattt t tt ataata
380 400 420 440
Tbru_publi :
Tbrucei : 447
Tserpentis : 447
Tlewisi : 447
Trangeli : 447
Tconorhini : 447
Tdionisii : 447
Terneyi : 447
TcruCLBren : 447
TcruCL : 447
TcruTula : 447
TcruM6241 : 447
TcruCANIII : 447
TcruG : 447
TcruSylvio : 447
TcruJRcl4 : 447
TcruEsmera : 447
TcruMarink : 447
Ltar_publi : 450
t at gc tttat gg tatgt ttaccatgtacaatgatgtc tattgagg ttaac gt tttag aatat
460 480 500 520
Tbru_publi :
Tbrucei : 522
Tserpentis : 522
Tlewisi : 522
Trangeli : 522
Tconorhini : 522
Tdionisii : 522
Terneyi : 522
TcruCLBren : 522
TcruCL : 522
TcruTula : 522
TcruM6241 : 522
TcruCANIII : 522
TcruG : 522
TcruSylvio : 522
TcruJRcl4 : 522
TcruEsmera : 522
TcruMarink : 522
Ltar_publi : 525
tagc ac gt cc t t gg cagtgatt tgttattgaatatg gg agtgaatttataaa gattttac

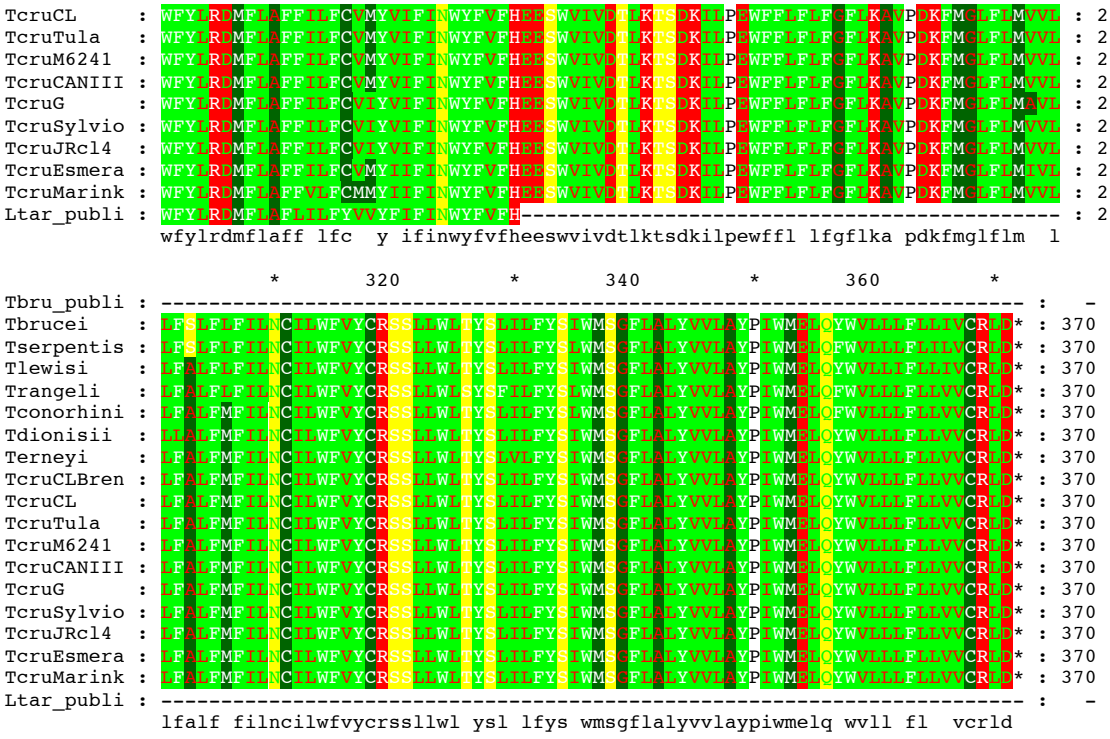

Tbru_publi
Tbrucei
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
Terneyi
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink
Ltar_publi





4.SI-S8a: GeneDoc CYb nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. The one T residue deleted in *T. cruzi* Sylvio X10 is shown as '*' and a gap has been inserted into all other sequences to maintain the frame. References are partial RNA sequence of *Tbru* (Tbru_publi) and partial cDNA of *Leishmania tarentolae* (Ltar_publi).





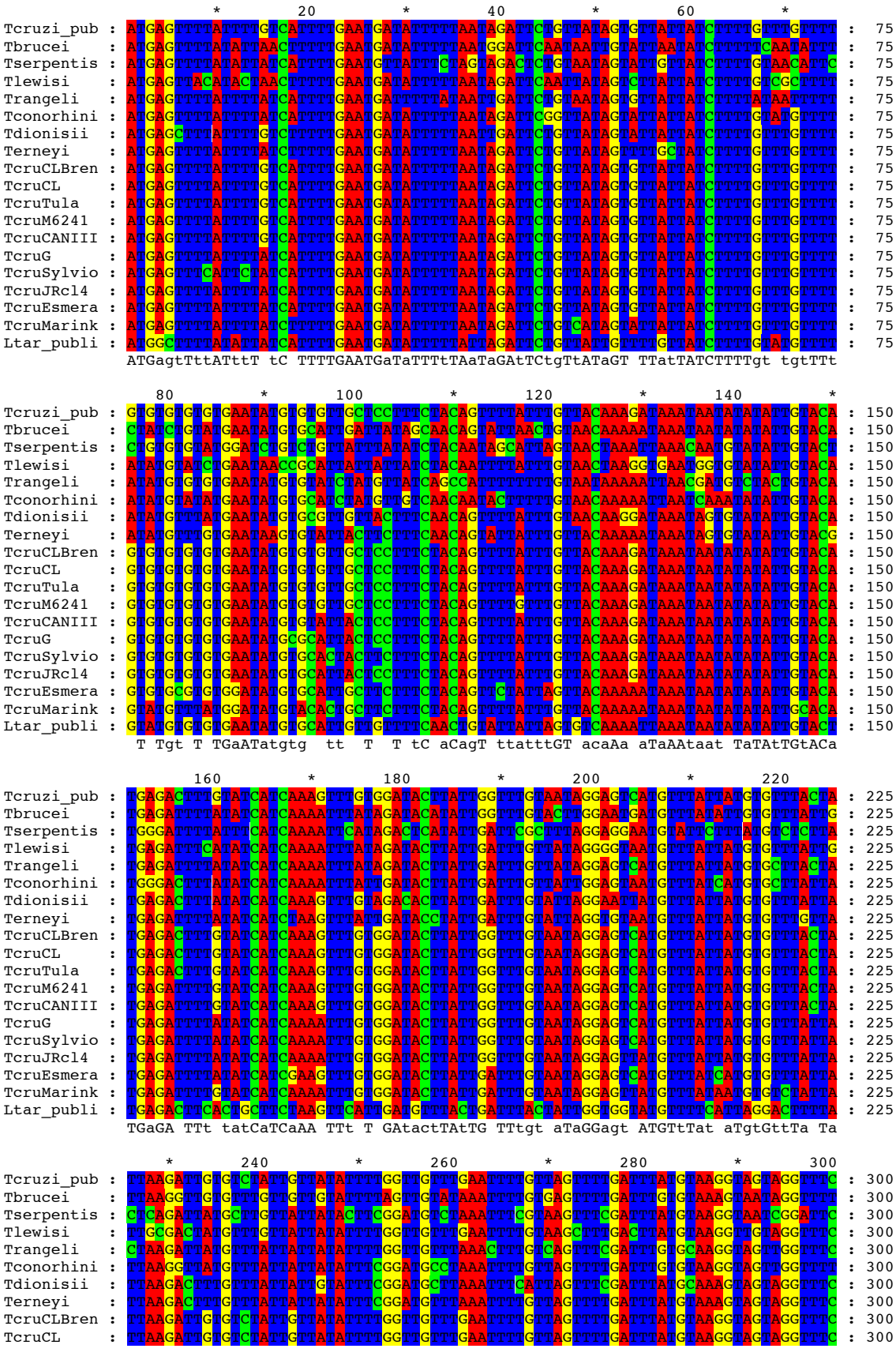
4.SI-S8b: GeneDoc CYb protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. References have also been translated.

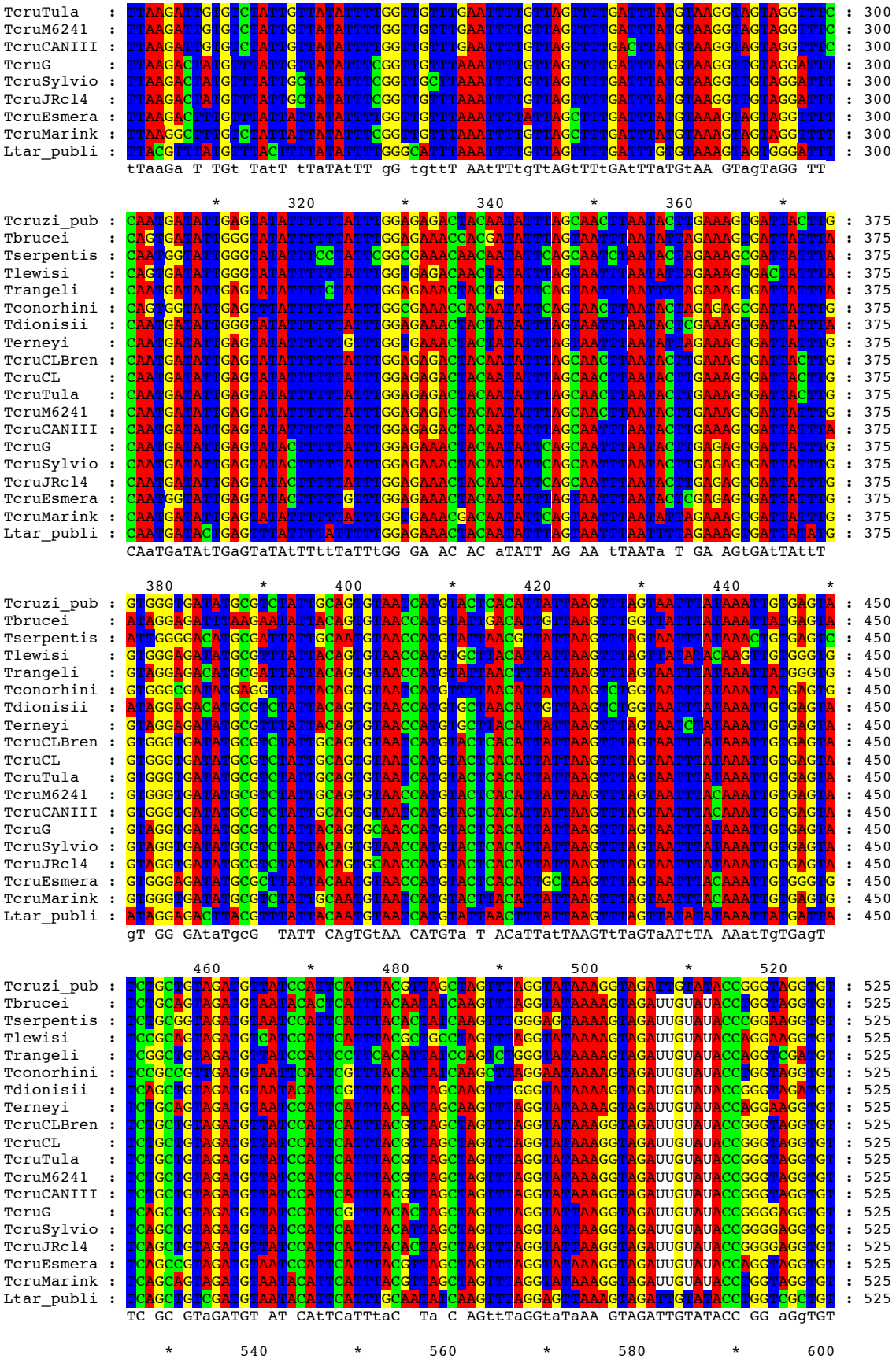
4.SI-S8c: CYb edited transcript percent identity matrix.

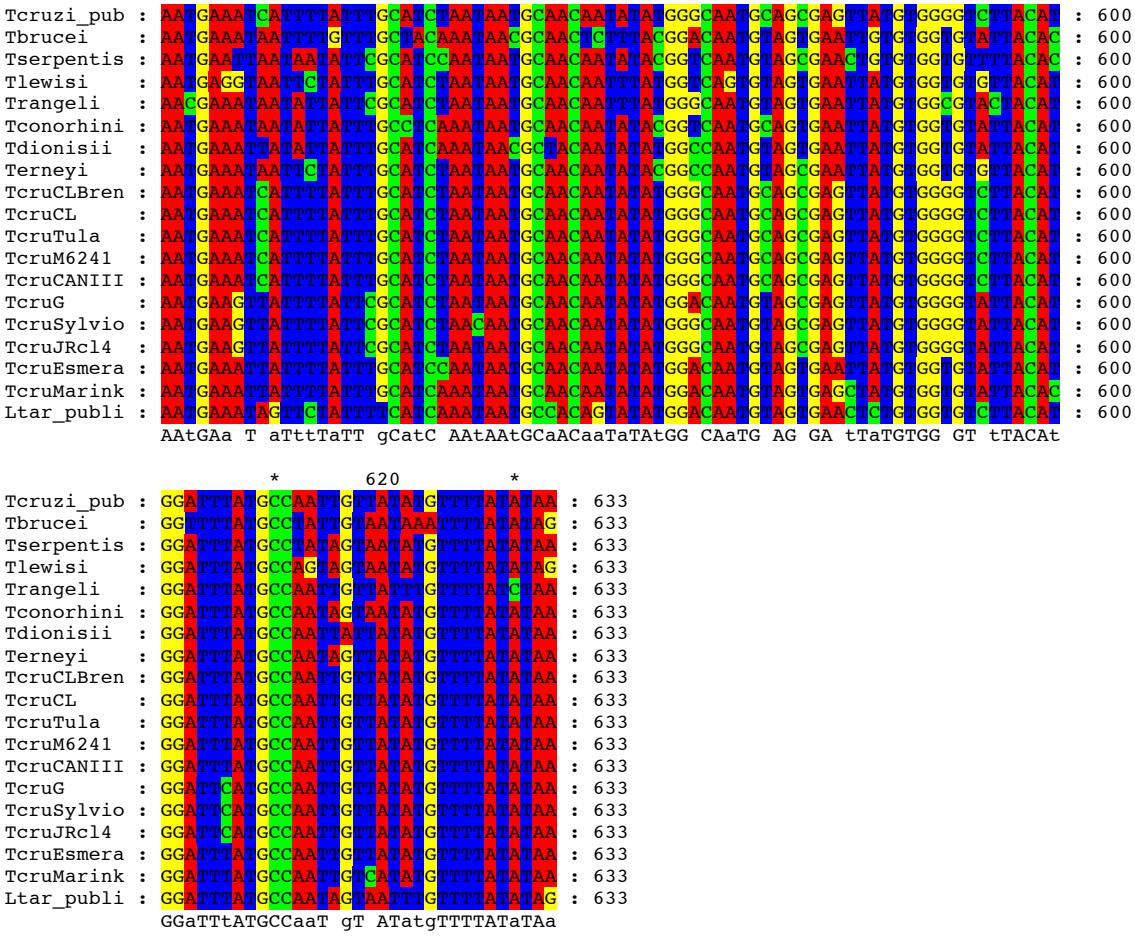
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	82	**100**															
Tlew	87	81	**100**														
Tran	86	81	87	**100**													
Tcon	85	82	86	87	**100**												
Tdio	86	81	87	88	86	**100**											
Tern	86	81	89	88	86	90	**100**										
TcCLB	83	80	85	86	84	88	88	**100**									
TcCL	83	80	85	86	84	88	88	100	**100**								
TcTula	83	80	85	86	84	88	88	100	100	**100**							
TcM62	83	80	85	86	85	88	88	100	100	100	**100**						
TcCAN	84	80	85	86	84	88	89	99	99	99	99	**100**					
TcG	84	80	84	85	84	87	88	94	94	94	94	95	**100**				
TcSyl	84	80	84	85	84	87	87	94	94	94	94	95	98	**100**			
TcJR	84	80	85	85	84	87	87	94	94	94	95	95	99	98	**100**		
TcEsm	83	80	86	85	85	88	88	92	92	92	92	92	91	91	91	**100**	
TcMar	84	82	87	87	85	88	91	90	90	90	90	91	89	89	89	90	**100**

4.SI-S8d: CYb protein percent identity matrix.

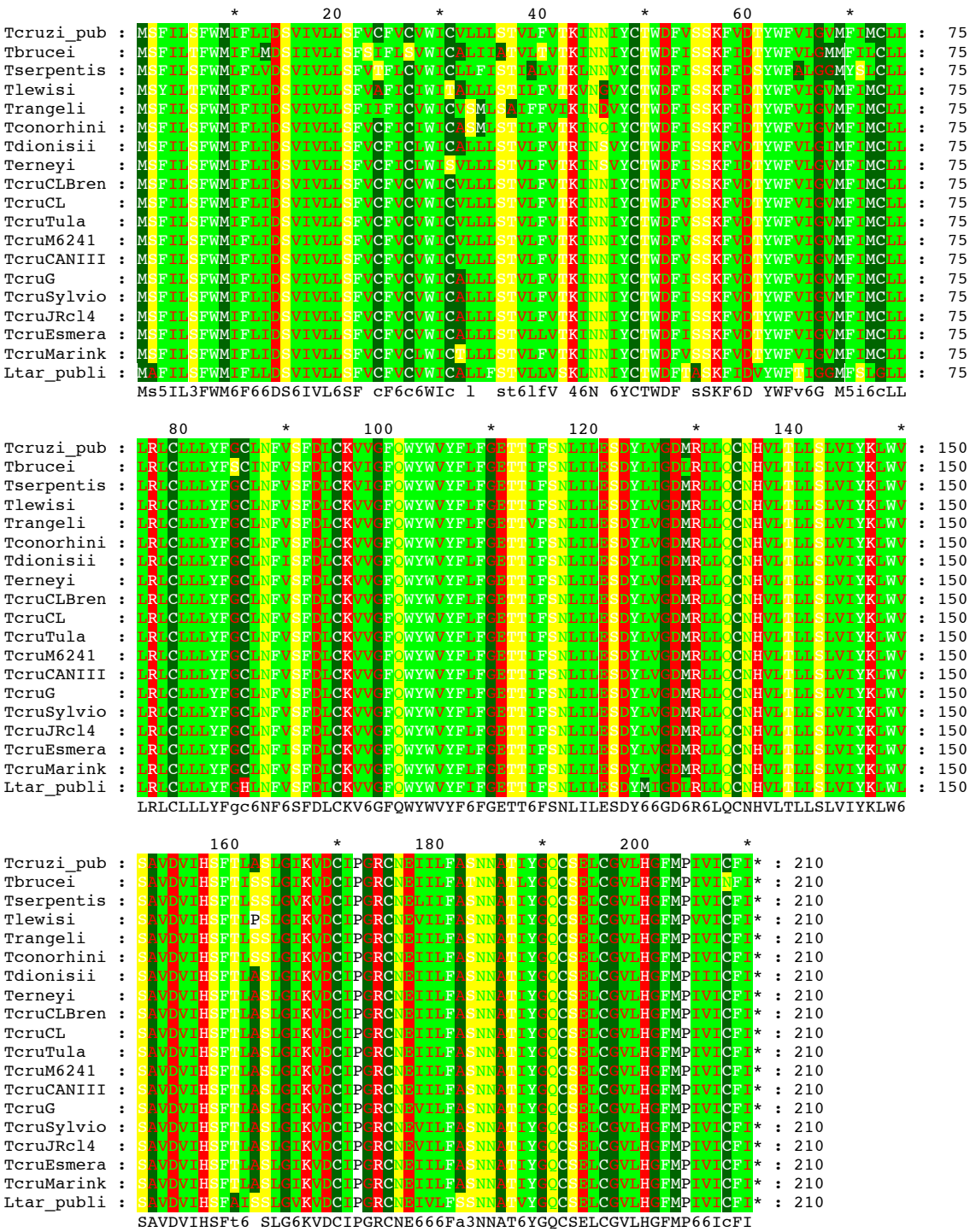
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	93	**100**															
Tlew	94	94	**100**														
Tran	91	94	96	**100**													
Tcon	92	95	96	97	**100**												
Tdio	92	94	97	97	98	**100**											
Tern	93	94	96	96	96	98	**100**										
TcCLB	91	92	94	96	96	96	95	**100**									
TcCL	91	92	94	96	96	96	95	100	**100**								
TcTula	91	92	94	96	96	96	95	100	100	**100**							
TcM62	91	92	94	96	96	96	95	100	100	100	**100**						
TcCAN	91	92	94	95	95	96	95	100	100	100	100	**100**					
TcG	91	92	93	94	95	96	95	98	98	98	98	98	**100**				
TcSyl	91	91	93	94	95	95	94	98	98	98	98	98	99	**100**			
TcJR	91	92	94	95	95	96	95	98	98	98	98	99	100	99	**100**		
TcEsm	91	93	95	95	96	97	96	97	97	97	97	97	97	96	97	**100**	
TcMar	91	93	95	96	96	97	96	98	98	98	98	97	97	96	97	97	**100**




TcruTula : 300
TcruM6241 : 300
TcruCANIII : 300
TcruG : 300
TcruSylvio : 300
TcruJRcl4 : 300
TcruEsmera : 300
TcruMarink : 300
Ltar_publi : 300
tTaaGa T TGt TatT tTaTAtTT gG tgttT AAtTTtgTtAGtTTtGAtTTaTGtAA GTagTaGG TT
320 340 360
Tcruzi_pub : 375
Tbrucei : 375
Tserpentis : 375
Tlewisi : 375
Trangeli : 375
Tconorhini : 375
Tdionisii : 375
Terneyi : 375
TcruCLBren : 375
TcruCL : 375
TcruTula : 375
TcruM6241 : 375
TcruCANIII : 375
TcruG : 375
TcruSylvio : 375
TcruJRcl4 : 375
TcruEsmera : 375
TcruMarink : 375
Ltar_publi : 375
CAaTGaTAtTGaGTaTAtTTttTaTTtGG GA AC AC aTATT AG AA tTAATa T GA AGtGAtTAttT
380 400 420 440
Tcruzi_pub : 450
Tbrucei : 450
Tserpentis : 450
Tlewisi : 450
Trangeli : 450
Tconorhini : 450
Tdionisii : 450
Terneyi : 450
TcruCLBren : 450
TcruCL : 450
TcruTula : 450
TcruM6241 : 450
TcruCANIII : 450
TcruG : 450
TcruSylvio : 450
TcruJRcl4 : 450
TcruEsmera : 450
TcruMarink : 450
Ltar_publi : 450
gT GG GAtaTgcG TATT CAgTGtAA CATGTa T ACaTTatTAAGTtTaGTaATtTA AAatTgTGagT
460 480 500 520
Tcruzi_pub : 525
Tbrucei : 525
Tserpentis : 525
Tlewisi : 525
Trangeli : 525
Tconorhini : 525
Tdionisii : 525
Terneyi : 525
TcruCLBren : 525
TcruCL : 525
TcruTula : 525
TcruM6241 : 525
TcruCANIII : 525
TcruG : 525
TcruSylvio : 525
TcruJRcl4 : 525
TcruEsmera : 525
TcruMarink : 525
Ltar_publi : 525
TC GC GTaGATGT AT CAtTCaTTtaC Ta C AGttTaGGtaTaAA GTAGATTGTATACC GG aGgTGT
540 560 580 600



4.SI-S9a: GeneDoc COII nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. References are cDNAs of *Tcru* (Tcruzi_pub) and *L. tarentolae* (Ltar_publi).



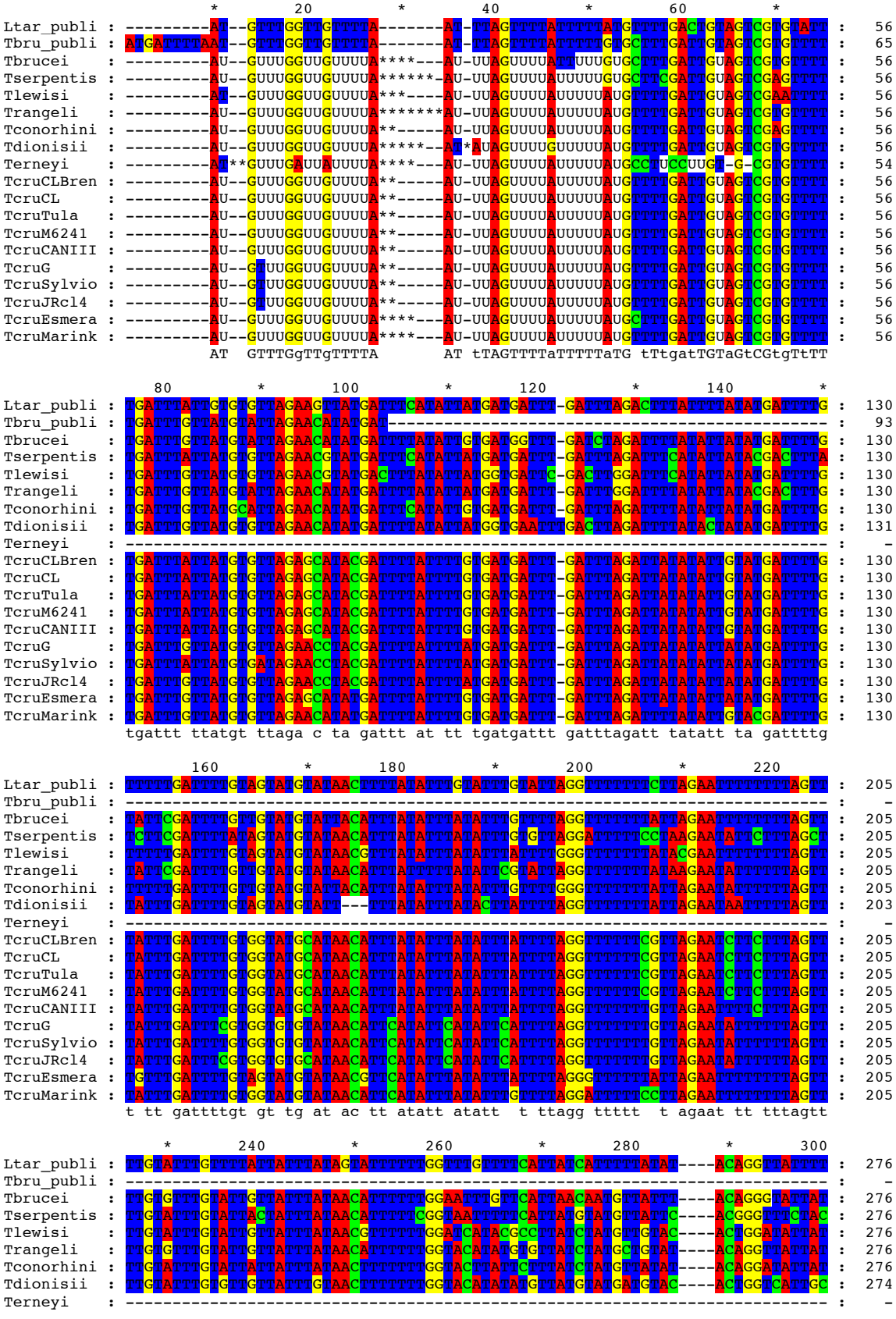
4.SI-S9b: GeneDoc COII protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. References have also been translated.

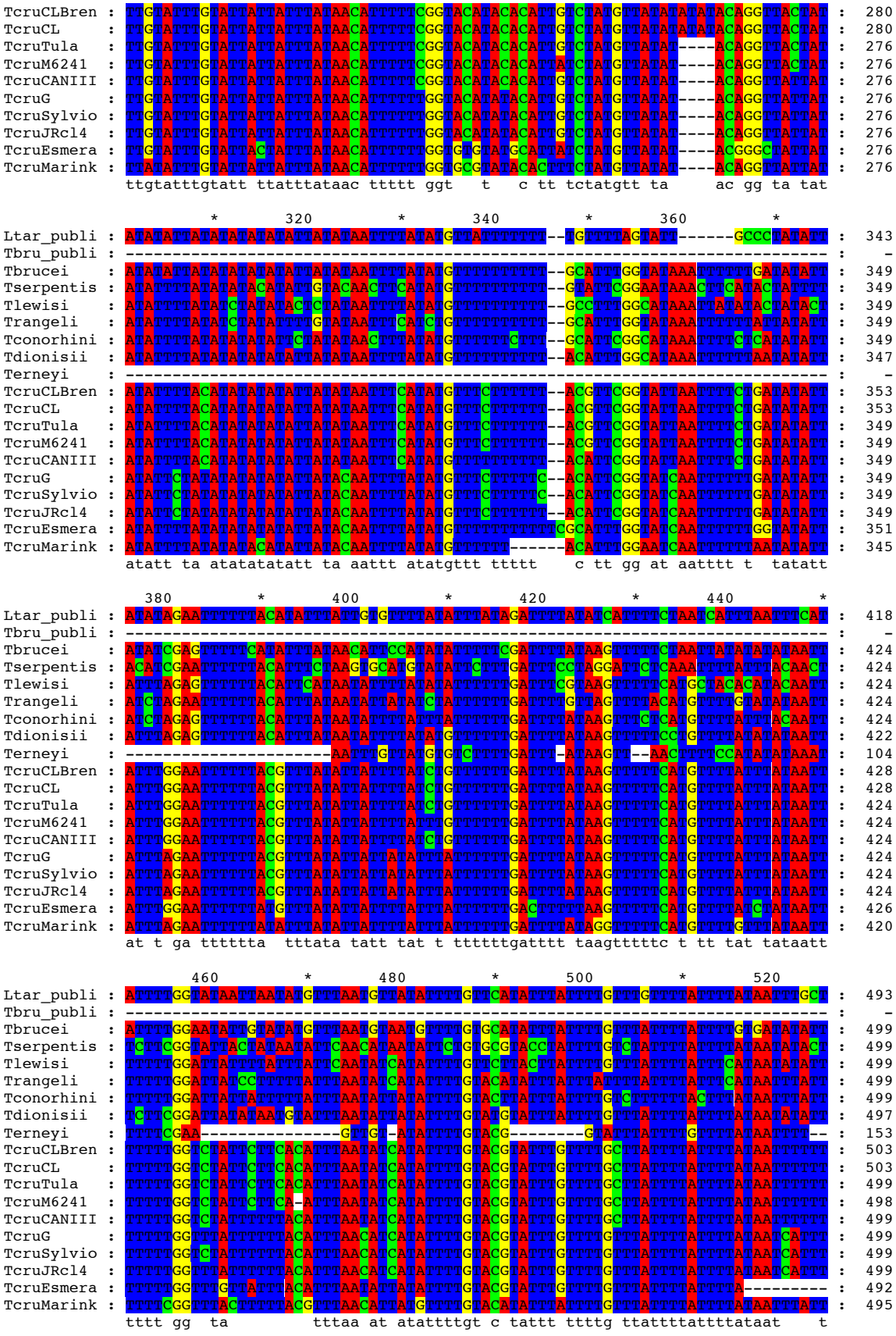
4.SI-S9c. COII edited trancript percent identity matrix.

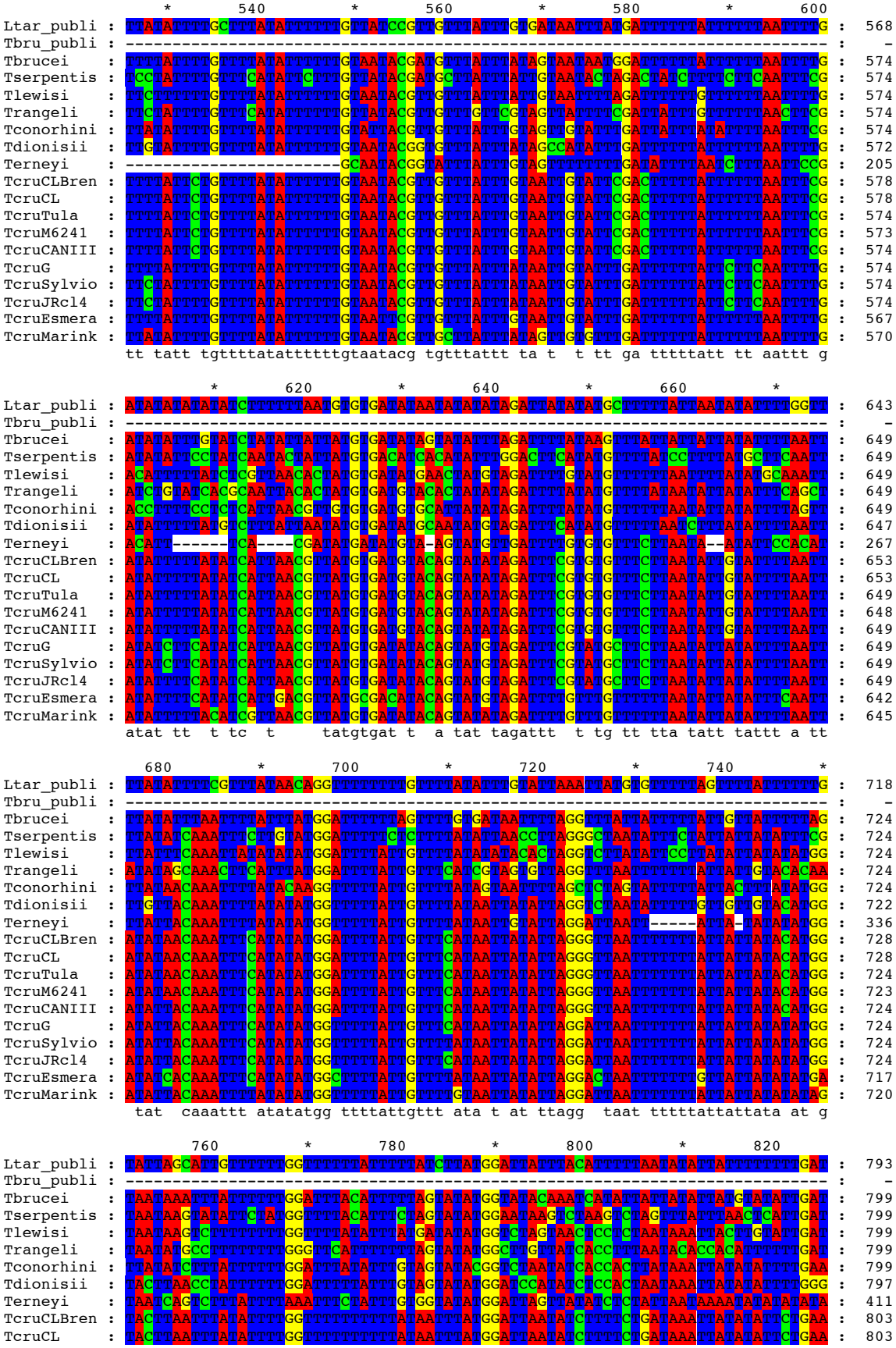
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	78	**100**															
Tlew	81	78	**100**														
Tran	80	79	82	**100**													
Tcon	83	80	82	85	**100**												
Tdio	83	79	82	84	85	**100**											
Tern	83	81	87	84	86	89	**100**										
TcCLB	81	78	83	84	84	88	89	**100**									
TcCL	81	78	83	84	84	88	89	100	**100**								
TcTula	81	78	83	84	84	88	89	100	100	**100**							
TcM62	81	78	84	84	84	88	89	99	99	99	**100**						
TcCAN	81	79	84	85	84	87	89	99	99	99	99	**100**					
TcG	81	79	84	84	85	88	89	94	94	94	94	94	**100**				
TcSyl	81	79	84	84	84	87	89	93	93	93	94	94	98	**100**			
TcJR	81	79	84	84	85	88	89	94	94	94	94	94	99	99	**100**		
TcEsm	82	79	84	84	86	88	90	91	91	91	92	91	92	92	92	**100**	
TcMar	83	79	84	82	86	88	89	91	91	91	91	91	91	91	91	92	**100**

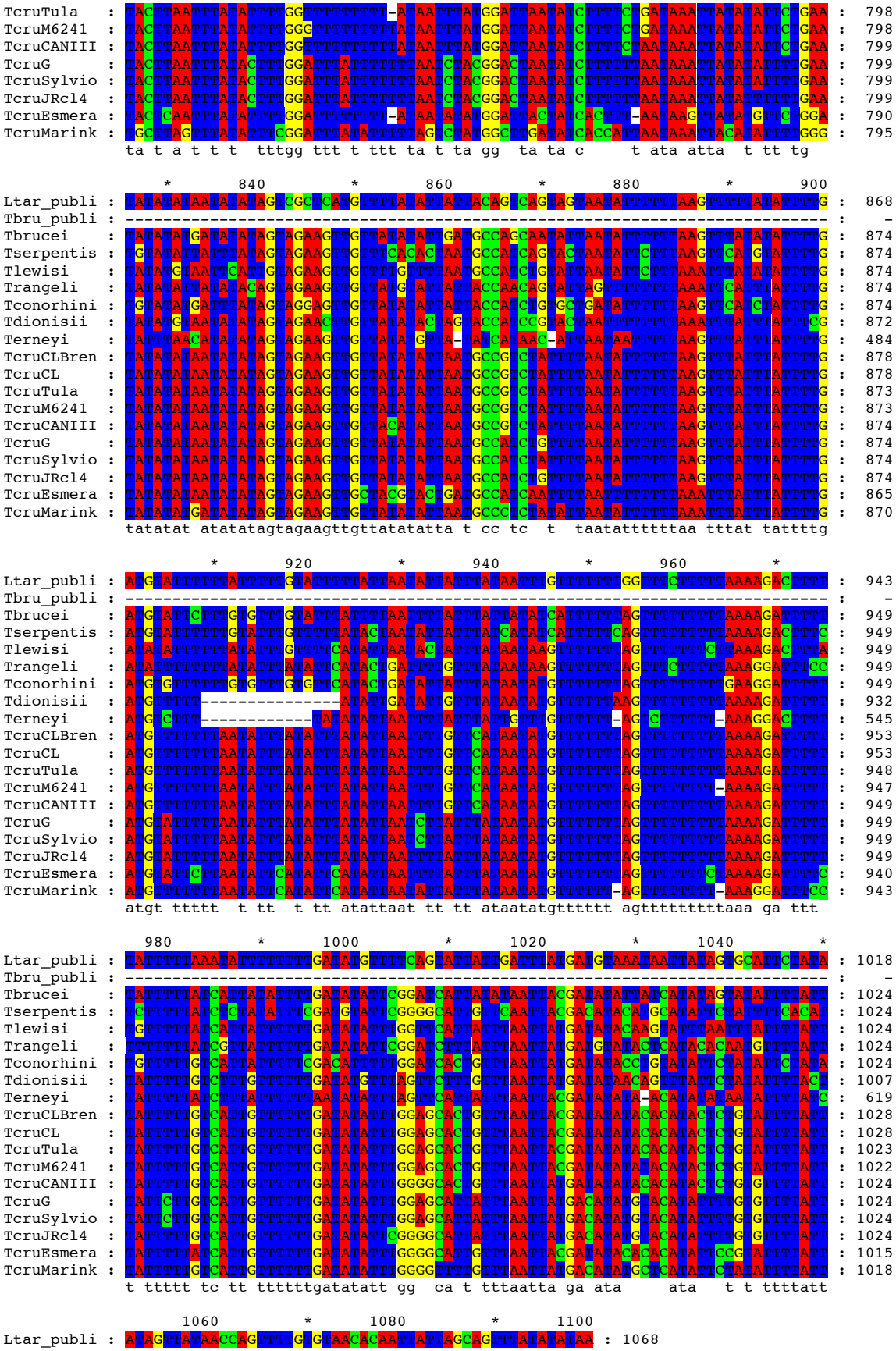
4.SI-S9d. COII protein percent identity matrix.

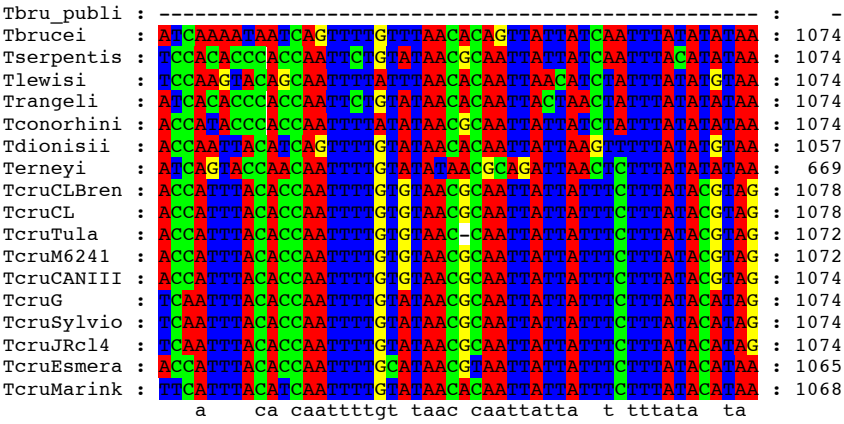
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	85	**100**															
Tlew	84	86	**100**														
Tran	84	86	90	**100**													
Tcon	86	88	94	95	**100**												
Tdio	85	87	91	90	93	**100**											
Tern	86	88	93	93	95	96	**100**										
TcCLB	86	87	92	92	95	94	96	**100**									
TcCL	86	87	92	92	95	94	96	100	**100**								
TcTula	86	87	92	92	95	94	96	100	100	**100**							
TcM62	86	87	92	92	95	94	96	100	100	100	**100**						
TcCAN	86	87	92	92	95	94	96	100	100	100	100	**100**					
TcG	87	88	93	92	96	95	96	99	99	99	99	99	**100**				
TcSyl	87	88	93	92	96	95	96	99	99	99	99	99	100	**100**			
TcJR	87	88	93	92	96	95	96	99	99	99	99	99	100	100	**100**		
TcEsm	87	88	92	91	95	95	95	98	98	98	98	98	99	99	99	**100**	
TcMar	86	87	92	91	95	95	96	99	99	99	99	99	98	98	98	98	**100**



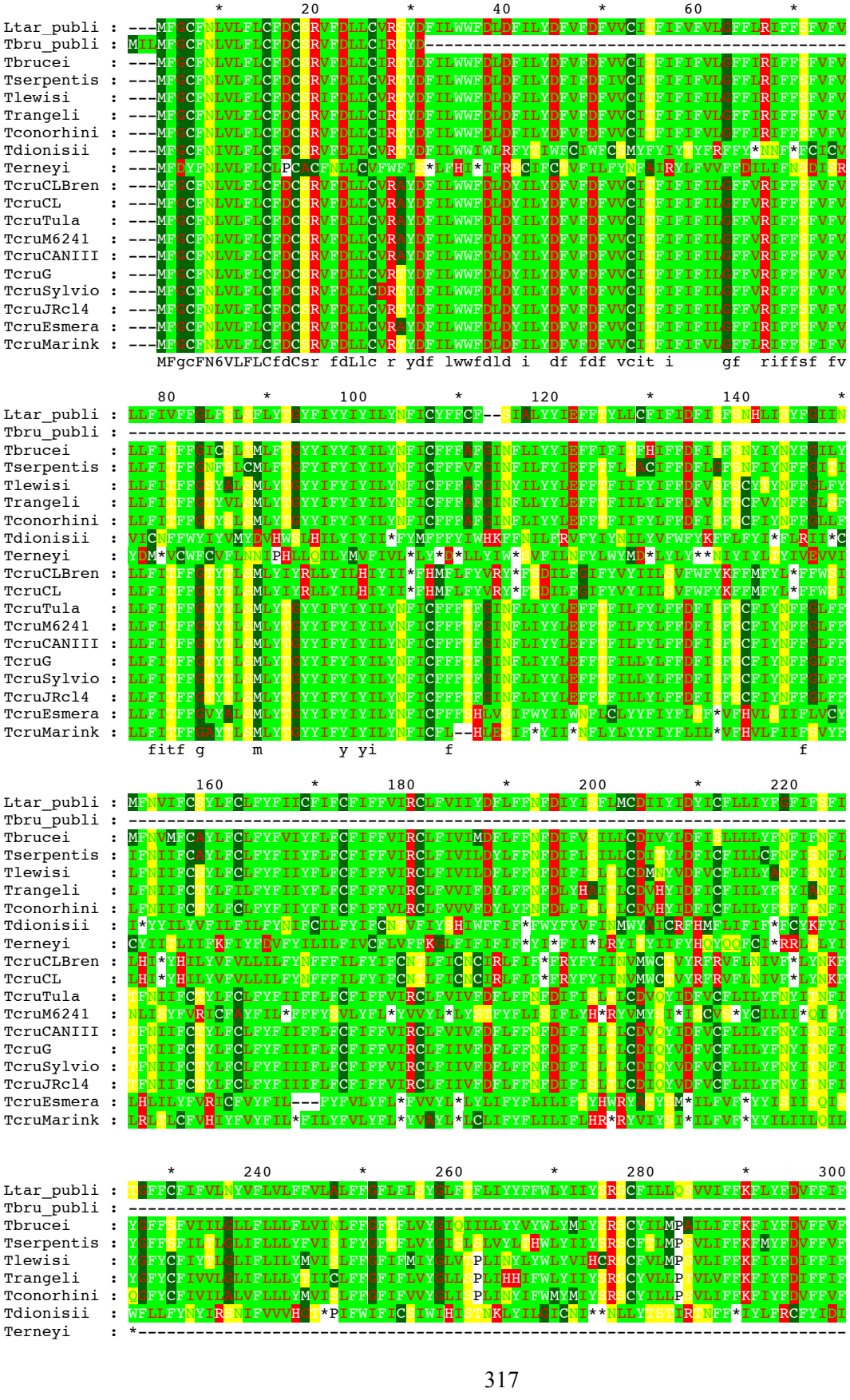








4.SI-S10a: GeneDoc MURF2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. References are cDNAs of *L. tar* (Ltar_publi) and *Tbru* (Tbru_publi).



4.SI-S10b: GeneDoc MURF2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. This alignment accounts for the frameshifts caused by the indels [shown in the corresponding nucleotide sequence alignment above (S10a)] and shows the resultant truncated proteins in the alternate frame. '*' indicates stop codon in the alignment. Several stop codons are seen in the altered frame of TcruCLBren, TcruCL, TcruTula, TcruM6241, TcruEsmera, TcruMarink, Tdionisii and Terneyi. References have also been translated.

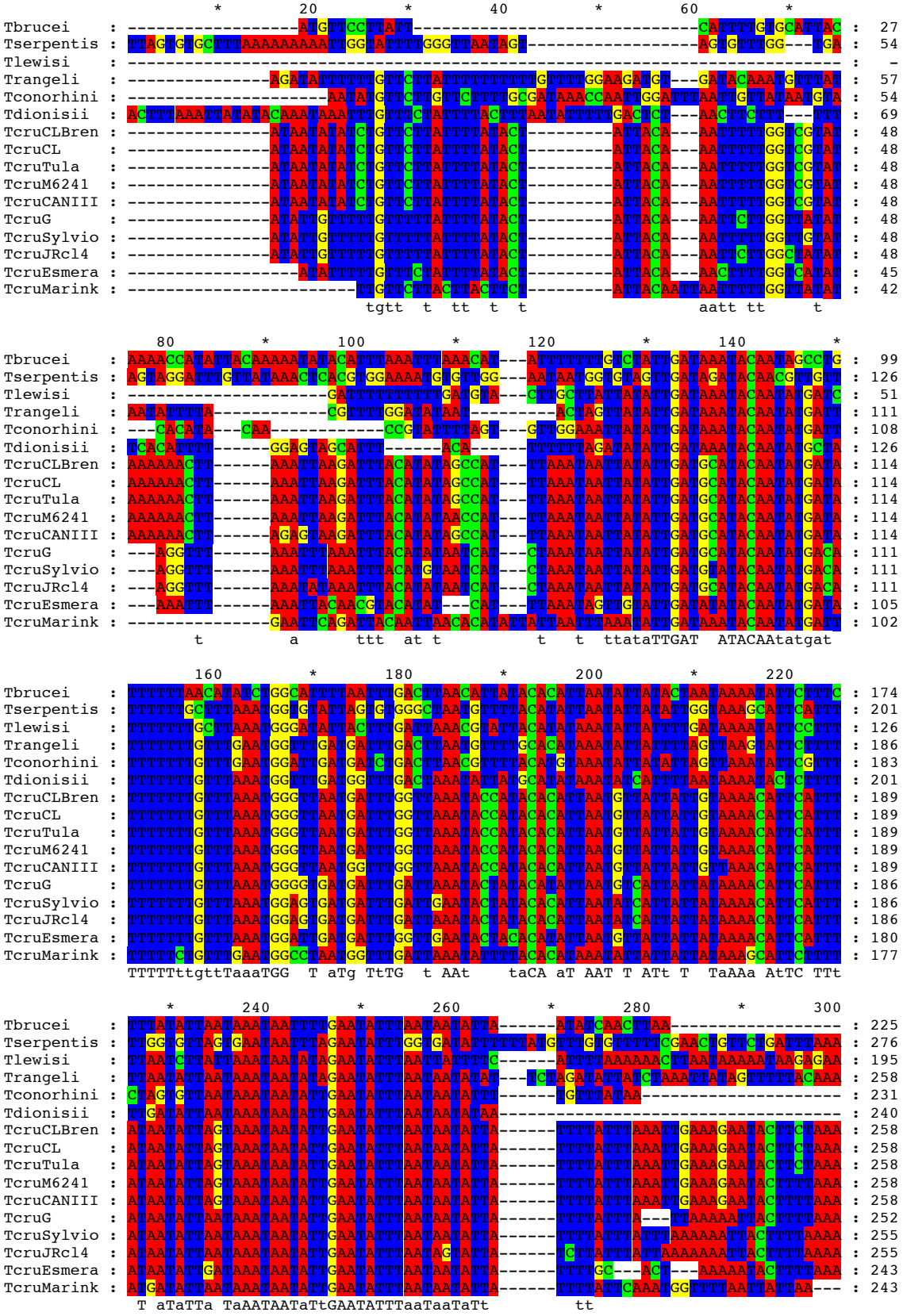
4.SI-S10c: MURF2 edited transcript percent identity matrix.

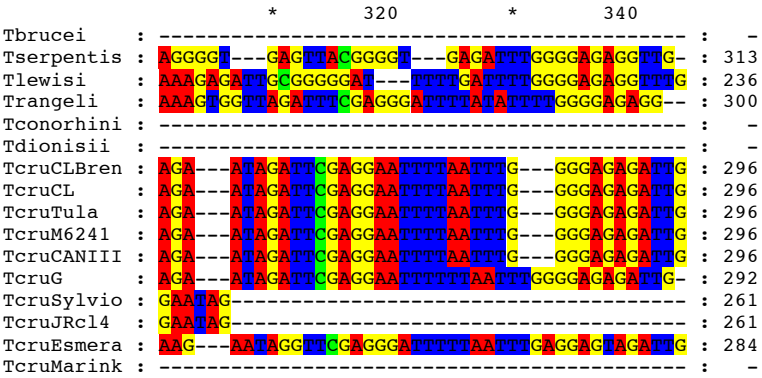
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	77	**100**															
Tlew	80	75	**100**														
Tran	80	75	80	**100**													
Tcon	82	77	82	83	**100**												
Tdio	81	74	83	81	84	**100**											
Tern	73	67	74	74	74	76	**100**										
TcCLB	81	75	80	81	84	83	78	**100**									
TcCL	81	75	80	81	84	83	78	100	**100**								
TcTula	81	75	80	81	83	83	78	100	100	**100**							
TcM62	81	75	80	81	84	84	78	100	100	100	**100**						
TcCAN	81	75	80	81	84	84	78	99	99	99	98	**100**					
TcG	81	76	81	81	84	84	77	92	92	92	92	93	**100**				
TcSyl	81	76	81	81	84	84	77	92	92	92	92	93	99	**100**			
TcJR	81	76	81	81	84	84	78	92	92	92	92	93	99	99	**100**		
TcEsm	81	75	81	81	83	83	77	90	90	90	90	90	90	90	90	**100**	
TcMar	82	76	81	83	84	85	77	88	88	88	88	89	90	90	90	89	**100**

4.SI-S10d: MURF2 protein percent identity matrix.

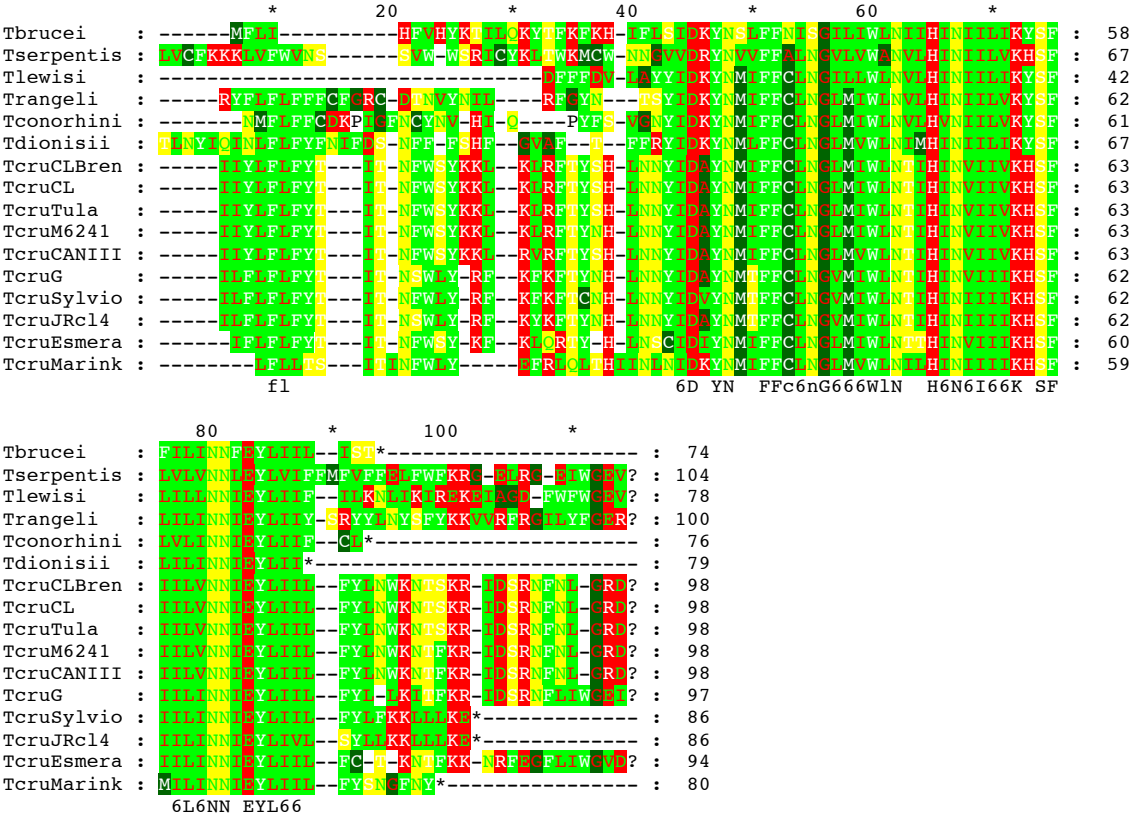
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	80	**100**															
Tlew	76	76	**100**														
Tran	76	76	80	**100**													
Tcon	79	77	82	87	**100**												
Tdio	24	22	23	23	23	**100**											
Tern	16	15	18	18	18	21	**100**										
TcCLB	32	32	34	34	32	18	17	**100**									
TcCL	32	32	34	34	32	18	17	100	**100**								
TcTula	60	59	65	65	67	22	18	33	33	**100**							
TcM62	44	43	46	47	48	23	19	35	35	67	**100**						
TcCAN	76	74	80	82	86	23	18	34	34	76	53	**100**					
TcG	76	76	82	82	85	24	18	34	34	74	51	96	**100**				
TcSyl	77	75	81	82	85	24	18	33	33	74	51	96	99	**100**			
TcJR	76	76	82	82	85	24	18	34	34	74	51	96	100	99	**100**		
TcEsm	54	55	55	56	57	20	18	35	35	40	39	62	60	60	60	**100**	
TcMar	47	48	46	48	48	21	18	34	34	40	40	50	50	50	50	72	**100**

This matrix shows the identities of the truncated MURF2 proteins.





4.SI-S11a: GeneDoc MURF5 gene alignment: Abbreviations for the names of the organisms, color code, gaps and consensus for the figure are as mentioned in S1. A slight shift in frame is noticeable in some sequences; this is because the transcript alignment reflects the protein alignment shown below.



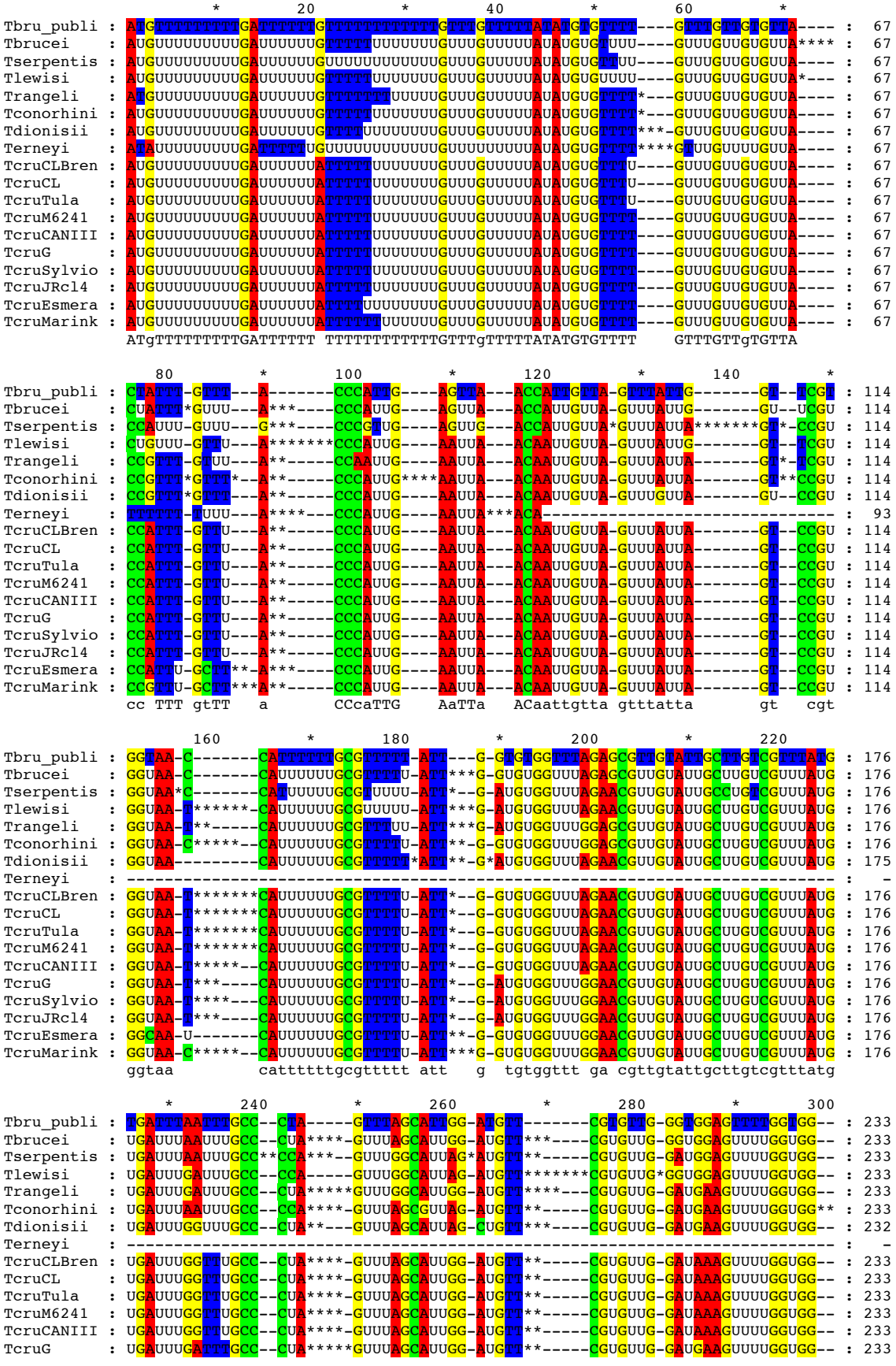
4.SI-S11b: GeneDoc MURF5 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. "?" indicates incomplete or lack of information.

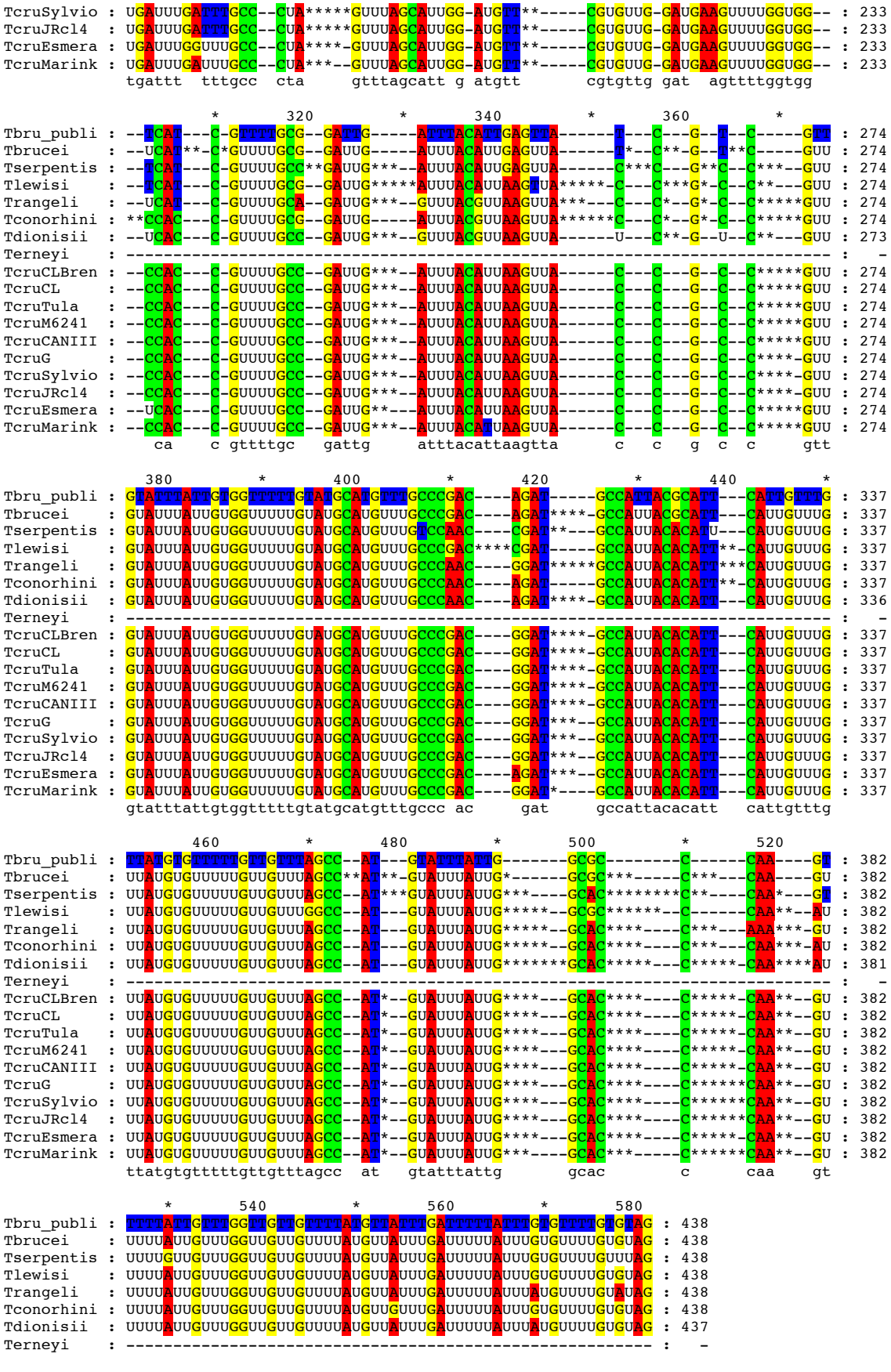
4.SI-S11c: MURF5 nucleotide percent identity matrix.

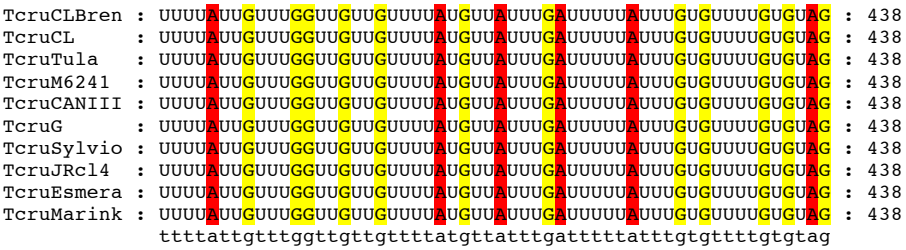
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	56	**100**														
Tlew	68	62	**100**													
Tran	66	62	73	**100**												
Tcon	62	60	74	72	**100**											
Tdio	68	56	76	73	71	**100**										
TcCLB	65	60	70	67	69	72	**100**									
TcCL	65	60	70	67	69	72	100	**100**								
TcTula	65	60	70	67	69	72	100	100	**100**							
TcM62	66	60	70	68	69	72	99	99	99	**100**						
TcCAN	63	60	69	67	69	73	98	98	98	98	**100**					
TcG	67	60	71	70	70	74	90	90	90	91	89	**100**				
TcSyl	67	59	69	67	69	74	88	88	88	89	87	97	**100**			
TcJR	67	58	68	66	69	73	86	86	86	87	85	97	96	**100**		
TcEsm	65	59	69	66	70	75	88	88	88	88	87	89	89	86	**100**	
TcMar	69	60	69	70	70	74	78	78	78	78	78	77	78	76	75	**100**

4.SI-S11d: MURF5 protein percent identity matrix.

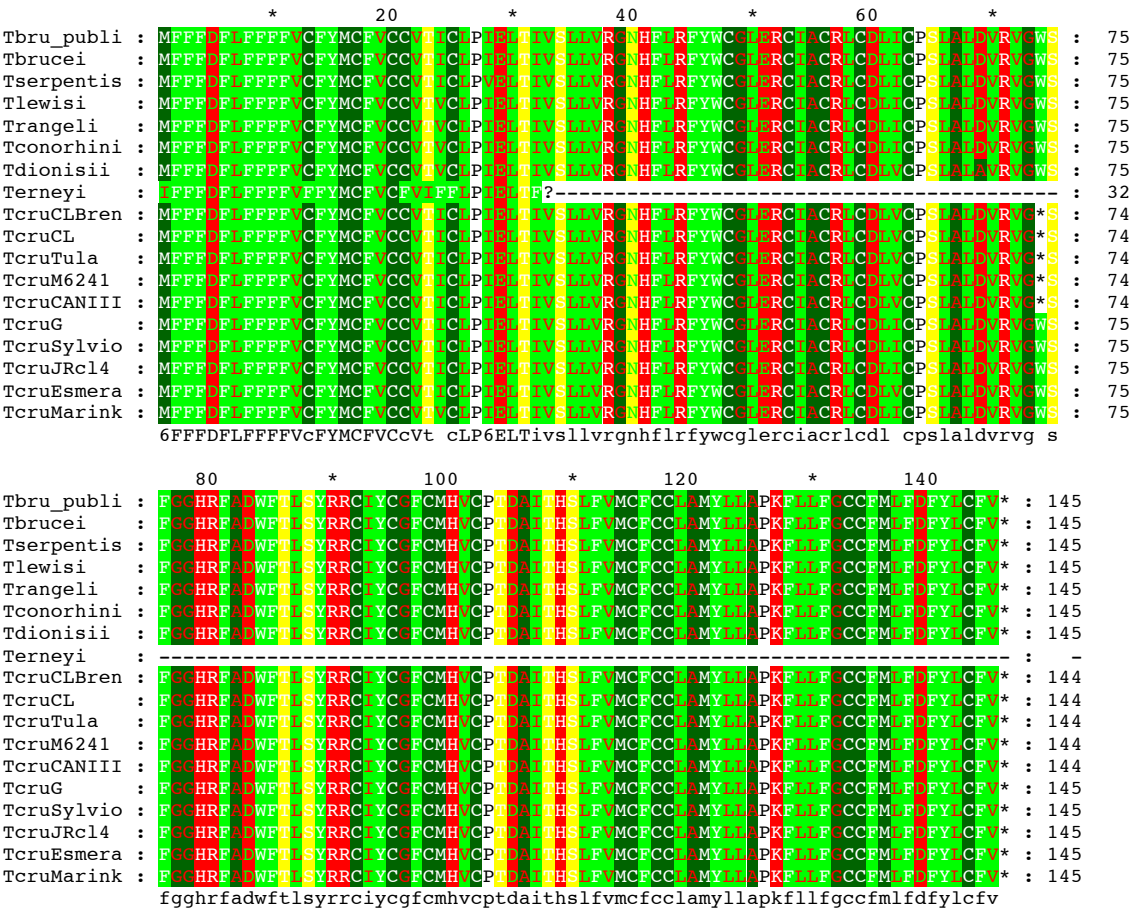
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	40	**100**														
Tlew	62	46	**100**													
Tran	53	39	58	**100**												
Tcon	48	48	73	67	**100**											
Tdio	59	41	76	65	66	**100**										
TcCLB	50	40	51	53	59	61	**100**									
TcCL	50	40	51	53	59	61	100	**100**								
TcTula	50	40	51	53	59	61	100	100	**100**							
TcM62	50	41	51	54	59	61	98	98	98	**100**						
TcCAN	47	40	51	52	58	63	96	96	96	96	**100**					
TcG	52	43	54	54	56	61	79	79	79	81	78	**100**				
TcSyl	55	42	55	56	56	64	76	76	76	77	73	89	**100**			
TcJR	54	39	53	57	56	61	74	74	74	76	72	89	92	**100**		
TcEsm	52	42	51	57	57	65	76	76	76	77	74	75	76	73	**100**	
TcMar	57	44	59	59	58	69	68	68	68	68	69	66	68	63	69	**100**



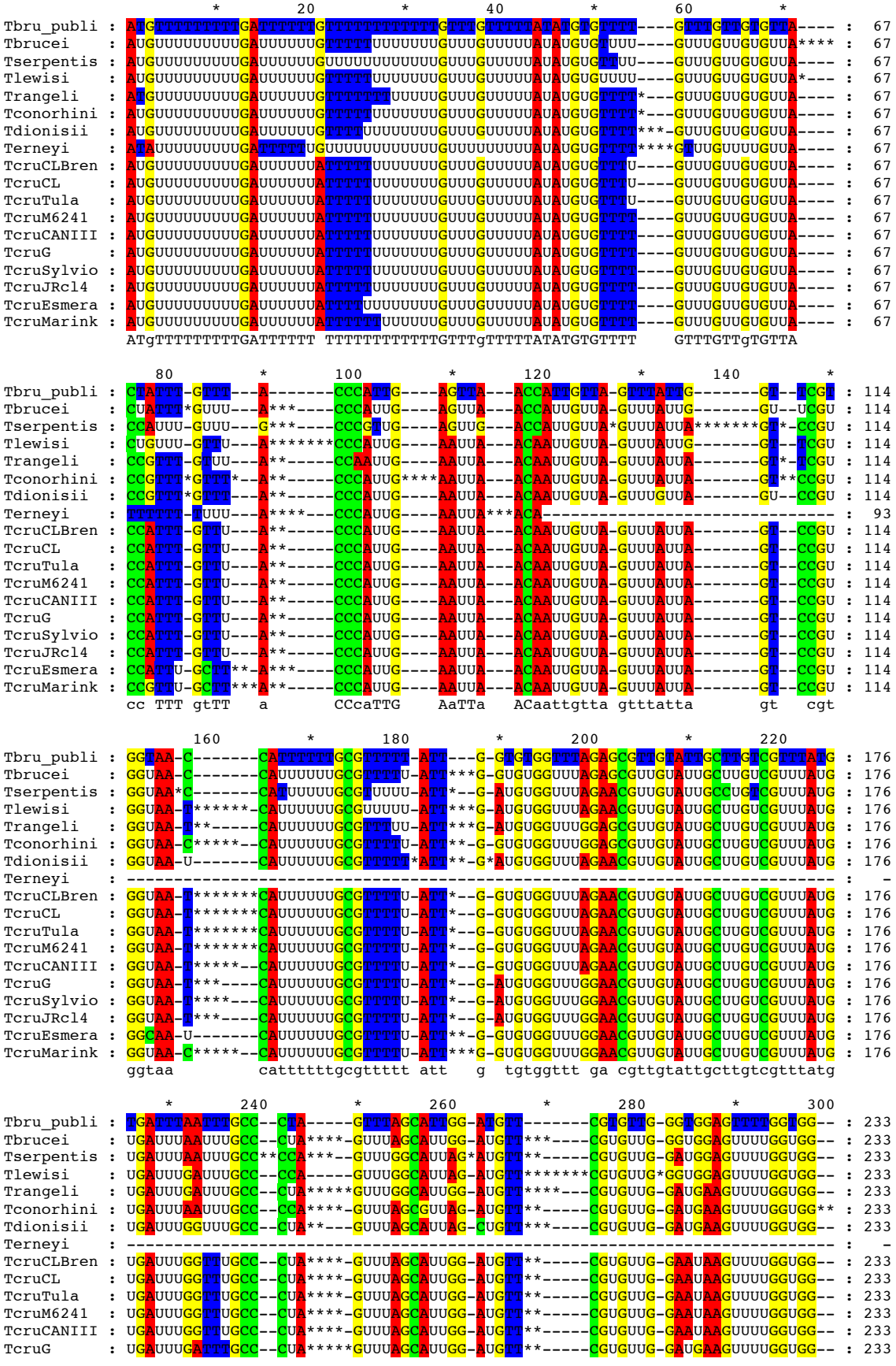


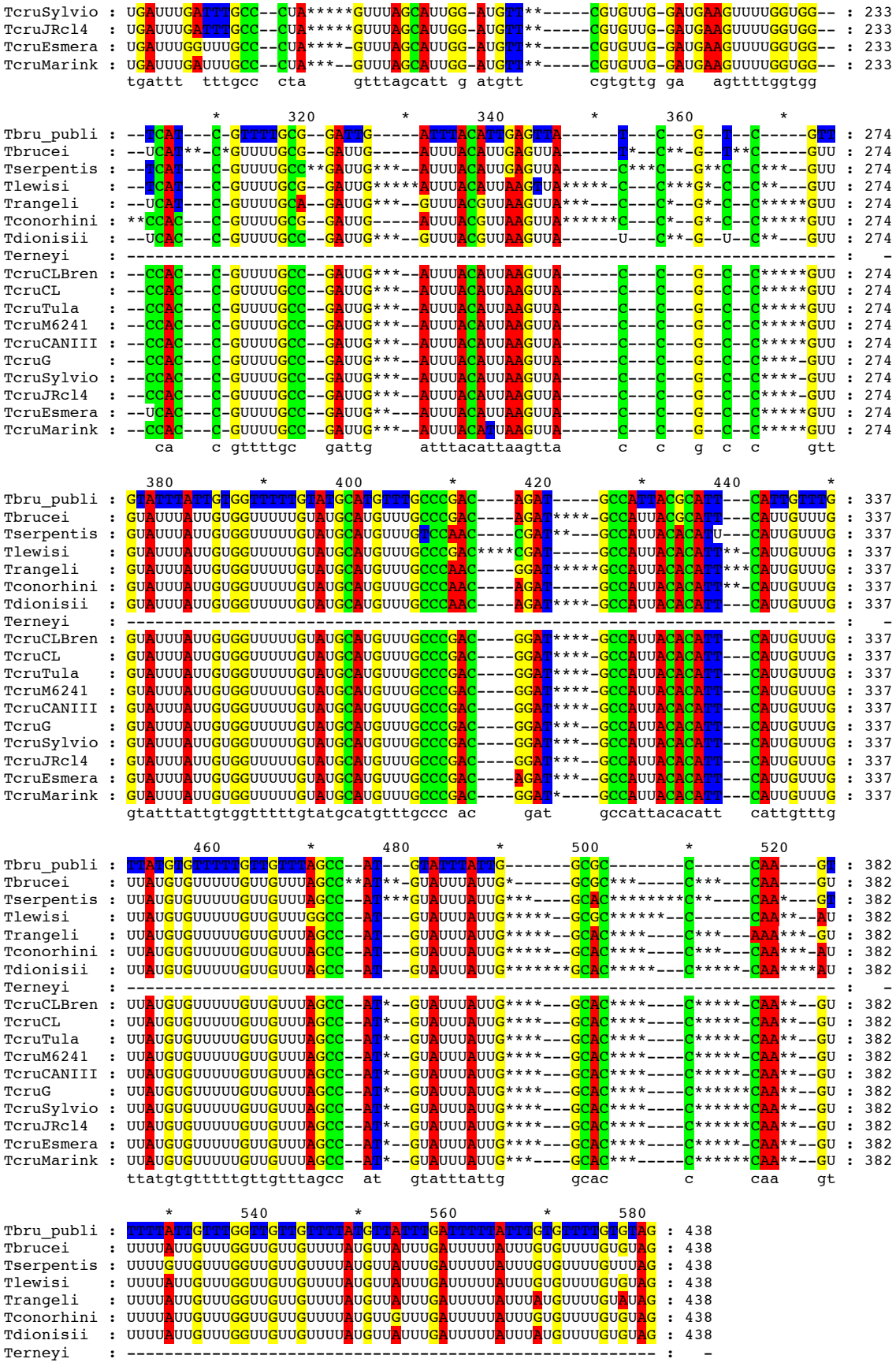


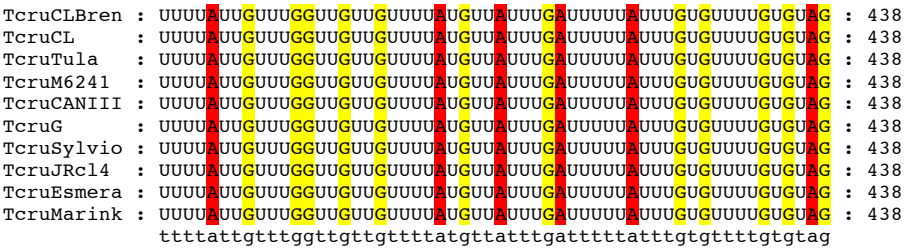
4.SI-S12a: GeneDoc ND8 / G1 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is cDNA of *T. bru* (Tbru_publi). "?" denotes incomplete sequence information in *Terneyi*.



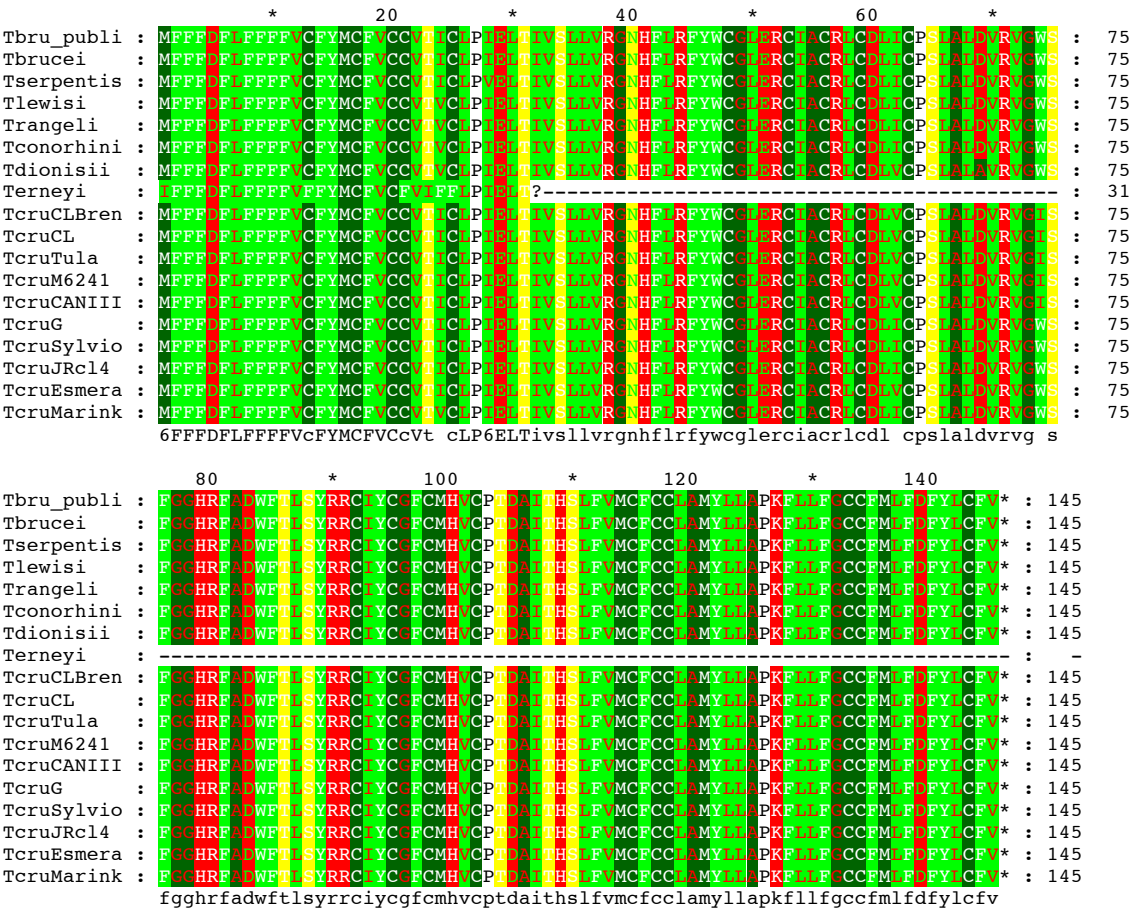
4.SI-S12b: GeneDoc ND8/ G1 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Note the premature stop codon at the 74th amino acid residue in members of DTUs III, IV and VI. Reference has also been translated. "?" denotes incomplete sequence information in *Terneyi.*







4.SI-S12c: GeneDoc ND8 / G1 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Please note the altered U insertion at position 221 (instead of 220) to generate the codon 'AUA' instead of stop codon 'UAA'. Reference is cDNA of *T. bru* (Tbru_publi). "?" denotes incomplete sequence information in *Terneyi*.



4.SI-S12d: GeneDoc ND8/ G1 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Note that the premature stop codon at the 74th amino acid residue in members of DTUs III, IV and VI has been remedied by the slight alteration of U insertion. Reference has also been translated. "?" denotes incomplete sequence information in *Terneyi.*

4.SI-S12e. ND8 edited transcript percent identity matrix.

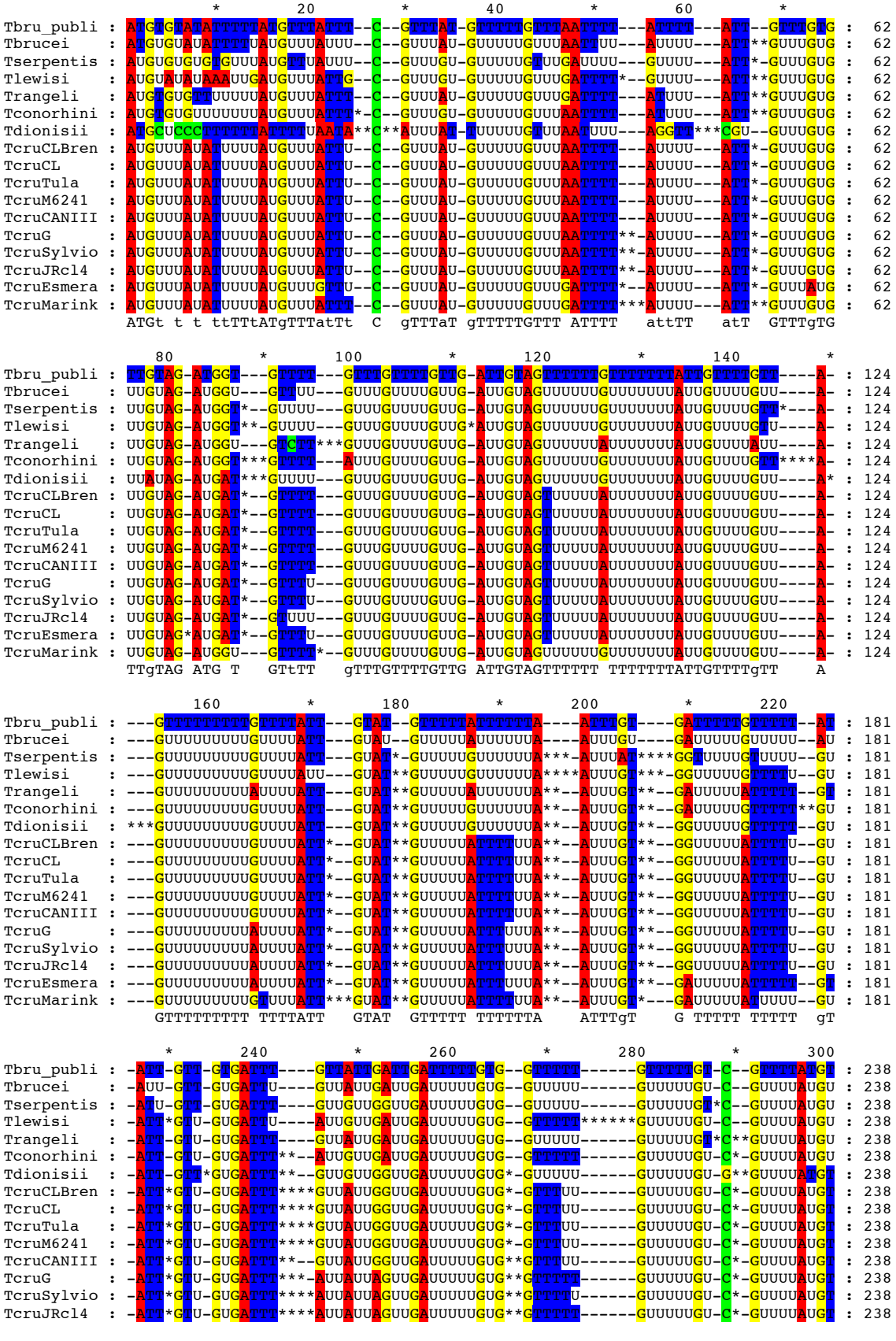
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	95	**100**															
Tlew	96	95	**100**														
Tran	94	94	95	**100**													
Tcon	95	95	95	96	**100**												
Tdio	94	94	95	96	96	**100**											
Tern*	91	87	94	91	92	92	**100**										
TcCLB	95	94	95	95	96	96	91	**100**									
TcCL	95	94	95	95	96	96	91	100	**100**								
TcTula	95	94	95	95	96	96	91	100	100	**100**							
TcM62	95	94	95	95	96	96	91	100	100	100	**100**						
TcCAN	95	94	95	95	96	96	91	100	100	100	100	**100**					
TcG	95	95	96	97	97	96	91	99	99	99	99	99	**100**				
TcSyl	95	95	96	97	97	96	91	99	99	99	99	99	100	**100**			
TcJR	95	95	96	97	97	96	91	99	99	99	99	99	100	100	**100**		
TcEsm	95	94	95	96	96	96	90	98	98	98	98	98	99	99	99	**100**	
TcMar	95	95	95	96	97	96	90	98	98	98	98	98	99	99	99	99	**100**

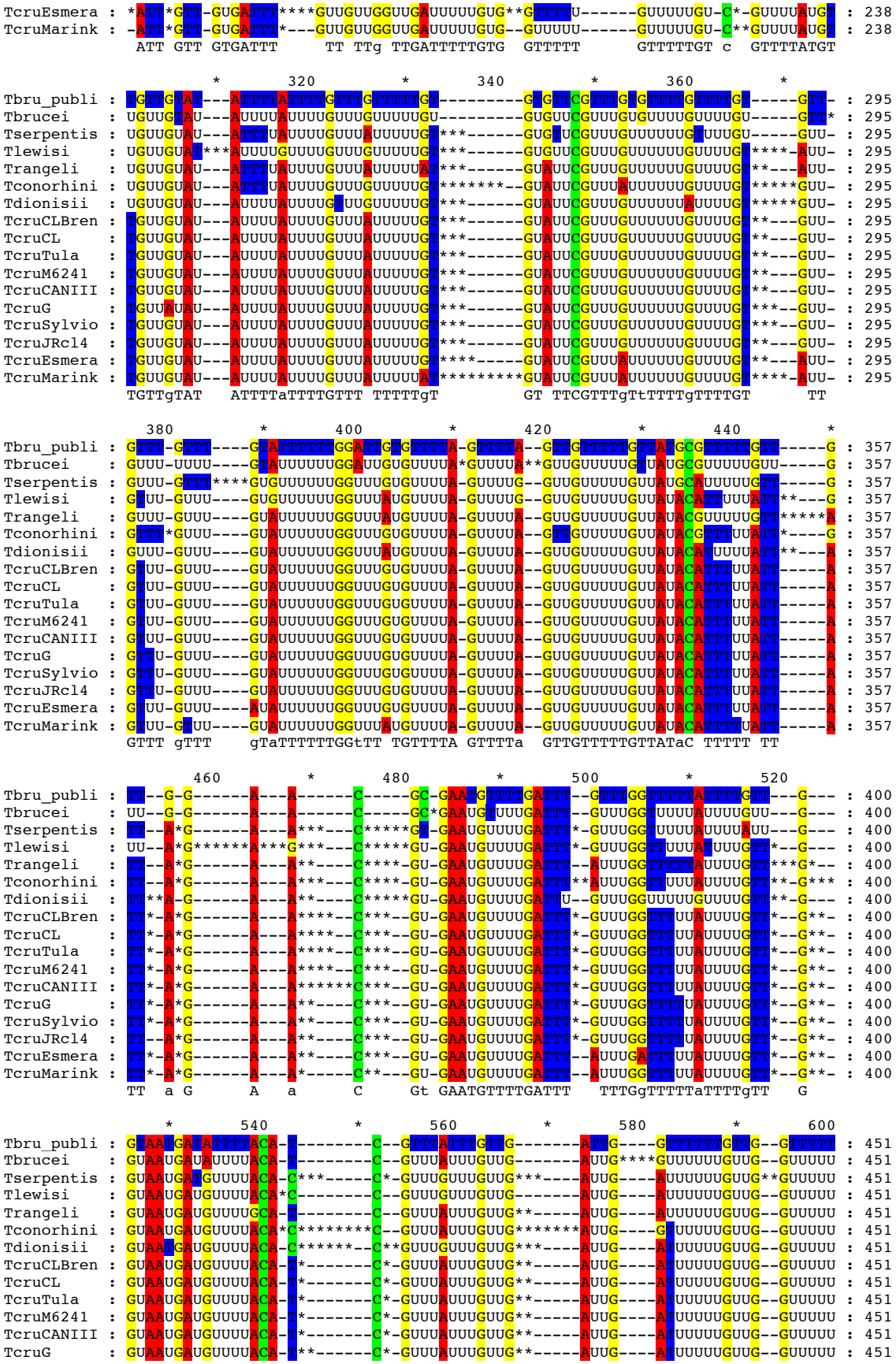
***** For the available length of *Tern* transcript.

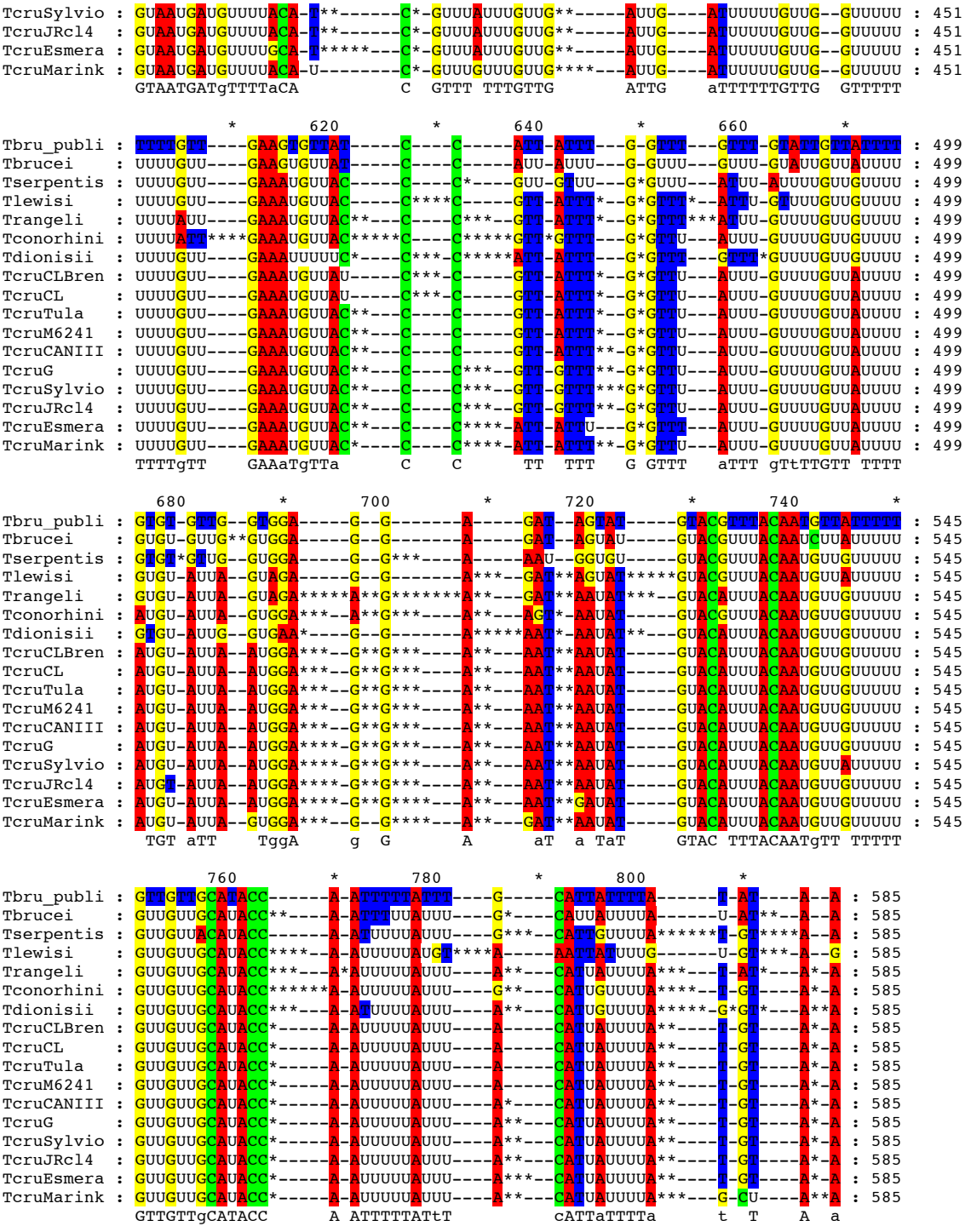
4.SI-S12f. ND8 protein percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern**	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	99	**100**															
Tlew	99	99	**100**														
Tran	99	99	100	**100**													
Tcon	99	99	100	100	**100**												
Tdio	98	97	99	99	99	**100**											
Tern*	81	77	81	81	81	81	**100**										
TcCLB	99	98	98	98	98	98	81	**100**									
TcCL	99	98	98	98	98	98	81	100	**100**								
TcTula	99	98	98	98	98	98	81	100	100	**100**							
TcM62	99	98	98	98	98	98	81	100	100	100	**100**						
TcCAN	99	98	98	98	98	98	81	100	100	100	100	**100**					
TcG	100	99	99	99	99	98	81	99	99	99	99	99	**100**				
TcSyl	100	99	99	99	99	98	81	99	99	99	99	99	100	**100**			
TcJR	100	99	99	99	99	98	81	99	99	99	99	99	100	100	**100**		
TcEsm	99	99	99	99	99	99	81	99	99	99	99	99	99	99	99	**100**	
TcMar	99	99	100	100	100	99	81	98	98	98	98	98	99	99	99	99	**100**

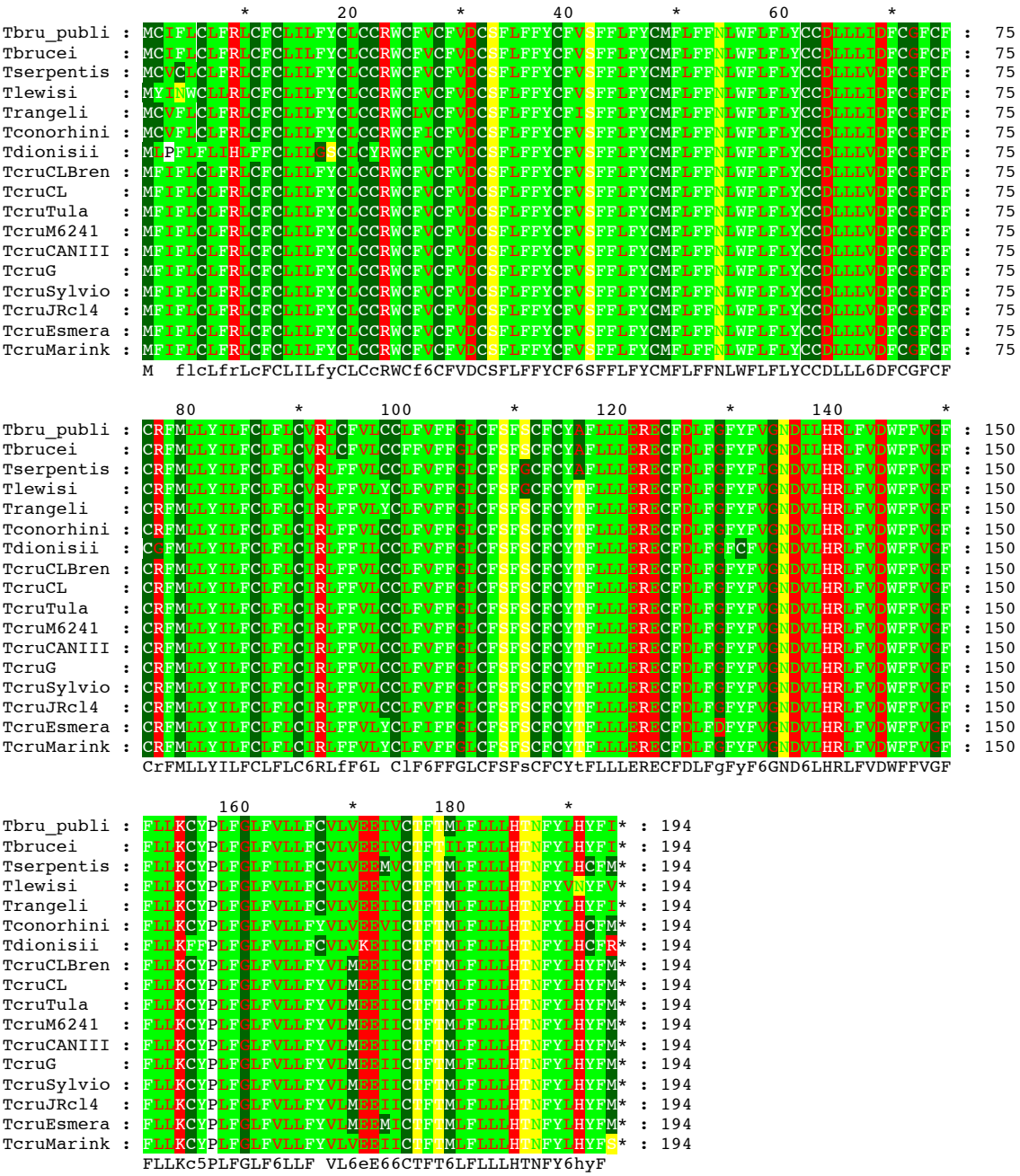
***** For the available length of *Tern* protein.







4.SI-S13a: GeneDoc ND9 / G2 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is cDNA of *T. bru* (Tbru_publi). *T. erneyi* has been left out due to lack of sequence information.



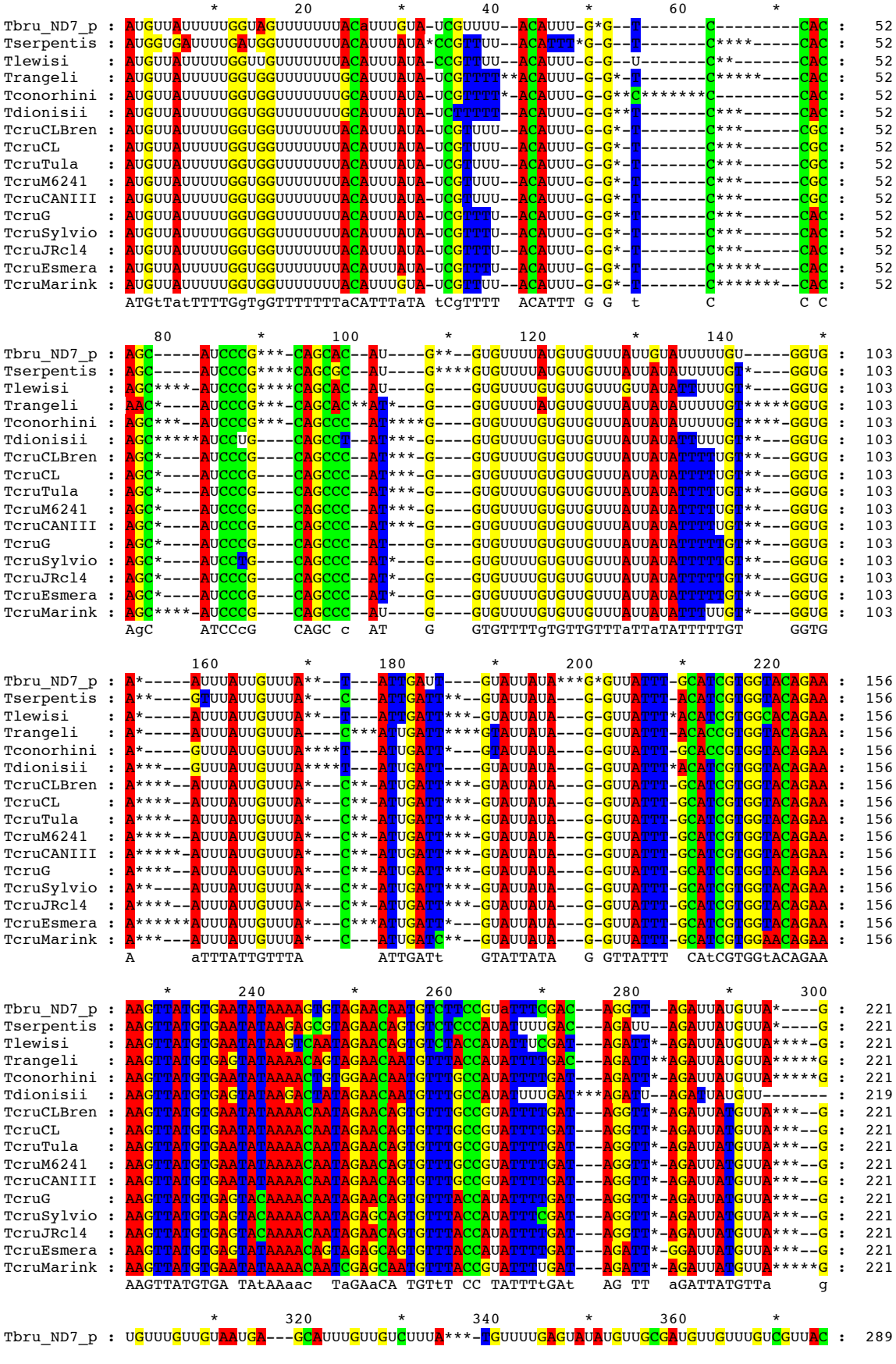
4.SI-S13b: GeneDoc ND9/ G2 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Reference have also been translated. *T. erneyi* has been left out due to lack of sequence information.

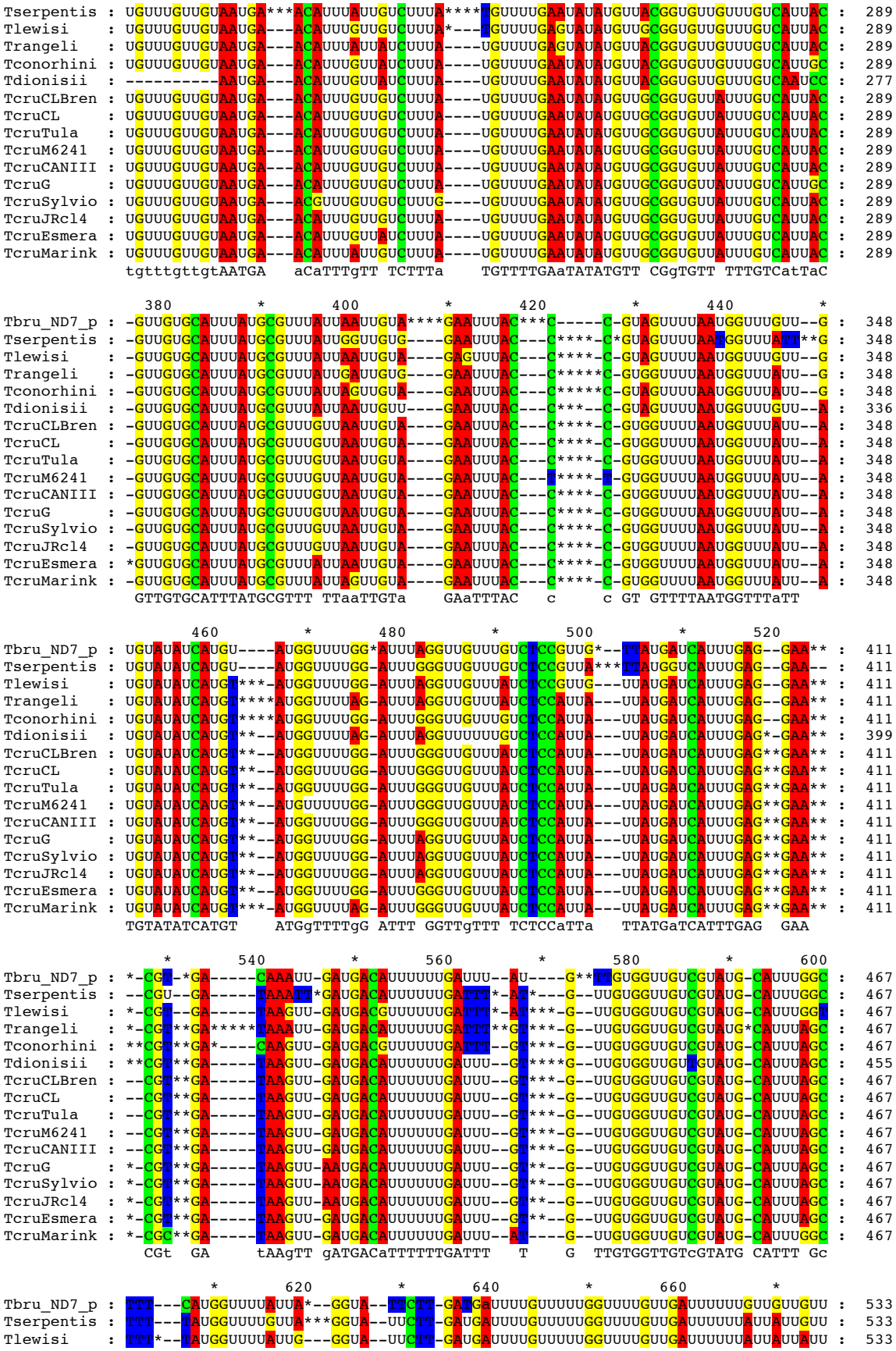
4.SI-S13c. ND9 edited transcript identiy matrix.

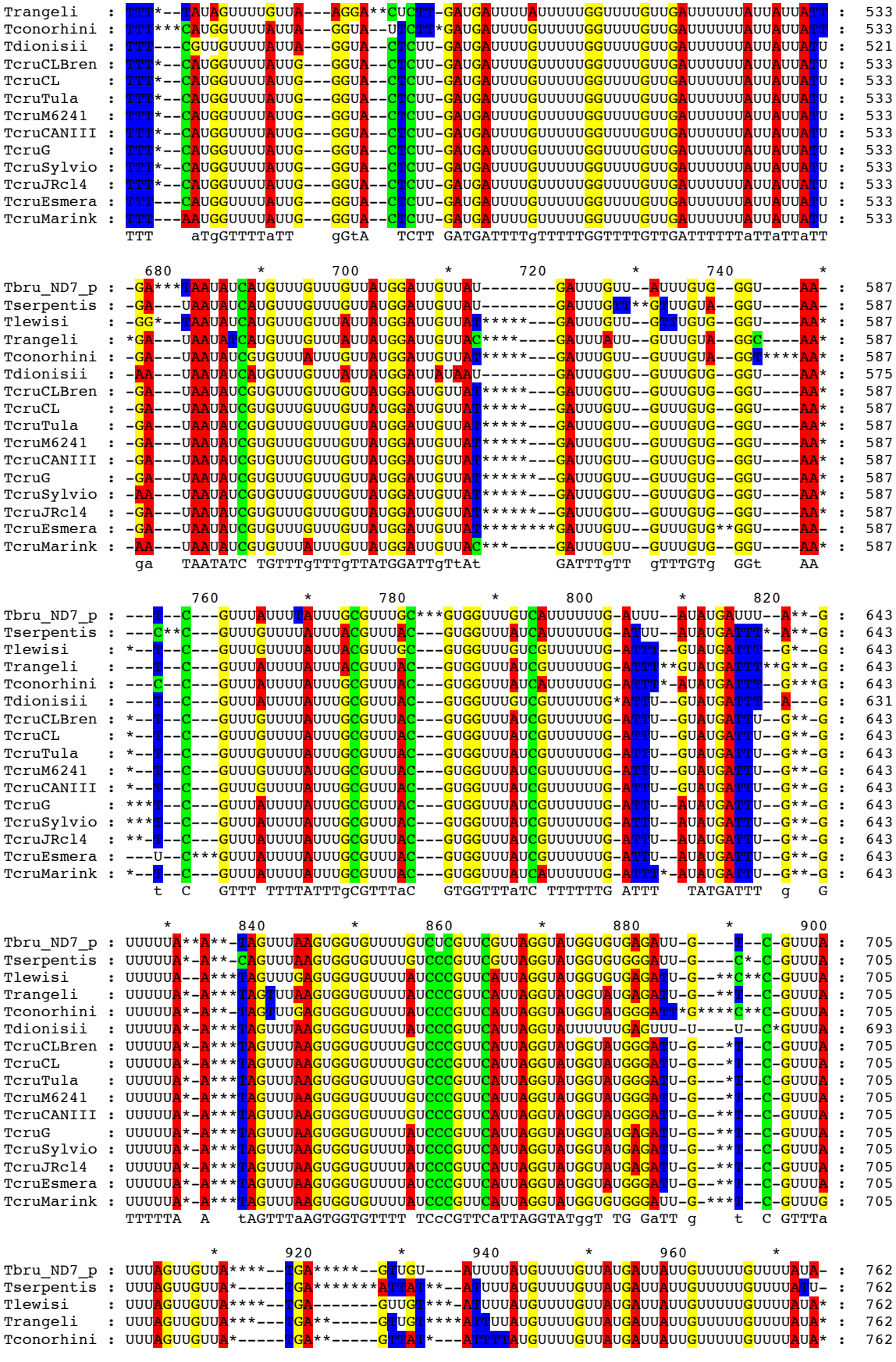
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	93	**100**														
Tlew	92	94	**100**													
Tran	93	92	93	**100**												
Tcon	94	95	94	95	**100**											
Tdio	91	92	91	91	92	**100**										
TcCLB	94	94	93	96	95	94	**100**									
TcCL	94	94	93	96	95	94	100	**100**								
TcTula	94	94	94	96	95	94	100	100	**100**							
TcM62	94	94	94	96	95	94	100	100	100	**100**						
TcCAN	94	94	94	96	95	94	100	100	100	100	**100**					
TcG	93	93	93	95	95	93	99	99	99	99	99	**100**				
TcSyl	94	93	93	95	95	93	99	99	99	99	99	100	**100**			
TcJR	94	94	93	95	95	93	99	99	99	99	99	100	100	**100**		
TcEsm	92	93	92	95	94	92	97	97	98	98	98	97	97	97	**100**	
TcMar	94	93	94	96	95	93	97	97	97	97	97	96	96	97	97	**100**

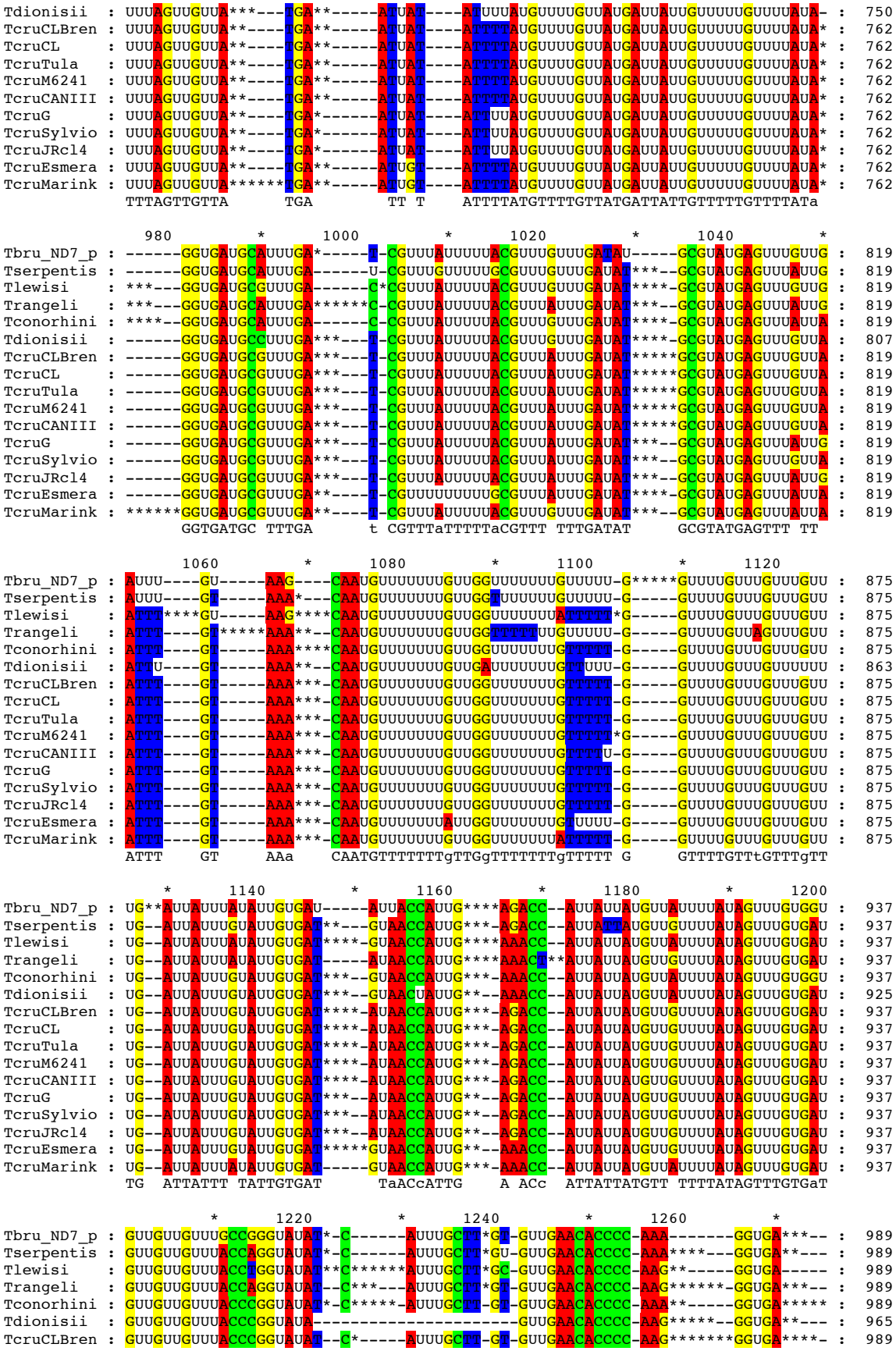
4.SI-S13d. ND9 protein identiy matrix.

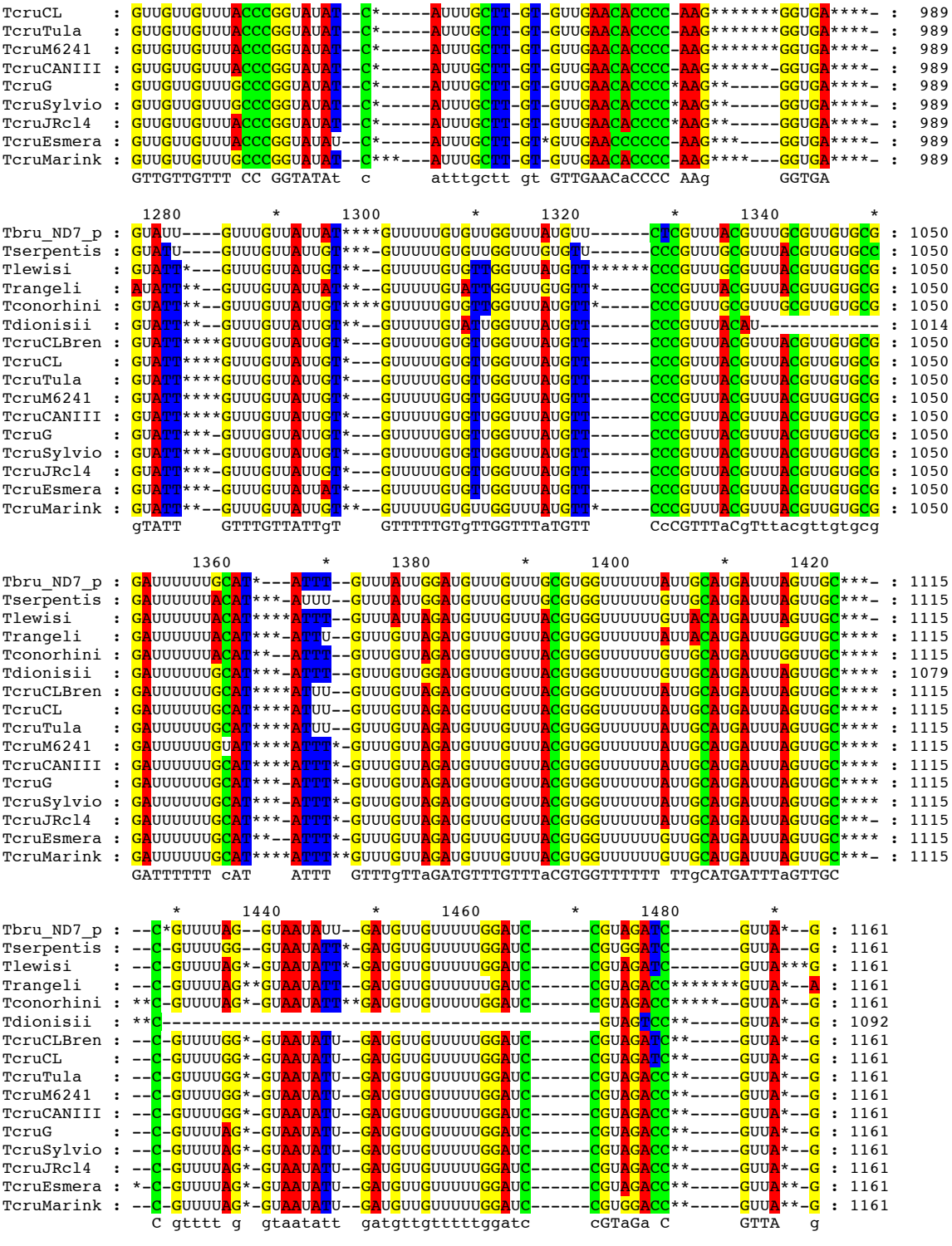
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	94	**100**														
Tlew	94	92	**100**													
Tran	95	93	93	**100**												
Tcon	94	94	92	96	**100**											
Tdio	88	88	87	89	90	**100**										
TcCLB	95	93	93	95	96	90	**100**									
TcCL	95	93	93	95	96	90	100	**100**								
TcTula	95	93	93	95	96	90	100	100	**100**							
TcM62	95	93	93	95	96	90	100	100	100	**100**						
TcCAN	95	93	93	95	96	90	100	100	100	100	**100**					
TcG	95	93	93	95	96	90	100	100	100	100	100	**100**				
TcSyl	95	93	93	95	96	90	100	100	100	100	100	100	**100**			
TcJR	95	93	93	95	96	90	100	100	100	100	100	100	100	**100**		
TcEsm	93	92	92	94	95	88	98	98	98	98	98	98	98	98	**100**	
TcMar	95	93	94	96	96	90	98	98	98	98	98	98	98	98	97	**100**



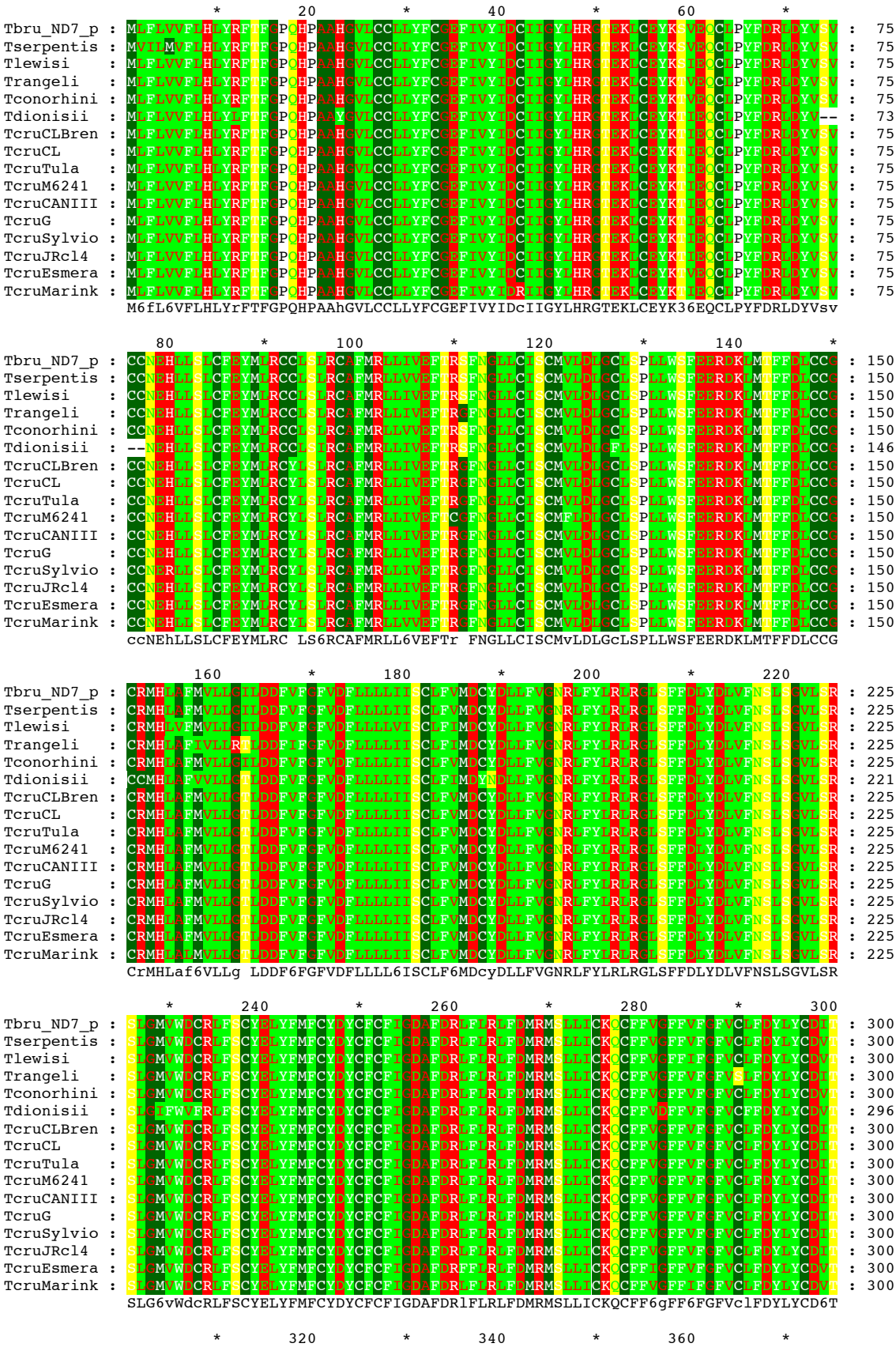


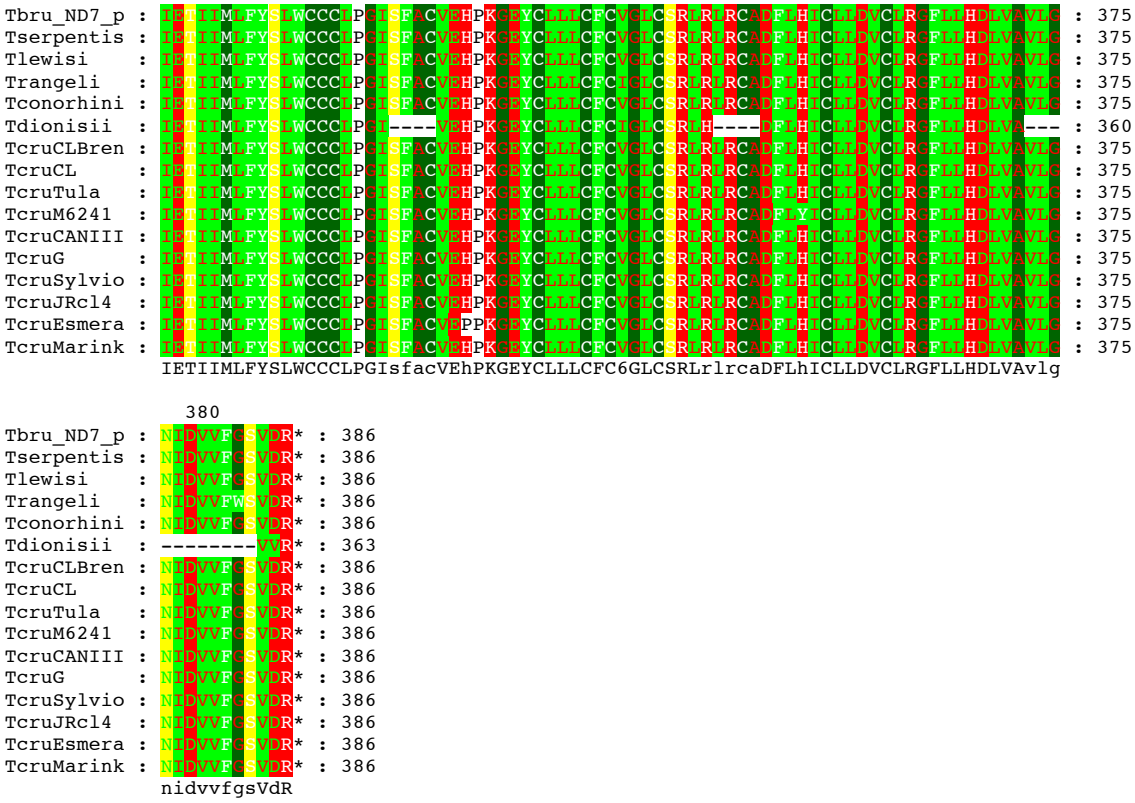

Trangeli : 533
Tconorhini : 533
Tdionisii : 521
TcruCLBren : 533
TcruCL : 533
TcruTula : 533
TcruM6241 : 533
TcruCANIII : 533
TcruG : 533
TcruSylvio : 533
TcruJRcl4 : 533
TcruEsmera : 533
TcruMarink : 533
680
700
720
740
Tbru_ND7_p : 587
Tserpentis : 587
Tlewisi : 587
Trangeli : 587
Tconorhini : 587
Tdionisii : 575
TcruCLBren : 587
TcruCL : 587
TcruTula : 587
TcruM6241 : 587
TcruCANIII : 587
TcruG : 587
TcruSylvio : 587
TcruJRcl4 : 587
TcruEsmera : 587
TcruMarink : 587
760
780
800
820
Tbru_ND7_p : 643
Tserpentis : 643
Tlewisi : 643
Trangeli : 643
Tconorhini : 643
Tdionisii : 631
TcruCLBren : 643
TcruCL : 643
TcruTula : 643
TcruM6241 : 643
TcruCANIII : 643
TcruG : 643
TcruSylvio : 643
TcruJRcl4 : 643
TcruEsmera : 643
TcruMarink : 643
840
860
880
900
Tbru_ND7_p : 705
Tserpentis : 705
Tlewisi : 705
Trangeli : 705
Tconorhini : 705
Tdionisii : 693
TcruCLBren : 705
TcruCL : 705
TcruTula : 705
TcruM6241 : 705
TcruCANIII : 705
TcruG : 705
TcruSylvio : 705
TcruJRcl4 : 705
TcruEsmera : 705
TcruMarink : 705
920
940
960
Tbru_ND7_p : 762
Tserpentis : 762
Tlewisi : 762
Trangeli : 762
Tconorhini : 762





4.SI-S14a: GeneDoc ND7 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is RNA sequence of *Tbru* (Tbru_ND7_p).


Tbru_ND7_p
Tserpentis
Tlewisi
Trangeli
Tconorhini
Tdionisii
TcruCLBren
TcruCL
TcruTula
TcruM6241
TcruCANIII
TcruG
TcruSylvio
TcruJRcl4
TcruEsmera
TcruMarink



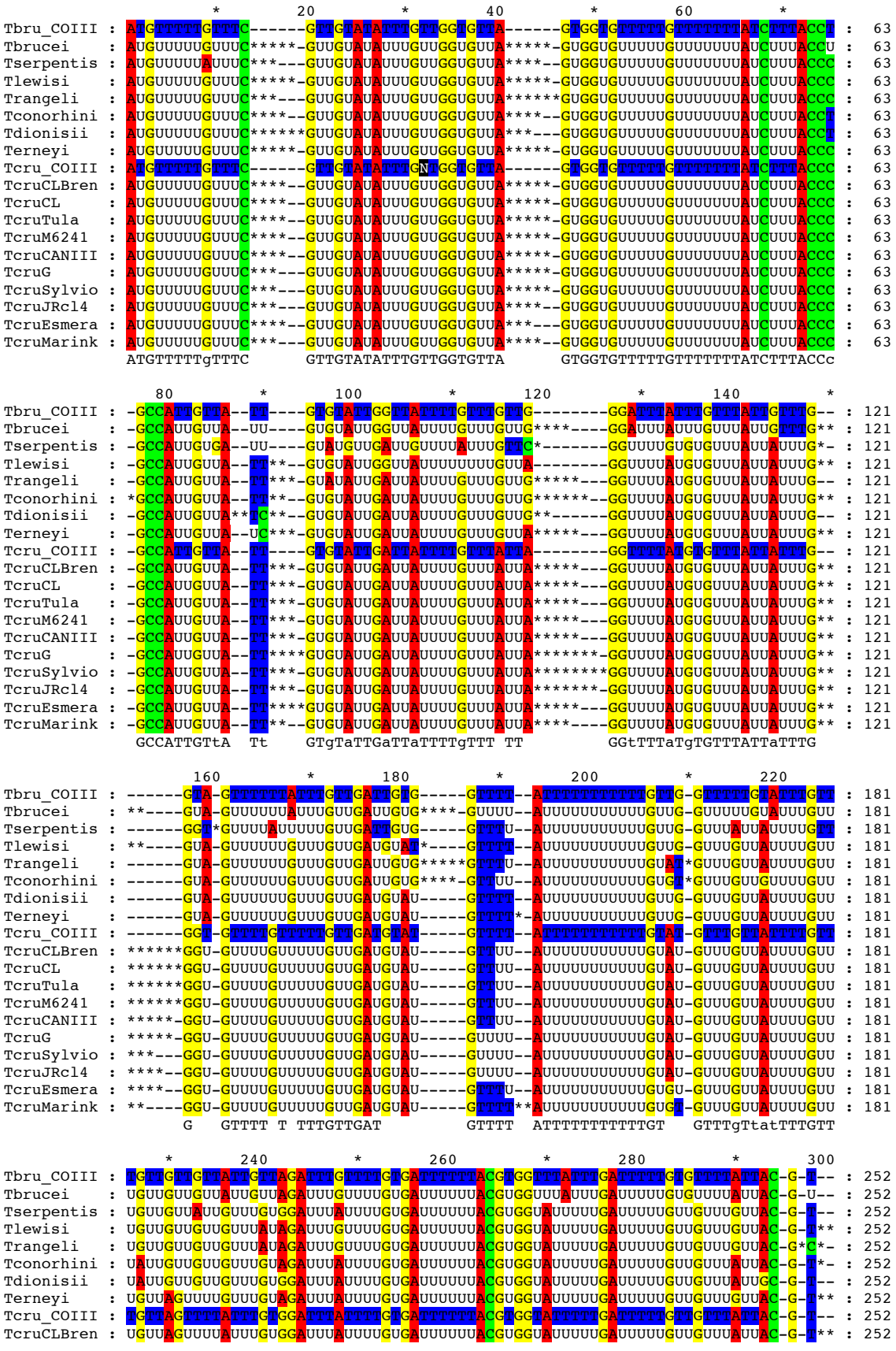
4.SI-S14b: GeneDoc ND7 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Reference has also been translated.

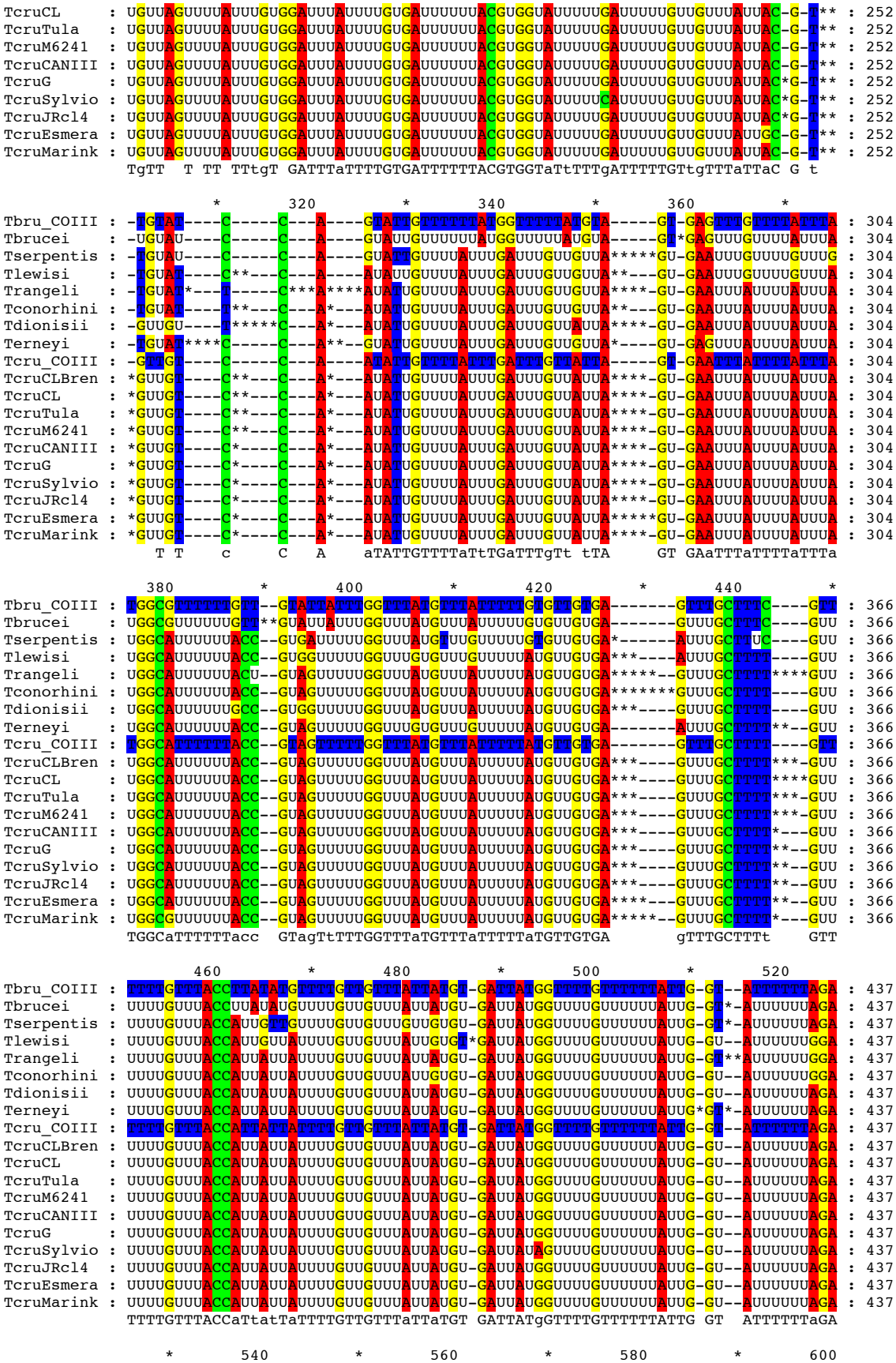
4.SI-S14c. ND7 edited transcript percent identity matrix.

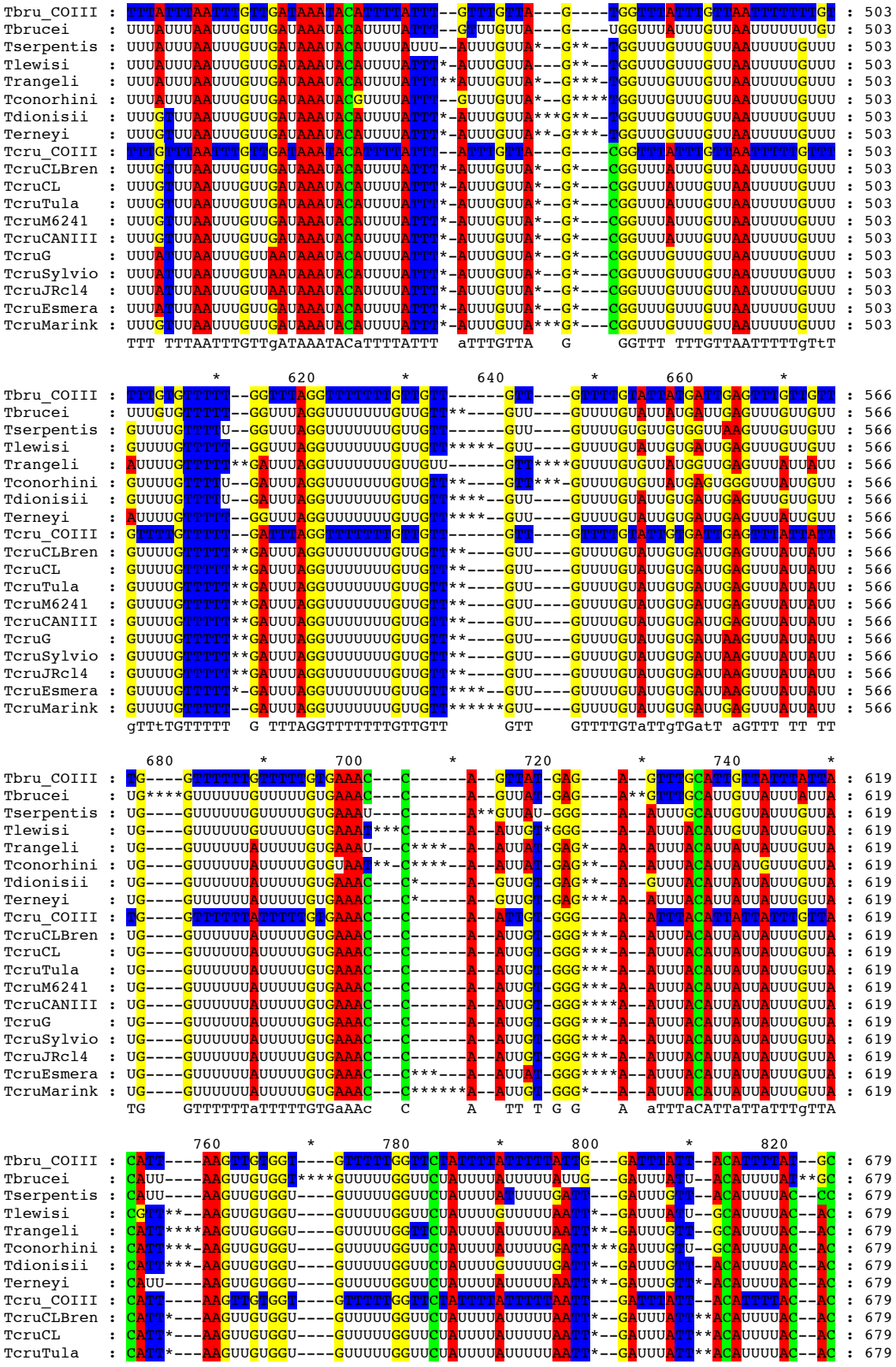
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	94	**100**														
Tlew	94	94	**100**													
Tran	93	93	94	**100**												
Tcon	94	94	94	94	**100**											
Tdio	92	92	93	93	94	**100**										
TcCLB	94	94	95	94	96	95	**100**									
TcCL	94	94	95	94	96	95	100	**100**								
TcTula	94	94	95	94	96	95	100	100	**100**							
TcM62	94	94	94	94	95	94	100	100	100	**100**						
TcCAN	94	94	95	94	96	95	100	100	100	100	**100**					
TcG	94	94	95	95	96	95	99	99	99	98	99	**100**				
TcSyl	94	93	94	94	95	95	98	98	98	98	98	99	**100**			
TcJR	94	94	95	95	96	95	99	99	99	98	99	100	99	**100**		
TcEsm	93	94	94	95	96	95	98	98	98	98	98	98	98	98	**100**	
TcMar	94	93	94	94	96	94	97	97	97	96	97	97	97	97	97	**100**

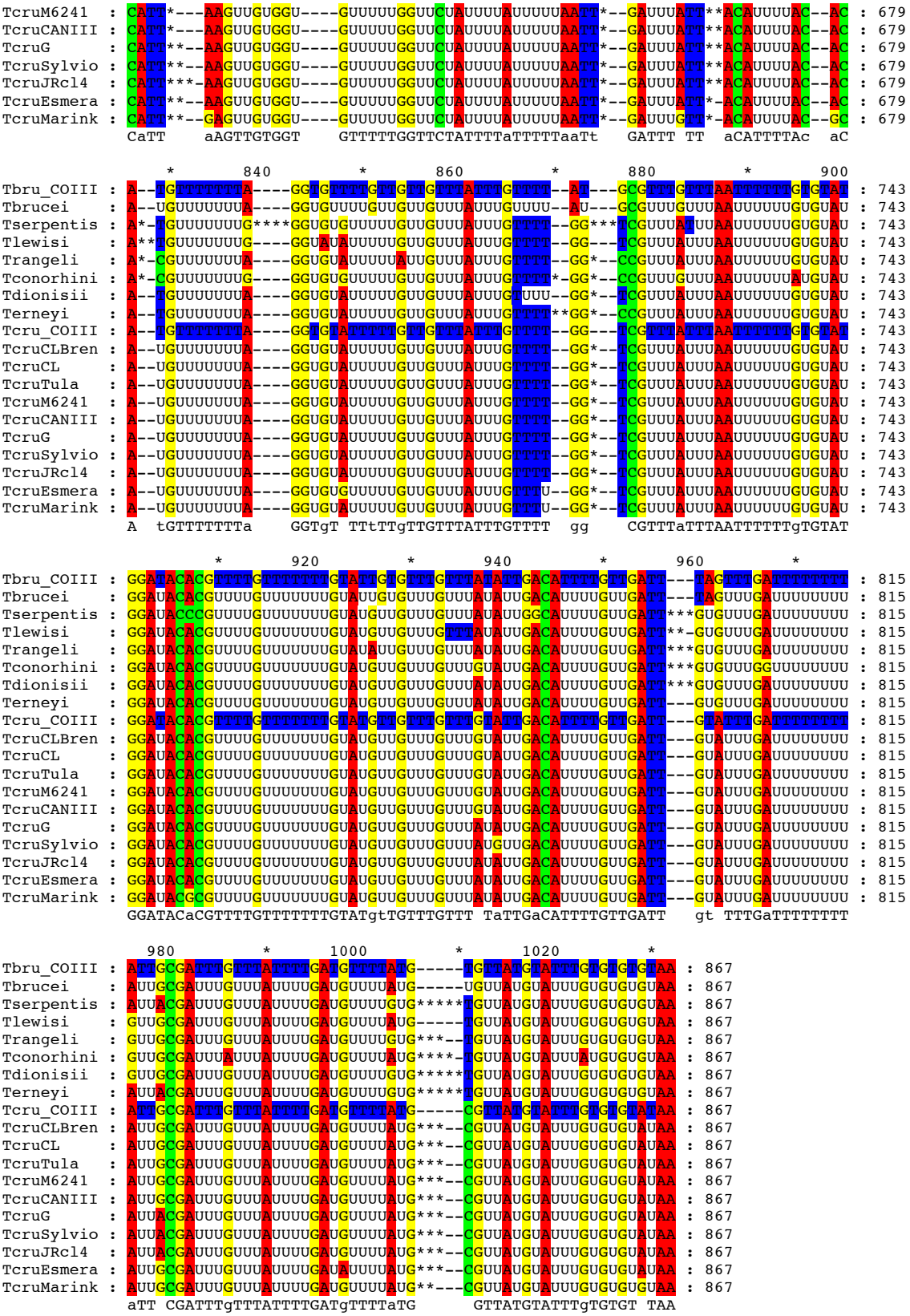
4.SI-S14d. ND7 protein percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	99	**100**														
Tlew	98	98	**100**													
Tran	97	96	96	**100**												
Tcon	99	99	98	97	**100**											
Tdio	94	93	94	94	94	**100**										
TcCLB	99	97	98	98	98	94	**100**									
TcCL	99	97	98	98	98	94	100	**100**								
TcTula	99	97	98	98	98	94	100	100	**100**							
TcM62	98	97	97	97	98	93	99	99	99	**100**						
TcCAN	99	97	98	98	98	94	100	100	100	99	**100**					
TcG	99	97	98	98	98	94	100	100	100	99	100	**100**				
TcSyl	98	97	97	97	98	94	100	100	100	99	100	100	**100**			
TcJR	99	97	98	98	98	94	100	100	100	99	100	100	100	**100**		
TcEsm	98	97	97	97	98	93	99	99	99	98	99	99	98	99	**100**	
TcMar	97	97	97	96	98	93	99	99	99	98	99	99	98	99	98	**100**

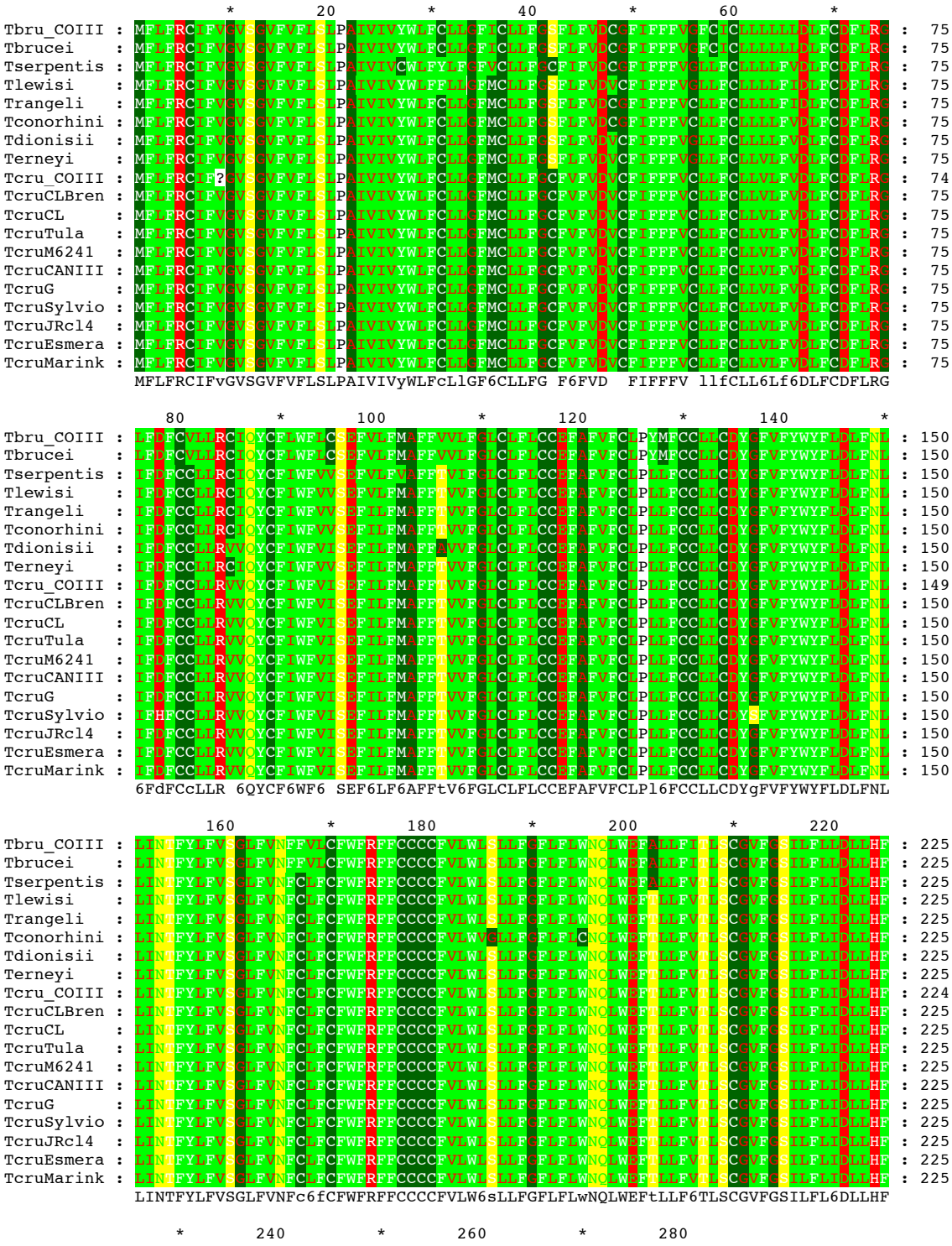


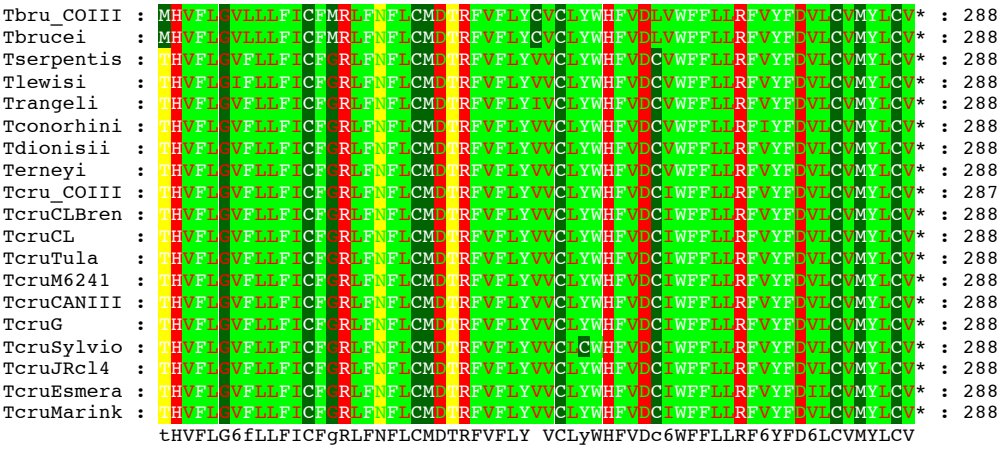






4.SI-S15a: GeneDoc COIII nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. References are cDNAs of *Tbru* (Tbru_COIII) and *Tcru* (Tcru_COIII).





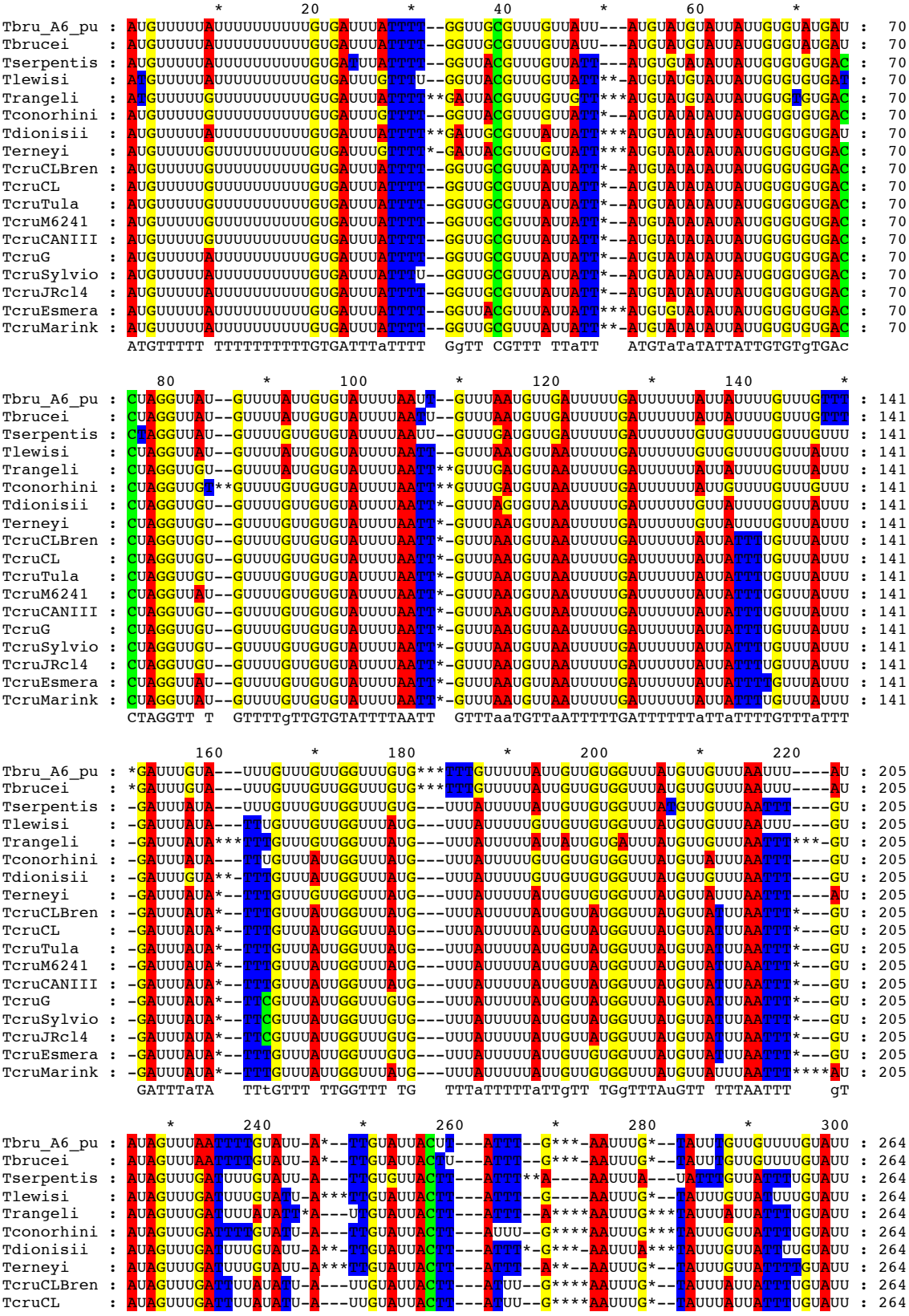
4.SI-S15b: GeneDoc COIII protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. References have also been translated. '?' in Tcru_COIII indicates lack of amino acid information due to 'N' in the corresponding nucleotide sequence above (S15a)

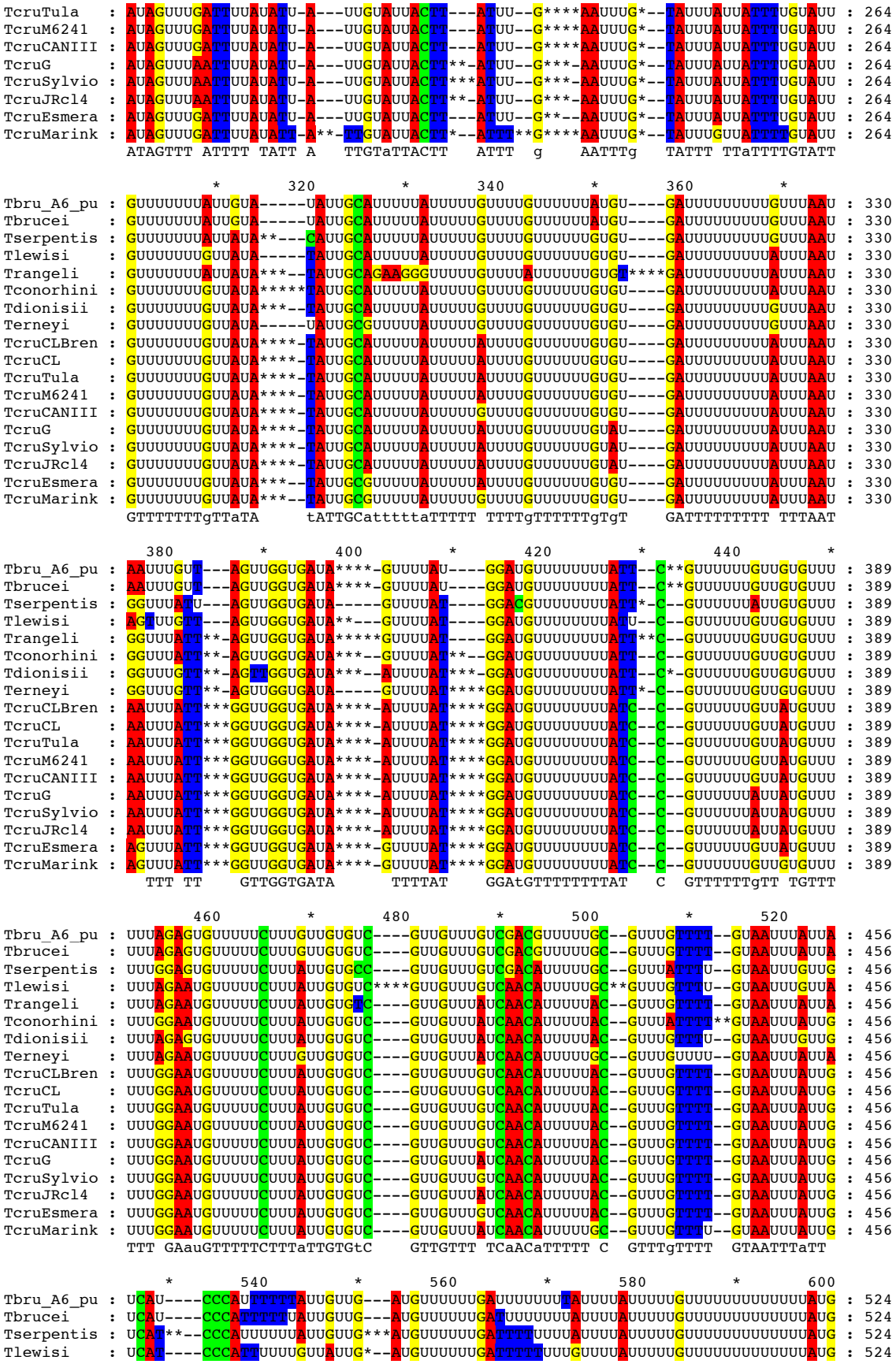
4.SI-S15c. COIII edited transcript percent identity matrix.

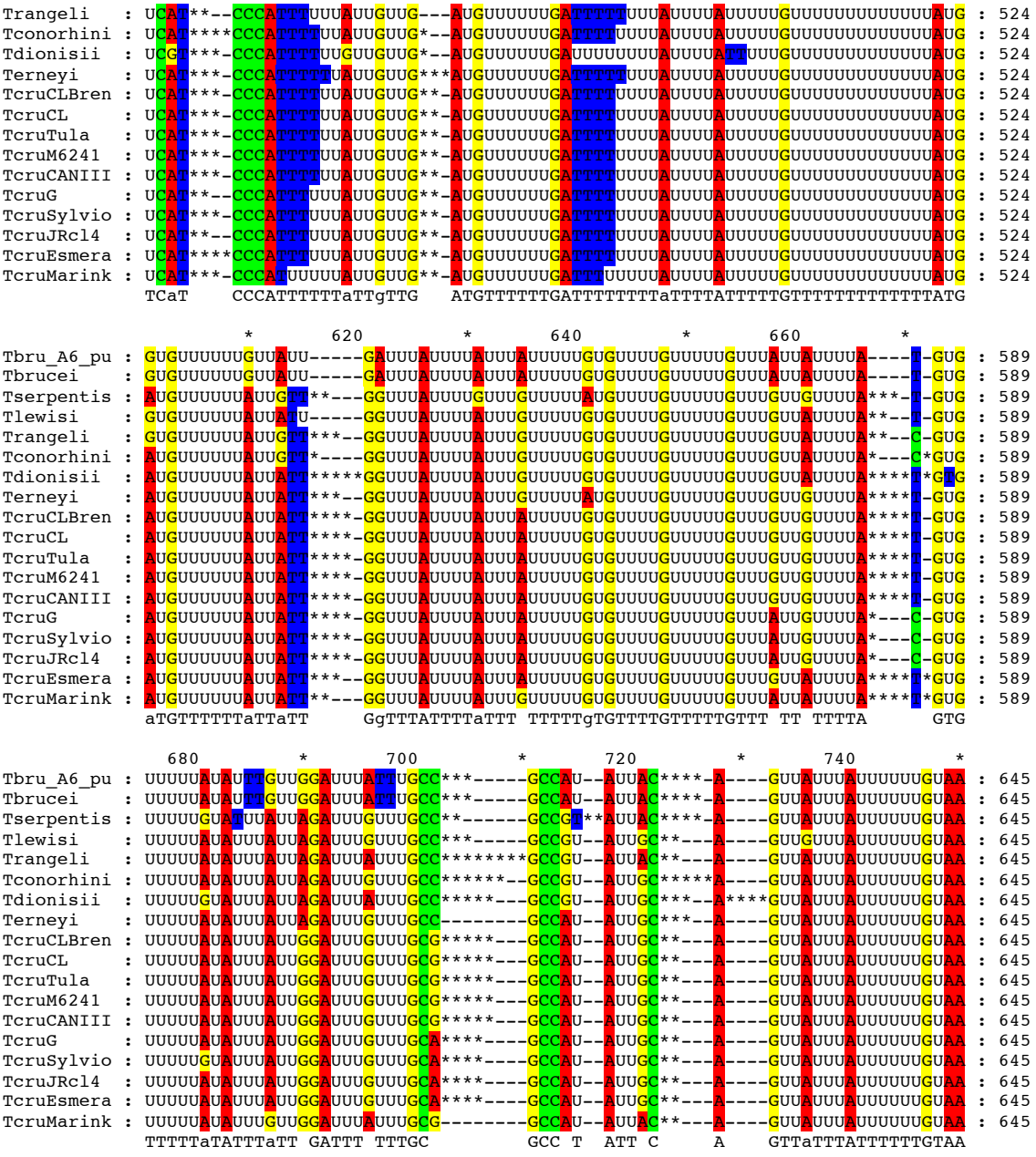
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	89	**100**															
Tlew	90	94	**100**														
Tran	90	93	95	**100**													
Tcon	90	92	94	96	**100**												
Tdio	90	93	95	95	95	**100**											
Tern	90	94	96	96	94	97	**100**										
TcCLB	89	92	94	94	93	96	96	**100**									
TcCL	89	92	94	94	93	96	96	100	**100**								
TcTula	89	92	94	94	93	96	96	100	100	**100**							
TcM62	89	92	94	94	93	96	96	100	100	100	**100**						
TcCAN	89	92	94	94	93	96	96	100	100	100	100	**100**					
TcG	89	93	94	94	93	96	96	99	99	99	99	99	**100**				
TcSyl	88	92	94	94	93	95	96	99	99	99	99	99	100	**100**			
TcJR	89	93	94	94	93	96	96	99	99	99	99	99	100	100	**100**		
TcEsm	89	93	94	94	94	96	96	99	99	99	99	99	99	99	99	**100**	
TcMar	89	92	94	94	93	96	96	99	99	99	99	99	99	98	99	99	**100**

4.SI-S15d. COIII protein percent identity matrix.

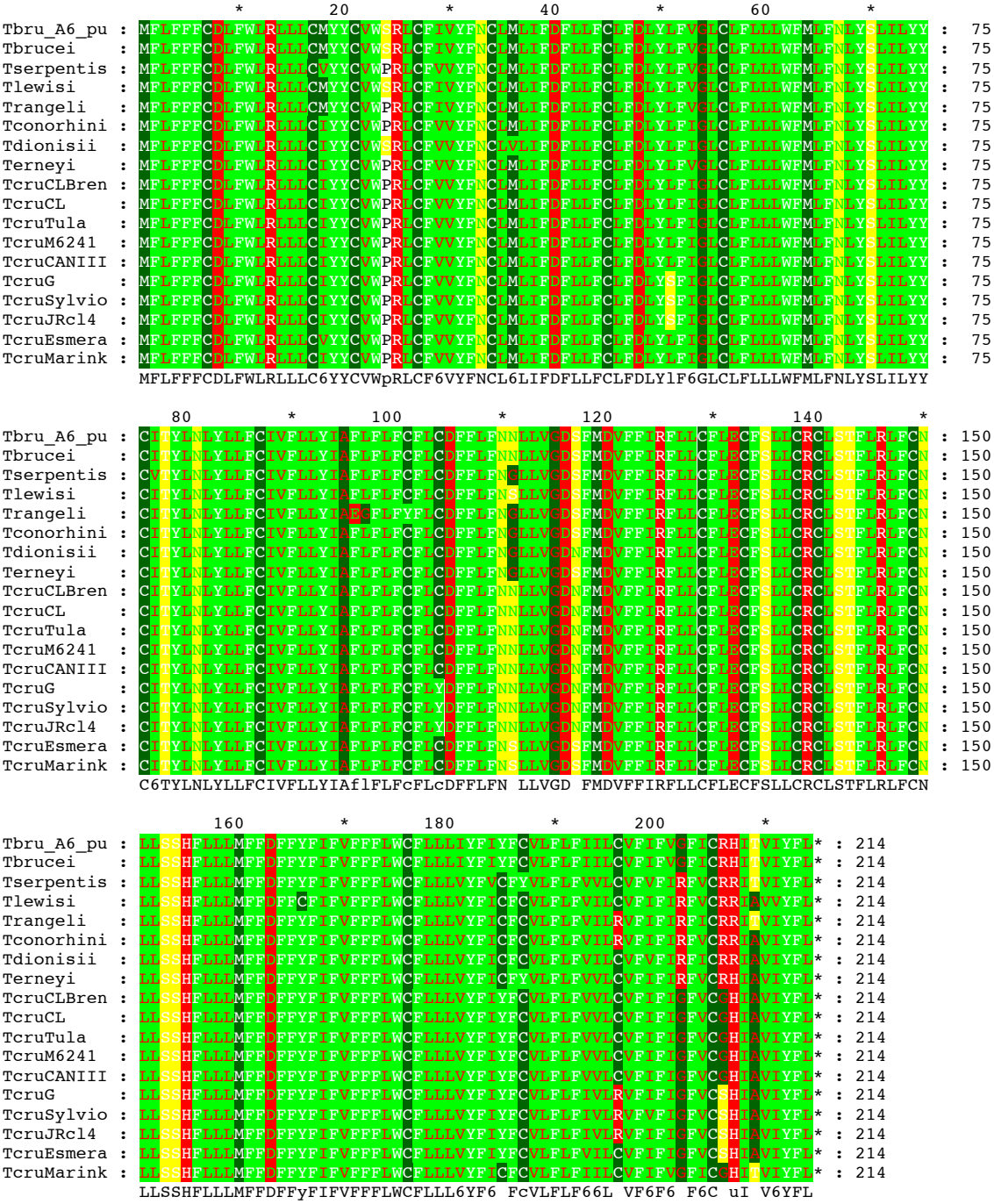
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	89	**100**															
Tlew	90	95	**100**														
Tran	90	95	98	**100**													
Tcon	89	95	96	98	**100**												
Tdio	89	94	97	97	96	**100**											
Tern	90	95	98	98	97	98	**100**										
TcCLB	88	94	96	96	95	98	98	**100**									
TcCL	88	94	96	96	95	98	98	100	**100**								
TcTula	88	94	96	96	95	98	98	100	100	**100**							
TcM62	88	94	96	96	95	98	98	100	100	100	**100**						
TcCAN	88	94	96	96	95	98	98	100	100	100	100	**100**					
TcG	88	94	96	96	95	98	98	100	100	100	100	100	**100**				
TcSyl	86	93	95	95	94	97	97	100	99	99	99	99	99	**100**			
TcJR	88	94	96	96	95	98	98	100	100	100	100	100	100	99	**100**		
TcEsm	87	94	95	96	95	98	97	100	100	100	100	100	100	99	100	**100**	
TcMar	88	94	96	96	95	98	98	100	100	100	100	100	100	99	100	100	**100**







4.SI-S16a: GeneDoc ATPase 6 / MURF4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position to maintain the frame. Reference is RNA of *Tbru* (Tbru_A6_pu).



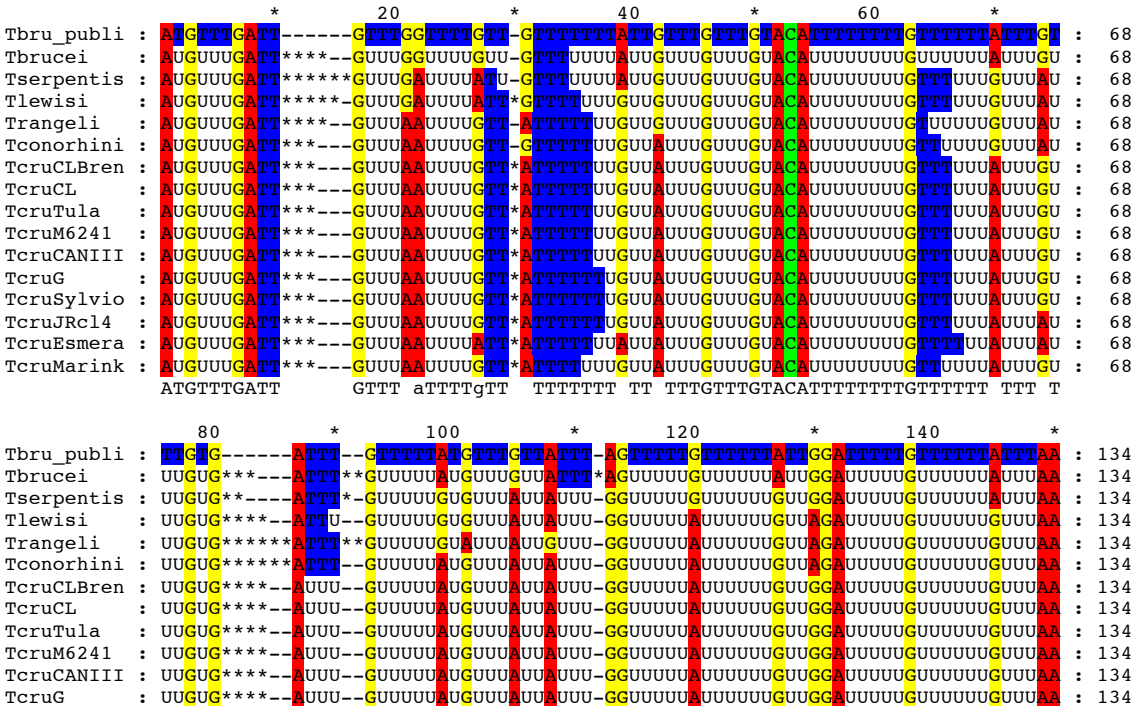
4.SI-S16b: GeneDoc ATPase 6 / MURF4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. References have also been translated and shown.

4.SI-S16c. A6 edited transcript percent identity matrix.

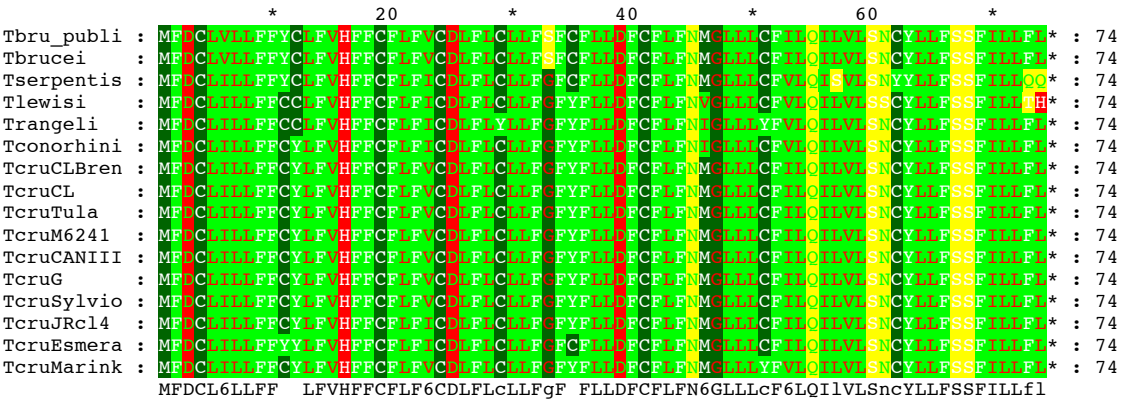
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	93	**100**															
Tlew	94	94	**100**														
Tran	92	93	94	**100**													
Tcon	93	95	96	96	**100**												
Tdio	93	94	96	94	96	**100**											
Tern	94	95	96	95	97	96	**100**										
TcCLB	93	93	94	94	96	95	95	**100**									
TcCL	93	93	94	94	96	95	95	100	**100**								
TcTula	93	93	94	94	96	95	95	100	100	**100**							
TcM62	93	93	94	93	96	95	95	100	100	100	**100**						
TcCAN	93	93	94	94	96	95	96	100	100	100	100	**100**					
TcG	93	92	93	93	95	94	94	98	98	98	98	98	**100**				
TcSyl	93	93	93	93	95	94	94	98	98	98	98	98	100	**100**			
TcJR	93	92	93	93	95	94	94	98	98	98	98	98	100	100	**100**		
TcEsm	93	94	95	94	96	95	95	98	98	98	98	98	98	98	98	**100**	
TcMar	94	93	95	94	96	95	96	97	97	97	97	97	97	96	97	98	**100**

4.SI-S16d. A6 protein percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	93	**100**															
Tlew	95	94	**100**														
Tran	94	94	95	**100**													
Tcon	93	96	96	96	**100**												
Tdio	93	95	95	94	97	**100**											
Tern	94	97	96	95	98	96	**100**										
TcCLB	94	94	94	93	96	95	97	**100**									
TcCL	94	94	94	93	96	95	97	100	**100**								
TcTula	94	94	94	93	96	95	97	100	100	**100**							
TcM62	94	94	94	93	96	95	97	100	100	100	**100**						
TcCAN	94	94	94	93	96	95	97	100	100	100	100	**100**					
TcG	93	92	92	92	95	93	95	98	98	98	98	98	**100**				
TcSyl	93	93	92	91	95	94	94	97	97	97	97	97	100	**100**			
TcJR	93	92	92	92	95	93	95	98	98	98	98	98	100	100	**100**		
TcEsm	95	94	95	93	97	95	96	98	98	98	98	98	96	96	96	**100**	
TcMar	96	93	94	94	96	95	95	96	96	96	96	96	95	94	95	97	**100**



4.SI-S17a: GeneDoc G3 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is the cDNA sequence of *Tbru* (Tbru_publi).



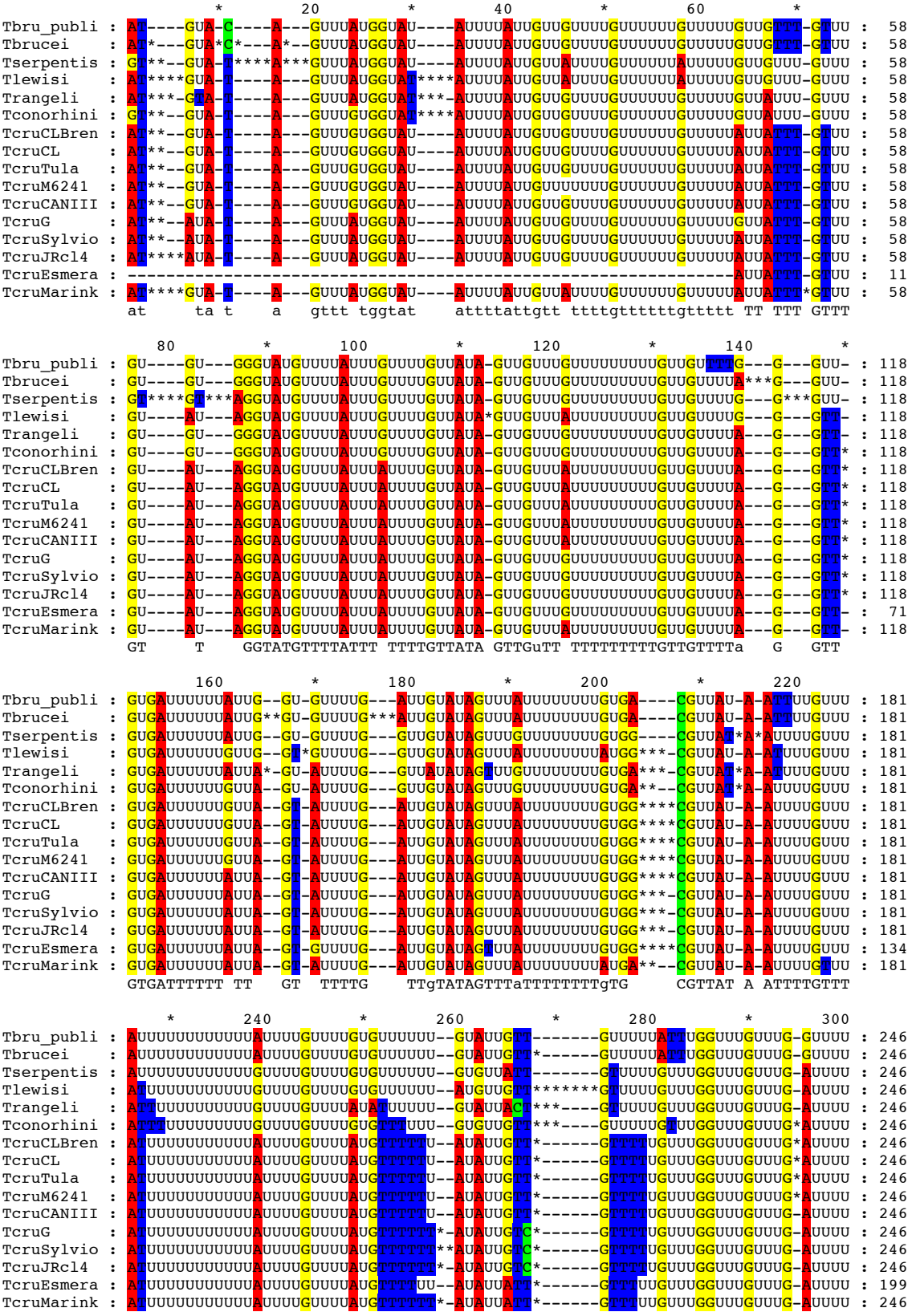
4.SI-S17b: GeneDoc G3 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Reference has also been translated.

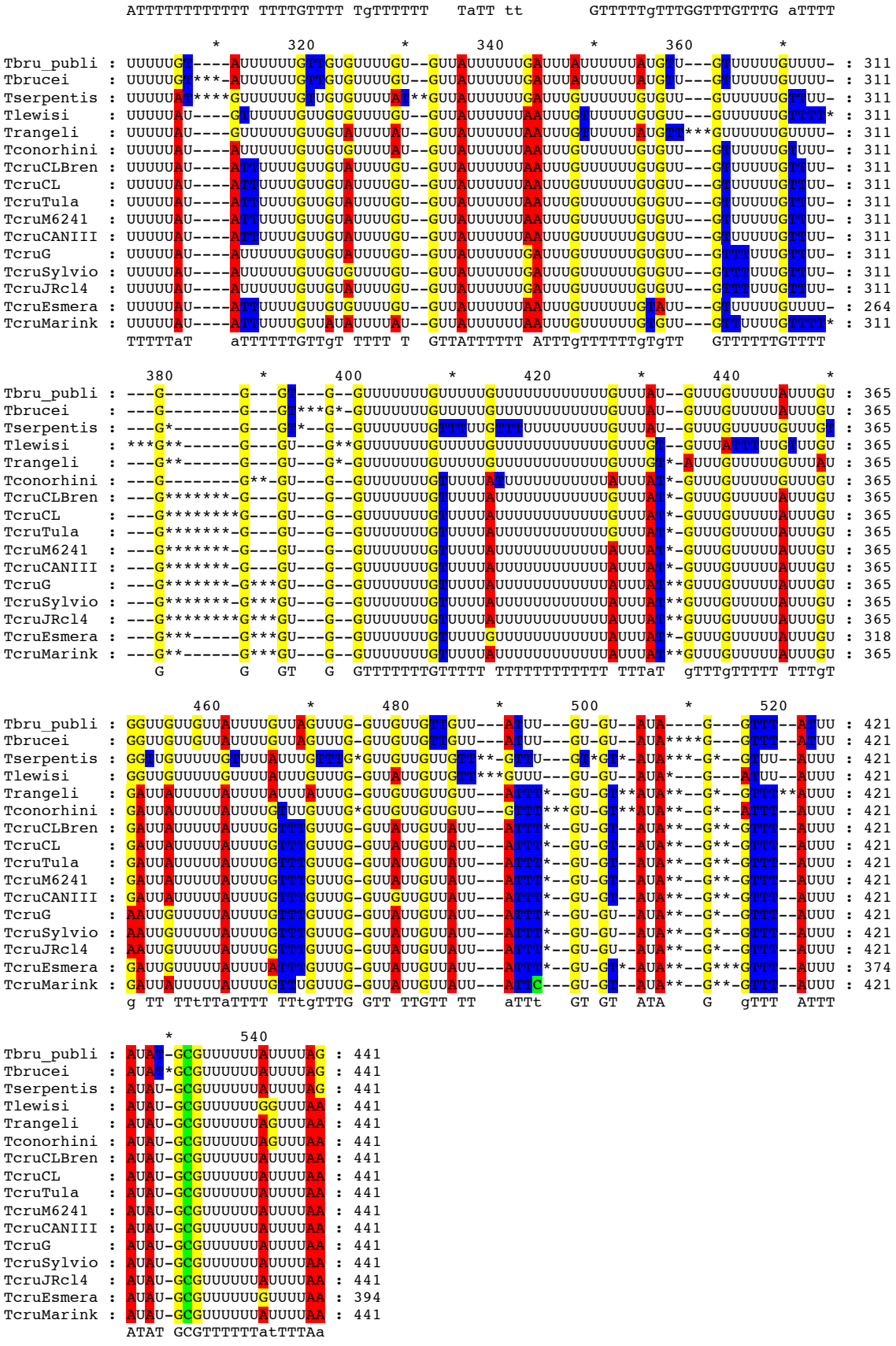
4.SI-S17c. G3 edited transcript percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**														
Tser	92	**100**													
Tlew	88	92	**100**												
Tran	90	89	92	**100**											
Tcon	91	90	92	94	**100**										
TcCLB	94	90	90	94	95	**100**									
TcCL	94	90	90	94	95	100	**100**								
TcTula	94	90	90	94	95	100	100	**100**							
TcM62	94	90	90	94	95	100	100	100	**100**						
TcCAN	94	90	90	94	95	100	100	100	100	**100**					
TcG	94	90	90	94	95	100	100	100	100	100	**100**				
TcSyl	94	90	90	94	95	100	100	100	100	100	100	**100**			
TcJR	94	91	90	94	96	100	100	100	100	100	100	100	**100**		
TcEsm	94	92	90	93	94	98	98	98	98	98	98	98	99	**100**	
TcMar	93	89	90	94	96	98	98	98	98	98	98	98	98	96	**100**

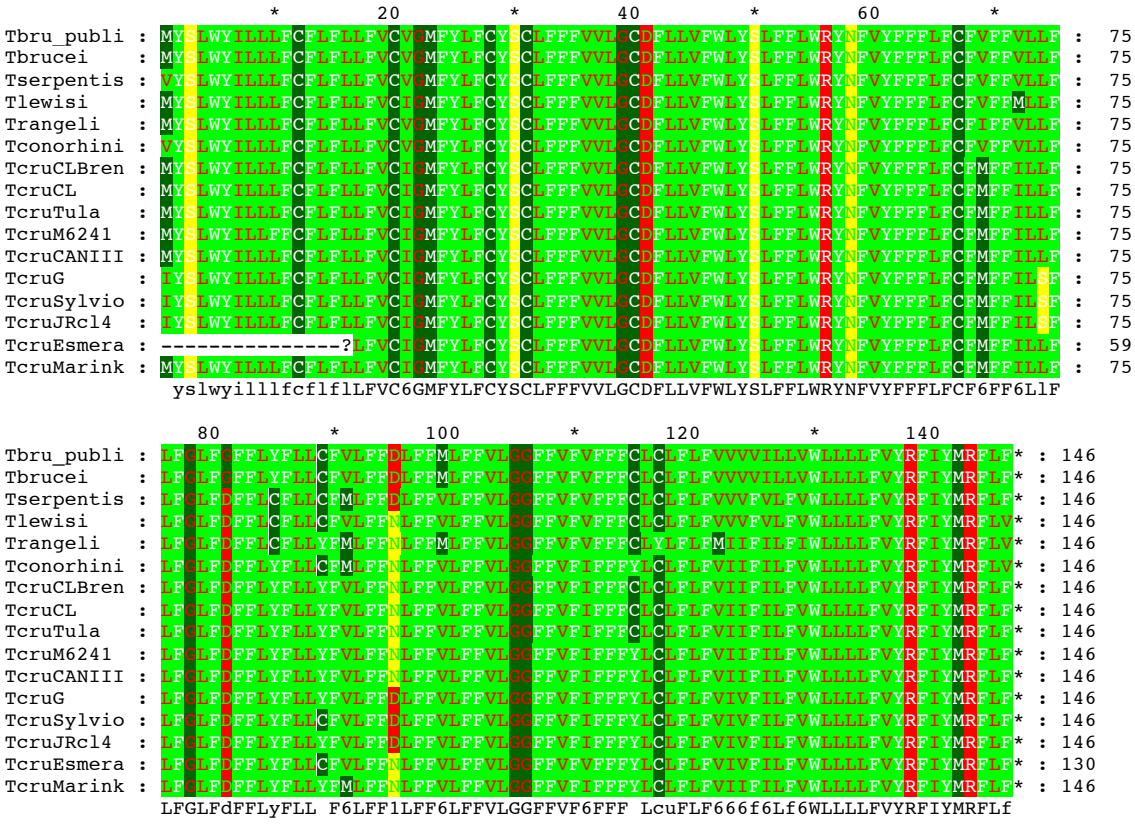
4.SI-S17d. G3 protein percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**														
Tser	89	**100**													
Tlew	86	89	**100**												
Tran	88	88	92	**100**											
Tcon	89	89	93	96	**100**										
TcCLB	93	88	91	92	96	**100**									
TcCL	93	88	91	92	96	100	**100**								
TcTula	93	88	91	92	96	100	100	**100**							
TcM62	93	88	91	92	96	100	100	100	**100**						
TcCAN	93	88	91	92	96	100	100	100	100	**100**					
TcG	93	88	91	92	96	100	100	100	100	100	**100**				
TcSyl	93	88	91	92	96	100	100	100	100	100	100	**100**			
TcJR	92	89	92	93	97	99	99	99	99	99	99	99	**100**		
TcEsm	95	92	89	91	95	96	96	96	96	96	96	96	97	**100**	
TcMar	91	88	91	95	96	97	97	97	97	97	97	97	96	93	**100**





4.SI-S18a: GeneDoc G4 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is RNA sequence of *Tbru* (Tbru_publi).



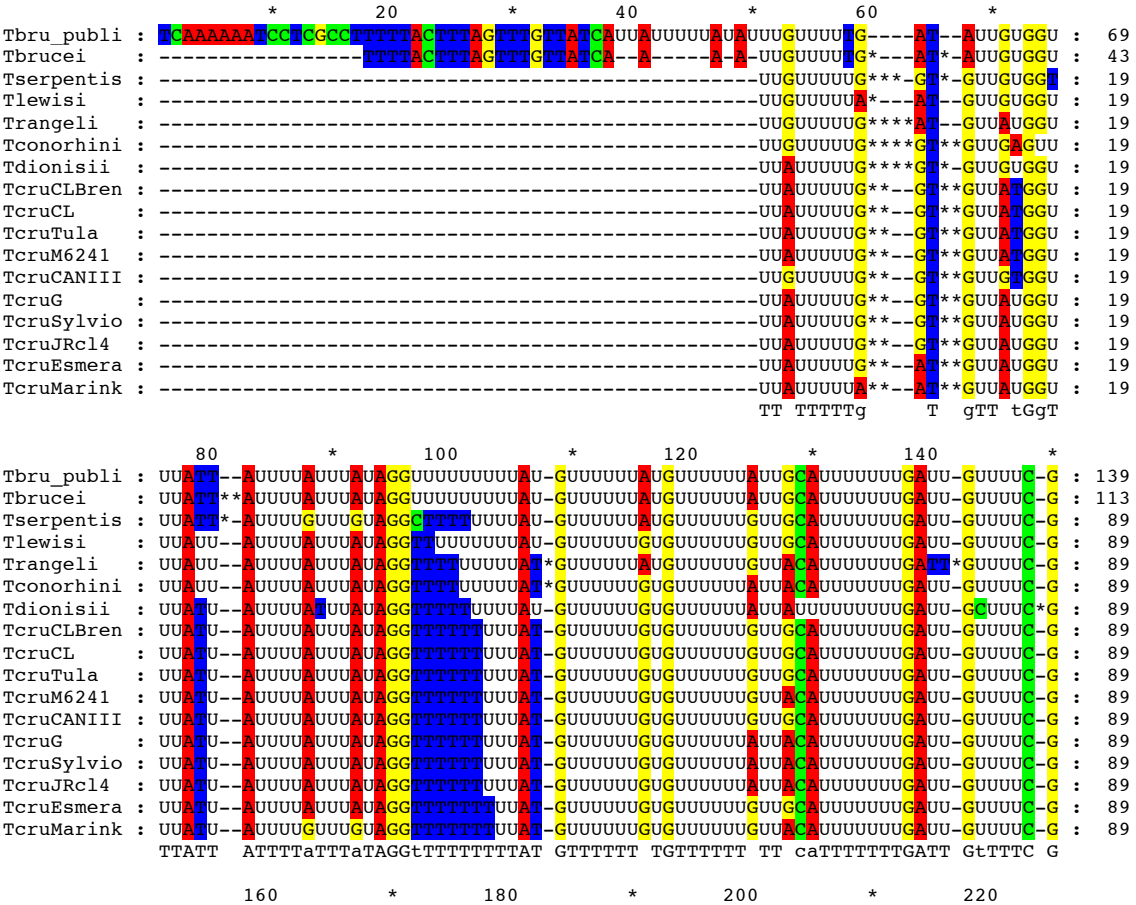
4.SI-S18b: GeneDoc G4 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. '?' in TcruEsmera refers to lack of codon information due to 5' deletion of the G4 gene in it. Reference has also been translated.

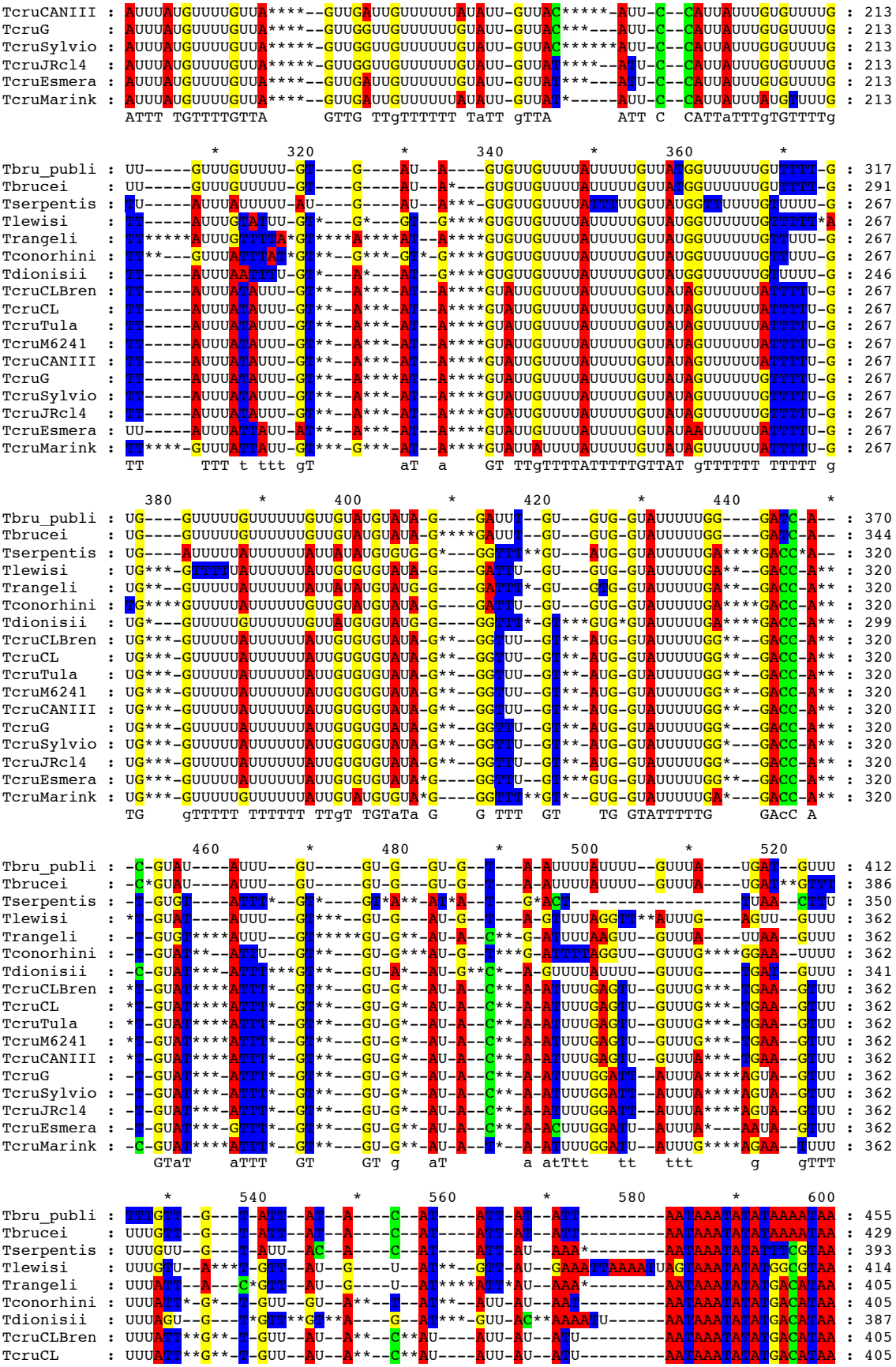
4.SI-S18c. G4 edited transcript percent identiy matrix.

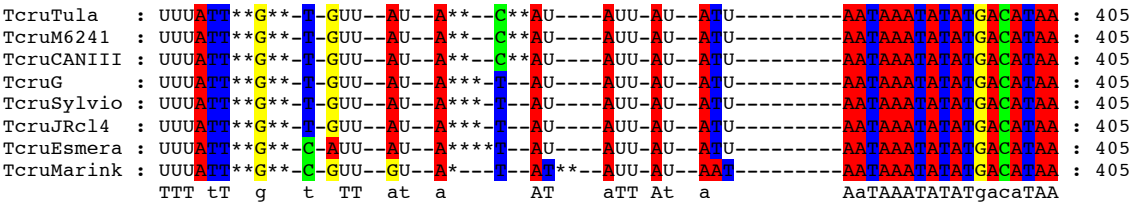
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**														
Tser	94	**100**													
Tlew	92	96	**100**												
Tran	93	94	92	**100**											
Tcon	93	95	94	95	**100**										
TcCLB	93	92	93	93	95	**100**									
TcCL	93	92	93	93	95	100	**100**								
TcTula	93	92	93	93	95	100	100	**100**							
TcM62	93	92	93	93	95	100	100	100	**100**						
TcCAN	94	93	93	94	95	99	99	99	99	**100**					
TcG	94	93	93	93	94	98	98	98	98	98	**100**				
TcSyl	94	93	93	93	94	98	98	98	98	98	100	**100**			
TcJR	94	93	92	93	94	98	98	98	98	98	100	100	**100**		
TcEsm	93	94	94	93	93	97	97	97	97	97	97	98	97	**100**	
TcMar	93	92	93	93	94	98	98	98	98	98	97	97	97	97	**100**

4.SI-S18d. G4 protein percent identiy matrix.

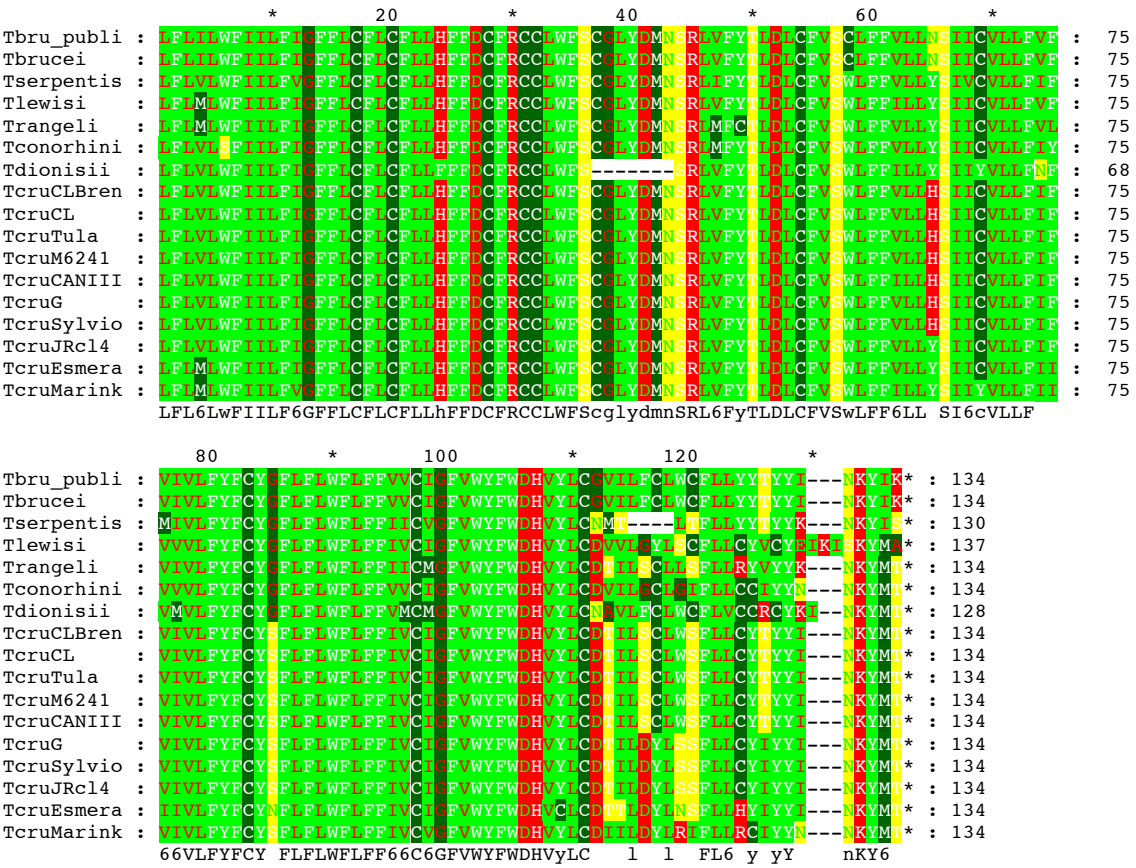
	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**														
Tser	95	**100**													
Tlew	93	96	**100**												
Tran	90	92	92	**100**											
Tcon	92	95	93	93	**100**										
TcCLB	92	92	94	92	95	**100**									
TcCL	92	92	94	92	95	100	**100**								
TcTula	92	92	94	92	95	100	100	**100**							
TcM62	90	90	92	91	95	99	99	99	**100**						
TcCAN	91	91	93	92	95	99	99	99	99	**100**					
TcG	91	92	92	89	93	97	97	97	97	97	**100**				
TcSyl	92	92	92	88	94	96	96	96	96	97	99	**100**			
TcJR	91	92	92	89	93	97	97	97	97	97	100	99	**100**		
TcEsm	92	93	95	90	95	97	97	97	98	98	97	98	97	**100**	
TcMar	90	92	92	92	96	99	99	99	99	99	97	96	97	97	**100**







4.SI-S19a: GeneDoc ND3 / G5 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Alignments start at the proposed non-canonical start codon (UUA or UUG, encoding Leucine). Reference is RNA sequence of *T. bru.*



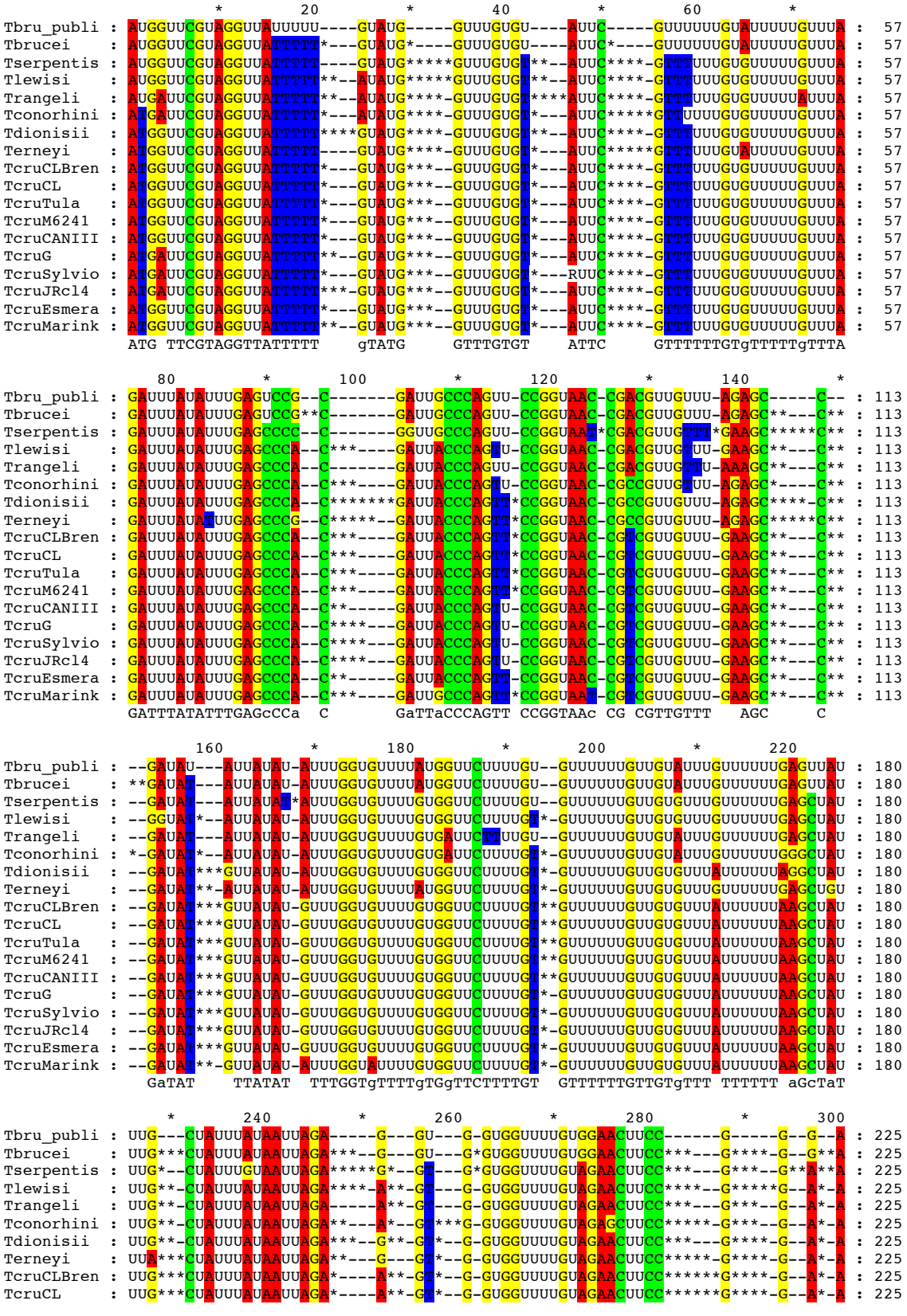
4.SI-S19b: GeneDoc ND3 / G5 protein alignment of the conserved motif: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Reference has also been translated.

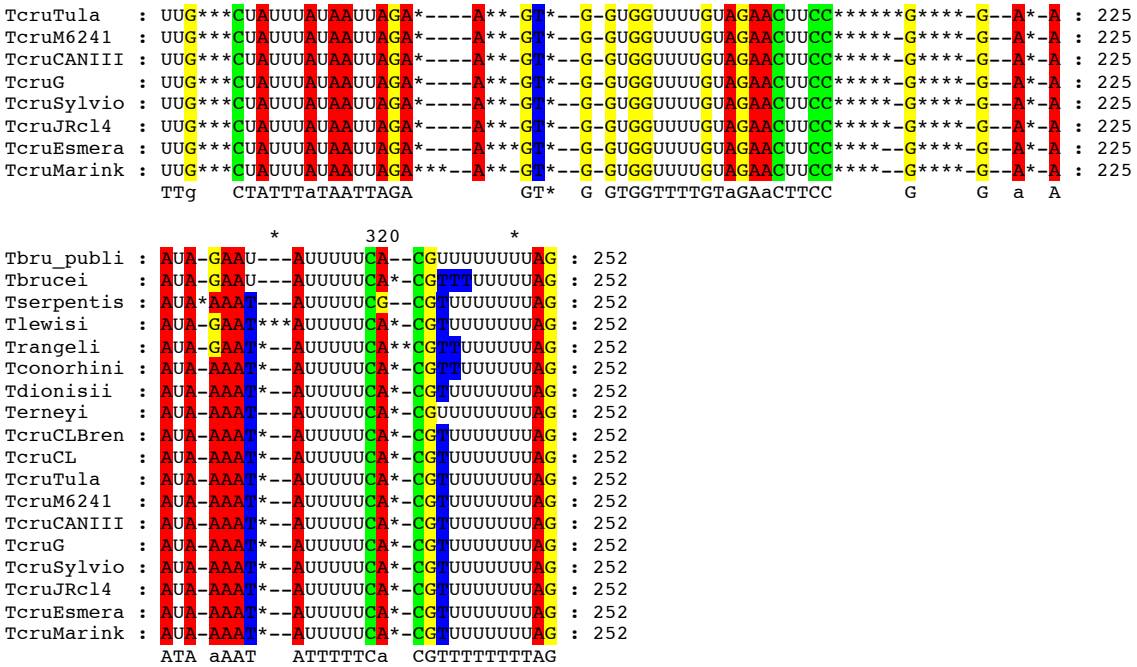
4.SI-S19c. ND3 edited transcript percent identiy matrix (for the conserved stretch)

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	75	**100**														
Tlew	79	78	**100**													
Tran	78	80	84	**100**												
Tcon	76	79	83	87	**100**											
Tdio	77	77	81	83	84	**100**										
TcCLB	78	80	83	83	84	84	**100**									
TcCL	78	80	83	83	84	84	100	**100**								
TcTula	78	80	83	83	84	84	100	100	**100**							
TcM62	78	80	82	84	85	85	100	100	100	**100**						
TcCAN	79	81	84	84	85	85	99	99	99	98	**100**					
TcG	78	80	84	84	86	85	96	96	96	96	96	**100**				
TcSyl	78	80	84	84	86	85	96	96	96	96	96	100	**100**			
TcJR	78	80	84	85	86	85	95	95	95	96	95	100	100	**100**		
TcEsm	76	79	82	83	83	82	93	93	93	93	93	94	94	95	**100**	
TcMar	76	79	82	83	85	84	89	89	89	89	89	90	90	91	91	**100**

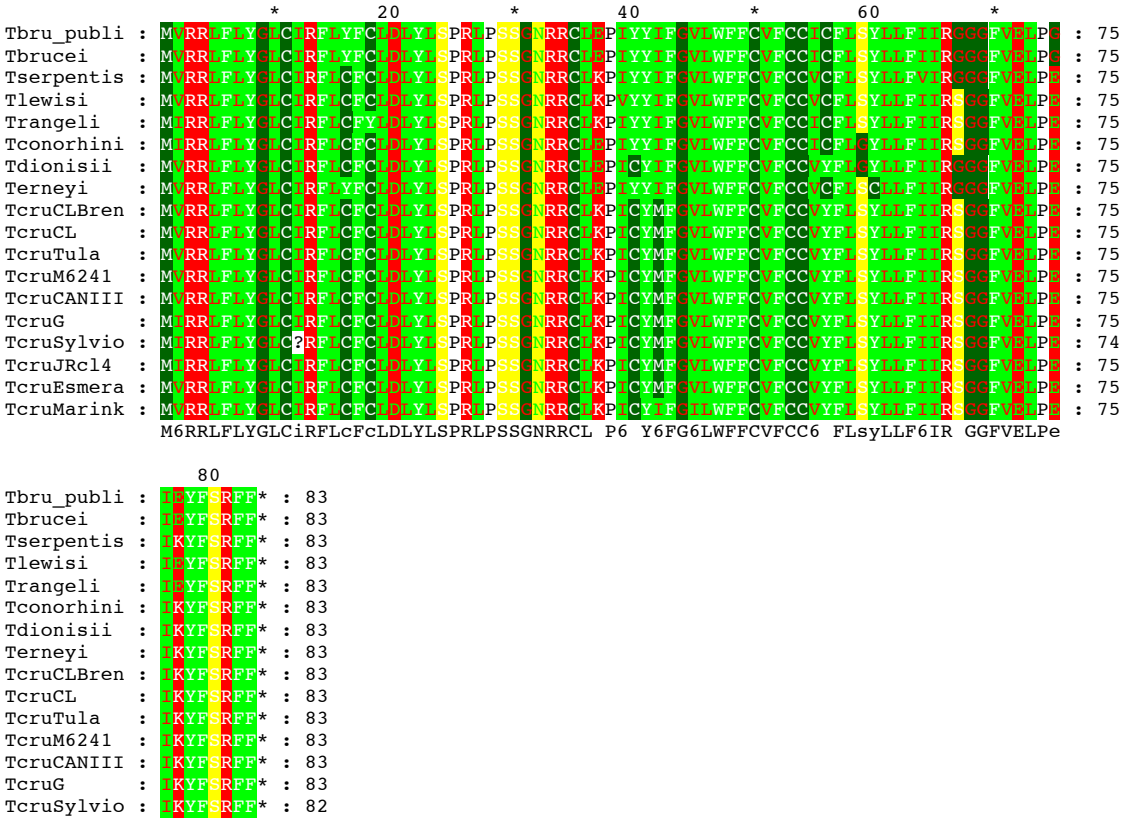
4.SI-S19d. ND3 protein percent identiy matrix (for the conserved motif)

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**															
Tser	86	**100**														
Tlew	87	83	**100**													
Tran	86	87	87	**100**												
Tcon	87	84	87	87	**100**											
Tdio	83	81	85	83	84	**100**										
TcCLB	90	87	87	90	90	84	**100**									
TcCL	90	87	87	90	90	84	100	**100**								
TcTula	90	87	87	90	90	84	100	100	**100**							
TcM62	90	87	87	90	90	84	100	100	100	**100**						
TcCAN	90	86	88	90	89	85	99	99	99	99	**100**					
TcG	88	86	89	89	90	83	97	97	97	97	96	**100**				
TcSyl	88	86	89	89	90	83	97	97	97	97	96	100	**100**			
TcJR	88	87	90	90	90	83	96	96	96	96	96	99	99	**100**		
TcEsm	85	85	87	88	87	80	91	91	91	91	90	93	93	94	**100**	
TcMar	83	85	86	87	88	83	88	88	88	88	89	90	90	91	90	**100**





4.SI-S20a: GeneDoc RPS12 nucleotide alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. U insertions are indicated as "U" against a white background in the alignment. Deleted Ts have been denoted by "*" in the alignment and a gap has been incorporated for the other organisms at that position. Reference is RNA of *Tbru* (Tbru_publi)



```
TcruJRcl4  : IKYFSRFF* : 83
TcruEsmera : IKYFSRFF* : 83
TcruMarink : IKYFSRFF* : 83
             IkYFSRFF
```

4.SI-S20b: GeneDoc RPS12 protein alignment: Abbreviations for the names of the organisms, gaps and consensus for the figure are as mentioned in S1. '*' indicates stop codon in the alignment. Reference has also been translated. '?' in TcruSylvio indicates lack of amino acid information due to 'R' in the corresponding nucleotide sequence above (S20a).

4.SI-S20c. RPS12 edited transcript percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	94	**100**															
Tlew	94	96	**100**														
Tran	94	94	98	**100**													
Tcon	94	94	96	98	**100**												
Tdio	94	95	96	95	96	**100**											
Tern	96	95	95	94	95	96	**100**										
TcCLB	93	95	97	95	95	98	95	**100**									
TcCL	93	95	97	95	95	98	95	100	**100**								
TcTula	93	95	97	95	95	98	95	100	100	**100**							
TcM62	93	95	97	95	95	98	95	100	100	100	**100**						
TcCAN	93	95	97	95	95	98	95	100	100	100	100	**100**					
TcG	92	95	96	96	95	97	94	100	100	100	100	100	**100**				
TcSyl	92	95	96	96	95	97	94	100	100	100	100	100	100	**100**			
TcJR	92	95	96	96	95	97	94	100	100	100	100	100	100	100	**100**		
TcEsm	93	95	97	95	95	98	95	100	100	100	100	100	100	100	100	**100**	
TcMar	93	96	96	94	94	97	94	98	98	98	98	98	98	98	98	98	**100**

4.SI-S20d. RPS12 protein percent identity matrix.

	Tbru	*Tser*	*Tlew*	*Tran*	*Tcon*	*Tdio*	*Tern*	*TcCLB*	*TcCL*	*TcTula*	*TcM62*	*TcCAN*	*TcG*	*TcSyl*	*TcJR*	*TcEsm*	*TcMar*
Tbru	**100**																
Tser	93	**100**															
Tlew	93	95	**100**														
Tran	93	93	95	**100**													
Tcon	93	93	93	95	**100**												
Tdio	92	94	92	89	94	**100**											
Tern	95	95	93	90	93	94	**100**										
TcCLB	89	94	94	92	92	95	92	**100**									
TcCL	89	94	94	92	92	95	92	100	**100**								
TcTula	89	94	94	92	92	95	92	100	100	**100**							
TcM62	89	94	94	92	92	95	92	100	100	100	**100**						
TcCAN	89	94	94	92	92	95	92	100	100	100	100	**100**					
TcG	88	93	93	93	93	94	90	99	99	99	99	99	**100**				
TcSyl	88	93	93	93	93	94	90	99	99	99	99	99	100	**100**			
TcJR	88	93	93	93	93	94	90	99	99	99	99	99	100	100	**100**		
TcEsm	89	94	94	92	92	95	92	100	100	100	100	100	99	99	99	**100**	
TcMar	89	94	94	92	92	95	92	98	98	98	98	98	96	96	96	98	**100**

5. REFERENCES

Abrahamsen, M.S., Templeton, T.J., Enomoto, S., Abrahante, J.E., Zhu, G., Lancto, C.A., Deng, M., Liu, C., Widmer, G., Tzipori, S., et al. (2004). Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science *304*, 441–445.

Acosta, I. da C.L., da Costa, A.P., Nunes, P.H., Gondim, M.F.N., Gatti, A., Rossi, J.L., Gennari, S.M., and Marcili, A. (2013). Morphological and molecular characterization and phylogenetic relationships of a new species of trypanosome in Tapirus terrestris (lowland tapir), Trypanosoma terrestris sp. nov., from Atlantic Rainforest of southeastern Brazi. Parasit. Vectors *6*, 349.

Adhya, S. (2008). Leishmania mitochondrial tRNA importers. Int. J. Biochem. Cell Biol. *40*, 2681–2685.

Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. J. Eukaryot. Microbiol. *59*, 429–493.

Alfonzo, J.D., and Söll, D. (2009). Mitochondrial tRNA import--the challenge to understand has just begun. Biol. Chem. *390*, 717–722.

Alfonzo, J.D., Thiemann, O., and Simpson, L. (1997). The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. Nucleic Acids Res. *25*, 3751–3759.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Alves, J.M.P., Voegtly, L., Matveyev, A.V., Lara, A.M., da Silva, F.M., Serrano, M.G., Buck, G.A., Teixeira, M.M.G., and Camargo, E.P. (2011). Identification and phylogenetic analysis of heme synthesis genes in trypanosomatids and their bacterial endosymbionts. PloS One *6*, e23518.

Alves, J.M.P., Serrano, M.G., Maia da Silva, F., Voegtly, L.J., Matveyev, A.V., Teixeira, M.M.G., Camargo, E.P., and Buck, G.A. (2013a). Genome evolution and phylogenomic analysis of Candidatus Kinetoplastibacterium, the betaproteobacterial endosymbionts of Strigomonas and Angomonas. Genome Biol. Evol. *5*, 338–350.

Alves, J.M.P., Klein, C.C., da Silva, F.M., Costa-Martins, A.G., Serrano, M.G., Buck, G.A., Vasconcelos, A.T.R., Sagot, M.-F., Teixeira, M.M.G., Motta, M.C.M., et al. (2013b). Endosymbiosis in trypanosomatids: the genomic cooperation between bacterium and host in the synthesis of essential amino acids is heavily influenced by multiple horizontal gene transfers. BMC Evol. Biol. *13*, 190.

Andrade, L.O., and Andrews, N.W. (2005). The Trypanosoma cruzi-host-cell interplay: location, invasion, retention. Nat. Rev. Microbiol. *3*, 819–823.

Aphasizhev, R., and Aphasizheva, I. (2008). Terminal RNA uridylyltransferases of trypanosomes. Biochim. Biophys. Acta *1779*, 270–280.

Aphasizhev, R., Aphasizheva, I., Nelson, R.E., Gao, G., Simpson, A.M., Kang, X., Falick, A.M., Sbicego, S., and Simpson, L. (2003). Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria. EMBO J. *22*, 913–924.

Aphasizheva, I., Maslov, D.A., and Aphasizhev, R. (2013). Kinetoplast DNA-encoded ribosomal protein S12: a possible functional link between mitochondrial RNA editing and translation in Trypanosoma brucei. RNA Biol. *10*, 1679–1688.

Aravin, A.A., Yurchenko VYu, null, Merzlyak, E.M., and Kolesnikov, A.A. (1998). The mitochondrial ND8 gene from Crithidia oncopelti is not pan-edited. FEBS Lett. *431*, 457–460.

Arnoult F, Akarid K, Grodet A, Petit P, Estaquier J, Ameisen J. On the evolution of programmed cell death: apoptosis of the unicellular eukaryote Leishmania major involves cysteine proteinase activation and mitochondrion permeabilization. Cell Death & Differentiation [serial on the Internet]. (2002, Jan); 9(1): 65.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res. *40*, W597–W603.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B.P., Carrington, M., Depledge, D.P., Fischer, S., Gajria, B., Gao, X., et al. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. *38*, D457–D462.

Bacchi, C.J., Nathan, H.C., Hutner, S.H., McCann, P.P., and Sjoerdsma, A. (1980). Polyamine metabolism: a potential therapeutic target in trypanosomes. Science *210*, 332–334.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. Science *290*, 972–977.

Benne, R. (1990). RNA editing in trypanosomes: is there a message? Trends Genet. TIG *6*, 177–181.

Benne, R., De Vries, B.F., Van den Burg, J., and Klaver, B. (1983). The nucleotide sequence of a segment of Trypanosoma brucei mitochondrial maxi-circle DNA that contains the gene for apocytochrome b and some unusual unassigned reading frames. Nucleic Acids Res. *11*, 6925–6941.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., and Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell *46*, 819–826.

Bernards, A., Michels, P.A., Lincke, C.R., and Borst, P. (1983). Growth of chromosome ends in multiplying trypanosomes. Nature *303*, 592–597.

Beverley, S.M. (1996). Hijacking the cell: parasites in the driver's seat. Cell *87*, 787–789.

Bhat, G.J., Koslowsky, D.J., Feagin, J.E., Smiley, B.L., and Stuart, K. (1990). An extensively edited mitochondrial transcript in kinetoplastids encodes a protein homologous to ATPase subunit 6. Cell *61*, 885–894.

Bhat, G.J., Myler, P.J., and Stuart, K. (1991). The two ATPase 6 mRNAs of Leishmania tarentolae differ at their 3' ends. Mol. Biochem. Parasitol. *48*, 139–149.

Bhattacharyya, S.N., and Adhya, S. (2004). The complexity of mitochondrial tRNA import. RNA Biol. *1*, 84–88.

Bienen, E.J., Webster, P., and Fish, W.R. (1991). Trypanosoma (Nannomonas) congolense: changes in respiratory metabolism during the life cycle. Exp. Parasitol. *73*, 403–412.

Blom, D., de Haan, A., van den Berg, M., Sloof, P., Jirku, M., Lukes, J., and Benne, R. (1998). RNA editing in the free-living bodonid Bodo saltans. Nucleic Acids Res. *26*, 1205–1213.

Blum, B., and Simpson, L. (1990). Guide RNAs in kinetoplastid mitochondria have a nonencoded 3' oligo(U) tail involved in recognition of the preedited region. Cell *62*, 391–397.

Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. Cell *60*, 189–198.

Borst, P. (1991). Why kinetoplast DNA networks? Trends Genet. TIG *7*, 139–141.

Bridgen, P.J., Cross, G.A., and Bridgen, J. (1976). N-terminal amino acid sequences of variant-specific surface antigens from Trypanosoma brucei. Nature *263*, 613–614.

Bringaud, F., Rivière, L., and Coustou, V. (2006). Energy metabolism of trypanosomatids: adaptation to available carbon sources. Mol. Biochem. Parasitol. *149*, 1–9.

Brisse, S., Barnabé, C., and Tibayrenc, M. (2000). Identification of six Trypanosoma cruzi phylogenetic lineages by random amplified polymorphic DNA and multilocus enzyme electrophoresis. Int. J. Parasitol. *30*, 35–44.

Brisse, S., Verhoef, J., and Tibayrenc, M. (2001). Characterisation of large and small subunit rRNA and mini-exon genes further supports the distinction of six Trypanosoma cruzi lineages. Int. J. Parasitol. *31*, 1218–1226.

Brisse, S., Henriksson, J., Barnabé, C., Douzery, E.J.P., Berkvens, D., Serrano, M., De Carvalho, M.R.C., Buck, G.A., Dujardin, J.-C., and Tibayrenc, M. (2003). Evidence for genetic exchange and hybridization in Trypanosoma cruzi based on nucleotide sequences and molecular karyotype. Infect. Genet. Evol. *2*, 173–183.

Burleigh, B.A. (2005). Host cell signaling and Trypanosoma cruzi invasion: do all roads lead to lysosomes? Sci. STKE Signal Transduct. Knowl. Environ. *2005*, pe36.

Caballero, Z.C., Costa-Martins, A.G., Ferreira, R.C., P Alves, J.M., Serrano, M.G., Camargo, E.P., Buck, G.A., Minoprio, P., and G Teixeira, M.M. (2015). Phylogenetic and syntenic data support a single horizontal transference to a Trypanosoma ancestor of a prokaryotic proline racemase implicated in parasite evasion from host defences. Parasit. Vectors *8*, 222.

Camargo, E.P. (1964). GROWTH AND DIFFERENTIATION IN TRYPANOSOMA CRUZI. I. ORIGIN OF METACYCLIC TRYPANOSOMES IN LIQUID MEDIA. Rev. Inst. Med. Trop. São Paulo *6*, 93–100.

Carranza, J.C., Kowaltowski, A.J., Mendonça, M.A.G., de Oliveira, T.C., Gadelha, F.R., and Zingales, B. (2009). Mitochondrial bioenergetics and redox state are unaltered in Trypanosoma cruzi isolates with compromised mitochondrial complex I subunit genes. J. Bioenerg. Biomembr. *41*, 299–308.

Carrillo, R., Thiemann, O.H., Alfonzo, J.D., and Simpson, L. (2001). Uridine insertion/deletion RNA editing in Leishmania tarentolae mitochondria shows cell cycle dependence. Mol. Biochem. Parasitol. *113*, 175–181.

Cavalier-Smith, T. (1997). Cell and genome coevolution: facultative anaerobiosis, glycosomes and kinetoplastan RNA editing. Trends Genet. *13*, 6–9.

Cavazzana Jr., M., Marcili, A., Lima, L., da Silva, F.M., Junqueira, Â.C.V., Veludo, H.H., Viola, L.B., Campaner, M., Nunes, V.L.B., Paiva, F., et al. (2010). Phylogeographical, ecological and biological patterns shown by nuclear (ssrRNA and gGAPDH) and mitochondrial (Cyt b) genes of trypanosomes of the subgenus Schizotrypanum parasitic in Brazilian bats. Int. J. Parasitol. *40*, 345–355.

Cazzulo, J.J. (1992). Aerobic fermentation of glucose by trypanosomatids. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *6*, 3153–3161.

Cazzulo, J.J. (1994). Intermediate metabolism in Trypanosoma cruzi. J. Bioenerg. Biomembr. *26*, 157–165.

Cazzulo, J.J., Franke de Cazzulo, B.M., Engel, J.C., and Cannata, J.J. (1985). End products and enzyme levels of aerobic glucose fermentation in trypanosomatids. Mol. Biochem. Parasitol. *16*, 329–343.

Chance, B., and Williams, G.R. (1956). The respiratory chain and oxidative phosphorylation. Adv. Enzymol. Relat. Subj. Biochem. *17*, 65–134.

Chaudhuri, M., Ajayi, W., and Hill, G.C. (1998). Biochemical and molecular properties of the Trypanosoma brucei alternative oxidase. Mol. Biochem. Parasitol. *95*, 53–68.

Chaudhuri, M., Ott, R.D., and Hill, G.C. (2006). Trypanosome alternative oxidase: from molecule to function. Trends Parasitol. *22*, 484–491.

Chaumont, F., Schanck, A.N., Blum, J.J., and Opperdoes, F.R. (1994). Aerobic and anaerobic glucose metabolism of Phytomonas sp. isolated from Euphorbia characias. Mol. Biochem. Parasitol. *67*, 321–331.

Chen, J., Rauch, C.A., White, J.H., Englund, P.T., and Cozzarelli, N.R. (1995). The topology of the kinetoplast DNA network. Cell *80*, 61–69.

Chen, M., Bennedsen, M., Zhai, L., and Kharazmi, A. (2001). Purification and enzymatic activity of an NADH-fumarate reductase and other mitochondrial activities of Leishmania parasites. APMIS Acta Pathol. Microbiol. Immunol. Scand. *109*, 801–808.

Chen, Q., Vazquez, E.J., Moghaddas, S., Hoppel, C.L., and Lesnefsky, E.J. (2003). Production of reactive oxygen species by mitochondria: central role of complex III. J. Biol. Chem. *278*, 36027–36031.

Chicharro, C., and Alvar, J. (2003). Lower trypanosomatids in HIV/AIDS patients. Ann. Trop. Med. Parasitol. *97 Suppl 1*, 75–78.

Clarkson, A.B., Bienen, E.J., Pollakis, G., and Grady, R.W. (1989). Respiration of bloodstream forms of the parasite Trypanosoma brucei brucei is dependent on a plant-like alternative oxidase. J. Biol. Chem. *264*, 17770–17776.

Combs, T.P., Nagajyothi, null, Mukherjee, S., de Almeida, C.J.G., Jelicks, L.A., Schubert, W., Lin, Y., Jayabalan, D.S., Zhao, D., Braunstein, V.L., et al. (2005). The adipocyte as an important target cell for Trypanosoma cruzi infection. J. Biol. Chem. *280*, 24085–24094.

Corell, R.A., Feagin, J.E., Riley, G.R., Strickland, T., Guderian, J.A., Myler, P.J., and Stuart, K. (1993). Trypanosoma brucei minicircles encode multiple guide RNAs which can direct editing of extensively overlapping sequences. Nucleic Acids Res. *21*, 4313–4320.

Corell, R.A., Myler, P., and Stuart, K. (1994). Trypanosoma brucei mitochondrial CR4 gene encodes an extensively edited mRNA with completely edited sequence only in bloodstream forms. Mol. Biochem. Parasitol. *64*, 65–74.

Covello, P.S., and Gray, M.W. (1993). On the evolution of RNA editing. Trends Genet. TIG *9*, 265–268.

Cruz, V.F. de la, Neckelmann, N., and Simpson, L. (1984). Sequences of six genes and several open reading frames in the kinetoplast maxicircle DNA of Leishmania tarentolae. J. Biol. Chem. *259*, 15136–15147.

de la Cruz, V.F., Lake, J.A., Simpson, A.M., and Simpson, L. (1985a). A minimal ribosomal RNA: sequence and secondary structure of the 9S kinetoplast ribosomal RNA from Leishmania tarentolae. Proc. Natl. Acad. Sci. U. S. A. *82*, 1401–1405.

de la Cruz, V.F., Simpson, A.M., Lake, J.A., and Simpson, L. (1985b). Primary sequence and partial secondary structure of the 12S kinetoplast (mitochondrial) ribosomal RNA from

Leishmania tarentolae: conservation of peptidyl-transferase structural elements. Nucleic Acids Res. *13*, 2337–2356.

Das, B.B., Sengupta, T., Ganguly, A., and Majumder, H.K. (2006). Topoisomerases of kinetoplastid parasites: why so fascinating? Mol. Microbiol. *62*, 917–927.

Da Silva, F.M., Noyes, H., Campaner, M., Junqueira, A.C.V., Coura, J.R., Añez, N., Shaw, J.J., Stevens, J.R., and Teixeira, M.M.G. (2004). Phylogeny, taxonomy and grouping of Trypanosoma rangeli isolates from man, triatomines and sylvatic mammals from widespread geographical origin based on SSU and ITS ribosomal sequences. Parasitology *129*, 549–561.

Dedet, J.P., and Pratlong, F. (2000). Leishmania, Trypanosoma and monoxenous trypanosomatids as emerging opportunistic agents. J. Eukaryot. Microbiol. *47*, 37–39.

Deschamps, P., Lara, E., Marande, W., López-García, P., Ekelund, F., and Moreira, D. (2011). Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. Mol. Biol. Evol. *28*, 53–58.

Diosque, P., Tomasini, N., Lauthier, J.J., Messenger, L.A., Monje Rumi, M.M., Ragone, P.G., Alberti-D'Amato, A.M., Pérez Brandán, C., Barnabé, C., Tibayrenc, M., et al. (2014). Optimized Multilocus Sequence Typing (MLST) Scheme for Trypanosoma cruzi. PLoS Negl. Trop. Dis. *8*.

Durrant, J.D., Hall, L., Swift, R.V., Landon, M., Schnaufer, A., and Amaro, R.E. (2010). Novel Naphthalene-Based Inhibitors of Trypanosoma brucei RNA Editing Ligase 1. PLoS Negl. Trop. Dis. *4*, e803.

Edgar, R.C. (2004a). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Edgar, R.C. (2004b). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Engman, D.M., Kirchhoff, L.V., and Donelson, J.E. (1989). Molecular cloning of mtp70, a mitochondrial member of the hsp70 family. Mol. Cell. Biol. *9*, 5163–5168.

Eperon, I.C., Janssen, J.W., Hoeijmakers, J.H., and Borst, P. (1983). The major transcripts of the kinetoplast DNA of Trypanosoma brucei are very small ribosomal RNAs. Nucleic Acids Res. *11*, 105–125.

Estévez, A.M., and Simpson, L. (1999). Uridine insertion/deletion RNA editing in trypanosome mitochondria--a review. Gene *240*, 247–260.

Fang, J., and Beattie, D.S. (2002). Novel FMN-containing rotenone-insensitive NADH dehydrogenase from Trypanosoma brucei mitochondria: isolation and characterization. Biochemistry (Mosc.) *41*, 3065–3072.

Fang, J., and Beattie, D.S. (2003). Identification of a gene encoding a 54 kDa alternative NADH dehydrogenase in Trypanosoma brucei. Mol. Biochem. Parasitol. *127*, 73–77.

Fang, J., Wang, Y., and Beattie, D.S. (2001). Isolation and characterization of complex I, rotenone-sensitive NADH: ubiquinone oxidoreductase, from the procyclic forms of Trypanosoma brucei. Eur. J. Biochem. FEBS *268*, 3075–3082.

Feagin, J.E., Jasmer, D.P., and Stuart, K. (1985). Apocytochrome b and other mitochondrial DNA sequences are differentially expressed during the life cycle of Trypanosoma brucei. Nucleic Acids Res. *13*, 4577–4596.

Feagin, J.E., Jasmer, D.P., and Stuart, K. (1987). Developmentally regulated addition of nucleotides within apocytochrome b transcripts in Trypanosoma brucei. Cell *49*, 337–345.

Feagin, J.E., Abraham, J.M., and Stuart, K. (1988a). Extensive editing of the cytochrome c oxidase III transcript in Trypanosoma brucei. Cell *53*, 413–422.

Feagin, J.E., Shaw, J.M., Simpson, L., and Stuart, K. (1988b). Creation of AUG initiation codons by addition of uridines within cytochrome b transcripts of kinetoplastids. Proc. Natl. Acad. Sci. U. S. A. *85*, 539–543.

Ferguson, M.A., and Cross, G.A. (1984). Myristylation of the membrane form of a Trypanosoma brucei variant surface glycoprotein. J. Biol. Chem. *259*, 3011–3015.

Ferguson, M.A., Haldar, K., and Cross, G.A. (1985). Trypanosoma brucei variant surface glycoprotein has a sn-1,2-dimyristyl glycerol membrane anchor at its COOH terminus. J. Biol. Chem. *260*, 4963–4968.

Fernandes, A.P., Nelson, K., and Beverley, S.M. (1993). Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. Proc. Natl. Acad. Sci. U. S. A. *90*, 11608–11612.

Fidalgo, L.M., and Gille, L. (2011). Mitochondria and trypanosomatids: targets and drugs. Pharm. Res. *28*, 2758–2770.

Flegontov, P., Votýpka, J., Skalický, T., Logacheva, M.D., Penin, A.A., Tanifuji, G., Onodera, N.T., Kondrashov, A.S., Volf, P., Archibald, J.M., et al. (2013). Paratrypanosoma is a novel early-branching trypanosomatid. Curr. Biol. CB *23*, 1787–1793.

Flegontov, P.N., Guo, Q., Ren, L., Strelkova, M.V., and Kolesnikov, A.A. (2006). Conserved repeats in the kinetoplast maxicircle divergent region of Leishmania sp. and Leptomonas seymouri. Mol. Genet. Genomics MGG *276*, 322–333.

Flores-López, C.A., and Machado, C.A. (2011). Analyses of 32 loci clarify phylogenetic relationships among Trypanosoma cruzi lineages and support a single hybridization prior to human contact. PLoS Negl. Trop. Dis. *5*, e1272.

Flynn, I.W., and Bowman, I.B. (1973). The metabolism of carbohydrate by pleomorphic African trypanosomes. Comp. Biochem. Physiol. B *45*, 25–42.

Franzén, O., Talavera-López, C., Ochaya, S., Butler, C.E., Messenger, L.A., Lewis, M.D., Llewellyn, M.S., Marinkelle, C.J., Tyler, K.M., Miles, M.A., et al. (2012). Comparative genomic analysis of human infective Trypanosoma cruzi lineages with the bat-restricted subspecies T. cruzi marinkellei. BMC Genomics *13*, 531.

de Freitas, J.M., Augusto-Pinto, L., Pimenta, J.R., Bastos-Rodrigues, L., Gonçalves, V.F., Teixeira, S.M.R., Chiari, E., Junqueira, Â.C.V., Fernandes, O., Macedo, A.M., et al. (2006). Ancestral Genomes, Sex, and the Population Structure of Trypanosoma cruzi. PLoS Pathog. *2*.

Gabriel, A., and Boeke, J.D. (1991). Reverse transcriptase encoded by a retrotransposon from the trypanosomatid Crithidia fasciculata. Proc. Natl. Acad. Sci. U. S. A. *88*, 9794–9798.

Gao, G., Kapushoc, S.T., Simpson, A.M., Thiemann, O.H., and Simpson, L. (2001). Guide RNAs of the recently isolated LEM125 strain of Leishmania tarentolae: an unexpected complexity. RNA N. Y. N *7*, 1335–1347.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D., and Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. *31*, 3784–3788.

Gaunt, M.W., Yeo, M., Frame, I.A., Stothard, J.R., Carrasco, H.J., Taylor, M.C., Mena, S.S., Veazey, P., Miles, G.A.J., Acosta, N., et al. (2003). Mechanism of genetic exchange in American trypanosomes. Nature *421*, 936–939.

Gerasimov, E.S., Kostygov, A.Y., Yan, S., and Kolesnikov, A.A. (2012a). From cryptogene to gene? ND8 editing domain reduction in insect trypanosomatids. Eur. J. Protistol. *48*, 185–193.

Gerasimov, E.S., Efimova, N.S., and Kolesnikov, A.A. (2012b). Three patterns of trypanosomatid cryptogene structural organization. Mol. Biol. *46*, 548–556.

Gibson, W., and Garside, L. (1990). Kinetoplast DNA minicircles are inherited from both parents in genetic hybrids of Trypanosoma brucei. Mol. Biochem. Parasitol. *42*, 45–53.

Gibson, W., Crow, M., and Kearns, J. (1997). Kinetoplast DNA minicircles are inherited from both parents in genetic crosses of Trypanosoma brucei. Parasitol. Res. *83*, 483–488.

Gille, L., and Nohl, H. (2001). The ubiquinol/bc1 redox couple regulates mitochondrial oxygen radical formation. Arch. Biochem. Biophys. *388*, 34–38.

Gilroy, F.V., Edwards, M.R., Norton, R.S., and O'Sullivan, W.J. (1988). Metabolic studies of the protozoan parasite, Crithidia luciliae, using proton nuclear magnetic resonance spectroscopy. Mol. Biochem. Parasitol. *31*, 107–115.

González, A., Rosales, J.L., Ley, V., and Díaz, C. (1990). Cloning and characterization of a gene coding for a protein (KAP) associated with the kinetoplast of epimastigotes and amastigotes of Trypanosoma cruzi. Mol. Biochem. Parasitol. *40*, 233–243.

González, C.I., Thomas, M.C., Martín, F., Alcami, J., Alonso, C., and López, M.C. (1997). Reverse transcriptase-like activity in Trypanosoma cruzi. Acta Trop. *63*, 117–126.

Gorbat, A., Maslov, D.A., Peters, L.S., Gaviernik, P., Viustenkhagen, T., and Kolesnikov, A.A. (1990). [Analysis of the sequence of repeats in divergent regions of maxi-circular DNA from kinetoplasts of Crithidia oncopelti]. Mol. Biol. (Mosk.) *24*, 1539–1548.

Gottlieb, R.A. (2003). Mitochondrial signaling in apoptosis: mitochondrial daggers to the breaking heart. Basic Res. Cardiol. *98*, 242–249.

Grevelink, S.A., and Lerner, E.A. (1996). Leishmaniasis. J. Am. Acad. Dermatol. *34*, 257–272.

Guler, J.L., Kriegova, E., Smith, T.K., Lukes, J., and Englund, P.T. (2008). Mitochondrial fatty acid synthesis is required for normal mitochondrial morphology and function in Trypanosoma brucei. Mol. Microbiol. *67*, 1125–1142.

Hajduk, S.L., Siqueira, A.M., and Vickerman, K. (1986). Kinetoplast DNA of Bodo caudatus: a noncatenated structure. Mol. Cell. Biol. *6*, 4372–4378.

Hamilton, P.B., Stevens, J.R., Gaunt, M.W., Gidley, J., and Gibson, W.C. (2004). Trypanosomes are monophyletic: evidence from genes for glyceraldehyde phosphate dehydrogenase and small subunit ribosomal RNA. Int. J. Parasitol. *34*, 1393–1404.

Hamilton, P.B., Cruickshank, C., Stevens, J.R., Teixeira, M.M.G., and Mathews, F. (2012). Parasites reveal movement of bats between the New and Old Worlds. Mol. Phylogenet. Evol. *63*, 521–526.

Hancock, K., and Hajduk, S.L. (1990). The mitochondrial tRNAs of Trypanosoma brucei are nuclear encoded. J. Biol. Chem. *265*, 19208–19215.

Hannaert, V., Bringaud, F., Opperdoes, F.R., and Michels, P.A. (2003). Evolution of energy metabolism and its compartmentation in Kinetoplastida. Kinetoplastid Biol. Dis. *2*, 11.

Hart, D.T., and Coombs, G.H. (1982). Leishmania mexicana: energy metabolism of amastigotes and promastigotes. Exp. Parasitol. *54*, 397–409.

van Hellemond, J.J., Opperdoes, F.R., and Tielens, A.G.M. (2005). The extraordinary mitochondrion and unusual citric acid cycle in Trypanosoma brucei. Biochem. Soc. Trans. *33*, 967–971.

van Hellemond, J.J., Hoek, A., Wichgers Schreur, P., Chupin, V., Ozdirekcan, S., Geysen, D., van Grinsven, K.W.A., Koets, A.P., Van den Bossche, P., Geerts, S., et al. (2007). Energy metabolism of bloodstream form Trypanosoma theileri. Eukaryot. Cell *6*, 1693–1696.

Hensgens, L.A., Brakenhoff, J., De Vries, B.F., Sloof, P., Tromp, M.C., Van Boom, J.H., and Benne, R. (1984). The sequence of the gene for cytochrome c oxidase subunit I, a frameshift containing gene for cytochrome c oxidase subunit II and seven unassigned reading frames in Trypanosoma brucei mitochondrial maxi-circle DNA. Nucleic Acids Res. *12*, 7327–7344.

von der Heyden, S., Chao, E.E., Vickerman, K., and Cavalier-Smith, T. (2004). Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. J. Eukaryot. Microbiol. *51*, 402–416.

Hill, G.C., and Cross, G.A. (1973). Cyanide-resistant respiration and a branched cytochrome system in Kinetoplastidae. Biochim. Biophys. Acta *305*, 590–596.

Holder, A.A., and Cross, G.A. (1981). Glycopeptides from variant surface glycoproteins of Trypanosoma Brucei. C-terminal location of antigenically cross-reacting carbohydrate moieties. Mol. Biochem. Parasitol. *2*, 135–150.

Horváth, A., Berry, E.A., and Maslov, D.A. (2000). Translation of the edited mRNA for cytochrome b in trypanosome mitochondria. Science *287*, 1639–1640.

Hughes, A.L., and Piontkivska, H. (2003a). Phylogeny of Trypanosomatidae and Bodonidae (Kinetoplastida) based on 18S rRNA: evidence for paraphyly of Trypanosoma and six other genera. Mol. Biol. Evol. *20*, 644–652.

Hughes, A.L., and Piontkivska, H. (2003b). Molecular phylogenetics of Trypanosomatidae: contrasting results from 18S rRNA and protein phylogenies. Kinetoplastid Biol. Dis. *2*, 15.

Jan Arts, G., and Benne, R. (1996). Mechanism and evolution of RNA editing in kinetoplastida. Biochim. Biophys. Acta BBA - Gene Struct. Expr. *1307*, 39–54.

Janssen, R.J.R.J., Nijtmans, L.G., van den Heuvel, L.P., and Smeitink, J.A.M. (2006). Mitochondrial complex I: structure, function and pathology. J. Inherit. Metab. Dis. *29*, 499–515.

Jiménez, M.I., López-Vélez, R., Molina, R., Cañavate, C., and Alvar, J. (1996). HIV co-infection with a currently non-pathogenic flagellate. Lancet Lond. Engl. *347*, 264–265.

Johnson, P.J., Kooter, J.M., and Borst, P. (1987). Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene. Cell *51*, 273–281.

Jukes, T.H., and Osawa, S. (1993). Evolutionary changes in the genetic code. Comp. Biochem. Physiol. B *106*, 489–494.

Kable, M.L., Seiwert, S.D., Heidmann, S., and Stuart, K. (1996). RNA editing: a mechanism for gRNA-specified uridylate insertion into precursor mRNA. Science *273*, 1189–1195.

Kannan, S., and Burger, G. (2008). Unassigned MURF1 of kinetoplastids codes for NADH dehydrogenase subunit 2. BMC Genomics *9*, 455.

Kim, K.S., Teixeira, S.M., Kirchhoff, L.V., and Donelson, J.E. (1994). Transcription and editing of cytochrome oxidase II RNAs in Trypanosoma cruzi. J. Biol. Chem. *269*, 1206–1211.

Kolesnikov, A.A., Merzliak, E.M., Bessolitsyna, E.A., Fediakov, A.V., and Shoenian, G. (2003). [Reduction of the edited domain of the mitochondrial A6 gene for ATPase subunit 6 in Trypanosomatidae]. Mol. Biol. (Mosk.) *37*, 637–642.

Koslowsky, D.J., Bhat, G.J., Perrollaz, A.L., Feagin, J.E., and Stuart, K. (1990). The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. Cell *62*, 901–911.

Lai, D.-H., Hashimi, H., Lun, Z.-R., Ayala, F.J., and Lukes, J. (2008). Adaptations of Trypanosoma brucei to gradual loss of kinetoplast DNA: Trypanosoma equiperdum and Trypanosoma evansi are petite mutants of T. brucei. Proc. Natl. Acad. Sci. U. S. A. *105*, 1999–2004.

Lake, J.A., Cruz, V.F. de la, Ferreira, P.C., Morel, C., and Simpson, L. (1988). Evolution of parasitism: kinetoplastid protozoan history reconstructed from mitochondrial rRNA gene sequences. Proc. Natl. Acad. Sci. *85*, 4779–4783.

Lambert, A.J., and Brand, M.D. (2007). Research on mitochondria and aging, 2006-2007. Aging Cell *6*, 417–420.

Landweber, L.F. (1992). The evolution of RNA editing in kinetoplastid protozoa. Biosystems *28*, 41–45.

Landweber, L.F., and Gilbert, W. (1993). RNA editing as a source of genetic variation. Nature *363*, 179–182.

Landweber, L.F., and Gilbert, W. (1994). Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. Proc. Natl. Acad. Sci. U. S. A. *91*, 918–921.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Lee, S.H., Stephens, J.L., Paul, K.S., and Englund, P.T. (2006). Fatty acid synthesis by elongases in trypanosomes. Cell *126*, 691–699.

Lewis, M.D., Llewellyn, M.S., Yeo, M., Acosta, N., Gaunt, M.W., and Miles, M.A. (2011). Recent, Independent and Anthropogenic Origins of Trypanosoma cruzi Hybrids. PLoS Negl Trop Dis *5*, e1363.

Li, C.J., Hwa, K.Y., and Englund, P.T. (1995). A DNase from the trypanosomatid Crithidia fasciculata. Nucleic Acids Res. *23*, 4426–4433.

Li, F., Ge, P., Hui, W.H., Atanasov, I., Rogers, K., Guo, Q., Osato, D., Falick, A.M., Zhou, Z.H., and Simpson, L. (2009). Structure of the core editing complex (L-complex) involved in uridine insertion/deletion RNA editing in trypanosomatid mitochondria. Proc. Natl. Acad. Sci. U. S. A. *106*, 12306–12310.

Lima, L., Silva, F.M. da, Neves, L., Attias, M., Takata, C.S.A., Campaner, M., de Souza, W., Hamilton, P.B., and Teixeira, M.M.G. (2012). Evolutionary insights from bat trypanosomes: morphological, developmental and phylogenetic evidence of a new species, Trypanosoma (Schizotrypanum) erneyi sp. nov., in African bats closely related to Trypanosoma (Schizotrypanum) cruzi and allied species. Protist *163*, 856–872.

Lima, L., Espinosa-Álvarez, O., Hamilton, P.B., Neves, L., Takata, C.S.A., Campaner, M., Attias, M., de Souza, W., Camargo, E.P., and Teixeira, M.M.G. (2013). Trypanosoma livingstonei: a new species from African bats supports the bat seeding hypothesis for the Trypanosoma cruzi clade. Parasit. Vectors *6*, 221.

Lima, L., Espinosa-Álvarez, O., Ortiz, P.A., Trejo-Varón, J.A., Carranza, J.C., Pinto, C.M., Serrano, M.G., Buck, G.A., Camargo, E.P., and Teixeira, M.M.G. (2015). Genetic diversity of Trypanosoma cruzi in bats, and multilocus phylogenetic and phylogeographical analyses supporting Tcbat as an independent DTU (discrete typing unit). Acta Trop.

Liu, B., Liu, Y., Motyka, S.A., Agbo, E.E.C., and Englund, P.T. (2005). Fellowship of the rings: the replication of kinetoplast DNA. Trends Parasitol. *21*, 363–369.

Lukes, J., Arts, G.J., van den Burg, J., de Haan, A., Opperdoes, F., Sloof, P., and Benne, R. (1994). Novel pattern of editing regions in mitochondrial transcripts of the cryptobiid Trypanoplasma borreli. EMBO J. *13*, 5086–5098.

Lukes, J., Jirkû, M., Dolezel, D., Kral'ová, I., Hollar, L., and Maslov, D.A. (1997). Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. J. Mol. Evol. *44*, 521–527.

Lukes, J., Lys Guilbride, D., Votypka, J., Zikova, A., Benne, R., and Englund, P.T. (2002). Kinetoplast DNA Network: Evolution of an Improbable Structure. Eukaryot. Cell *1*, 495–502.

Lukes, J., Hashimi, H., and Zíková, A. (2005). Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. Curr. Genet. *48*, 277–299.

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

Machado, C.A., and Ayala, F.J. (2001). Nucleotide sequences provide evidence of genetic exchange among distantly related lineages of Trypanosoma cruzi. Proc. Natl. Acad. Sci. U. S. A. *98*, 7396–7401.

Madej, M.J., Alfonzo, J.D., and Hüttenhofer, A. (2007). Small ncRNA transcriptome analysis from kinetoplast mitochondria of Leishmania tarentolae. Nucleic Acids Res. *35*, 1544–1554.

Maia da Silva, F., Marcili, A., Lima, L., Cavazzana, M., Ortiz, P.A., Campaner, M., Takeda, G.F., Paiva, F., Nunes, V.L.B., Camargo, E.P., et al. (2009). Trypanosoma rangeli isolates of bats from Central Brazil: genotyping and phylogenetic analysis enable description of a new lineage using spliced-leader gene sequences. Acta Trop. *109*, 199–207.

Marande, W., Lukes, J., and Burger, G. (2005). Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot. Cell *4*, 1137–1146.

Marcili, A., Lima, L., Cavazzana, M., Junqueira, A.C.V., Veludo, H.H., Maia Da Silva, F., Campaner, M., Paiva, F., Nunes, V.L.B., and Teixeira, M.M.G. (2009). A new genotype of Trypanosoma cruzi associated with bats evidenced by phylogenetic analyses using SSU rDNA, cytochrome b and Histone H2B genes and genotyping based on ITS1 rDNA. Parasitology *136*, 641–655.

Marini, J.C., Levene, S.D., Crothers, D.M., and Englund, P.T. (1983). A bent helix in kinetoplast DNA. Cold Spring Harb. Symp. Quant. Biol. *47 Pt 1*, 279–283.

Marsden, P.D. (1990). Mucocutaneous leishmaniasis. BMJ *301*, 656–657.

Maslov, D.A. (2010). Complete set of mitochondrial pan-edited mRNAs in Leishmania mexicana amazonensis LV78. Mol. Biochem. Parasitol. *173*, 107–114.

Maslov, D.A., and Simpson, L. (1992). The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. Cell *70*, 459–467.

Maslov, D.A., and Simpson, L. (1994). RNA editing and mitochondrial genomic organization in the cryptobiid kinetoplastid protozoan Trypanoplasma borreli. Mol. Cell. Biol. *14*, 8174–8182.

Maslov, D.A., Kolesnikov, A.A., and Zaitseva, G.N. (1984). Conservative and divergent base sequence regions in the maxicircle kinetoplast DNA of several trypanosomatid flagellates. Mol. Biochem. Parasitol. *12*, 351–364.

Maslov, D.A., Sturm, N.R., Niner, B.M., Gruszynski, E.S., Peris, M., and Simpson, L. (1992). An intergenic G-rich region in Leishmania tarentolae kinetoplast maxicircle DNA is a pan-edited cryptogene encoding ribosomal protein S12. Mol. Cell. Biol. *12*, 56–67.

Maslov, D.A., Avila, H.A., Lake, J.A., and Simpson, L. (1994a). Evolution of RNA editing in kinetoplastid protozoa. Nature *368*, 345–348.

Maslov, D.A., Thiemann, O., and Simpson, L. (1994b). Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of Leishmania tarentolae. Mol. Biochem. Parasitol. *68*, 155–159.

Maslov, D.A., Hollar, L., Haghighat, P., and Nawathean, P. (1998). Demonstration of mRNA editing and localization of guide RNA genes in kinetoplast-mitochondria of the plant trypanosomatid Phytomonas serpens. Mol. Biochem. Parasitol. *93*, 225–236.

Maslov, D.A., Nawathean, P., and Scheel, J. (1999). Partial kinetoplast-mitochondrial gene organization and expression in the respiratory deficient plant trypanosomatid Phytomonas serpens. Mol. Biochem. Parasitol. *99*, 207–221.

Maslov, D.A., Votýpka, J., Yurchenko, V., and Lukeš, J. (2013). Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. Trends Parasitol. *29*, 43–52.

Mather, M.W., Henry, K.W., and Vaidya, A.B. (2007). Mitochondrial drug targets in apicomplexan parasites. Curr. Drug Targets *8*, 49–60.

Mehta, A., and Shaha, C. (2004). Apoptotic Death in Leishmania donovani Promastigotes in Response to Respiratory Chain Inhibition COMPLEX II INHIBITION RESULTS IN INCREASED PENTAMIDINE CYTOTOXICITY. J. Biol. Chem. *279*, 11798–11813.

Minchin, E.A. (1908). Investigations on the development of trypanosomes in tsetse flies and other Diptera. Quarterly Journal of Microscopical Science (Royal microscopical society.).

Missel, A., and Göringer, H.U. (1994). Trypanosoma brucei mitochondria contain RNA helicase activity. Nucleic Acids Res. *22*, 4050–4056.

Moreira, D., López-García, P., and Vickerman, K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. Int. J. Syst. Evol. Microbiol. *54*, 1861–1875.

Morita, Y.S., Paul, K.S., and Englund, P.T. (2000). Specialized fatty acid synthesis in African trypanosomes: myristate for GPI anchors. Science *288*, 140–143.

Motta, M.C.M. (2008). Kinetoplast as a potential chemotherapeutic target of trypanosomatids. Curr. Pharm. Des. *14*, 847–854.

Motta, M.C.M., Catta-Preta, C.M.C., Schenkman, S., de Azevedo Martins, A.C., Miranda, K., de Souza, W., and Elias, M.C. (2010). The bacterium endosymbiont of Crithidia deanei undergoes coordinated division with the host cell nucleus. PloS One *5*, e12415.

Muhich, M.L., Neckelmann, N., and Simpson, L. (1985). The divergent region of the Leishmania tarentolae kinetoplast maxicircle DNA contains a diverse set of repetitive sequences. Nucleic Acids Res. *13*, 3241–3260.

Myler, P.J., Glick, D., Feagin, J.E., Morales, T.H., and Stuart, K.D. (1993). Structural organization of the maxicircle variable region of Trypanosoma brucei: identification of potential replication origins and topoisomerase II binding sites. Nucleic Acids Res. *21*, 687–694.

Myler, P.J., Audleman, L., deVos, T., Hixson, G., Kiser, P., Lemley, C., Magness, C., Rickel, E., Sisk, E., Sunkin, S., et al. (1999). Leishmania major Friedlin chromosome 1 has an unusual distribution of protein-coding genes. Proc. Natl. Acad. Sci. U. S. A. *96*, 2902–2906.

Nawathean, P., and Maslov, D.A. (2000). The absence of genes for cytochrome c oxidase and reductase subunits in maxicircle kinetoplast DNA of the respiration-deficient plant trypanosomatid Phytomonas serpens. Curr. Genet. *38*, 95–103.

Nicholas, K.B., Nicholas, H.B.J., and Deerfield, D.W. (1997). GeneDoc: analysis and visualization of genetic variation. Embnew News *4*.

Niemann, M., Schneider, A., and Cristodero, M. (2011). Mitochondrial translation in trypanosomatids: a novel target for chemotherapy? Trends Parasitol. *27*, 429–433.

Obungu, V.H., Kiaira, J.K., Njogu, R.M., and Olembo, N.K. (1999). Catabolism of proline by procyclic culture forms of Trypanosoma congolense. Comp. Biochem. Physiol. B Biochem. Mol. Biol. *123*, 59–65.

Ochs, D.E., Otsu, K., Teixeira, S.M., Moser, D.R., and Kirchhoff, L.V. (1996). Maxicircle genomic organization and editing of an ATPase subunit 6 RNA in Trypanosoma cruzi. Mol. Biochem. Parasitol. *76*, 267–278.

Ochsenreiter, T., and Hajduk, S.L. (2006). Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. EMBO Rep. *7*, 1128–1133.

Ochsenreiter, T., Anderson, S., Wood, Z.A., and Hajduk, S.L. (2008a). Alternative RNA Editing Produces a Novel Protein Involved in Mitochondrial DNA Maintenance in Trypanosomes. Mol. Cell. Biol. *28*, 5595–5604.

Ochsenreiter, T., Cipriano, M., and Hajduk, S.L. (2008b). Alternative mRNA Editing in Trypanosomes Is Extensive and May Contribute to Mitochondrial Protein Diversity. PLoS ONE *3*.

Opperdoes, F.R., and Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle in Trypanosoma brucei: the glycosome. FEBS Lett. *80*, 360–364.

Opperdoes, F.R., and Coombs, G.H. (2007). Metabolism of Leishmania: proven and predicted. Trends Parasitol. *23*, 149–158.

Opperdoes, F.R., and Michels, P.A.M. (2008). Complex I of Trypanosomatidae: does it exist? Trends Parasitol. *24*, 310–317.

Opperdoes, F.R., Markoš, A., and Steiger, R.F. (1981). Localization of malate dehydrogenase, adenylate kinase and glycolytic enzymes in glycosomes and the threonine pathway in the mitochondrion of cultured procyclic trypomastigotes of Trypanosoma brucei. Mol. Biochem. Parasitol. *4*, 291–309.

Osato, D., Rogers, K., Guo, Q., Li, F., Richmond, G., Klug, F., and Simpson, L. (2009). Uridine insertion/deletion RNA editing in trypanosomatid mitochondria: In search of the editosome. RNA N. Y. N *15*, 1338–1344.

Osawa, S., Jukes, T.H., Watanabe, K., and Muto, A. (1992). Recent evidence for evolution of the genetic code. Microbiol. Rev. *56*, 229–264.

Pays, E., Vanhamme, L., and Pérez-Morga, D. (2004). Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries. Curr. Opin. Microbiol. *7*, 369–374.

Podlipaev, S.A., Sturm, N.R., Fiala, I., Fernandes, O., Westenberger, S.J., Dollet, M., Campbell, D.A., and Lukes, J. (2004). Diversity of insect trypanosomatids assessed from the spliced leader RNA and 5S rRNA genes and intergenic regions. J. Eukaryot. Microbiol. *51*, 283–290.

Porcel, B.M., Denoeud, F., Opperdoes, F., Noel, B., Madoui, M.-A., Hammarton, T.C., Field, M.C., Da Silva, C., Couloux, A., Poulain, J., et al. (2014). The Streamlined Genome of Phytomonas spp. Relative to Human Pathogenic Kinetoplastids Reveals a Parasite Tailored for Plants. PLoS Genet. *10*.

Ramesh, V., and Mukherjee, A. (1995). Post-kala-azar dermal leishmaniasis. Int. J. Dermatol. *34*, 85–91.

Read, L.K., Fish, W.R., Muthiani, A.M., and Stuart, K. (1993). Maxicircle DNA and edited mRNA sequences of closely related trypanosome species: implications of kRNA editing for evolution of maxicircle genomes. Nucleic Acids Res. *21*, 4073.

Read, L.K., Wilson, K.D., Myler, P.J., and Stuart, K. (1994). Editing of Trypanosoma brucei maxicircle CR5 mRNA generates variable carboxy terminal predicted protein sequences. Nucleic Acids Res. *22*, 1489–1495.

Reguera, R.M., Redondo, C.M., Gutierrez de Prado, R., Pérez-Pertejo, Y., and Balaña-Fouce, R. (2006). DNA topoisomerase I from parasitic protozoa: a potential target for chemotherapy. Biochim. Biophys. Acta *1759*, 117–131.

Roberts, C.W., McLeod, R., Rice, D.W., Ginger, M., Chance, M.L., and Goad, L.J. (2003). Fatty acid and sterol metabolism: potential antimicrobial targets in apicomplexan and trypanosomatid parasitic protozoa. Mol. Biochem. Parasitol. *126*, 129–142.

Rosenzweig, D., Smith, D., Opperdoes, F., Stern, S., Olafson, R.W., and Zilberstein, D. (2008). Retooling Leishmania metabolism: from sand fly gut to human macrophage. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *22*, 590–602.

Rota-Stabelli, O., Yang, Z., and Telford, M.J. (2009). MtZoa: A general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol. Phylogenet. Evol. *52*, 268–272.

Ruvalcaba-Trejo, L.I., and Sturm, N.R. (2011). The Trypanosoma cruzi Sylvio X10 strain maxicircle sequence: the third musketeer. BMC Genomics *12*, 58.

Ryley, J.F. (1951). Studies on the metabolism of the protozoa. I. Metabolism of the parasitic flagellate, Trypanosoma lewisi. Biochem. J. *49*, 577–585.

Ryley, J.F. (1956). Studies on the metabolism of the Protozoa. 7. Comparative carbohydrate metabolism of eleven species of trypanosome. Biochem. J. *62*, 215–222.

Sanchez-Moreno, M., Lasztity, D., Coppens, I., and Opperdoes, F.R. (1992). Characterization of carbohydrate metabolism and demonstration of glycosomes in a Phytomonas sp. isolated from Euphorbia characias. Mol. Biochem. Parasitol. *54*, 185–199.

Santhamma, K.R., and Bhaduri, A. (1995). Characterization of the respiratory chain of Leishmania donovani promastigotes. Mol. Biochem. Parasitol. *75*, 43–53.

Schägger, H. (2001). Respiratory chain supercomplexes. IUBMB Life *52*, 119–128.

Schnaufer, A., Domingo, G.J., and Stuart, K. (2002). Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. Int. J. Parasitol. *32*, 1071–1084.

Schneider, A., and Maréchal-Drouard, L. (2000). Mitochondrial tRNA import: are there distinct mechanisms? Trends Cell Biol. *10*, 509–513.

Sen, N., and Majumder, H.K. (2008). Mitochondrion of protozoan parasite emerges as potent therapeutic target: exciting drugs are on the horizon. Curr. Pharm. Des. *14*, 839–846.

Sen, N., Banerjee, B., Gupta, S.S., Das, B.B., Ganguly, A., and Majumder, H.K. (2007). Leishmania donovani: Dyskinetoplastid cells survive and proliferate in the presence of pyruvate and uridine but do not undergo apoptosis after treatment with camptothecin. Exp. Parasitol. *115*, 215–219.

Shapiro, T.A., and Englund, P.T. (1995). The structure and replication of kinetoplast DNA. Annu. Rev. Microbiol. *49*, 117–143.

Shapiro, T.A., and Showalter, A.F. (1994). In vivo inhibition of trypanosome mitochondrial topoisomerase II: effects on kinetoplast DNA maxicircles. Mol. Cell. Biol. *14*, 5891–5897.

Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell *53*, 401–411.

Shaw, J.M., Campbell, D., and Simpson, L. (1989). Internal frameshifts within the mitochondrial genes for cytochrome oxidase subunit II and maxicircle unidentified reading frame 3 of Leishmania tarentolae are corrected by RNA editing: evidence for translation of the edited cytochrome oxidase subunit II mRNA. Proc. Natl. Acad. Sci. U. S. A. *86*, 6220–6224.

Shioiri, C., and Takahata, N. (2001). Skew of Mononucleotide Frequencies, Relative Abundance of Dinucleotides, and DNA Strand Asymmetry. J. Mol. Evol. *53*, 364–376.

Shlomai, J. (2004). The structure and replication of kinetoplast DNA. Curr. Mol. Med. *4*, 623–647.

Silva Paes, L., Suarez Mantilla, B., Julia Barison, M., Wrenger, C., and Mariano Silber, A. (2011). The Uniqueness of the Trypanosoma cruzi Mitochondrion: Opportunities to Target New Drugs Against Chagas' Disease. Curr. Pharm. Des. *17*, 2074–2099.

Silvestro, D., and Michalak, I. (2012). raxmlGUI: a graphical front-end for RAxML. Org. Divers. Evol. *12*, 335–337.

Simpson, A.G.B., and Roger, A.J. (2004). Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol. Phylogenet. Evol. *30*, 201–212.

Simpson, L., and Maslov, D.A. (1994a). Ancient origin of RNA editing in kinetoplastid protozoa. Curr. Opin. Genet. Dev. *4*, 887–894.

Simpson, L., and Maslov, D.A. (1994b). RNA editing and the evolution of parasites. Science *264*, 1870–1871.

Simpson, L., and Maslov, D.A. (1999). Evolution of the U-Insertion/Deletion RNA Editing in Mitochondria of Kinetoplastid Protozoa. Ann. N. Y. Acad. Sci. *870*, 190–205.

Simpson, A.G.B., Lukes, J., and Roger, A.J. (2002). The evolutionary history of kinetoplastids and their kinetoplasts. Mol. Biol. Evol. *19*, 2071–2083.

Simpson, A.G.B., Gill, E.E., Callahan, H.A., Litaker, R.W., and Roger, A.J. (2004a). Early evolution within kinetoplastids (euglenozoa), and the late emergence of trypanosomatids. Protist *155*, 407–422.

Simpson, A.G.B., Stevens, J.R., and Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. *22*, 168–174.

Simpson, A.M., Neckelmann, N., de la Cruz, V.F., Muhich, M.L., and Simpson, L. (1985). Mapping and 5' end determination of kinetoplast maxicircle gene transcripts from Leishmania tarentolae. Nucleic Acids Res. *13*, 5977–5993.

Simpson, L., Neckelmann, N., de la Cruz, V.F., Simpson, A.M., Feagin, J.E., Jasmer, D.P., and Stuart, K. (1987). Comparison of the maxicircle (mitochondrial) genomes of Leishmania tarentolae and Trypanosoma brucei at the level of nucleotide sequence. J. Biol. Chem. *262*, 6182–6196.

Simpson, L., Wang, S.H., Thiemann, O.H., Alfonzo, J.D., Maslov, D.A., and Avila, H.A. (1998). U-insertion/deletion Edited Sequence Database. Nucleic Acids Res. *26*, 170–176.

Simpson, L., Thiemann, O.H., Savill, N.J., Alfonzo, J.D., and Maslov, D.A. (2000). Evolution of RNA editing in trypanosome mitochondria. Proc. Natl. Acad. Sci. U. S. A. *97*, 6986–6993.

Simpson, L., Sbicego, S., and Aphasizhev, R. (2003). Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business. RNA N. Y. N *9*, 265–276.

Simpson, L., Aphasizhev, R., Gao, G., and Kang, X. (2004b). Mitochondrial proteins and complexes in Leishmania and Trypanosoma involved in U-insertion/deletion RNA editing. RNA N. Y. N *10*, 159–170.

Sloof, P., de Haan, A., Eier, W., van Iersel, M., Boel, E., van Steeg, H., and Benne, R. (1992). The nucleotide sequence of the variable region in Trypanosoma brucei completes the sequence analysis of the maxicircle component of mitochondrial kinetoplast DNA. Mol. Biochem. Parasitol. *56*, 289–299.

Sloof, P., Arts, G.J., van den Burg, J., van der Spek, H., and Benne, R. (1994). RNA editing in mitochondria of cultured trypanosomatids: translatable mRNAs for NADH-dehydrogenase subunits are missing. J. Bioenerg. Biomembr. *26*, 193–203.

Souza, A.E., Myler, P.J., and Stuart, K. (1992). Maxicircle CR1 transcripts of Trypanosoma brucei are edited and developmentally regulated and encode a putative iron-sulfur protein homologous to an NADH dehydrogenase subunit. Mol. Cell. Biol. *12*, 2100–2107.

Souza, A.E., Shu, H.H., Read, L.K., Myler, P.J., and Stuart, K.D. (1993). Extensive editing of CR2 maxicircle transcripts of Trypanosoma brucei predicts a protein with homology to a subunit of NADH dehydrogenase. Mol. Cell. Biol. *13*, 6832–6840.

de Souza, W. (2008). An introduction to the structural organization of parasitic protozoa. Curr. Pharm. Des. *14*, 822–838.

de Souza, W., and Motta, M.C. (1999). Endosymbiosis in protozoa of the Trypanosomatidae family. FEMS Microbiol. Lett. *173*, 1–8.

de Souza, W., Attias, M., and Rodrigues, J.C.F. (2009). Particularities of mitochondrial structure in parasitic protists (Apicomplexa and Kinetoplastida). Int. J. Biochem. Cell Biol. *41*, 2069–2080.

Srivastava, I.K., Rottenberg, H., and Vaidya, A.B. (1997). Atovaquone, a broad spectrum antiparasitic drug, collapses mitochondrial membrane potential in a malarial parasite. J. Biol. Chem. *272*, 3961–3966.

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

Stephens, J.L., Lee, S.H., Paul, K.S., and Englund, P.T. (2007). Mitochondrial fatty acid synthesis in Trypanosoma brucei. J. Biol. Chem. *282*, 4427–4436.

Stevens, J.R. (2014). Free-living bodonids and derived parasitic trypanosomatids: but what lies in between? Trends Parasitol. *30*, 113–114.

Stevens, J.R., Teixeira, M.M., Bingle, L.E., and Gibson, W.C. (1999). The taxonomic position and evolutionary relationships of Trypanosoma rangeli. Int. J. Parasitol. *29*, 749–757.

Stevens, J.R., Noyes, H.A., Schofield, C.J., and Gibson, W. (2001). The molecular evolution of Trypanosomatidae. Adv. Parasitol. *48*, 1–56.

Stoco, P.H., Wagner, G., Talavera-Lopez, C., Gerber, A., Zaha, A., Thompson, C.E., Bartholomeu, D.C., Lückemeyer, D.D., Bahia, D., Loreto, E., et al. (2014). Genome of the Avirulent Human-Infective Trypanosome—Trypanosoma rangeli. PLoS Negl Trop Dis *8*, e3176.

Stuart, K., Brun, R., Croft, S., Fairlamb, A., Gürtler, R.E., McKerrow, J., Reed, S., and Tarleton, R. (2008). Kinetoplastids: related protozoan pathogens, different diseases. J. Clin. Invest. *118*, 1301–1310.

Stuart, K.D., Schnaufer, A., Ernst, N.L., and Panigrahi, A.K. (2005). Complex management: RNA editing in trypanosomes. Trends Biochem. Sci. *30*, 97–105.

Sturm, N.R., and Campbell, D.A. (2010). Alternative lifestyles: The population structure of Trypanosoma cruzi. Acta Trop. *115*, 35–43.

Sturm, N.R., and Simpson, L. (1990a). Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from Leishmania tarentolae mitochondria: RNA editing intermediates. Cell *61*, 871–878.

Sturm, N.R., and Simpson, L. (1990b). Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. Cell *61*, 879–884.

Sturm, N.R., Maslov, D.A., Blum, B., and Simpson, L. (1992). Generation of unexpected editing patterns in Leishmania tarentolae mitochondrial mRNAs: misediting produced by misguiding. Cell *70*, 469–476.

Sturm, N.R., Vargas, N.S., Westenberger, S.J., Zingales, B., and Campbell, D.A. (2003). Evidence for multiple hybrid groups in Trypanosoma cruzi. Int. J. Parasitol. *33*, 269–279.

Sutherland, D.V., Barns, A.M., and Ross, C.A. (1995). Trypanosoma evansi: measurement of pyruvate production as an indicator of the drug sensitivity of isolates in vitro. Trop. Med. Parasitol. Off. Organ Dtsch. Tropenmedizinische Ges. Dtsch. Ges. Für Tech. Zusammenarbeit GTZ *46*, 93–98.

Sutton, R.E., and Boothroyd, J.C. (1986). Evidence for Trans splicing in trypanosomes. Cell *47*, 527–535.

Talavera, G., and Castresana, J. (2007). Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst. Biol. *56*, 564–577.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. *9*, 678–687.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. *30*, 2725–2729.

Tarleton, R.L. (2003). Chagas disease: a role for autoimmunity? Trends Parasitol. *19*, 447–451.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. *17*, 57–86.

Teixeira, M.M.G., Borghesan, T.C., Ferreira, R.C., Santos, M.A., Takata, C.S.A., Campaner, M., Nunes, V.L.B., Milder, R.V., de Souza, W., and Camargo, E.P. (2011). Phylogenetic validation of the genera Angomonas and Strigomonas of trypanosomatids harboring bacterial endosymbionts with the description of new species of trypanosomatids and of proteobacterial symbionts. Protist *162*, 503–524.

Thiemann, O.H., Maslov, D.A., and Simpson, L. (1994). Disruption of RNA editing in Leishmania tarentolae by the loss of minicircle-encoded guide RNA genes. EMBO J. *13*, 5689–5700.

Tibayrenc, M. (2003). Genetic subdivisions within Trypanosoma cruzi (Discrete Typing Units) and their relevance for molecular epidemiology and experimental evolution. Kinetoplastid Biol. Dis. *2*, 12.

Tibayrenc, M., and Ayala, F.J. (2002). The clonal theory of parasitic protozoa: 12 years on. Trends Parasitol. *18*, 405–410.

Tielens, A.G.M., and van Hellemond, J.J. (2009). Surprising variety in energy metabolism within Trypanosomatidae. Trends Parasitol. *25*, 482–490.

Tielens, A.G., and Van Hellemond, J.J. (1998). Differences in energy metabolism between trypanosomatidae. Parasitol. Today Pers. Ed *14*, 265–272.

Tomasini, N., and Diosque, P. (2015). Evolution of Trypanosoma cruzi: clarifying hybridisations, mitochondrial introgressions and phylogenetic relationships between major lineages. Mem. Inst. Oswaldo Cruz *110*, 403–413.

Torri, A.F., and Englund, P.T. (1992). Purification of a mitochondrial DNA polymerase from Crithidia fasciculata. J. Biol. Chem. *267*, 4786–4792.

Truc, P., Büscher, P., Cuny, G., Gonzatti, M.I., Jannin, J., Joshi, P., Juyal, P., Lun, Z.-R., Mattioli, R., Pays, E., et al. (2013). Atypical Human Infections by Animal Trypanosomes. PLoS Negl. Trop. Dis. *7*.

Turrens, J.F. (1987). Possible role of the NADH-fumarate reductase in superoxide anion and hydrogen peroxide production in Trypanosoma brucei. Mol. Biochem. Parasitol. *25*, 55–60.

Tzfati, Y., Abeliovich, H., Kapeller, I., and Shlomai, J. (1992). A single-stranded DNA-binding protein from Crithidia fasciculata recognizes the nucleotide sequence at the origin of replication of kinetoplast DNA minicircles. Proc. Natl. Acad. Sci. U. S. A. *89*, 6891–6895.

Vaidya, A.B. (2004). Mitochondrial and plastid functions as antimalarial drug targets. Curr. Drug Targets Infect. Disord. *4*, 11–23.

Van Hellemond, J.J., Simons, B., Millenaar, F.F., and Tielens, A.G. (1998). A gene encoding the plant-like alternative oxidase is present in Phytomonas but absent in Leishmania spp. J. Eukaryot. Microbiol. *45*, 426–430.

Vasil'eva, M.A., Bessolitsyna, E.A., Merzliak, E.M., and Kolesnikov, A.A. (2004). [The identification of 12S rRNA gene promotor in Leptomonas seymouri mitochondrial DNA]. Mol. Biol. (Mosk.) *38*, 985–989.

Verner, Z., Cermáková, P., Skodová, I., Kriegová, E., Horváth, A., and Lukes, J. (2011). Complex I (NADH:ubiquinone oxidoreductase) is active in but non-essential for procyclic Trypanosoma brucei. Mol. Biochem. Parasitol. *175*, 196–200.

Vickerman, K. (1994). The evolutionary expansion of the trypanosomatid flagellates. Int. J. Parasitol. *24*, 1317–1331.

Viola, L.B., Attias, M., Takata, C.S.A., Campaner, M., De Souza, W., Camargo, E.P., and Teixeira, M.M.G. (2009). Phylogenetic analyses based on small subunit rRNA and glycosomal glyceraldehyde-3-phosphate dehydrogenase genes and ultrastructural characterization of two snake Trypanosomes: Trypanosoma serpentis n. sp. from Pseudoboa nigra and Trypanosoma cascavelli from Crotalus durissus terrificus. J. Eukaryot. Microbiol. *56*, 594–602.

Votýpka, J., Klepetková, H., Jirků, M., Kment, P., and Lukeš, J. (2012). Phylogenetic relationships of trypanosomatids parasitising true bugs (Insecta: Heteroptera) in sub-Saharan Africa. Int. J. Parasitol. *42*, 489–500.

de Vries, B.F., Mulder, E., Brakenhoff, J.P., Sloof, P., and Benne, R. (1988). The variable region of the Trypanosoma brucei kinetoplast maxicircle: sequence and transcript analysis of a repetitive and a non-repetitive fragment. Mol. Biochem. Parasitol. *27*, 71–82.

van Weelden, S.W.H., Fast, B., Vogt, A., van der Meer, P., Saas, J., van Hellemond, J.J., Tielens, A.G.M., and Boshart, M. (2003). Procyclic Trypanosoma brucei do not use Krebs cycle activity for energy generation. J. Biol. Chem. *278*, 12854–12863.

van Weelden, S.W.H., van Hellemond, J.J., Opperdoes, F.R., and Tielens, A.G.M. (2005). New functions for parts of the Krebs cycle in procyclic Trypanosoma brucei, a cycle not operating as a cycle. J. Biol. Chem. *280*, 12451–12460.

Westenberger, S.J., Barnabé, C., Campbell, D.A., and Sturm, N.R. (2005). Two Hybridization Events Define the Population Structure of Trypanosoma cruzi. Genetics *171*, 527–543.

Westenberger, S.J., Cerqueira, G.C., El-Sayed, N.M., Zingales, B., Campbell, D.A., and Sturm, N.R. (2006). Trypanosoma cruzi mitochondrial maxicircles display species- and strain-specific variation and a conserved element in the non-coding region. BMC Genomics *7*, 60.

Xu, C., and Ray, D.S. (1993). Isolation of proteins associated with kinetoplast DNA networks in vivo. Proc. Natl. Acad. Sci. U. S. A. *90*, 1786–1789.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. *11*, 367–372.

Yatawara, L., Le, T.H., Wickramasinghe, S., and Agatsuma, T. (2008). Maxicircle (mitochondrial) genome sequence (partial) of Leishmania major: gene content, arrangement and composition compared with Leishmania tarentolae. Gene *424*, 80–86.

Yurchenko, V.Y., Lukes, J., Jirku, M., and Maslov, D.A. (2009). Selective recovery of the cultivation-prone components from mixed trypanosomatid infections: a case of several novel species isolated from Neotropical Heteroptera. Int. J. Syst. Evol. Microbiol. *59*, 893–909.

Zavala-Castro, J.E., Acosta-Viana, K., Guzmán-Marín, E., Rosado-Barrera, M.E., and Rosales-Encina, J.L. (2000). Stage specific kinetoplast DNA-binding proteins in Trypanosoma cruzi. Acta Trop. *76*, 139–146.

Zijlstra, E.E., Musa, A.M., Khalil, E. a. G., el-Hassan, I.M., and el-Hassan, A.M. (2003). Post-kala-azar dermal leishmaniasis. Lancet Infect. Dis. *3*, 87–98.

Zingales, B., Andrade, S.G., Briones, M.R.S., Campbell, D.A., Chiari, E., Fernandes, O., Guhl, F., Lages-Silva, E., Macedo, A.M., Machado, C.R., et al. (2009). A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Mem. Inst. Oswaldo Cruz *104*, 1051–1054.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.

Zuker, M., and Jacobson, A.B. (1998). Using reliability information to annotate RNA secondary structures. RNA *4*, 669–679.

6. VITA

Preethi Ranganathan Iyengar was born on 18th May 1987, in Mumbai, India, and is an Indian citizen. She graduated from Swamy Vivekanand High School and Junior College, Mumbai, in 2004. She received her Bachelor of Science degree in Biotechnology from V. E. S. College of Arts, Science and Commerce, University of Mumbai, in 2007, and stood first in University of Mumbai in the Biotechnology stream (aggregate, out of ~ 1225 students; and second in Biotechnology component). Subsequently, she received her Master of Science degree in Biotechnology (again with a Distinction) from Ramniranjan Jhunjhunwala College, University of Mumbai in 2009. During her M. Sc., she was a student trainee at Bhabha Atomic Research Center [Molecular Biology Division (Molecular Genetics Section)], Mumbai, under the guidance of Dr. Hari S. Misra, for four months from May 2008 to Sep 2008. This was her first brush with Molecular Biology, and she worked extensively on a putative DNA double-strand break repair protein in *Deinococcus radiodurans* R1, a radioresistant bacterium. After receiving her Masters degree, she worked at Tata Institute of Fundamental Research (Department of Biology), Mumbai under the guidance of Dr. Basuthkar J. Rao, for six months from Dec 2009 to May 2010. At Dr. Rao's lab, she worked with eukaryotic cells (algae) for the first time, and her research focused on standardization of reverse genetics techniques to develop Rad51 and Dad1 mutant libraries in *Chlamydomonas reinhardtii*. She joined Virginia Commonwealth University (School of Medicine) in Aug 2010, and started working towards her Ph.D. in Microbiology and Immunology, with a concentration in Molecular Biology and Genetics. She joined Dr. Gregory A. Buck's lab in June 2011, and has since been working towards her dissertation under his mentorship. Upon successful completion of her doctoral degree, she will move to Atlanta

(Georgia, USA), to start her post-doctoral training in Clinical and Translational Oncology, under the mentorship of Dr. Ritu Aneja, at Georgia State University.