PERVASIVE ALTERNATIVE RNA EDITING IN *TRYPANOSOMA BRUCEI*

By

Laura Elizabeth Kirby

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Microbiology and Molecular Genetics—Doctor of Philosophy

2019

ABSTRACT

PERVASIVE ALTERNATIVE RNA EDITING IN *TRYPANOSOMA BRUCEI*

By

Laura Elizabeth Kirby

Trypanosoma brucei is a single celled eukaryote that utilizes a complex RNA editing system to render many of its mitochondrial genes translatable. Editing of these genes requires multiple small RNAs called guide RNAs to direct the insertion and deletion of uridines. These gRNAs act sequentially, each generating the anchor binding site for the next gRNA. This sequential dependence should render the process quite fragile, and mutations in the gRNAs should not be tolerated. In the examination of the gRNA transcriptome of *T. brucei*, many gRNAs were identified that are capable of generating alternative mRNA sequences, and potentially disrupting the editing process. In this work, the effects of alternative editing are characterized. This analysis revealed the role of gRNAs in developmental regulation of gene expression, showing a correlation between the abundance of the initiating gRNAs across two different points in the life cycle of *T. brucei* and their expression. This study also revealed the existence of mitochondrial dual-coding genes, which provide protection for genetic material that is not under selection at all points of the life cycle of *T. brucei*. The examination of these dual-coding genes showed that RNA editing patterns can shift between cell lines and under different energetic conditions. Examining the gRNAs involved in these editing pathways revealed that there is a high amount of mismatching base pairs that are tolerated for editing to function, and that gRNA abundance is not a reliable predictor for editing preference. Finally, a

reexamination of the gRNA transcriptome revealed that many gRNAs are still unidentified and most likely are generating new alternatively edited sequences.

ACKNOWLEDGEMENTS

I would first like to acknowledge my family, without whose support I could not have completed the work I have done. I am forever grateful for their continuous encouragement and constant faith in me.

I gratefully acknowledge Dr. Donna Koslowsky, who has been an excellent mentor. She not only provided her extensive insights on the field of RNA editing, but also aided and advised me in my professional development. She had shaped me as a researcher and as a person, and I was very fortunate in being able to work with her.

I would also like to thank the small undergraduate army I have had the priveledge to work with. They taught me a great deal about mentoring and leadership, and their many hours greatly aided me in the completion of this work.

For their assistance in my computational education, I would like to thank Dr. Yanni Sun and Dr. Arend Hintze. Without their assistance, my research would not have been possible.

For serving as my graduate committee, I would like to thank Dr. Shannon Manning, Dr. Charles Hoogstraten, Dr. Chris Adami and Dr. Yanni Sun. Their many contributions have shaped and refined my research.

I would also like to thank Dr. Cori Fata-Hartley, for serving as my teaching mentor and helping me design and conduct an education research project.

I would like to thank the friends I have gained during my time at Michigan State who have supported me and been wonderful colleagues: Alexis Weber, Sandy Olenic, Ahrom Kim, and Shreya Saha.

Finally, I would like to thank the Department of Microbiology and Molecular Genetics, the College of Natural Science, the Elenor L. Gilmore Endowment, the Frank Peabody Microbiology Student Research Fund, the Russell B. DuVall Endowment, the Berttina Wentworth Fellowship, and the Marvis A. Richardson Endowed Fellowship for their support of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Kinetoplastids

Trypanosoma brucei is a member of the Kinetoplastea, a group of protozoans characterized by a large network of DNA in their mitochondria known as the kinetoplast that is physically attached to the flagellum [1]. While not all kinetoplastids are parasites, the group encompasses some of the most successful parasites in existence, inhabiting an incredibly wide range of hosts from plants to invertebrates to vertebrates [2,3]. The dixenous members cycle between two distinct hosts and can encounter different environments with distinct metabolic constraints. These environmental shifts require rapid and extensive changes in gene expression. This is particularly interesting considering the kinetoplastids' bizarre and complicated use of RNA editing for their mitochondrial gene expression.

Kinetoplastid RNA Editing

RNA editing is one of several unique genetic features found in the mitochondria of these parasites. RNA editing creates open reading frames in "cryptogenes" by insertion and deletion of uridylate residues at specific sites within the mRNA. The U-insertions/deletions are directed by small guide RNAs (gRNAs) and can repair frameshifts, generate start and stop codons and more than double the size of the transcript [4].

The kinetoplast DNA (kDNA) consists of two types of DNA molecules, maxicircles and minicircles. Maxicircles are large circular DNA molecules that contain the genes for two ribosomal RNAs, 12S and 9S, and the protein coding genes [5]. While some of the protein-coding genes do not require RNA editing prior to translation, most require extensive editing

before they can be translated [6,7]. The sequence changes are guided by small complementary RNA molecules (the gRNAs) that are encoded on the minicircles [8]. Minicircles make up the bulk of the kinetoplastid network with each minicircle encoding 1–5 gRNAs.

This effectively means that the genetic information for the edited mitochondrial mRNAs is dispersed between the mRNA cryptogenes on the maxicircles and as many as 10,000 gRNA encoding minicircles. In *T. brucei*, the extensive editing of a single transcript can require more than 40 gRNAs and hundreds of editing events [9]. The gRNAs act as templates for the large multi-subunit protein complex known as the editosome [4,6]. The editosome cleaves the mRNA, inserts or deletes the correct number of uridines and then re-ligates the mRNA in an energy intensive process. This is repeated until the mRNA is complementary to the small gRNA. The initiating gRNA interacts with the 3' end of the pre-edited transcript and generates the anchor binding region for the next gRNA. In fact, all subsequent gRNAs anchor to the edited sequence created by the preceding gRNA. Editing proceeds from the 3' end to the 5' end of the mRNA transcript with the terminating gRNA either creating the start codon or bringing an existing start codon into frame. Because each gRNA directs editing that generate the anchor region for the next gRNA, the RNA editing process is sequentially dependent on correct editing by each gRNA. As a result, the process is incredibly fragile.

Trypanosoma brucei

Trypanosoma brucei is the causative agent of Human African trypanosomiasis (HAT) and one agent of Animal African Trypanosomiasis (AAT). Each year, 10,000 new cases of HAT are reported, and 3 million cattle are killed, severely impacting the lives and livelihood of those in infected areas [10,11]. The trypanosomes live in two distinct environments: the animal host

and the insect vector, the tsetse fly. These environments are distinct in temperature and nutrient composition, providing a challenge to *T. brucei* as it cycles between hosts. While in the mammalian host, *T. brucei* lives entirely extracellularly in the bloodstream. It is frequently subject to attacks by the host's adaptive immune system, and the population evades these attacks through antigenic variation [12]. This part of the life cycle can be quite long, with the longest known infection lasting 29 years [13]. In the bloodstream, the bulk of the trypanosome population exists in the actively dividing slender form. The slender form is optimized to utilize its glucose rich environment, using glycolysis to generate energy [14,15]. During this stage of the life cycle, the mitochondrion is down-regulated, lacking both Krebs cycle enzymes and a functional electron transport chain (ETC) [16]. While the activity of the mitochondrion is relatively low during the bloodstream stage (BS), expression of the mitochondrial genome is still essential [17,18]. Once the population reaches an optimum density, a small portion of the population transitions into stumpy form trypanosomes. The stumpy form is nondividing and appears to be transitional, activating mitochondrial genes in preparation for uptake in a blood meal by its tsetse fly vector and subsequent transfer to a harsher environment [14]. Successful transition to the fly vector requires activation of the ETC, and ATP synthesis via oxidative phosphorylation. Once inside the tsetse fly, the parasite utilizes proline as its primary energy source while residing and actively dividing in the midgut [19–21]. This stage of the life cycle is followed by a dramatic bottleneck when the trypanosomes transition from the midgut to the salivary glands of the tsetse fly, with as few as 1-5 trypanosomes completing the transition [22,23]. From the salivary glands, trypanosomes are then refluxed into their next mammalian

host during a bloodmeal. In order to adapt to these sudden changes in environment, *T. brucei* must vastly alter its gene expression, most notably, in its mitochondria.

The 22 kb maxicircle of *T. brucei* encodes several genes involved in the mitochondrial ETC and oxidative phosphorylation, NADH dehydrogenase (ND) subunits 1-5 and 7-9, cytochrome oxidase (CO) subunits I-III, cytochrome b (CYb), ATP synthase subunit 6 (A6), as well as genes encoding the ribosomal protein small subunit 12 (RPS12), 12S and 9S rRNAs, and some genes with unknown functions: C-rich regions (CR) 3 and 4, and Maxicircle unidentified reading frames (Murf) 2 and 5 [5]. Twelve of these genes require some amount of RNA editing to be translatable, with some requiring only one or two gRNAs (COII, CYb, MurfII), and others requiring editing across the span of the transcript (ND3, ND7, ND8, ND9, COIII, A6, RPS12, CR3, and CR4) [4,6,7].

Distinct differences in mitochondrial transcript abundance, polyadenylation and the extent of RNA editing are observed during the complex life cycle (Table 1). The pattern of differential RNA editing observed is especially interesting. For example, the CYb and COII mRNAs are edited during the insect stage, but are primarily unedited in bloodstream forms [24,25]. In contrast, editing of the NADH dehydrogenase subunit transcripts (ND3, ND7, ND8 and ND9) and RPS12 appears to occur preferentially in bloodstream forms [5,26–30]. Other transcripts, COIII and A6 are edited equally in both life cycle stages [31,32].

Aside from the genes encoded on the maxicircle, there are the minicircle gRNAs. The minicircles range in abundance from 5,000–10,000 present in each network, and are ~1kb in size, with each minicircle encoding 2–5 gRNAs. In *T. brucei*, there are more than 200 different minicircle sequence classes (~1200 gRNAs) [8,33]. While the minicircles make up a bulk of the

Table 1. Differences in mitochondrial transcript abundance, polyadenylation and the extent of RNA editing in two life cycle stages of *T. brucei*.

Gene	No. of uridines		Edited size (nt)	Stage Edited	Relative level of mature RNA			PolyA tail length		Number of PC major classes[g]	References
	Added	Deleted			Long Slender	Short Stumpy	Procyclic	Bloodstream	Procyclic		
12S	2-17 (tail)	0	1149	N.D.	0.04	1.3	1.0	N.A.	N.A.	N.A.[j]	[24,34,35]
9S	7(tail)	0	611	N.D.	0.07	1.4	1.0	N.A.	N.A.	N.A.	[24,34,35]
CYb	34	0	1,151	P[a]	~0	0.5	1.0	Short (UE[h] & E[i])	Short (E) & Long (E)	11	[24,25,36,37]
A6	448	28	821	P/BS[b]	1.0	N.D.[f]	1.0	Short (E) & Long (E)	Short (E) & Long (E)	81	[31,36]
COI	0	0	1,647	NE[c]	0.07	0.4	1.0	Short	Short & Long	N.A.	[24,37,38]
COII	4	0	663	P	~0	0.5	1.0	Short	Short (UE) & Long (E)	N.A.[k]	[24,37,39]
COIII	547	41	969	P/BS	1.0	N.D.	1.0	Short (E)	Short(E) & Long(E)	151	[32]
ND1	0	0	960	NE	~1	N.D.	~1	Short & Long	Short & Long	N.A.	[38,40,41]
ND2	0	0	1,343	NE	>1.0	N.D.	1.0	Short & Long	Short & Long	N.A.	[37,42,43]
ND3	210	13	452	P/BS	>1.0	N.D.	1.0	Short (E)	Short (E)	34	[26]
ND4	0	0	1,314	NE	~1.0	N.D.	~1.0	Short & Long	Short & Long	N.A.	[37,38,44,45]
ND5	0	0	1,779	NE	0.5	0.8	1.0	N.D.	N.D.	N.A.	[24,38]
ND7	553	89	1,238	5'P/BS, 3'BS[d]	~10	N.D.	1.0	Short(E)	Short(UE)	129	[5,27]
ND8	259	46	574	BS[e]	~20	N.D.	1.0	Short (E) & Long (E)	Short (E)	70	[5,28,37]
ND9	345	20	649	BS	>1.0	N.D.	1.0	Short (PE) & Long (E)	Short (PE)	39	[29]
RPS12	132	28	325	BS	>1.0	N.D.	1.0	Short(UE & E) & Long (E)	Short(UE & E) & Long (E)	50	[30]
Murf 2	26	4	1,111	P/BS	~1	~1	~1	Short & Long	Short & Long	1	[40,46]
Murf 5	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.A.	N.D.
CR3	148	13	299	BS	>1.0	N.D.	1.0	N.D.	N.D.	37	[47]
CR4	325	40	567	BS	1.0	N.D.	~0	Short (E) & Long (E)	Short (UE)	41	[48]

[a]P, transcript is edited only in the procyclic (insect) developmental stage.
[b]P/BS, transcript is edited in both bloodstream and procyclic stages.
[c]NE, never edited, editing of these transcripts has not been reported.
[d]The ND7 transcript is differentially edited in the procyclic and bloodstream stages.
[e]BS These transcripts are only fully edited in the bloodstream developmental stage [49].
[f]N.D. Values have not yet been determined.
[g]All data comes from EATRO 164 procyclic gRNA transcriptome previously published [9].
[h]UE, unedited, the transcripts which carried these tails were typically unedited.
[i]E, edited, the transcripts carrying these tails were typically edited.
[j]N.A., Not applicable.
[k]COII is a cis-edited transcript. Poly-A tails listed as short are between 10 and 50 nts long and tails listed as long are between 150 and 200 nts long.

mitochondrial DNA, early studies using both Northern blot and primer extension analyses on a limited number of gRNAs indicate that gRNAs are present in both insect and bloodstream forms, suggesting that the regulation of RNA editing is not at the level of gRNA availability [28,50,51].

Trypanosoma vivax

Like *T. brucei*, *Trypanosoma vivax* is a causative agent of AAT. Its kinetoplast DNA is also very similar. The maxicircles in *T. vivax* possess the same genes as the maxicircle of *T. brucei*, and the genes that are edited in *T. brucei* are also edited in *T. vivax* to the same extent [52]. The minicircles of *T. vivax* vary significantly in size, from 300-1100 bp, encoding 1-3 gRNAs [52]. The life cycle of *T. vivax* is highly similar to that of *T. brucei*. In its mammalian host, it lives extracellularly in the bloodstream, primarily metabolizing glucose. When it is taken up by the tsetse fly in a bloodmeal, it initially resides in the proventriculus and foregut. From there, cells migrate to the proboscis and propagate, preparing to be deposited into the next mammalian host [53].

Trypanosoma cruzi

T. cruzi is known for causing Chagas disease in South and Central America and is carried between hosts by triatomine bugs. Like *T. brucei* and *T. vivax*, it has highly similar kDNA, with the maxicircle containing the same genes in the same order. The edited genes of *T. brucei* and *T. vivax* are also edited in *T. cruzi* to the same extent [54]. Like *T. brucei*, *T. cruzi* spends most of its insect stage in the nutrient depleted midgut of its host, metabolizing amino acids for survival [55,56]. Once cells have propagated, they migrate to the hindgut and are excreted from the fly.

Transmission to the mammalian host occurs by contact with a wound or mucous membrane. Once inside the mammalian host, *T. cruzi* invades many different types of nucleated cells by using the microtubule cytoskeleton of the host cell to recruit lysosomes to create vacuolar compartments where *T. cruzi* resides [57]. Once inside the vacuole, *T. cruzi* metabolizes glucose as its primary energy source [58,59].

Leishmania spp.

Leishmania spp. are found on almost every continent in the world, and infect 700,000 to 1.2 million people annually [60]. *Leishmania spp.* are transmitted by the phlebotomine sand fly. Unlike the trypanosomes, while in the midgut of the sand fly, *Leishmania spp.* primarily metabolize glucose, because of the fly's frequent sap meals. *Leishmania spp.* are transmitted to their mammalian hosts by the bite of the sand fly. Once inside the host, they are phagocytosed by host cells. Inside macrophages, they replicate within lysosome like compartments, and it is believed that these compartments are not nutrient restrictive [61,62]. The maxicircle of the *Leishmania spp.* parasites has the same genes as the trypanosomes, but their editing patterns significantly vary. The pan-edited genes in *Leishmania spp.* are ND3, ND8, ND9, RPS12, CR3, and CR4, and the partially edited genes are A6, COII, COIII, MurfII, and ND7 [63].

Phytomonas spp.

Phytomonas spp. parasitize plants, utilizing their sucrose and polysaccharides as energy sources. Their insect vectors maintain a highly sap rich diet that allows the parasites to continually metabolize carbohydrates, unlike the *Trypanosoma spp*. Possibly as a result of this,

several metabolic pathways are incomplete. The pathways for beta oxidation of fatty acids or oxidation of amino acids are missing key enzymes, but the pathways for the synthesis of these metabolites are more complete [64]. The ETC of the mitochondria is also affected. Genes for all cytochromes are missing from nuclear and kDNA, and the cytochrome oxidase genes normally present on the maxicircle are also missing [64,65]. The other maxicircle genes are present, and ND3, ND8, ND9, RPS12, CR3 and CR4 are pan-edited, while ND7, A6 and MurfII are only partially edited, as with *Leishmania spp.*

Procyclic gRNA Transcriptome

The gRNA transcriptome of insect stage (procyclic) *T. brucei* was previously sequenced [50]. This library was generated from the EATRO 164 cell line, grown in SDM79 medium, the most commonly used medium when culturing procyclic trypanosomes. As no reference genome exists for the minicircles, gRNAs could only be identified based on their function. Using a longest common substring algorithm, gRNAs were identified based on their complementarity to previously determined fully edited mRNA sequences. The RNA editing system tolerates G:U base pairs, so this program allows these base pairs in alignments, but these base pairs do not contribute to the overall alignment score as much as canonical Watson-Crick base pairs (1 point for G:U and 2 points Watson-Cricks). Guide RNAs with scores higher than 45 points were identified as editing gRNAs. Due to the fact that trypanosomes post transcriptionally add a poly-uridine (poly-U) tail is to gRNAs, the sequences generated in this transcriptome possess the poly-U tail as well [66]. This program ignores the transcript's poly-U tail in the alignment, and it does not contribute to the score. Using this program, full complements of gRNAs were found for A6, COIII, CR4, CYb, and RPS12, and near full complements were identified for the

other edited genes. This study found that multiple different sequence classes of gRNAs (major classes) edited the same region of an mRNA (this group of gRNAs is called a population). Major classes within a population had many transition mutations, primarily A-G mutations, which appeared to be due to the editing system's toleration of G-U base pairs. Interestingly, populations of gRNAs varied extremely in transcript abundance, with abundance varying from <10 to >350,000 reads.

The gRNAs identified in this study possessed common characteristics. 64% of transcripts had 38-48 nucleotides (nt) of complementarity to their target mRNA. 84% of transcripts had 6 or fewer non-complementary nts at 5' end, and most transcripts had 0 nts non-complementary at 3' end prior to the poly-U tail. Conservation was observed in the gRNA transcription start site, with 74% of transcripts starting with 5'-ATATA-3'. Interestingly, a large proportion of transcripts had 5'-AAAAA-3' transcription start sites as well.

Beyond the identification of gRNAs directing the conventional mRNA edits, this study identified a number of gRNAs that could generate alternative edits. Most of these edits caused minor changes to the predicted mRNA and protein sequences, by either changing a single amino acid (ND8) or changing no amino acids at all (A6). However, some gRNAs were identified dramatically altered the mRNA or protein sequence. One edit in the essential A6 gene caused a frameshift that would alter and shorten the C-terminus of the protein [17]. Another generated a dramatically different sequence at the 3' end of CR3, and no gRNA was identified capable of editing that sequence. Interestingly, another study identified an alternative edit in COIII, that linked an open reading frame in the unedited 5' end of the transcript with the reading frame in

the edited 3' end of the transcript [67,68]. The gRNA required to generate this alternative edit was not identified in the procyclic gRNA transcriptome of *T. brucei*.

Evolution and retention of RNA editing in kinetoplastids

The kinetoplastid RNA editing system is energetically expensive and, due to the system's sequential dependence, should be highly fragile. This means that with even high accuracy rates for each gRNA, the overall fidelity of the process is astonishingly low. Even a single point mutation could drastically change the editing pattern, and stop the editing process, aborting expression of the protein. A major question in the field has been why this fragile and metabolically expensive system of RNA editing would evolve and persist.

Initially, it was proposed that U-insertion/deletion editing (kRNA editing) was one of many RNA editing processes that were in fact relics of the RNA world. However, the very different mechanisms of the RNA editing systems in existence and their very limited distribution within specific groups of organisms indicate that they are more likely derived traits that evolved later in evolution [69,70]. The sheer complexity of the kRNA editing process, with no obvious selective advantage, led to the proposal that insertion/deletion editing arose via a constructive neutral evolution (CNE) pathway [71]. RNA editing in trypanosomes is always mentioned in support of CNE as an example of how seemingly non-advantageous, complex processes can arise [72,73]. More recently however, it has been hypothesized that RNA editing co-evolved with G-quadruplex structures found in the pre-edited mRNAs [74]. These structures can help regulate transcription in order to promote DNA replication and prevent kDNA loss, and thus provide an advantage to the organism. However, they must be removed by the RNA editing system prior to translation [74]. Another prominent hypothesis is that RNA editing is

advantageous because it is a mechanism by which an organism can fragment and scatter essential genetic information throughout a genome [75,76]. Kinetoplast DNA is far less stable than chromosomal DNA, and loss of minicircles due to asymmetric division of the kDNA network have been frequently observed, particularly in laboratory cultures of *Leishmania tarentolae* [77,78]. Buhrman et al. [76] suggest that the scattering of essential gRNA genes throughout the DNA network would prevent fast growing deletion mutants from outcompeting more metabolically versatile parasites during growth in the mammalian host. Using a mathematical model of gene fragmentation in changing environments (absence of functional selection), they showed a distinct advantage for gene fragmentation. In their model, the number of tolerable generations under periods of relaxed selective pressure was increased by more than 40% before loss of the ability to move to the next life cycle stage.

One mechanism for protecting small asexual populations is by increasing the severity of the mutations that can occur. If mutations severely impact fitness, deleterious mutations are selected out, preventing their fixation [79]. This phenomenon increases the 'drift robustness' of a population. One study modeled the acquisition of drift robustness mathematically and computationally [80]. This study showed that in a simulated environment, small populations evolved a lower fitness than large populations, but when the most common genotypes from these populations were placed in a scenario with extremely high genetic drift, the genotypes evolved from small populations experienced a smaller decline in fitness than the genotypes evolved in a large population. Furthermore, they examined the types of mutations in the fitness landscape nearest to the peaks that each simulated population had fixed on and found that in smaller populations, there was an excess of mutations possible that were neutral,

beneficial or strongly deleterious, whereas in larger populations, there were more small-effect deleterious mutations possible. As the RNA editing process may be operating as a proof-reading system to weed out mutations by making them lethal, these findings strongly support the hypothesis that RNA editing is beneficial to the trypanosomes by providing a level of drift robustness to the population as a whole.

While these hypotheses do address the evolution and retention of the RNA editing system itself, they do not address another key issue with the system as a whole: maintaining genetic material used in RNA editing while that material is not under selection. During the life cycle of *T. brucei*, trypanosomes undergo a severe bottleneck as they transition through the tsetse fly and into the mammalian host, and then within the bloodstream, they undergo multiple bottlenecks at each antigenic switch, as they evade the host immune system [22]. Such bottlenecks create additional forces of genetic drift, where genes can be lost even if their deleterious fitness effect is considerable. This life cycle should make *T. brucei* particularly sensitive to genetic drift, especially for those genes which are not under selection (Krebs cycle and ETC) and should make them extremely vulnerable to Muller's ratchet (the gradual increase of mutational load that eventually leads to extinction) [81–84].

During a reexamination of the EATRO 164 procyclic gRNA transcriptome, a number of gRNAs were identified capable of shifting the open reading frame of their respective transcripts. These gRNAs acted at the 5' end of edited mRNAs and either shifted the position of an existing start codon or generated a new start codon that would allow that transcript to be translated in an alternative reading frame. Surprisingly, the alternative reading frames spanned the full or nearly full length of their transcripts, suggesting that these transcripts were capable

of generating two distinctly different protein products. Based on these observations, we hypothesize that trypanosomes use dual-coding genes to protect genetic information by essentially hiding a gene not under selection (i.e. ETC genes) within one that remains under selection. Thus, the ability to access overlapping reading frames may be one explanation for how genetic material that is unused in one life cycle stage may be preserved while it is not under selection.

Dual-coding and dual-function genes

Dual-coding genes are defined as a stretch of DNA containing overlapping open reading frames (ORFs) [85,86]. Overlapping reading frames are common in viruses and are thought to persist due to strong genome size constraints [87,88]. More recently however, overlapping genes have been identified in mammalian and bacterial genomes [89–92]. In these organisms, size is not an issue and the potential advantage of overlapping genes is less clear. Maintaining dual-coding genes is costly, as it constrains the flexibility of the amino acid composition of both proteins, constraining the ability of each protein to become optimally adapted [93]. As this constraint can be alleviated by gene duplication, it is thought that dual-coding regions can survive long evolutionary spans only if the overlap provides a selective advantage. In mammals, many of the identified dual-coding genes produce two proteins that bind and regulate each other [94,95]. For these proteins, dual-coding may be advantageous for the tight co-expression needed. An alternative model suggests that under high mutation rates, the overlapping of critical nucleotide residues is advantageous because it may reduce the target size for lethal mutations [96].

The use of genetic information with more than one function is not a new idea in *T. brucei*. The nuclear encoded α-ketoglutarate dehydrogenase E2 (α-KDE2) is known to be a dual-function protein, in that it plays important roles in both the Krebs cycle and in mitochondrial DNA inheritance [97]. RNAi knockdowns of this gene in bloodstream form (BF) trypanosomes also show a pronounced reduction in cell growth. Similarly, the Krebs cycle enzyme α-ketoglutarate decarboxylase (α-KDE1) is a dual-function protein with overlapping targeting signals that allow it to be localized to both the mitochondrion and glycosomes [98]. RNAi knockdowns of α-KDE1 in BF trypanosomes is lethal, suggesting that, in addition to its enzymatic role in the Krebs cycle, it plays an essential role in glycosomal function in *T. brucei* [98].

Another example of this was identified in the RNA editing system. Alternative editing of COIII is reported to generate a novel DNA-binding protein, Alternatively Edited Protein-1 (AEP-1), that functions in mitochondrial DNA maintenance [67,68]. In this transcript, one alternative gRNA generates sequence changes at two sites that links an open reading frame (ORF) found in the pre-edited 5' end, to the 3' transmembrane domains found in the COIII edited ORF. This was the first indication that one cryptogene could contain information for more than one protein. It has been previously suggested that both alternative editing and dual-function proteins are important mechanisms for expanding the functional diversity of proteins found in trypanosomes [67,97–99]. We hypothesize that because trypanosomes live exclusively extracellularly in their mammalian host, they are more sensitive to genetic drift, and an equally important role for these dual-coding/function genes may be the protection of genetic information.

Project Summary

The goal of this work is to examine the impact of alternative editing on the protein diversity, editing efficiency, developmental regulation, and genetic integrity of *Trypanosoma brucei*. This investigation began with the generation of the gRNA transcriptome of bloodstream form EATRO 164 *T. brucei*. This analysis identified near full complements of gRNAs for the edited genes, as was discovered in the procyclic transcriptome. A detailed comparison of the gRNAs identified in both datasets revealed conserved characteristic, such as anchor length, length of complementarity, and transcription start site sequences, even though very few identical sequences existed between the two transcriptomes. Additionally, an interesting correlation was found that suggests a relationship between the relative abundance of initiating gRNAs between stages and the developmental pattern of mRNA editing.

During this comparison of the two transcriptomes, a number of alternative editing gRNAs were identified. Notably, three of these gRNAs were capable of shifting translation of ND7, RPS12, and CR3 into alternative reading frames. This discovery prompted the analysis of the mitochondrially encoded transcripts to determine which of the genes had the capacity to be dual-coding. Using mutational bias analysis, we show that as many as six cryptogenes in addition to the previously discovered COIII/AEP-1, encode more than one protein, and that RNA editing allows access to both reading frames.

In order to determine if mRNA transcripts with access to multiple open reading frames exist within the mitochondrial transcriptome, we deep sequenced the transcript populations of three putative dual coding genes: RPS12, the 5' editing domain of ND7 (ND7 5'), and CR3. Using the previously generated gRNA transcriptomes, we constructed detailed editing pathways for

each of these genes. We found evidence that CR3 and ND7 5' are dual-coding genes, based on the identification of transcripts that would translate into different reading frames. This study indicates that RNA editing can be used to access multiple open reading frames using two different methods: in ND7 5', different gRNAs bring alternate start codons into frame and in CR3, different gRNAs can shift the reading frame of the existing start codon. In addition, CR3 showed incredible editing diversity. In two different cell lines, highly divergent editing patterns were characterized, with the two cell lines using different sets of gRNAs to edit the CR3 cryptogene. This suggests that the use of a gRNA-guided editing system can also dramatically increase protein diversity in spite of an incredibly rigid and mutationally fragile system.

With a more complete understanding of the existing edits found in the mRNA transcriptomes of RPS12, ND7 5' and CR3, we used this knowledge to analyze the RNA editing system's ability to tolerate noise. Reexamining the procyclic gRNA transcriptome revealed many previously unidentified gRNAs, potentially capable of generating alternative edits or disrupting the editing system. Using a new program called ACORNS (Assemble Clusters Of Related Nucleotide Sequence), the gRNAs were grouped into clusters based on sequence homology. This allowed us to determine which unidentified gRNAs were related to previously identified gRNAs. This analysis showed that more than half of the unidentified gRNAs were not related to any gRNA of known function, suggesting that many more alternative edits are waiting to be discovered. In order to analyze the impact of the gRNAs that were related to previously identified gRNAs, another new program, GUIDE (gRNA Uridine Insertion/Deletion Editor), was created. This program is able to analyze the functionality of gRNA clusters generated by ACORNS, by simulating the editing process. Combining this data with the mRNA transcriptomes

previously generated, we conducted a detailed analysis of each population of gRNAs capable of editing these three genes. This analysis revealed a surprisingly high tolerance for mismatches and gaps in mRNA/gRNA alignments in the editing system, most notably in the editing of the essential RPS12 [100].

This project found that not only is alternative editing present in *T. brucei*, but that it is pervasive, and the system, as a whole, is surprisingly robust. We propose the hypothesis that the RNA editing system does in fact promote the genetic robustness of *T. brucei* through the facilitation of dual-coding genes, as well as the introduction of alternative edits that increase protein diversity and allow the editing system to continue to evolve.

CHAPTER 2: ANALYSIS OF THE *TRYPANOSOMA BRUCEI* EATRO 164 BLOODSTREAM GUIDE RNA TRANSCRIPTOME

Abstract

The mitochondrial genome of *Trypanosoma brucei* contains many cryptogenes that must be extensively edited following transcription. The RNA editing process is directed by guide RNAs (gRNAs) that encode the information for the specific insertion and deletion of uridylates required to generate translatable mRNAs. We have deep sequenced the gRNA transcriptome from the bloodstream form of the EATRO 164 cell line. Using conventionally accepted fully edited mRNA sequences, ~1 million gRNAs were identified. In contrast, over 3 million reads were identified in our insect stage gRNA transcriptome. A comparison of the two life cycle transcriptomes show an overall ratio of procyclic to bloodstream gRNA reads of 3.5:1. This ratio varies significantly by gene and by gRNA populations within genes. The variation in the abundance of the initiating gRNAs for each gene, however, displays a trend that correlates with the developmental pattern of edited gene expression. A comparison of related major classes from each transcriptome revealed a median value of ten single nucleotide variations per gRNA. Nucleotide variations were much less likely to occur in the consecutive Watson-Crick anchor region, indicating a very strong bias against G:U base pairs in this region. This work indicates that gRNAs are expressed during both life cycle stages, and that differential editing patterns observed for the different mitochondrial mRNA transcripts are not due to the presence or absence of gRNAs. However, the abundance of certain gRNAs may be important in the developmental regulation of RNA editing.

Author Summary

Trypanosoma brucei is the causative agent of African sleeping sickness, a disease that threatens millions of people in sub-Saharan Africa. During its life cycle, *Trypanosoma brucei* lives in either its mammalian host or its insect vector. These environments are very different, and the transition between these environments is accompanied by changes in parasite energy metabolism, including distinct changes in mitochondrial gene expression. In trypanosomes, mitochondrial gene expression involves a unique RNA editing process, where U-residues are inserted or deleted to generate the mRNA's protein code. The editing process is directed by a set of small RNAs called guide RNAs. Our lab has previously deep sequenced the gRNA transcriptome of the insect stage of *T. brucei*. In this paper, we present the gRNA transcriptome of the bloodstream stage. Our comparison of these two transcriptomes indicates that most gRNAs are present in both life cycle stages, even though utilization of the gRNAs differs greatly during the two life-cycle stages. These data provide unique insight into how RNA systems may allow for rapid adaptation to different environments and energy utilization requirements.

Introduction

The life cycle of *Trypanosoma brucei* involves two distinct environments, the animal host and the insect vector. These environments are distinct in temperature and nutrient composition, providing a unique challenge to *T. brucei* as it cycles between hosts. In the bloodstream, trypanosomes exist in two forms, the actively dividing slender form and the non-dividing stumpy form. The slender form is optimized to utilize its glucose rich environment, using glycolysis to generate energy [14]. The stumpy form appears to be transitional, activating

mitochondrial genes in preparation for uptake in a blood meal by its tsetse fly vector and subsequent transfer to a harsher environment [14]. Once inside the tsetse fly, the parasite utilizes proline to drive oxidative phosphorylation and ATP production in the mitochondrion [19]. While the activity of the mitochondrion is relatively low during the bloodstream stage (BS), expression of the mitochondrial genome is still essential [17,18]. In *T. brucei*, the mitochondrial genome consists of two types of DNA molecules, maxicircles and minicircles. Maxicircles are 22kb circular DNA that contain the genes for two ribosomal RNAs, 12S and 9S, and eighteen mRNA genes [5]. While some of the protein-coding genes do not require RNA editing prior to translation, most require extensive editing before they can be translated [6,7]. This process involves the insertion of hundreds of uridylates (U)s and less frequently deletion of Us, often doubling the size of the transcript. The sequence changes are guided by small complementary RNA molecules (the guide RNAs) that are encoded on the minicircles [8]. Minicircles make up the bulk of the kinetoplastid network (anywhere from 5,000–10,000 present in each network) with each minicircle encoding 3–5 gRNAs. In *T. brucei*, there are more than 200 different minicircle sequence classes (~1200 gRNAs) [8].

Distinct differences in mitochondrial transcript abundance, polyadenylation and the extent of RNA editing are observed during the complex life cycle (Table 1). The pattern of differential RNA editing observed is especially interesting. For example, the cytochrome b (CYb) and cytochrome oxidase II (COII) mRNAs are edited during the insect stage, but are primarily unedited in bloodstream forms [24,25]. In contrast, editing of the NADH dehydrogenase subunit transcripts (ND3, ND7, ND8 and ND9) and editing of the ribosomal protein subunit 12 transcript (RPS12) appears to occur preferentially in bloodstream forms [5,26–30]. Other

transcripts, cytochrome oxidase III (COIII) and ATPase subunit 6 (A6) are edited in both life cycle stages [31,32]. Early studies using both Northern blot and primer extension analyses on a limited number of gRNAs indicate that gRNAs are present in both insect and bloodstream forms, suggesting that the regulation of RNA editing is not at the level of gRNA availability [28,50,51]. Our lab has previously published deep sequencing results of the gRNA transcriptome of the *T. brucei* EATRO 164 procyclic form [9]. Here we present the deep sequencing data for the gRNA transcriptome of a bloodstream form of EATRO 164. A total of 211 populations of gRNAs were identified. We define a population as a group of gRNAs that may vary in sequence, but direct the editing of the same or near same region of the mRNA. Because kinetoplastid RNA editing allows G:U base pairing, most populations contain multiple sequence classes that can guide the generation of the same mRNA sequence. While the number of populations identified was similar to the number identified in the procyclic gRNA transcriptome (214 populations), the total number of gRNAs identified was much reduced and the coverage was less complete; full complements of gRNAs were only identified for COIII and CYb. In spite of the reduced number of gRNAs, an interesting correlation was found that suggests a relationship between the relative abundance of initiating gRNAs between stages and the developmental pattern of mRNA editing.

Materials and Methods

Parasites, isolation of mitochondria and RNA extraction

T. brucei brucei clone IsTar from stock EATRO 164 were grown in rats and isolated as previously described [101]. Bloodstream forms were virtually all long-slender forms isolated after 4 days of infection. Parasites were used immediately for isolation of mitochondria using

differential centrifugation as previously described or stored frozen at -80°C until RNA extraction [9]. Both total RNA from whole parasites and mitochondrial RNA (mtRNA) from purified mitochondria were isolated by the acid guanidinium-phenol-chloroform method [102].

Ethics statement

Rats were raised according to the animal husbandry guidelines established by Michigan State University. All vertebrate animal use procedures were approved by MSU's Institutional Animal Care and Use Committee (Application 03/11-051-00). MSU has filed with the Office of Laboratory Animal Welfare (OLAW) an assurance document that commits the university to compliance with NIH policy and the Guide for the Care and Use of laboratory Animals.

Library preparation and Illumina sequencing

Samples of mtRNA and total RNA were both treated with DNAse RQI and size fractioned on a polyacrylamide gel as previously described [20]. Guide RNAs were extracted from the gel and prepped for sequencing using the Illumina 'Small RNA' protocol as previously described [9]. Libraries from both mtRNA and total RNA samples were deep sequenced on Illumina GAIIx. Reads were then processed and trimmed as previously described [9]. Data with two or more Ns, shorter than 20nts after trimming or with an overall mean Q-score < 25 were discarded. Redundant reads were then removed, while maintaining the number of redundant reads and reads containing fewer than 4 consecutive Ts were removed.

Identification of gRNAs

To identify gRNAs, each transcript read was aligned to the conventionally edited mRNAs based on known base pairing rules (canonical Watson-Crick base pairs and the G-U base pair). In the initial screen, no gaps were allowed in the alignment, allowing the formulation of the

gRNA-mRNA alignment as an extended longest common substring (LCS) problem as previously described [25]. Matched gRNAs were then scored (two points for G:C and A:U base pairs and one point for G:U base pairs). gRNAs with scores >45 were identified as guiding a specific region based on the identified mRNA fully edited sequence. Additional searches with reduced stringency (scores >30) were performed on regions with low gRNA coverage. The matched gRNAs were sorted into populations based on their guiding positions, and the populations analyzed and sorted into major sequence classes.

Results

Much of the initial characterization of RNA editing in *T. brucei* was done using the EATRO 164 strain. These experiments suggested that RNA editing was developmentally regulated in that certain genes were shown to be more fully edited in some stages than others (Table 1) [24–32,46]. It was also reported that the developmental regulation was not controlled by gRNA availability, as gRNAs were found in both life cycle stages [28,50,51]. In these early studies, however, only a small number of gRNAs were investigated. In this study, we used deep sequencing to compare the gRNA transcriptomes of a bloodstream form to a procyclic form of *T. brucei* EATRO 164. The EATRO 164 strain was isolated in 1960 from *Alcephalus lichtensteini* and maintained in the lab of Dr. K. Vickerman until being obtained by Dr. Stuart in 1966 [103]. Dr. Stuart derived the procyclic form from the Bloodstream culture in 1979 [103]. Both cell lines have been maintained in separate culture since that time.

Trypanosomes from the EATRO 164 strain were grown in Wistar rats to a parasitemia of 1–2 x 109 trypanosomes per mL and isolated using DEAE cellulose columns. Mitochondria and gRNAs were purified as previously described [9]. Libraries were generated using gRNAs isolated

from whole cell RNA and gRNAs isolated from mitochondrial RNA. Both bloodstream gRNA libraries were searched using conventionally accepted fully edited mRNA sequences, and a total of 1,024,604 gRNA reads were identified. Surprisingly, the library generated using gRNAs isolated from whole cell RNA had more than twice as many identified gRNA reads as the data generated using gRNAs isolated from mitochondrial RNA. To insure sufficient abundance and gRNA coverage, the two data sets were combined for the analyses presented here. In contrast, over 3 million gRNA reads were identified in our procyclic gRNA transcriptome generated from gRNAs isolated from mitochondrial RNA. Of the 1,024,604 reads identified from the bloodstream transcriptomes, 982,450 reads were sorted into major sequence classes.

The overall ratio of identified procyclic gRNA reads to BS gRNA reads was 3.5:1. This ratio varies significantly by gene (Table 2), and by populations within genes (APPENDIX A) and, except for the initiating gRNA, no apparent trend relating gRNA abundance and developmental editing pattern was observed. Interestingly, for the initiating gRNA, mRNAs that are fully edited in the procyclic stage only, or are fully edited in both life cycle stages had initiating gRNAs with more reads in the procyclic data set (Figure 1) [31,32,40,46]. In contrast, mRNAs that are only fully edited or are more abundant in the BS, had more initiating gRNAs reads in the BS data set (Figure 1) [26–30].

Because the identified gRNAs from the BS cells were less abundant, the rule used to identify major gRNA sequence classes was relaxed. Instead of using a strict cut off for the minimum number of reads required, the cut off was assessed on a case-by-case basis. For example, if the total population only had 100 reads, a sequence class with only 10 reads would still be identified as a major sequence class. Once all major classes were identified, 657

Table 2. Number of gRNA transcripts in procyclic and bloodstream major classes and ratio of procyclic transcripts to bloodstream transcripts for each gene.

Gene	Bloodstream gRNA Reads	Procyclic gRNA Reads	Ratio of PC to BS Reads
A6	41,628	266,532	6.40
COIII	371,139	948,845	2.56
CR3	13,316	236,808	17.78
CR4	25,753	51,979	2.02
CYb	11,022	31,622	2.87
MurfII	157	2,605	16.59
ND3	13,567	75,739	5.58
ND7	291,927	702,061	2.40
ND8	112,868	584,639	5.18
ND9	83,924	72,027	0.86
RPS12	17,191	403,131	23.45
Total	982,492	3,375,988	3.44

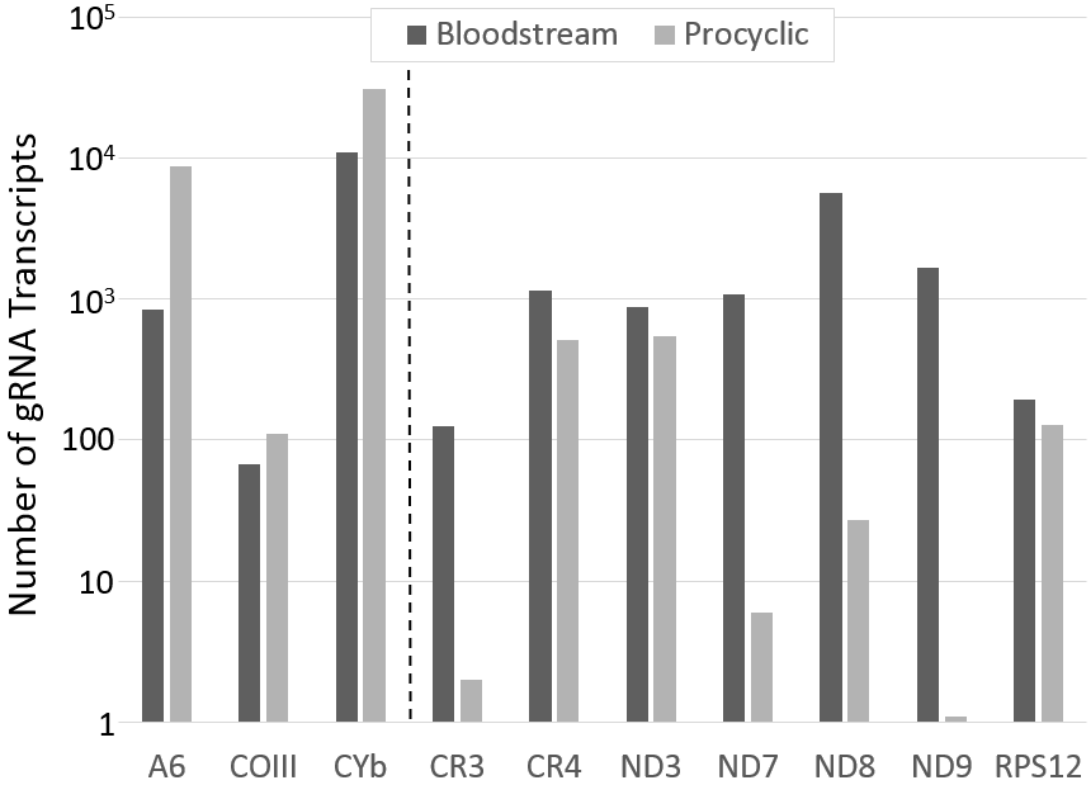


Figure 1. The abundance of the initiating gRNA of all edited mRNAs in each stage. mRNAs to the left of the dashed line are constitutively edited or are edited only in the procyclic stage [31,32,40,46]. mRNAs to the right of the dashed line are only fully edited or more abundant fully edited in the bloodstream stage [26–30].

sequence classes were identified that could be sorted into 211 populations (Table 3). Although the overall gRNA numbers were down in comparison to the procyclic data set, most of the populations found in that stage (214 gRNA populations) were also identified in the BS transcriptome. However, there were a number of populations that were unique to either the procyclic or BS stage.

Table 3. Summary of the gRNA data coverage for each gene.

Gene	Populations		Unique Populations		Average* gRNA Overlap (nts)		Gaps		Weak Overlaps	
	BS	PC	BS	PC	BS	PC	BS	PC	BS	PC
A6	29	28	1	0	18	20	1	0	0	0
COIII	42	39	4	1	19	22	0	0	0	0
CR3	9	9	0	0	19	14	1	0	0	2
CR4	16	18	0	2	17	18	2	0	0	0
CYb	2	2	0	0	12	14	0	0	0	0
MurfII	1	1	0	0	N.A.	N.A.	1	1	0	0
ND3	12	12	0	0	15	15	1	1	0	0
ND7	45	48	2	5	17	21	7	2	4	1
ND8	20	21	2	3	17	21	2	1	1	0
ND9	24	23	1	0	16	16	1	0	0	2
RPS12	11	13	0	2	17	21	2	0	1	0
Total	211	213	10	13	17	19	18	5	6	5

[a]The average gRNA overlaps were determined excluding any regions where neighboring gRNAs shared no overlap.

Surprisingly, when the bloodstream and procyclic data sets were compared, only 37 identical major sequence classes were found in both. However, distinctly related sequence classes could be identified when comparing the BS and procyclic populations. Comparing the related major classes from each transcriptome (BS vs procyclic) revealed a median value of ten single nucleotide variations per gRNA. Interestingly, nt variations were much less likely to occur in the consecutive Watson-Crick anchor region of the gRNA than in the rest of the gRNA indicating a very strong bias against G:U base pairs in this region (Figure 2).

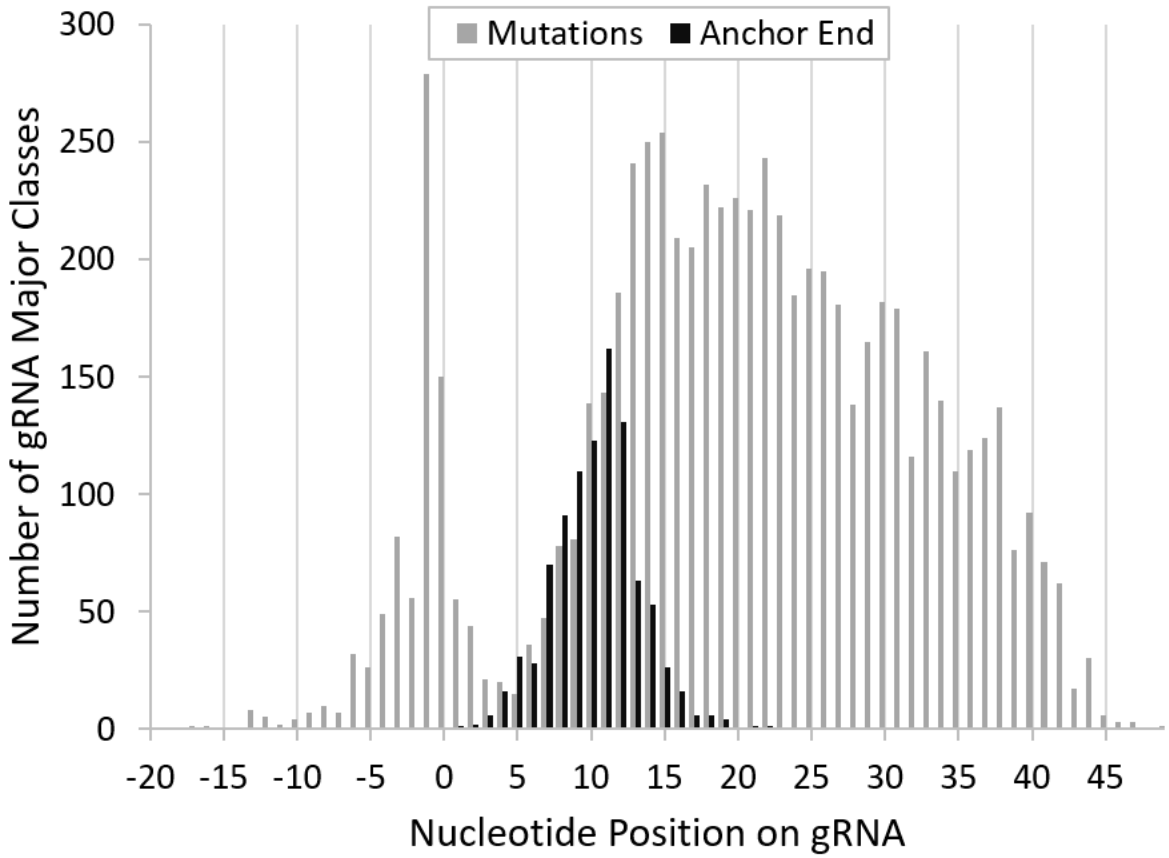


Figure 2. The frequency of nt variations versus nucleotide position in the gRNA. Gray bars indicate the number of gRNA sequence classes with an identified nt difference between related procyclic and bloodstream gRNAs. Nucleotide numbering for each gRNA was normalized by setting the start of the Watson-Crick anchor region to zero. Black bars indicate the number of gRNA sequence classes whose contiguous Watson-Crick anchors end at that position (start of Watson-Crick = zero, so this is an indication of the length of the contiguous Watson-Crick region).

The Watson-Crick anchors (defined as the number of consecutive nts in the 5' region with only G:C and A:U base pairs) had a median length of eleven nucleotides and anchor length did not vary between the two forms. The vast majority of major classes of gRNAs had consecutive Watson-Crick anchors greater than seven nts long (92.5%). In addition, most gRNAs with Watson-Crick anchors shorter than eight nts were not an abundant major class for their respective populations. Consistent with observations made from the procyclic data set, most gRNAs had zero non-base pairing nucleotides 5' to the poly-uridine tail and 4 to 6 non-base pairing nucleotides 5' to the anchor region (Figure 3). Also consistent with procyclic data, most

of the gRNAs (59%) had 38 to 48 nts of complementarity (including anchor regions) with their respective mRNAs (Figure 4). Transcription start sites also did not vary, as preference for an RYAYA start site was observed (Table 4).

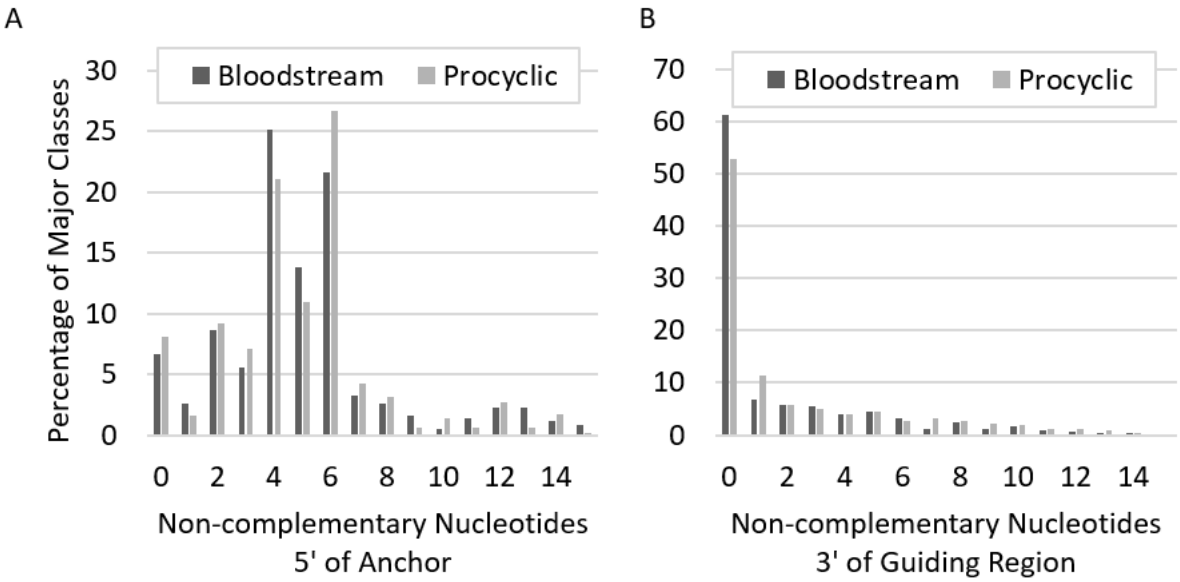


Figure 3. Comparing the number of non-complementary nucleotides 5' of the anchoring region (A) or 3' of the guiding region (excluding the U-tail) (B) in procyclic and bloodstream gRNAs.

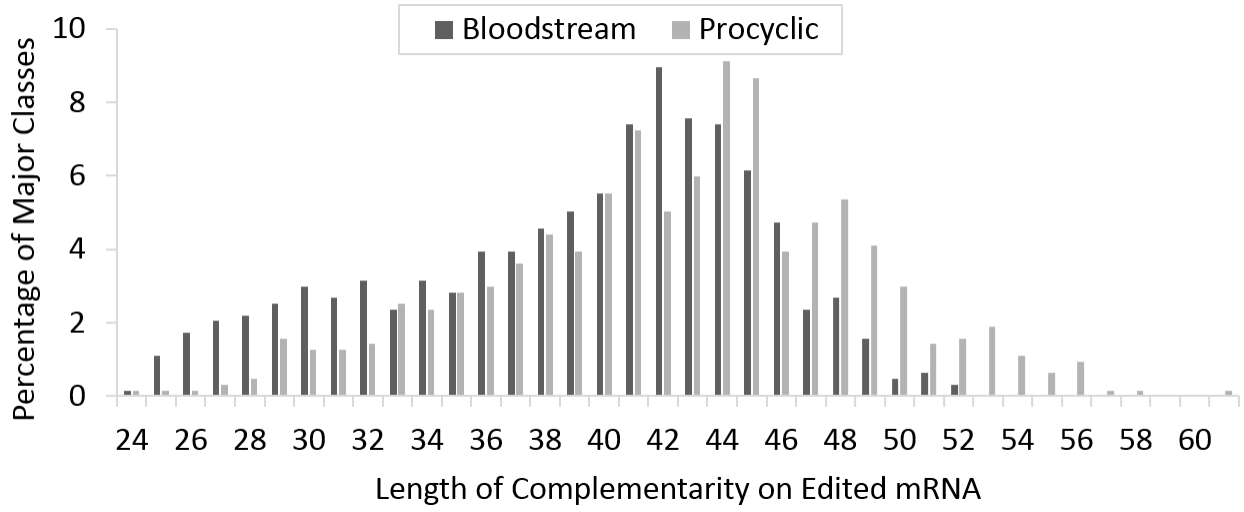


Figure 4. Length of gRNA complementarity (including anchors) to fully edited mRNAs for both bloodstream and procyclic gRNAs.

Table 4. Most common gRNA transcription start sites in procyclic and bloodstream data.

Initiating Sequence	Stage	Percentage of Sequence Classes	Percentage of Transcripts
ATATAT	Bloodstream	32.20%	33.60%
	Procyclic	35.20%	37.40%
ATATAA	Bloodstream	20.00%	17.90%
	Procyclic	21.10%	24.70%
AAAAAA	Bloodstream	3.60%	1.60%
	Procyclic	4.60%	1.30%
ATATAC	Bloodstream	3.80%	1.50%
	Procyclic	4.30%	3.80%
ATACAA	Bloodstream	2.40%	1.30%
	Procyclic	2.80%	1.70%
ATATTA	Bloodstream	4.90%	13.40%
	Procyclic	2.60%	0.90%
ATATAG	Bloodstream	2.20%	0.70%
	Procyclic	2.60%	7.60%
ATAAAT	Bloodstream	2.50%	1.00%
	Procyclic	2.60%	0.70%
ATACAT	Bloodstream	2.70%	3.20%
	Procyclic	2.20%	2.20%
ATAAAA	Bloodstream	1.70%	0.20%
	Procyclic	2.10%	1.10%

Coverage and gaps

In order to determine if the BS gRNA transcriptome contained a full complement of guide RNAs, the gRNA populations were aligned to the fully edited mRNAs (APPENDIX B). We note, that for an mRNA to be fully edited, not only must all editing sites on the mRNA be covered by a gRNA, the downstream gRNA must generate the anchor binding site for the subsequent gRNA. Therefore, adjacent gRNAs must overlap. Overall, there was an average of 17 nts of overlap between adjacent gRNAs, with the average overlap varying slightly by gene (Table 3). As the median Watson-Crick Anchor is 11 nts, in most cases, the overlap extends beyond the Watson-Crick anchor of the subsequent gRNA. However, we did observe a number

of regions where the overlap is minimal. Currently, there is no data that stipulates the minimum anchor needed for efficient editing. However, we postulate that similar to microRNAs, for an anchoring sequence to be sufficiently specific, it should be at least six nucleotides [104]. Indeed, when examining the overlaps between most gRNAs, there are only ten (four procyclic and six BS) that are less than six nucleotides (Figure 5).

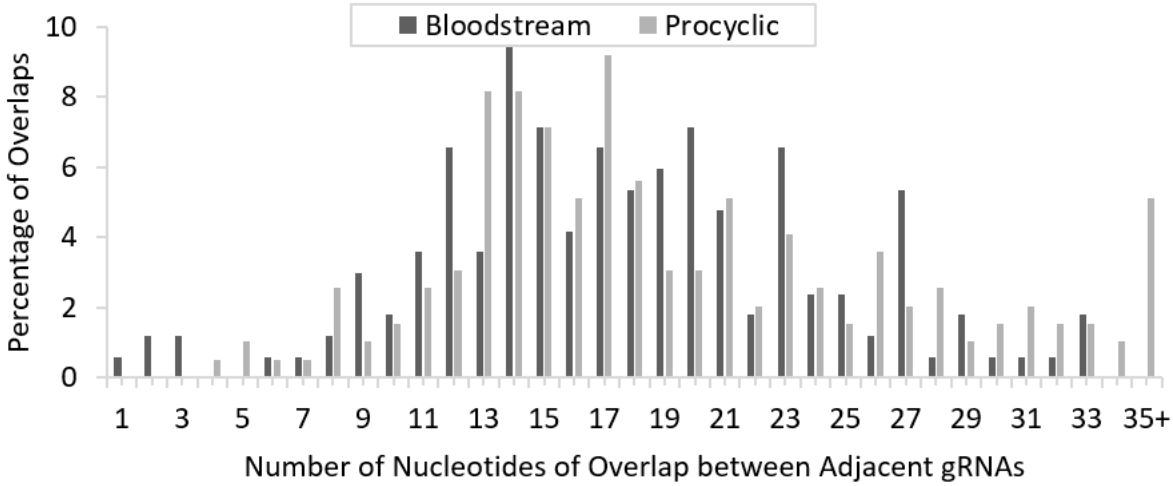


Figure 5. The percentage of different nucleotide overlaps found between adjacent gRNAs. gRNAs were aligned to their fully edited mRNA sequence and the number of mRNA nts with complementarity to both adjacent gRNAs determined.

We therefore used six nucleotides as a cut off to identify regions with potential missing guide RNAs for both life cycle stage transcriptomes. In contrast to the procyclic data, where full complements of gRNAs were identified for five of the mRNA transcripts (A6, COIII, CR4, CYb, and RPS12), in the BS transcriptome, a full complement of gRNAs was only identified for COIII and CYb. Overall, there are 12 edited regions where no gRNAs were identified, and five regions with weak gRNA overlaps in the BS data (Table 5). Of these 17 regions, seven belong to ND7 alone. Interestingly, nine of the 17 missing populations are in very low abundance in the procyclic data, having 100 or fewer reads. Because the number of reads in the BS data is ~3.5 fold less abundant, this could account for some of these regions of poor coverage. There are six

regions that lack gRNA coverage in both data sets. These are found in CR3, MurfII, ND3 and ND7 (Table 5). Interestingly, three of these regions are close to the 3' end of their respective genes. Regions of weak overlap (ND9(238–242), ND9(609–612)) and regions without gRNA coverage (CR3(278–292), ND8(541–553)) that are unique to the procyclic transcriptome were also observed. Interestingly, the regions of poor procyclic coverage are found in CR3, ND8 and ND9, all transcripts that are preferentially edited in the BS form [5,28,29,47].

Table 5. Identified gaps or weak overlaps (less than 6 nucleotides) between populations of gRNAs observed in both data sets.

Gene	Stage Missing Coverage	Range	Gap or Overlap	Abundance of Equivalent gRNA
A6	Bloodstream	669-670	2 nt Gap	39,063
CR3	BS/P[a]	233/226-230	1 nt G/5 nt O	Missing in Both Stages
CR3	Procyclic	278-292	15 nt Gap	125
CR4	Bloodstream	143-165	23 nt Gap	7,175
CR4	Bloodstream	302-306	5 nt Gap	643
MurfII	BS/P	80-85	6 nt Gap	Missing in Both Stages
ND3	BS/P	389-401	13 nt Gap	Missing in Both Stages
ND7	BS/P	92-94	3 nt Gap	Missing in Both Stages
ND7	Bloodstream	95-120	26 nt Gap	1
ND7	Bloodstream	292-293	2 nt Overlap	3,259
ND7	Bloodstream	325-326	2 nt Gap	888
ND7	BS/P	485-486	0 nt Overlap	Missing in Both Stages
ND7	Bloodstream	1000-1000	1 nt Overlap	101
ND7	Bloodstream	1079-1085	7 nt Gap	44
ND7	BS/P	1086/1086-1088	1 nt G/3 nt G	Missing in Both Stages
ND7	Bloodstream	1225-1232	8 nt Gap	123
ND7	Bloodstream	1269-1270	2 nt Overlap	63
ND8	Bloodstream	54-56	3 nt Overlap	1
ND8	Bloodstream	153-159	7 nt Gap	4
ND8	Bloodstream	386-389	4 nt Gap	2
ND8	Procyclic	541-553	13 nt Gap	413
ND9	Procyclic	238-242	5 nt Overlap	652
ND9	Bloodstream	340-342	3 nt Gap	36
ND9	Procyclic	609-612	4 nt Overlap	7
RPS12	Bloodstream	122-132	11 nt Gap	3
RPS12	Bloodstream	156-158	3 nt Overlap	62
RPS12	Bloodstream	337-349	13 nt Gap	128

Table 6. Summary of populations found in both data sets that have more reads in the bloodstream data set than in the procyclic data set.

Gene	PC and BS Shared Populations	Populations more abundant in BS	Percentage of populations more abundant in BS
A6	28	6	21%
COIII	38	12	32%
CR3	9	5	56%
CR4	16	10	63%
CYb	2	0	0%
MurfII	1	0	0%
ND3	12	6	50%
ND7	43	20	47%
ND8	18	8	44%
ND9	23	16	70%
RPS12	11	4	36%
Total	201	87	43%

Gene specific gRNA characteristics

ATPase 6. In the BS gRNA transcriptome, a total of 29 gRNA populations containing 86 different major sequence classes were identified that could guide the editing of A6 (Table 3; APPENDIX C Part A). One population was identified that was unique to the BS transcriptome (gA6(281-329)). The gRNAs bordering this population share extensive overlap, so its absence in the procyclic transcriptome would not impact the editing process (APPENDIX C Part A). We note that two of the gRNAs identified have single nucleotide mismatches. The bloodstream gA6(640–668) has an identified mismatch (C:U) that disrupts the complementarity of the gA6(640–668) population (APPENDIX B Part A). The second mismatched gRNA (gA6(520–533)) would introduce a frameshift. Excluding these two mismatched regions, there is complete coverage of ATPase 6. In contrast to the procyclic data, where the conventional initiating gRNA and the gRNA immediately following it were extremely rare, both of these gRNAs, gA6(773–822), previously identified as gA6-14 and gA6(745–789), were fairly abundant, each having hundreds

of reads. The alternative initiating gRNA identified in the procyclic data set was not found. This finding is similar to that found in the *T. brucei* Lister strain 427 where authors identified alternative initiating gRNAs not found in the EATRO 164 procyclic gRNA transcriptome [105].

Another disparity between the two life cycle data sets was found when comparing the abundance of gRNAs implicated in a potential alternative edit. In the procyclic gRNA transcriptome, a gRNA was identified that would guide the insertion of 11 U-residues instead of the needed 12 between G555 and A568 [9]. This gRNA (pA6(557–593)) was 25-fold more abundant than the conventional gRNA (pA6(549–593)). In the BS data set however, more than 400 reads of the 12U gRNA were identified and only one read was found that would encode the alternative 11U edit. Surprisingly, while G555-A568 would be correctly edited (insertion of 12Us), the next editing site (A549-G555) is edited by bsA6(520–553), the gRNA that introduces the 1 nt frameshift. This frameshift would generate a predicted protein with nearly the same amino acid sequence as the procyclic 11U frameshift edit (two amino acid changes) (Figure 6).

```
160       170       180       190       200       210       220
Conv-FFDFFYFIFVFFFYGVFCYWFILFIFVFCFCLLFYVFLYLLDLFAAILQLFIFCNMILQLIMDFLLFLLFV
11U -FFDFFYFIFVFFFMVFFVIDLFYLFLCFVFVYYFMCFYICWIYLPPYYSYLFFVMWFCSW
4U  -FFDFFYFILFFFFMVFFVIDLFYLFLCFVFVYYFMCFYICWIYLPPYYSYLFFVMWFCSW
```

Figure 6. Alignment of conventional ATPase 6 protein sequence to hypothetical proteins generated by the 11U alternative edited mRNA and the 4U alternatively edited mRNA. Double underlined residues show where the alternative sequences differ from the conventional sequence. The shaded residues in the 4U sequence show where it differs from the 11U sequence.

Cytochrome oxidase subunit III. Forty-two gRNA populations, guiding the editing of COIII were identified in the BS transcriptome; three more than in the procyclic data set (APPENDIX C Part B). This disparity is caused by the presence of several unique populations. While the procyclic data set contained one unique population, the BS data contained four gRNA

populations not previously identified. Of these four unique populations, three of them are required for full overlapping coverage in the bloodstream. They are not however, required for full coverage in the procyclic stage. These three unique gRNA populations all span relatively small regions of weak overlap (Figure 7, APPENDIX B Part B).

Figure 7. Editing sites 420–489 of COIII aligned with the gRNAs identified for that region in the procyclic (grey) and bloodstream (black) data sets. The gRNA covering 443–474 was only found in the bloodstream data set.

An alternative edit of COIII has been described, involving distinct edits at two adjacent sites that links the open reading frame of the edited 3' end to an ORF found in the 5' pre-edited sequence [67]. The previously identified alternative gRNA that can generate the needed editing events was not found in either the BS or procyclic transcriptomes.

C-rich regions 3 and 4. In the BS data set, nine populations and 34 major sequence classes were identified that direct the editing of the CR3 transcript. The coverage of edited CR3 is nearly complete in the bloodstream data set with only a one nt gap in coverage (editing site 233) (APPENDIX B Part C). This is in contrast to the procyclic transcriptome, where gRNAs that matched the published sequence downstream of nt 196 were very rare (<10 copies) and no gRNAs were identified that could direct editing near the 3' end (nucleotides 275–292).

A full consensus sequence for edited CR4 has only been found in BS *T. brucei* [48]. Using this sequence, 16 gRNA populations, containing 62 major sequence classes were identified in

the BS transcriptome (APPENDIX C Part D). In contrast to the procyclic data, where a full complement of gRNAs were identified, there are two gaps in the BS coverage (Table 5).

Cytochrome b and maxicircle unidentified reading frame II. RNA editing in the Cytochrome b (CYb) transcript is limited to the 5' end and two gRNA populations are sufficient to guide the small number of edits needed to render the CYb transcript functional. Both populations were observed in both data sets, with a total of 6 major classes. Interestingly, in both data sets, the initiating gRNA is significantly more abundant than the second gRNA, being approximately 30 fold more abundant in the procyclic data set and approximately 200 fold more abundant in the bloodstream data set (APPENDIX C Part E). This is in contrast to most of the other transcripts where the initiating gRNAs are not very abundant. In addition, almost all of the CYb gRNA major classes have an A-run transcription start site, deviating from the common RYAYA initiation site pattern.

Editing in MurfII is also limited to the 5' end and requires only two gRNAs. One of these gRNAs (gMurfII(30–79)) is encoded on the maxicircle [106]. While this gRNA was observed in both data sets, the gRNAs identified were not identical. A purine-purine transition near the 3' end of the gRNA differentiates the procyclic and BS forms (APPENDICES B and C Parts F). An initiating gRNA is needed to generate the 3' most edits that create the anchor sequence for gMurfII (30–79). This gRNA was not found in either data set, despite additional searches with reduced search stringency.

NADH dehydrogenase subunits 3, 7, 8, and 9. In the initial characterization of RNA editing in *T. brucei* EATRO 164, fully edited ND subunit transcripts were only found in RNA isolated from the BS stage. We were therefore surprised to find that fewer ND gRNA

populations were identified in the BS transcriptome and a full complement of gRNAs was not identified for any of the ND subunits. The most complete coverage was found for ND3 and ND9. For ND3, the BS data set contained twelve populations and 41 major classes of gRNAs. One gap in coverage was observed, from 389–401. This region overlaps a region that has no clear consensus sequence, 375–395 [26]. ND9 is the only gene in this study whose bloodstream gRNA reads outnumber the procyclic gRNA reads identified (Table 2). Twenty-four bloodstream gRNA populations were identified with all edited nucleotides covered if gRNAs with a single base pair mismatch are taken into account (APPENDIX B Part J).

While 45 gRNA populations were identified for ND7 in the BS data set, the gRNA coverage was significantly worse when compared to the identified procyclic gRNAs (Table 5). Despite the poor coverage, two unique gRNA populations (bs gRNA (772–816) and (1128–1182)) were identified (APPENDICES B and C Parts H). ND8 also had poor gRNA coverage (Table 5). Interestingly, there are several populations in ND8 that contain highly abundant gRNA sequence classes with mismatches that shorten the complementarity of the gRNA. These usually have a single mismatch in the gRNA that would otherwise guide conventional editing (APPENDIX C Part I).

Ribosomal protein S12. The BS data set contained 11 populations and 26 major sequence classes that direct editing of RPS12 (Table 3). While the procyclic transcriptome contained a full complement of gRNAs, the BS RPS12 data contains one gap in coverage and one region of poor overlap, (Table 5). This was surprising, as RPS12 has been shown to be essential in both life cycle stages [100,107]. The region of the mRNA with poor coverage has a high percentage of C residues and gRNAs covering this region may utilize C:A base pairs. If this

is the case, some classes of gRNAs may not have been detected, as the program used to search for gRNAs does not allow for C:A base pairs (APPENDIX B Part K).

Discussion

This is the first comprehensive characterization of the mitochondrial gRNA transcriptome from the bloodstream stage of *Trypanosoma brucei brucei*. As we have previously characterized the insect stage gRNA transcriptome, these data allow the comparison of gRNA characteristics across the two main life cycle stages [9]. In the EATRO 164 BS gRNA transcriptome, gRNAs for every edited gene were identified. Interestingly, while the number of populations identified in this data set was only slightly lower than that reported in the procyclic data set, the total number of gRNA transcript reads identified was considerably lower despite the fact that multiple transcriptome libraries were combined. While this may be a reflection of the down regulation of mitochondrial transcription in the bloodstream stage (see Table 1), it is impossible to rule out technical problems in the generation and sequencing of the libraries. It has been previously reported that gRNA presence did not correlate with developmental RNA editing patterns in *T. brucei* and our data does not challenge this [50,51]. The data did however, show an interesting trend in the abundance of the initiating gRNAs as relates to their developmental editing patterns. It may be that the abundance of the initiating gRNAs is regulated in order to control editing of their target mRNAs. However, we cannot rule out the possibility that not all of the populations of initiating gRNAs were identified. For the pan-edited mRNAs, the initiating gRNAs direct sequence changes that are often downstream of the stop codon. Sequence changes in this region would be tolerated, as long as the anchor sequence for the next gRNA is maintained. This type of mutation was observed in the 3' end of ATPase 6 [19].

In addition, characterization of the initiating gRNAs in the Lister 427 *T. brucei* cell line identified several gRNAs that would direct an alternative editing pattern, suggesting a high tolerance for sequence changes near the mRNA 3' ends. [105].

As expected, general gRNA characteristics are conserved across the two life-cycle stages. Populations retain the general location of their anchors, there is relatively little shift in the location of populations, and the lengths of complementarity are very similar. We did observe that considerable nucleotide variations were found in the guiding regions of the gRNAs from the different life cycle strains of the EATRO 164 cells. This particular cell line dates back to 1960 when the BS form was originally acquired [103]. Procyclic cells were derived from the BS stock in 1979 and the two cell lines maintained separately since that date [103]. Mixed trypanosome genotypes are detected frequently in field isolates from both tsetse flies and mammals and it may be that separation into different culture conditions allowed different genotypes to predominate in each life cycle strain [22,108,109]. Because gRNAs utilize both canonical (Watson-Crick) as well as G:U base-pairing to direct the change in sequence, most transition mutations in the gRNA, would not lead to changes in the mRNA sequence and would not be selected against [33]. We do note however, that a very strong bias against A to G transitions is observed in the anchor regions of the gRNAs. This suggests that transition mutations in this region are not tolerated. This suggests that the editing machinery recognizes and selects for a conventional base-paired double helix in the initial gRNA/mRNA pairing. The ability to discriminate against G:U base-pairs in the initial interaction would greatly increase the accuracy of the gRNA targeting event. Considering the sequential nature of the overall editing process, this would be very advantageous.

Coverage

Surprisingly, complete gRNA coverage was observed only for the pan-edited COIII and for CYb, where editing is limited to the 5' end. The identification of the CYb gRNAs was expected, as it has been previously reported that the gRNAs are present in both life cycle stages even though editing of CYb is limited to the procyclic stage [8,24]. The full coverage of COIII was also not surprising, as COIII was shown to be fully edited and equally abundant in both stages [32]. However, we expected to see complete coverage of ATPase 6 and RPS12 as both of these transcripts have been shown to be essential in both life cycle stages [17,100,107,110]. For ATPase 6, we did identify a total of 29 gRNA populations that do cover all of the editing sites. However, one of the gRNAs (bsA6(643–667)) has a single nucleotide mismatch (C:U) and one would introduce a frameshift (bsA6(520–553)). The C:U mismatch occurs near the middle of the gRNA, placing the C:U mismatch in a region that is unusually high in Gs and Cs (APPENDIX B Part A). It may be that the G:C base pairs immediately upstream of the mismatch stabilize the gRNA/mRNA interaction, allowing it to be tolerated. The frameshift gRNA is also interesting, as it occurs just upstream (1 editing site) of another site where we had previously observed a frameshift sequence anomaly. Both frameshifts (the BS 4U and the Procyclic 11U) generate a predicted protein with nearly the same amino acid sequence. As the frameshifts occur downstream of the highly conserved amino acid region involved in proton translocation [31], it may be that this different carboxyl terminus is tolerated.

Near full coverage is also observed for RPS12. For this transcript, one BS identified gRNA (bsRPS12(96–121)) has an A-nt insertion that disrupts the gRNA complementarity. Surprisingly, the other mRNA transcript found with near complete coverage was ND9 (one gRNA has a single

nt mismatch). All of the other mitochondrially encoded Complex I members did have substantial gaps in coverage. Currently, there is considerable debate on the necessity of Complex I subunits for either stage of the trypanosome life cycle. Studies using RNAi and knockout cell lines of nuclear-encoded members of Complex I have shown that the complex is unnecessary for survival in either life cycle stage [111,112]. However, the nuclear encoded Complex I member genes are maintained [42], and while we not did identify full coverage for the ND transcripts, a vast majority of the gRNAs were found in both life cycle stages.

This study used high-throughput sequencing to characterize the gRNA transcriptome during the bloodstream stage of the trypanosome life cycle. This work suggests that gRNAs are expressed during both life cycle stages, and that differential editing patterns observed for the different mitochondrial mRNA transcripts are not due to the presence or absence of gRNAs.

Accession Numbers

SAMN04302078, SAMN04302079, SAMN04302080, and SAMN04302081 NCBI's Sequence Read Archive.

Acknowledgments

The authors dedicate this work in memory of David Judah, MS, DVM. He was a wonderful colleague.

We also acknowledge the work of Joshua Foster, Mark Johnson, James Rauschendorfer, Heather Tyler, Callie Vivian, and Alexis Weber who were involved in sorting and identifying gRNAs, the MSU RTSF for their contribution in deep sequencing and Ken Stuart and Jason

Carnes at the Center for Infectious Disease Research for supplying the *T. brucei* strains used in this study.

CHAPTER 3: MITOCHONDRIAL DUAL-CODING GENES IN *TRYPANOSOMA BRUCEI*

Abstract

Trypanosoma brucei is transmitted between mammalian hosts by the tsetse fly. In the mammal, they are exclusively extracellular, continuously replicating within the bloodstream. During this stage, the mitochondrion lacks a functional electron transport chain (ETC). Successful transition to the fly, requires activation of the ETC and ATP synthesis via oxidative phosphorylation. This life cycle leads to a major problem: in the bloodstream, the mitochondrial genes are not under selection and are subject to genetic drift that endangers their integrity. Exacerbating this, *T. brucei* undergoes repeated population bottlenecks as they evade the host immune system that would create additional forces of genetic drift. These parasites possess several unique genetic features, including RNA editing of mitochondrial transcripts. RNA editing creates open reading frames by the guided insertion and deletion of U-residues within the mRNA. A major question in the field has been why this metabolically expensive system of RNA editing would evolve and persist. Here, we show that many of the edited mRNAs can alter the choice of start codon and the open reading frame by alternative editing of the 5' end. Analyses of mutational bias indicate that six of the mitochondrial genes may be dual-coding and that RNA editing allows access to both reading frames. We hypothesize that dual-coding genes can protect genetic information by essentially hiding a non-selected gene within one that remains under selection. Thus, the complex RNA editing system found in the mitochondria of

trypanosomes provides a unique molecular strategy to combat genetic drift in non-selective conditions.

Author Summary

In African trypanosomes, many of the mitochondrial mRNAs require extensive RNA editing before they can be translated. During this process, each edited transcript can undergo hundreds of cleavage/ligation events as U-residues are inserted or deleted to generate a translatable open reading frame. A major paradox has been why this incredibly metabolically expensive process would evolve and persist. In this work, we show that many of the mitochondrial genes in trypanosomes are dual-coding, utilizing different reading frames to potentially produce two very different proteins. Access to both reading frames is made possible by alternative editing of the 5' end of the transcript. We hypothesize that dual-coding genes may work to protect the mitochondrial genes from mutations during growth in the mammalian host, when many of the mitochondrial genes are not being used. Thus, the complex RNA editing system may be maintained because it provides a unique molecular strategy to combat genetic drift.

Introduction

Trypanosomes are one of the most successful parasites in existence, inhabiting an incredibly wide range of hosts [2,3]. The dixenous members cycle between two distinct hosts and can encounter different environments with distinct metabolic constraints. These parasites are unique in that they all possess glycosomes (where glycolysis occurs) as well as mitochondria [16]. The salivarian trypanosomes (e.g. *T. brucei*, *T. vivax*) are especially interesting, because

they are exclusively extracellular in their mammalian hosts, continuously replicating within the bloodstream over periods of months. During this stage of the life cycle, the mitochondrion is down-regulated, lacking both Krebs cycle enzymes and a functional electron transport chain (ETC) [21]. Successful transition to the fly vector, requires activation of the ETC and ATP synthesis via oxidative phosphorylation. This unique lifecycle leads to a major problem: when the mitochondrial genes are unused, they are not under selection, hence the integrity of these genes are threatened by genetic drift [75,113]. Exacerbating this, salivarian trypanosomes undergo a severe bottleneck as they transition through the tsetse fly and into the mammalian host, and then within the bloodstream, they undergo multiple bottlenecks at each antigenic switch, as they evade the host immune system [22]. Such bottlenecks create additional forces of genetic drift, where genes can be lost even if their deleterious fitness effect is considerable. These parasites possess several unique genetic features, including RNA editing of the mitochondrial transcripts. RNA editing creates open reading frames in “cryptogenes” by insertion and deletion of uridylate residues at specific sites within the mRNA. The U-insertions/deletions are directed by small guide RNAs (gRNA) and can repair frameshifts, generate start and stop codons and more than double the size of the transcript (for review see [4]). While the mRNA cryptogenes are encoded on maxicircles (25±50 copies per DNA network), the guide RNAs are encoded on thousands of 1 kb minicircles, encoding 3±5 gRNA genes each [8]. This effectively means that the genetic information for the edited mitochondrial mRNAs is dispersed between the mRNA cryptogenes on the maxicircles and the thousands of gRNA coding minicircles. The extensive editing of a single transcript can require more than 40 gRNAs and hundreds of editing events [9]. While the initial gRNA can interact with the 3' end of the

pre-edited transcript, all subsequent gRNAs anchor to edited sequence created by the preceding gRNA. Hence, editing proceeds from the 3' end to the 5' end of the mRNA transcript with the terminal gRNA (last one in the cascade) often creating the start codon needed for translation. This sequential dependence means that with even high accuracy rates for each gRNA, the overall fidelity of the process is astonishingly low. A major question in the field has been why this fragile and metabolically expensive system of RNA editing would evolve and persist.

Another level of complexity in the kinetoplastids RNA editing process was the detection of an alternative editing event that leads to the production of a functionally discrete protein isoform. Alternative editing of Cytochrome Oxidase III (COIII) is reported to generate a novel DNA-binding protein, AEP-1, that functions in mitochondrial DNA maintenance [67,68]. In this transcript, one alternative gRNA generates sequence changes at two sites that links an open reading frame (ORF) found in the pre-edited 5' end, to the 3' transmembrane domains found in the COIII edited ORF. This was the first indication, that one cryptogene could contain information for more than one protein. Here, we show that as many as six additional cryptogenes also encode for more than one protein. Analyses of the terminal gRNA populations indicate that gRNA sequence variants exist that can alter the choice of the start codon and the open reading frame by alternative editing of the 5' end of the mRNA. Mutational bias analyses indicate that six of the mitochondrial genes may be dual-coding, with RNA editing allowing access to both reading frames. Dual-coding genes are defined as a stretch of DNA containing overlapping open reading frames (ORFs) [85,86]. Of particular interest are dual-coding genes that contain two ORFs read in the same direction: a canonical protein (normally annotated as

protein coding in the literature) and an alternative ORF. Maintaining dual-coding genes is costly, as it constrains the flexibility of the amino acid composition of both proteins. Hence, it is thought that dual-coding genes can survive long evolutionary spans only if the overlap is advantageous to the organism [93]. We hypothesize that trypanosomes use dual-coding genes to protect genetic information by essentially hiding a non-selected (ETC) gene within one that remains under selection. Thus, the ability to access overlapping reading frames may be added to a growing list of gene protective strategies made possible by the complex RNA editing process [74,75,113].

Materials and Methods

Trypanosome growth

T. brucei procyclic clones from IsTAR (EATRO 164), TREU 667 and TREU 927 cell lines were grown in SDM79 at 27°C and harvested at a cell density of 1-3x10^7. The TREU 667 cell line was originally isolated from a bovine host in 1966 in Uganda [114]. The TREU 927 cell line was originally isolated from *Glossina pallidipes* in 1970 in Kenya [115]. The EATRO 164 strain was isolated in 1960 from *Alcephalus lictensteini* and maintained in the lab of Dr. K. Vickerman until being obtained by Dr. Ken Stuart in 1966 [103]. Dr. Stuart derived the procyclic form from the bloodstream form culture in 1979.

Guide RNA isolation, preparation, and sequencing

Mitochondrial mRNAs and gRNAs were isolated as previously described [9]. All RNAs were treated with Promega DNAse RQI. In order to isolate gRNAs from TREU 667 and TREU 927 cells, RNAs were size fractionated on a polyacrylamide gel as previously described [9]. Guide

RNAs were then extracted and prepped for sequencing using the Illumina Small RNA protocol [9]. Libraries from TREU 667 and TREU 927 were deep sequenced on the Illumina GAIIx; reads were processed and trimmed as previously described [9].

Messenger RNA preparation and sequencing

In order to isolate target mRNAs, isolated TREU 667 mitochondrial RNAs were reverse transcribed using the Applied Biosystems High Capacity cDNA Reverse Transcription Kit. CR3 cDNAs were amplified via PCR using the following primers (underlined portions are gene specific and non-underlined portions are tag regions used in deep sequencing reaction):

CR3DS5'NEV:ACACTGACGACATGGTTCTACAAGAAATATAAATATGTGTATG

CR3DS3'170:TACGGTAGCAGAGACTTGGTCTCAATAAACCCATATTAAATAAAAAACAAAAATCC

After amplification, the products were purified using the QIAquick PCR Purification Kit, and paired end Illumina deep sequencing was performed on the Illumina Miseq (2x 250 bp paired end run). Low quality results were removed using FaQCs, adapters were removed using Trimmomatic and PEAR was used to merge paired end reads. Finally, Fastx was used to compile identical reads while maintaining the number of redundant reads. CR3 edited transcripts were identified by comparing sequence downstream of the 5' never edited region to the edited CR3 sequence. Guide RNAs were identified by using the mRNA sequences as queries against our existing gRNA databases, as previously described [9].

Mutational frequency and editing conservation analysis

Mitochondrial pan-edited genes were categorized as potentially dual-coding based on identification of extended alternative reading frames and/or presence of identified gRNAs that generate alternative 5' end sequences. These genes include CR3, CR4, ND3, the 5' editing

domain of ND7, ND9 and RPS12. Nondual-coding pan-edited genes include ATPase 6, COIII, ND8 and the 3' editing domain of ND7. Partially edited genes include CYb,Murf II and COII. Never edited genes include COI, ND1, ND2, ND4 and ND5. For all analyses, ND7 was considered as two separate coding regions: the 5' editing domain (ND7N) and the 3' editing domain (ND7C) [27]. As we hypothesize that only the 5' editing domain of ND7 is dual-coding, mutation calculations for ND7N was pooled with the dual-coding genes and ND7C was pooled with nondual-coding pan-edited genes. *T. brucei* and *T. vivax* mRNA sequences of mitochondrial encoded genes were aligned based on protein sequence using Clustal Omega [116]. Nucleotide sequence mutations were identified and their effects on the amino acid sequence were classified as silent, missense or nonsense mutations. Missense mutations were further divided into three groups based on the PAM 250 matrix where conversions with a value <0 were considered not conserved, conversions with a value 0x0.5 were considered modestly conserved, and conversions with a value >0.5 were considered strongly conserved [117]. Mutation frequencies were normalized for each gene using nucleotide sequence length. Frequencies were compared using unpaired t-tests.

The extent of editing conservation between *T. vivax* and *T. brucei* was calculated by aligning the pan-edited genes based on ACG sequence. For each alignment, each location between an A, C or G nucleotide where a U-residue was inserted or deleted in either sequence was considered an editing site. Editing sites were classified as identical in both sequences, altered in insertion or deletion length, having switched from an insertion site to a deletion site, or only occurring in one of the sequences. Percent editing conservation was based on total number of editing sites within each mRNA. Percentages were compared using unpaired t-tests.

A principal component analysis (PCA) was performed on all three reading frames of the pan-edited genes using the scikit-learn principal component analysis tool [118]. For this analysis, the predicted protein sequences for all three reading frames were aligned using Clustal Omega [116]. Missense, nonsense, and indel mutations were quantified. Missense mutations were further divided into three groups as described above. Each mutation type was quantified and the relative frequency of each mutation calculated based on protein amino acid length. The variables used in the PCA include the protein mutation frequencies and the percentage of identical editing sites in each mRNA. The first reading frame of each gene is defined as the ORF published in the literature.

Data availability

CR3 sequence accession number: SAMN06318039. TREU 927 gRNA sequence accession number: SAMN06318154. TREU 667 gRNA sequence accession number: SAMN06318153. NCBI's Sequence Read Archive.

Results

In *T. brucei*, analyses of the gRNA transcriptome for the pan-edited transcripts indicate that full editing involves a large number of gRNA populations [9,119]. In addition, most of the gRNA populations (population defined as guiding the same or near same region of the mRNA) contain multiple sequence classes. The sequence classes most often differ in R to R or Y to Y mutations, hence guide the generation of the same mRNA sequence (A:U and G:U base pairs allowed). During these analyses, we noted that the terminal gRNA population for Cytosine-rich Region 3 (CR3) (putative NADH dehydrogenase subunit 4L [120]), had 3' sequences that would extend editing beyond the previously identified translation start codon. In addition, this

population had several sequence variants that would generate different edited sequences in this region. The most abundant terminal gRNA would introduce a stop codon in-frame with two alternative AUG start codons found near the 5' end (Figure 8A). Other sequence classes however, would either bring the upstream AUGs into frame, or shift the reading frame. Intriguingly, the alternative +1 reading frame (ARF) did not contain any premature termination codons. In order to determine if these gRNAs were utilized, we used Illumina deep sequencing to identify the most abundant forms of fully edited CR3 transcripts. Surprisingly, we identified multiple forms of the mRNA (Figure 8A, 8B and 8C and APPENDIX D). The first was the fully edited sequence predicted by the most abundant gRNA identified (Figure 8A). The other transcripts however, had unique editing patterns at the 5' end (Figure 8B and 8C and APPENDIX D). Use of these 5' CR3 sequences allowed us to identify novel gRNAs. Predicted translation of these mRNA sequences indicate that they use the +1 reading frame, and that the protein generated would be the same length as the ORF previously identified. This suggests that CR3 is dual-coding, and that selection of the terminal gRNA determines which reading frame will be used. A re-examination of the terminal gRNAs for the pan-edited genes indicated that at least two other transcripts, NADH dehydrogenase subunit 7 (ND7) and ribosomal protein subunit 12 (RPS12), have identified gRNA sequence variants within the terminal gRNA population that allow access to alternative reading frames (Figure 8D and 8E and APPENDICES E and F). Interestingly, the alternative gRNA for ND7 generates a +2 frameshift with a 65 amino acid open reading frame. The ND7 transcript is differentially edited in two distinct domains separated by 59 nts that are not edited in the mature transcript (the HR3 region) [27]. Only the 5' domain is edited in both life cycle stages; full editing of the 3' domain was only found in

bloodstream form (BF) parasites. The stop codon for the +2 frameshift is found within the HR3 region, therefore this alternative protein would be generated by full editing of only the 5' domain. While the most abundant gRNA in the Eatro BF transcriptome (~50,000 reads) would generate a sequence utilizing the identified ND7 ORF (Figure 8D ORF), the most abundant gRNA (>100,000) in the Eatro 164 procyclic library is the +2 ARF gRNA (Figure 8D ARF and APPENDIX E). In RPS12, the alternative gRNA deletes an additional U-residue downstream of the existing start codon, shifting the reading frame into the +1 ARF (Figure 8E). Interestingly, in *Leishmania tarentolae*, a gRNA, gRPS12VIIIa, has been identified that would also shift the frame of the existing start codon into the +1 ARF [121].

The identification of gRNAs that could alter the reading frame led us to re-analyze the ORFs of the edited transcripts. In addition to CR3, we found extended ORFs in two different frames for Cytosine-rich Region 4 (CR4) and NADH Dehydrogenase (ND) subunit 9, while several others had shorter, but still significant ORFs in alternative frames (Figure 9). We do note that the original sequence publications for both CR4 and RPS12 (CR6) had indicated that the fully edited sequence contained extended ORFs in two different frames [30,48]. Additionally, NADH Dehydrogenase subunit 3 (ND3) was also considered to be potentially dual coding, based on mutational analysis described below.

As we did find potential ARFs in the edited transcripts, we analyzed the predicted ORFs for biases in their mutational pattern. Dual-coding genes often display an atypical codon mutation bias due to constraints imposed by the need to maintain protein function in both genes. In single-coding genes, changes in the third nucleotide of a codon give rise to synonymous amino acids, so this position (N3) is much less constrained. In contrast, in dual-

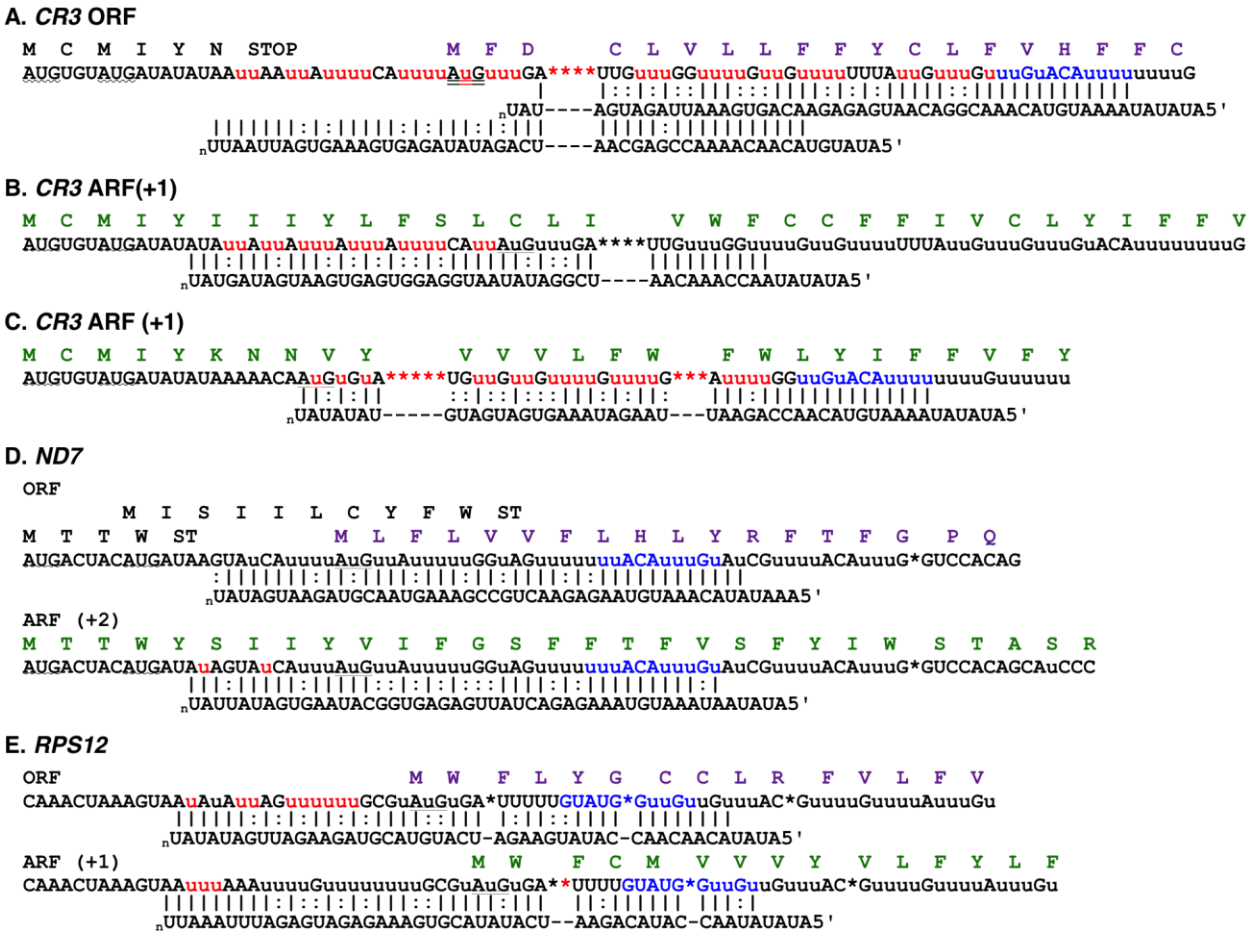


Figure 8. Alternative editing of the 5' end of pan-edited genes results in access to different reading frames. CR3 (A, B, C): Sequenced mRNA variants are aligned with gRNAs and predicted protein sequences. Inserted U-residues are lowercase while deleted U-residues are shown as asterisks. Canonical Watson-Crick base pairs (|); G:U base pairs (:). Previously identified start codons are doubled underlined. Potential upstream AUG start codons are indicated by wave underlines. Alternatively edited nucleotides are shown in red. Common anchor regions are shown in blue. ND7 (D), and RPS12 (E): Predicted mRNA and protein sequences, based on identified gRNAs.

coding genes, the N3 position in one frame is the N1 or N2 position in the alternative frame. Therefore, they have low rates of synonymous mutations [122]. This codon bias has been used to develop algorithms to detect novel overlapping genes [123,124]. These algorithms however, cannot be used in the analysis of our edited transcripts as the two-component genetic system (mRNAs created by gRNA editing) introduces another layer of mutational constraint [125]. In addition, the edited sequence of the transcripts is known for only a limited number of

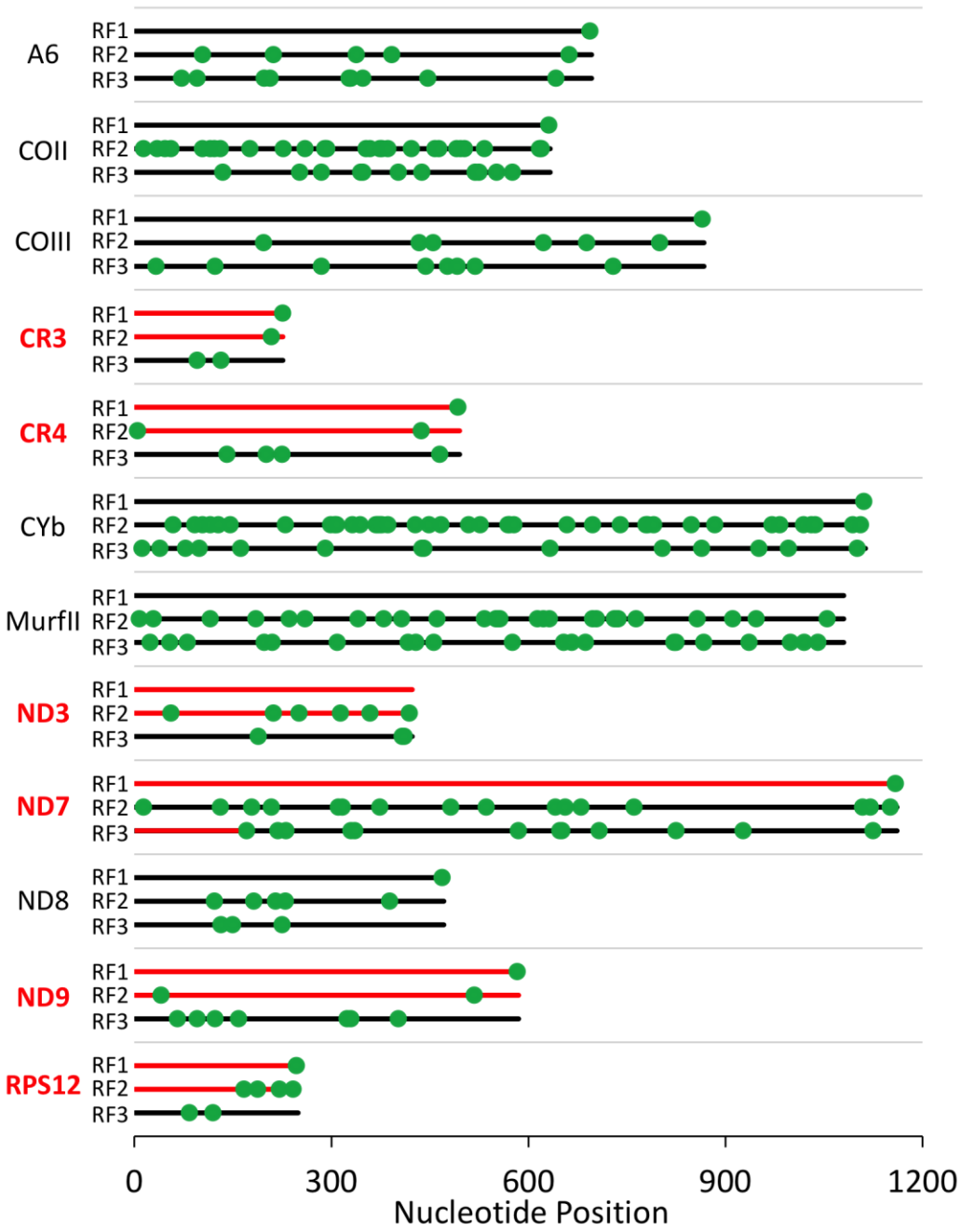


Figure 9. Positions of stop codons on all RFs of the edited genes in *T. brucei*. For each gene, reading frame 1 (RF1) is designated as the protein ORF previously identified. Hypothetical dual-coding reading frames are shown in red. A6 = ATPase 6; CO = Cytochrome Oxidase; CYb = Cytochrome b; Murf = Maxicircle unidentified reading frame; ND3 ±ND9 = NADH Dehydrogenase subunits.

kinetoplastids, and only the salivarian trypanosomes have the same general life cycle; other

kinetoplastids, like *Leishmania* and *T. cruzi*, have evolved different infective cycles and are

under very different selective pressures [113,126,127]. Fully edited sequences are known for *T. vivax*, the earliest branching salivarian trypanosome [52,128]. *T. vivax* differs from *T. brucei* in that they complete the insect phase of their life cycle entirely within the proboscis of the fly. This parasite has been described as an intermediate stage in the evolutionary pathway from mechanical transmission (ancestral) to full adaptation to the midgut and salivary glands of the tsetse fly [129]. Using the *T. vivax* sequence, we analyzed mutation patterns in all of the mitochondrially-encoded mRNAs (Figure 10). mRNA sequences were aligned by codons based on their protein alignments (Clustal Omega [116]). Mutated codons were identified and classified as silent, missense and nonsense mutations. Missense mutations were further divided into three groups based on the PAM 250 matrix [117]. These data clearly show that the RNA editing process significantly constrains the types of mutations tolerated within the mitochondrial genome. In comparison to the genes that are not edited (ND1, ND2, ND4, ND5, COI) or have limited editing (CYb,Murf II and COII), a distinct suppression of silent mutations and strongly conserved missense mutations were observed for all of the pan-edited genes, consistent with previous observations (Figure 10) [125]. A suppression of mutations that lead to moderately conserved amino acid replacements was also observed, but these were not as striking due to the low frequency of this type of mutation. No significant difference was observed in the frequency of not conserved missense mutations, though a trend towards a lower frequency of these mutations in the putative dual-coding genes (CR3, CR4, ND3, ND9, 5'ND7 and RPS12) was noted. This was complemented by a significant increase in the frequency of strongly conserved missense mutations in the putative dual-coding genes in comparison to the other pan-edited genes (3'ND7, ND8, A6 and COIII).

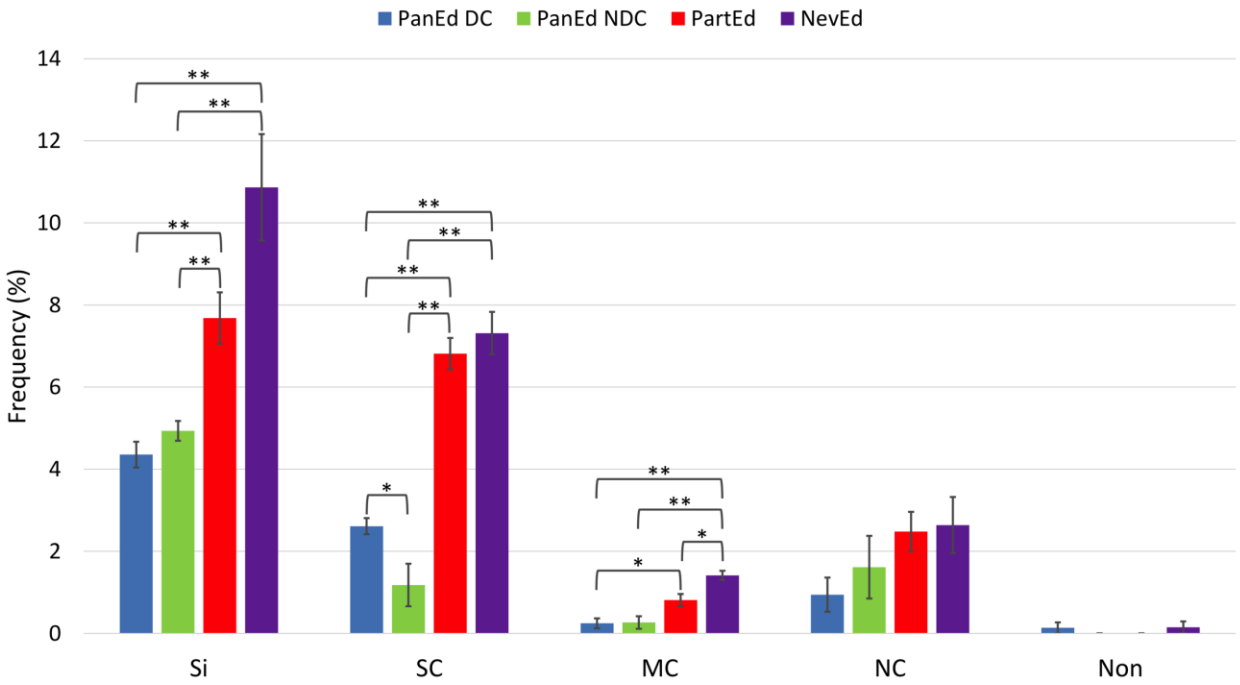


Figure 10. Mutational frequencies in mitochondrially encoded genes categorized by effect on amino acid sequence. *T. brucei* and *T. vivax* pan-edited dual-coding (PanEd DC), pan-edited nondual-coding (PanEd NDC), partially edited (PartEd), and never edited (NevEd) mRNA sequences were aligned based on their amino acid alignment (reading frame 1, defined as the reading frame encoding the gene product previously annotated in the literature) [116]. Mutations were categorized as silent (Si), strongly conserved (SC), modestly conserved (MC), not conserved (NC) or nonsense (Non). The amount of conservation was determined using the PAM 250 matrix, where conversions with a value $0<x\leq0.5$ were considered modestly conserved, and conversions with a value >0.5 were considered strongly conserved, and conversions with a value ≤0 were considered not conserved. Error bars depict standard error. * $p<0.05$, ** $p<0.01$ (unpaired t-test).

Surprisingly, while the overall mutational frequency of the fully edited pan-edited genes was similar, a comparison of the conservation of editing patterns did show a significant difference between the putative dual-coding and the other pan-edited genes (Figure 11). The dual-coding genes consistently had a lower conservation of their editing pattern. Upon further examination, we found that most changes in the editing pattern resulted from thymidine insertions and deletions within the maxicircle DNA sequence, which was then corrected by the editing machinery. These types of mutations do not result in a change to the final mRNA

sequence once edited. The *T. brucei* (Tb) dual-coding genes appeared to consistently insert more U-residues, while *T. vivax* (Tv) had more U-residues encoded within the DNA sequence. Indeed, comparisons of the length of the coding regions of Tb and Tv cryptogenes (unedited sequence) show that the putative dual-coding genes are almost 10% shorter in Tb. In contrast, the nondual coding cryptogenes are not significantly shorter (~2.5%). Some of the other changes in editing patterns did generate small internal frameshifts as previously described by Landweber and Gilbert [125]. However, the high prevalence of internal frameshifts reported for COIII by Landweber and Gilbert is reflected in our analysis only for COIII and A6.

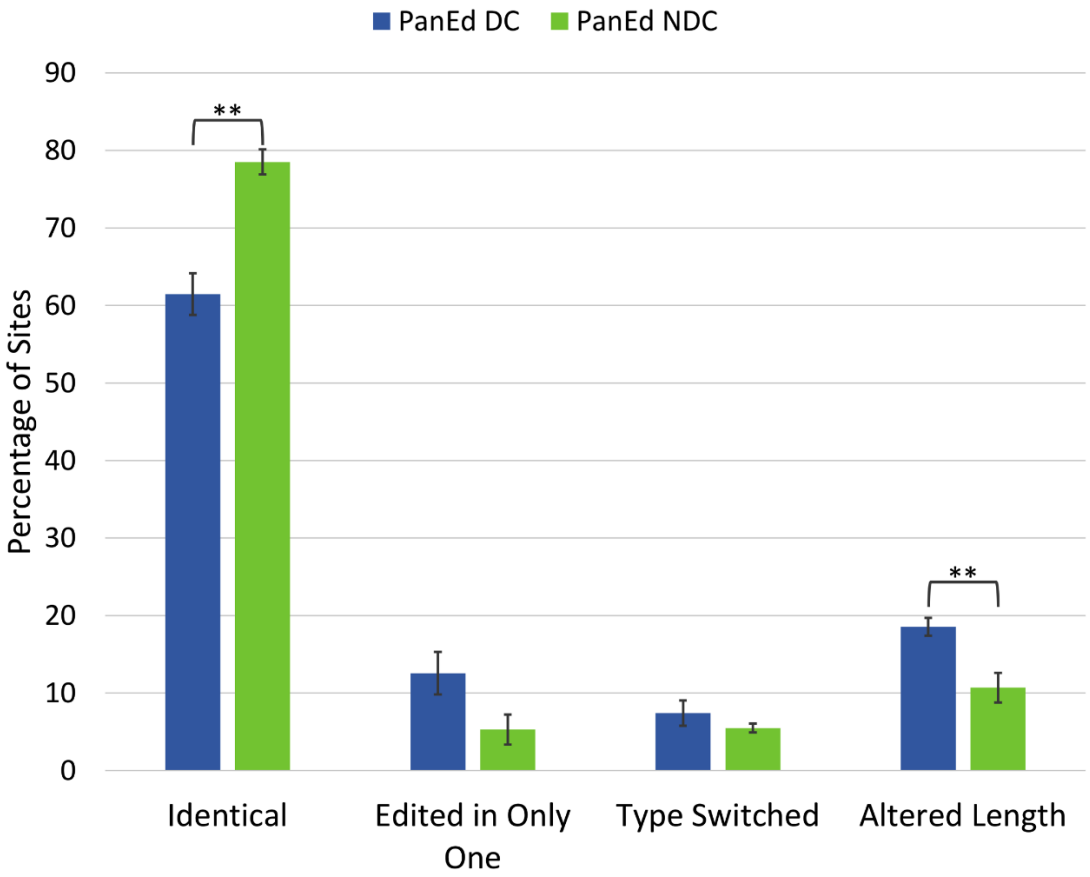


Figure 11. Percent conservation of editing patterns between *T. brucei* and *T. vivax*. Alignment of the fully edited mRNAs was based on ACG sequence (see APPENDIX G). Each editing site was defined as a site on at least one of the two aligned mRNAs where an editing event occurred. Sites were then classified as identical, only identified in one of the two sequences, type switched (one site is an insertion and the other is a deletion), or altered in length. Error bars depict standard error. * $p<0.05$, ** $p< 0.01$ (unpaired t-test).

Since differences in the types of amino acid mutations were observed, we performed a principal component analysis on the frequency of mutation types for all three reading frames of the pan-edited genes (Figure 12). In addition, we included the percentage of editing site conservation as a variable. This analysis clearly clustered the putative +1 dual-coding transcripts (reading frame 2). The first component (z-axis,) is strongly based on editing conservation, and separates the dual-coding genes from the other pan-edited genes as expected. While component 2 (x-axis) separated ORF1 and ORF3 from ORF2 of each gene, component 3 clearly separated the dual-coding ORF2s from nondual-coding ORF2s. The ND7N ORF3 was the only exception, and the gRNA data suggests that it is a dual-coding gene using the +2 (ORF3) reading frame. This suggests that an additional layer of mutational constraint beyond that imposed by the RNA editing process can be detected for six of the extensively edited transcripts.

Because dual-coding genes are often conserved in multiple species, we analyzed the available sequences of other kinetoplastids (*Leishmania tarentolae* (Lt), *Leishmania mexicana amazonensis* (Lma), *Phytomonas serpens* (Ps), *Perkinsela CCAP1560/4* (Pk)) to determine if they also contain multiple overlapping reading frames with homology to those found in *T. brucei*. Interestingly, many of the alternative reading frames did show some homology to the ARFs found in Tb. However, most of these ARFs are punctuated with stop codons (APPENDIX H). Extended alternative reading frames are found in CR3, 5'ND7 and RPS12 in Ps. However, the extended ARF in the Ps CR3 is in the +2 reading frame and the ND7 and RPS12 ARFs shows very little homology with the Tb/Tv ARF (APPENDIX H Parts A, D and F) [130]. Interestingly, while *Perkinsela* has lost many of the genes in the mitochondria, RPS12 was retained [131]. The Pk RPS12 ARF possesses a near full open reading frame with one stop codon three codons after an

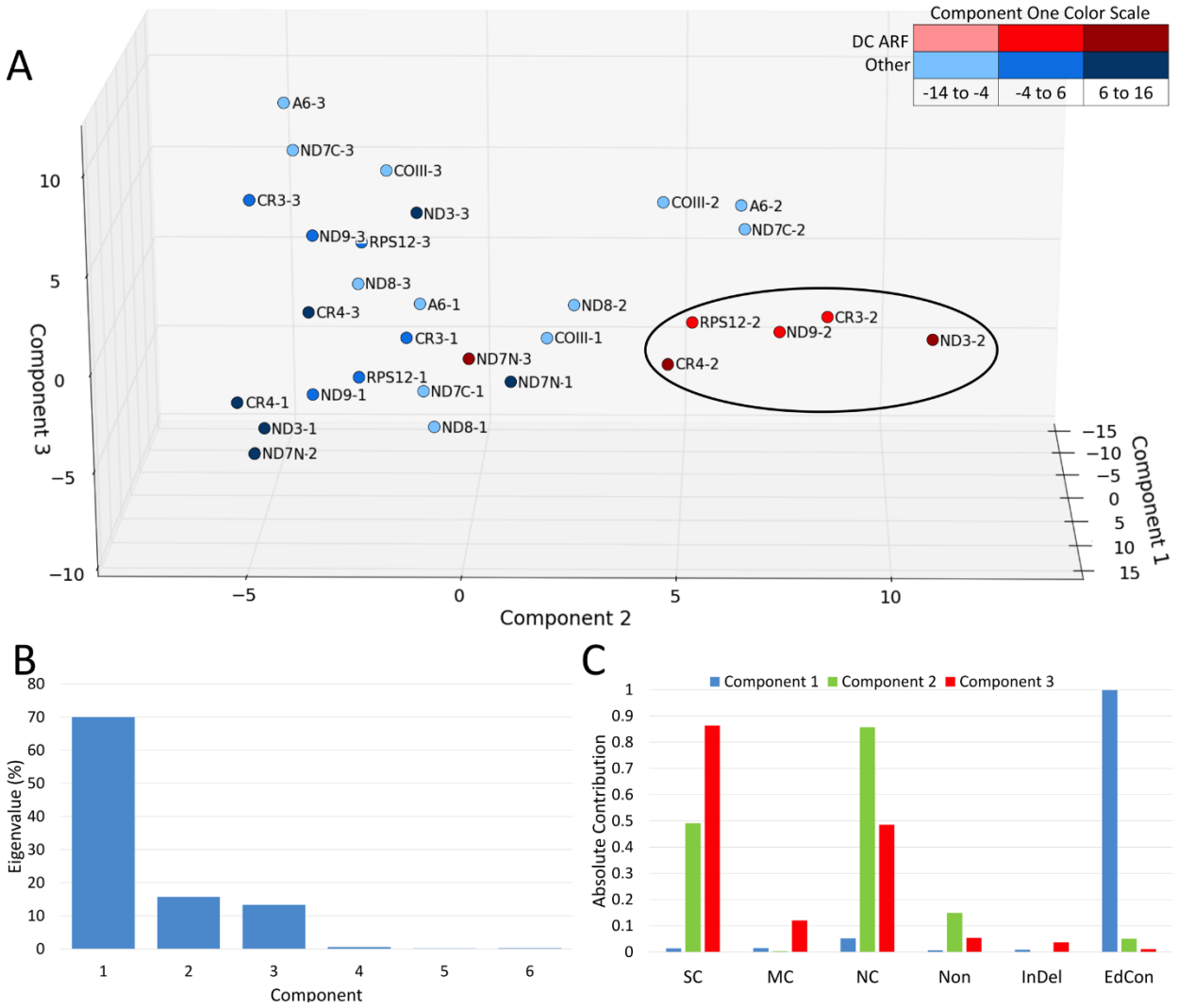


Figure 12. Principal component analysis of frequency of amino acid mutation types and editing conservation between *T. brucei* and *T. vivax* pan-edited transcripts. A. First factorial plan (z-axis: first component, x-axis: second component, y-axis: third component). ND7N = ND7 5' editing domain, ND7C = ND7 3' editing domain. ORF2 and ORF3 are defined as the +1 and +2 reading frames, respectively. B. Histogram of eigenvalues for first six components. Eigenvalues represent the amount of the variance accounted for by each component. C. Absolute contribution of each analyzed mutation frequency to components 1, 2, and 3. Amount of conservation was determined using the PAM 250 matrix as described in Figure 10. Mutation type: SC = strongly conserved, MC = moderately conserved, NC = Not conserved, Non = Nonsense, InDel = insertion or deletion. Editing conservation (EdCon) was determined using alignments of edited mRNAs (APPENDIX G). Aligned editing sites were characterized as identical or altered.

in frame start near the 5' end of the edited transcript. This pattern is reminiscent of the

conventionally edited sequences of CR3 and ND7, and could suggest that an alternative edit

may remove the stop codon, allowing access to the ARF. The *L. tarentolae* CR4 orthologue also has two extended ORFs. Interestingly, the published sequence for Lt CR4 appears to switch between the two ORFs (switch appears to occur in a stretch of 13 inserted Us) [77]. This may explain why only the carboxyl half of the published Lt CR4 showed good homology with Tb and Lma [132]. Translation of the Lt ARF does generate a protein with the N-terminus showing high homology to the conventional Tb and Lma CR4, while translation of the published ORF shows some homology to the Tb CR4 ARF (APPENDIX H Part B). These data are intriguing enough that these sequences should be re-examined. While most of the other pan-edited transcripts had multiple stop codons in the +1 and +2 reading frames, many did show good homology to the Tb ARF sequences. Particularly intriguing are the ND3, ND8 and ND9 alignments. While internal stop codons are found in Tv, Lt and Lma ND9 ARFs, they show strong homology to the Tb ND9 ARF throughout the protein (APPENDIX H Part E). In ND3, the amino ends of the ARFs show strong homology between all four of the *Trypanosoma* and *Leishmania* species (APPENDIX H Part C). This homology decreases after an internal stop codon found in the same position in 3 of the 4 species. As ND8 is the only other pan-edited gene in Lt, Lam and Ps, we also examined the conservation of the ND8 ORF and ARF, even though our mutational analyses did not tag the ND8 gene as dual-coding. While the ND8 ARFs were punctuated by multiple stop codons, they surprisingly also showed areas of strong homology between all 5 species, especially down stream of an internal methionine (APPENDIX H Part G). We do note that we cannot rule out the possibility that alternative editing can remove stop codons observed in the ARFs.

Analyses of the ARF predicted proteins suggest that they are all short transmembrane proteins with two or more predicted transmembrane alpha helices (Figure 13) [133,134]. While

functional homologues are often difficult to detect in trypanosomes, searches using the predicted protein sequence of each ARF did identify small molecule transport proteins with limited confidence. Using Phyre2, the ND7 ARF was identified as a homolog of the bacterial sugar transporter SemiSWEET (61.5% confidence) [135,136]. SemiSWEET, which forms homodimeric structures, is also a distant homolog of the yeast mitochondrial pyruvate carrier 1 (MPC1). This protein has two transmembrane alpha helices and forms a heterodimer with either of the other two pyruvate carrier proteins [137,138]. While still very speculative, it is intriguing that the small ARF proteins might oligomerize to form small mitochondrial membrane transporters.

CR3ARF: MCMIYKNNVYVVVLFWFWLYIFFVFYLFVICFYVCYLVFVFYWIFVFYLIWVYCCVYFFFLFYHLICCYHFYYCI (75 A.A.)

CR4ARF: FWLCIFFFFIWCVCVLCTVYGIFYCCFVFCFCCLFVWVCFICFVIVVCFFLLFWVVIFYWCFDCIVYFFCDVIILFIFFFYFVLCFL
YCCFYLVCLVFFCIFCCVLCYFLIYFLCCFLFWVVFLFLFFFFVYVCFYLWLLLFC (143 A.A.)*

ND3ARF: SKNPRLFTLVCYHYFYICFWYCGLLFYL**!**VFFYVFYVFYCIFLIVFVVVCGFRVVCMIWIHVWCFIHWIYVLLVVCFLYC (80 A.A.)*

ND7ARF: MTTWYSIIYVIFGSFFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFASWYRKVMWI (65 A.A.)

ND9ARF:FYFIVCVVDGVLFVLLIVVFCFFIVLLVFFCFIVCFYFLICDFCFYIVVICYWLIFVVFVFVVLCCCIFYFVCFCVFVCVLFCVVCLYFLDCV
LVLVVFVMRFCCWNANVLICLVFILLVMIFYIVYLLIGFLLVFFCWSVIHYLVCLYCYFVCWWRR (158 A.A.)*

RPS12: MWFCMVVVYVLFYLFYVIIWVRDCPVPVTDVYCMPYFIYIILFGCCVVFFVVLLV (55 A.A.)

Figure 13. Amino acid sequences of ARFs of dual-coding genes. Predicted transmembrane regions are shaded in gray [133,134]. *No start codon was identified and the amino acid sequence shown begins at the 5' end or after the first stop codon at the 5' end. Exclamation point indicates premature termination codon.

Discussion

The work presented here, suggests that as many as six of the extensively edited mRNAs in *T. brucei* are dual-coding and that it is alternative editing using different terminal gRNAs that allows access to the two different reading frames. Deep sequencing of the 5' end of CR3 indicates fully edited transcripts that have access to both reading frames are present in the mitochondrial transcriptome and gRNA analyses indicate that three different cell lines contain

gRNAs that can alternatively edit the 5' ends of CR3, RPS12 and ND7. In addition, analyses of the mutational bias in pan-edited genes suggest that an additional layer of mutational constraint is observed in the putative dual-coding genes. While the overall mutational frequency observed for the fully edited mRNAs is similar for all pan-edited genes, the types of amino acid changes that appear to be tolerated are significantly different. This is consistent with these genes having to maintain functional proteins in two different reading frames. Analyses of other trypanosomes, do show that some of the ARFs have intriguing homology to the ARFs identified in *T. brucei* and *T. vivax*. However, most of the ARFs are punctuated with stop codons. These data are difficult to interpret because we cannot rule out the possibility that the stop codons are removed by alternative editing events. In addition, the other trypanosome species have evolved very different infective life cycles and are under different selective pressures. For example, *P. serpens* is a pathogen that infects important crops and is transmitted by sap-feeding bugs. These parasites have glucose readily available in both life cycle stages and are unique in that they lack a fully functional respiratory electron transport chain [64,65]. For *Leishmania*, all life cycle stages possess an active Krebs cycle and ETC linked to the generation of ATP [61,62,139]. These unique adaptations to different hosts suggest that they may not be under the same evolutionary pressure to maintain dual-coding genes.

Overlapping reading frames are common in viruses, and are thought to persist due to strong genome size constraints [87,88]. More recently however, over-lapping genes have been identified in mammalian and bacterial genomes [89–92]. In these organisms, size is not an issue and the potential advantage of overlapping genes is less clear. For dual-coding genes, the need to maintain both ORFs constrains the ability of each protein to become optimally adapted [93].

As this constraint can be alleviated by gene duplication, it is thought that dual-coding regions can survive long evolutionary spans only if the overlap provides a selective advantage. In mammals, many of the identified dual-coding genes like Gnas1 and XBP1, produce two proteins that bind and regulate each other [94,95]. For these proteins, dual-coding may be advantageous for the tight co-expression needed. An alternative model, suggests that under high mutation rates, the overlapping of critical nucleotide residues is advantageous because it may reduce the target size for lethal mutations [96]. This may be particularly important for organisms that have evolved to exist in dual-metabolic environments (two hosts). We hypothesize that the trypanosome mitochondrial ARFs encode small metabolite transporters that provide a distinct growth advantage to bloodstream form parasites. The complete overlap of these small transporter genes with electron transport chain (ETC) genes would protect the integrity of the ETC genes that are required only in the insect host. Thus, in trypanosomes, dual-coding genes may be a mechanism to combat genetic drift during extended periods of growth in non-selective environments. In *T. brucei*, it is known that a number of bloodstream form essential proteins are functionally linked to Krebs cycle or ETC genes. While not a ªclassic º dual-coding gene in that production of the alternative protein does not involve overlapping reading frames, the pan-edited COIII gene does contain the information for two distinct proteins, COIII and AEP-1. AEP-1 is important for kinetoplastid DNA maintenance and overexpression of the DNA-binding domain results in a dominant negative phenotype including decreased cell growth and aberrant mitochondrial DNA structure [68]. The nuclear encoded α-ketoglutarate dehydrogenase E2 (α-KDE2) is known to be a dual-function protein, in that it plays important roles in both the Krebs cycle and in mitochondrial DNA inheritance [97]. RNAi

knockdowns of this gene in bloodstream form (BF) trypanosomes also show a pronounced reduction in cell growth. Similarly, the Krebs cycle enzyme α-ketoglutarate decarboxylase (α-KDE1) is also a dual-function protein with overlapping targeting signals that allow it to be localized to both the mitochondrion and glycosomes [98]. RNAi knockdowns of α-KDE1 in BF trypanosomes is lethal, suggesting that in addition to its enzymatic role in the Krebs cycle, it plays an essential role in glycosomal function in *T. brucei* [98]. It has been previously suggested that both alternative editing and dual-function proteins are important mechanisms for expanding the functional diversity of proteins found in trypanosomes [67,97–99]. We hypothesize, that in salivarian trypanosomes, an equally important role for these dual-coding/function genes may be the protection of genetic information.

The 'why' of the unique RNA editing process in kinetoplastids has been a long-standing paradox. The complex machinery and the sheer number of gRNAs required to direct the thousands of U-insertion/deletions indicate that this process is metabolically very costly. Initially, it was proposed that U-insertion/deletion editing (kRNA editing) was one of many RNA editing processes that were in fact relics of the RNA world. However, the very different mechanism of the RNA editing systems in existence, and their very limited distribution within specific groups of organisms indicate that they are more likely derived traits that evolved later in evolution [69,70]. The sheer complexity of the kRNA editing process, with no obvious selective advantage, led to the proposal that insertion/deletion editing arose via a constructive neutral evolution (CNE) pathway [71]. Indeed, RNA editing in trypanosomes is always mentioned in support of CNE as an example of how seemingly non-advantageous, complex processes can arise [72,73]. More recently however, it has been hypothesized that RNA editing

co-evolved with G-quadruplex structures found in the pre-edited mRNAs [74]. These structures are thought to be advantageous in that they can help regulate transcription in order to promote DNA replication and prevent kinetoplast DNA loss. However, they must be removed by the RNA editing system prior to translation [74]. Another prominent hypothesis is that RNA editing is advantageous because it is a mechanism by which an organism can fragment and scatter essential genetic information throughout a genome [75,76]. Kinetoplast DNA is far less stable than chromosomal DNA, and loss of minicircles due to asymmetric division of the kDNA network have been frequently observed, especially in laboratory cultures of *Leishmania* [76,77]. Buhrman et al. [76] suggest that the scattering of essential guide RNA genes throughout the DNA network, would prevent fast growing deletion mutants from outcompeting more metabolically versatile parasites during growth in the mammalian host. Using a mathematical model of gene fragmentation in changing environments (absence of functional selection), they showed a distinct advantage for gene fragmentation. In their model, the number of tolerable generations under periods of relaxed selective pressure was increased by more than 40% before loss of the ability to move to the next life cycle stage. If the dual-coding ARFs give BF trypanosomes a selective growth advantage similar to that observed by the COIII alternative protein AEP1, then the number of 'essential' gRNA genes would increase greatly. Currently, only AEP1, A6 and RPS12 mitochondrial genes have been experimentally shown to be essential [68,100,110]. In addition, the presence of alternative editing and dual-coding genes would complement the protection provided by gene fragmentation by also shielding the genes from deleterious point mutations within critical ETC genes. This suggests that the complex RNA editing system found in the mitochondria may therefore provide multiple molecular strategies

to increase genetic robustness. Protection of the mitochondrial genome during growth in the mammal would increase the capacity for successful transfer to an insect vector and maximize the parasites long-term survival and spread.

Acknowledgments

We thank the Ken Stuart Lab for trypanosome cell lines and Chris Adami for helpful discussions. We would also like to acknowledge the Dr. Marvis A. Richardson Endowed Fellowship Fund for their recognition of LEK.

CHAPTER 4: ANALYSIS OF THREE PAN-EDITED MRNAS REVEALS DUAL-CODING GENES AND COMPLEX MULTIPATH EDITING

Abstract

Trypanosoma brucei is a single celled eukaryote that possesses a highly complex RNA editing system. In this system, a large set of small RNAs, called guide RNAs direct the insertion and deletion of uridines in mitochondrial mRNAs. These changes extensively alter the target mRNAs, up to doubling them in length. Recently, mutational analysis showed that several of the edited genes possessed capacity to encode two different protein products. These overlapped reading frames could be accessed through alternative RNA editing, that shifts the translated reading frame. In this study, we analyzed the editing patterns of three putative dual-coding genes, ribosomal protein S12 (RPS12), the 5' editing domain of NADH dehydrogenase subunit 7 (ND7 5'), and C-rich region 3 (CR3). We found evidence that fully edited ND7 5' and CR3 are can translate in more than one reading frame. Moreover, we found that CR3 has a complex set of editing pathways that vary substantially between cell lines, and that changing available energy sources also alters the editing preferences of CR3 and ND7 5'. These findings suggest that editing patterns can be influenced by the current environment, and that alternative editing may be utilized by the trypanosomes to introduce variation within this fragile editing system.

Introduction

Trypanosoma brucei is a member of the Kinetoplastea, a group of protozoans characterized by a large network of DNA in their mitochondria known as the kinetoplast [1]. The kinetoplast is composed of two types of concatenated circular DNA molecules: maxicircles and minicircles. The maxicircles all encode mitochondrial ribosomal RNAs as well as 18 protein coding genes, most of which are components of the electron transport chain. The approximately 30-50 identical copies of the maxicircle make up a relatively small proportion of the kinetoplast [5]. Most of the DNA network is composed of 5,000 and 10,000 1 kb minicircles, each of which encodes 2-5 small non-coding guide RNAs (gRNAs) [8,33]. These gRNAs are used in the process of RNA editing. In *T. brucei,* RNA editing consists of specific uridine insertion and deletion events that render 12 of the 18 mitochondrially encoded mRNAs translatable [4]. The gRNAs act as templates for the large editosome complex which cleaves the mRNA, inserts or deletes the correct number of uridines and then re-ligates the mRNA in an energy intensive process. This is repeated until the mRNA is complementary to the small gRNA. Each gRNA directs edits that generate the anchor region for the next gRNA, thus the RNA editing process is sequentially dependent on correct editing by each gRNA. As editing of some of the extensively edited mRNAs can involve upwards of 40 gRNAs, this renders the process incredibly fragile. [140]. We hypothesize, that such an expensive and fragile process evolved in response to the unique life cycle of *T. brucei*.

T. brucei is a dixenous parasite, invading the bloodstream of a mammalian host and being transmitted between hosts by bite of a tsetse fly. Once taken up in a blood meal by the tsetse fly, *T. brucei* transitions into the replicating procyclic state in the midgut, and the energy

T. brucei requires for this replication is gained through metabolism of amino acids [19,20]. This is accomplished through use of a portion of the Krebs cycle and the electron transport chain (ETC), thus most of the ATP required is produced by the mitochondria [19–21]. This stage of the life cycle is followed by a dramatic bottleneck when the trypanosomes transition from the midgut to the salivary glands of the tsetse fly [22,23]. From the salivary glands, trypanosomes are then refluxed into their next mammalian host during a bloodmeal. Once the parasite is deposited into its mammalian host, it quickly transitions to utilizing glycolysis for its energy generation, removing the requirement for ATP production in the mitochondria [15]. While in the mammalian host, *T. brucei* lives entirely extracellularly. It is frequently subject to attacks by the host's adaptive immune system, and the population evades these attacks through antigenic variation [12]. This part of the life cycle can be quite long, with the longest known infection lasting 29 years [13]. This life cycle should make *T. brucei* particularly sensitive to genetic drift, especially for those genes which are not under selection (Krebs's cycle and ETC) and should make them extremely vulnerable to Muller's ratchet (the gradual increase of mutational load that eventually leads to extinction) [81–84]. One mechanism for protecting small asexual populations is by increasing the severity of the mutations that can occur. If mutations severely impact fitness, mutated individuals are selected out, preventing their fixation [79]. Recently, computer modeling studies suggest that small asexual populations can evolve this type of mechanism (termed "drift robustness") in order to maintain fitness [80]. The sequential dependence of the kRNA editing process implies that the system is inherently fragile to mutations. Even a single point mutation can drastically change the editing pattern, and stop the editing process, aborting expression of the protein. Hence, the RNA editing process may

operate as a proof-reading system to weed out mutations by making them lethal. This is effective however, only if the mitochondrial genes are under selection. Previously, we showed that many of the mitochondrially pan-edited genes have a distinct mutational bias that is suggestive of dual-coding genes (coding two proteins by overlapping reading frames) [141]. The overlapping of ETC genes not under selection in the bloodstream stage with genes that are under selection during this stage of the life cycle, would prevent the accumulation of mutations. As the extensively overlapped genes share most gRNAs, this strategy would ensure that almost all of the genetic material is protected.

Our analyses suggested that out of the twelve pan-edited genes in *T. brucei*, six are potentially dual coding, and that the RNA editing system is used to determine which reading frame is accessed. In order to determine if mRNA transcripts with access to multiple open Reading frames (ORFs) exist within the mitochondrial transcriptome, we deep sequenced the mRNA transcript populations of three putative dual coding genes: ribosomal protein S12 (RPS12), the 5' editing domain of NADH dehydrogenase subunit 7 (ND7 5'), and C-rich region 3 (CR3). Using the previously generated gRNA transcriptomes, we constructed detailed editing pathways for each of these genes. The editing pathway of RPS12 was primarily linear, reflecting the high degree of conservation required for a gene that is essential [100,107]. We found no evidence of utilization of the gRNA that provides access to the alternative reading frame [141]. In contrast, we did identify transcripts using different reading frames for both CR3 and ND7 5'. This study indicates that RNA editing can be used to access multiple open reading frames using two different methods: in ND7 5', different gRNAs bring alternate start codons into frame and in CR3, different gRNAs can shift the reading frame of the existing start codon. In addition, CR3

showed incredible editing diversity, in that two different cell lines showed very different editing patterns, using different sets of gRNAs to edit the CR3 cryptogene. This suggests that the use of a gRNA-guided editing system can also dramatically increase protein diversity in spite of an incredibly rigid and mutationally fragile system.

Materials and Methods

T. brucei culture and RNA Isolation

T. brucei clones from strains EATRO 164 and TREU 667 were grown in SDM79 and harvested as previously described [9]. EATRO 164 cells grown in SDM79 were then gradually transitioned to SDM80 using serial 1:3 dilutions when cells reached a density of at least $5x10^6$ cells/mL. SDM80 was prepared as described by Lamour et al. with the exception of using undialzyed FBS, and reducing the amount of FBS added by half [142]. This results in the final concentration of glucose being 0.5 mM instead of 0.15 mM. This concentration is still well below that of SDM79, which has a glucose concentration of 6 mM. Once cells had been acclimated to SDM80, cells were harvested as previously described [9]. Mitochondrial vesicles were isolated using differential spins and mitochondrial RNA was then isolated from vesicles as previously described [9].

Preparation, Sequencing, and Analysis of mRNAs

cDNAs were generated from isolated RNAs using the Applied Biosystems High Capacity cDNA Reverse Transcription Kit. CR3, RPS12, and ND7 5' editing domain cDNAs were amplified via PCR using the following primers (underlined sequences are gene specific and non-underlined sequences are tag regions used in deep sequencing reaction):

CR3 5': ACACTGACGACATGGTTCTACAAGAAATATAAATATGTG

CR3 3' Short: TACGGTAGCAGAGACTTGGTCTACAAAAATTATTTGCATACTT

CR3 3’ Extended: TACGGTAGCAGAGACTTGGTCTACAAAAATTATTTGCATACTTTTTT

RPS12 5': ACACTGACGACATGGTTCTACACTAATACACTTTTG

RPS12 3': TACGGTAGCAGAGACTTGGTCTAAAAACATATCTTAT

ND7 5’: ACACTGACGACATGGTTCTACAGATACAAAAAAACATGAC

ND7 3’: TACGGTAGCAGAGACTTGGTCTCTTTTATATTCACATAACTTTTCTGTAC

Amplified cDNAs from EATRO 164 cells grown in SDM79, EATRO 164 cells grown in SDM80, and TREU 667 cells grown in SDM79 were individually barcoded and combined in equal molar amounts. Samples were sequenced in a 2x250bp paired end format (PE250) using an Illumina MiSeq Standard flow cell and 500 cycle reagent cartridge, version 2. Sequence data was preprocessed as previously described [141].

Sequence data was then separated by cell line, growth media and gene. Sequence data was then analyzed using a new pipeline and program called SKETCH (Segmentation of Kinetoplast Edited Transcripts to Characterize editing Heterogeneity). This program allowed us to classify mRNAs at the block editing level and determine which editing patterns were most prominent.

For each set of sequences, SKETCH would remove low quality sequences whose sequences containing more than 5 mismatches to the unedited template, disregarding uridines. In order to classify the editing patterns observed in the mRNA transcripts, SKETCH requires a set of template sequences. Initially, the templates supplied to SKETCH were the conventional fully edited and unedited sequences for each of the three genes examined. These sequences were

then segmented based on the editing blocks previously defined by the locations of gRNA populations [9]. Each transcript was then classified by editing block, with each block being classified as matching the unedited sequence, matching the fully edited sequence or being unknown. After the initial characterization of the transcripts, the most abundant unknown sequences for each editing block were then added to the reference pool. Sequences were then reclassified by SKETCH based on the newly added reference sequences. This process was repeated until the most abundant forms of editing were identified. SKETCH code is available upon request. To validate the newly identified editing patterns as true alternatives, the new sequences were screened against the gRNA transcriptome as previously described [9]. Sequences with a gRNA match were then considered valid alternative edits.

Uridine deletion analysis

For RPS12 and ND7 5', once full editing pathways were characterized, editing sites with DNA encoded Us were identified. Each encoded U site was then characterized based on the proximity of the preceding gRNAs' 3' poly-U tail as well as whether the all uridines at the site in question were deleted in the final fully edited sequence. For each site, a window was defined consisting of the 6 sites upstream and 6 sites downstream of the site in question. Using these parameters, the mRNA transcripts of RPS12 and ND7 were analyzed at each deletion window. The window of each transcript for each encoded U was examined and classified as unedited, fully edited or partially edited. Partially edited sequences were then classified based on the editing state of the encoded U site. For sequences with total deletions, each editing sequence was then classified based on the states in the 3' end of the window as either matching the fully edited sequence or not. Code available upon request.

Results

In order to confirm that transcripts with access to two reading frames exist *in vivo*, we analyzed the mRNA transcriptomes for three of the putative dual-coding genes, RPS12, ND7 5' and CR3. This mRNA deep sequencing data was then used in combination with the sequenced gRNA transcriptomes, to generate precise editing pathway maps. In order to determine how robust the observed editing pathways were, we characterized editing in two different cells lines, TREU 667 and EATRO 164. In addition, we examined the effect of energy source on these editing pathways by using two different media, SDM79 and SDM80. SDM79 is the standard medium used to grow the procyclic stage parasite. However, it contains 6mM glucose, and experiments have shown that under these levels of glucose, the procyclic stage can grow in the absence of electron transport chain (ETC) activity [112,142–147]. The SDM80 medium was developed to more closely resemble insect gut conditions and has very low glucose concentrations [142]. Trypanosome growth in this medium requires ATP production using the ETC [142].

RPS12 is an essential component of the mitochondrial ribosome [30,100,107]. RPS12 is extensively edited (pan-edited) with 132 Us inserted and 28 Us deleted. Full editing is directed by 12 populations of gRNAs (defined as a group of gRNAs that edit the same region of an mRNA) [9,30]. In this analysis, we identify 10 populations, with three of the previously identified populations being combined with other populations that shared a very high amount of overlap. One new population (F) was identified through a search of the gRNA transcriptome under reduced stringency. Analyses of the canonical editing pattern indicate that there are two long ORFs, and mutational bias analyses indicate that both ORFs may be selected for [141]. The

longest ORF encodes the RPS12 protein and encompasses a second shorter ORF of unknown function [30]. Northern blots revealed that edited RPS12 mRNAs were found in both life cycle stages, however, edited mRNAs were more abundant in bloodstream form than procyclic form trypanosomes [30].

Because RPS12 is essential, we expected it to have a very robust editing pattern in both cell lines, as well as under both energy conditions. In contrast, neither ND7 or CR3 appear to be essential in the insect stage of the parasite [112]. The canonical ND7 has two separate editing domains that are edited independently [27]. Interestingly, while the 3' editing domain is fully edited only in the bloodstream life cycle stage, the 5' editing domain is edited in both life cycle stages [5,27]. In addition, the mutational bias analyses indicate that only the 5' editing domain has characteristics indicative of a dual coding gene. The canonically edited CR3 is also a putative Complex I member (ND4L) and is preferentially edited in the BS stage [47,120]. Complex I has been shown to be non-essential in both life cycle stages, and other mitochondrially encoded complex I subunits, ND3, ND8, and ND9, have been shown to be preferentially edited in the bloodstream stage [5,26,28,29,111,112].

RNA seq data was generated by reverse transcribing all mtRNAs using random primers. For both RPS12 and CR3, transcripts were then selectively amplified using sequence specific primers targeted to the terminal 5' and 3' never edited regions as to not bias against any possible editing pattern. For ND7, the 5' editing domain was selectively amplified using sequence specific primers targeted to the 5' never edited region and the homology region 3 (HR3) that separates the 5' and 3' editing domains [27]. The HR3 is a span of 59 nts that is also never edited, hence should not bias the analysis. The targeted transcriptome libraries were

generated from TREU 667 cells grown in SDM79 and EATRO 164 cells grown in SDM79 and SDM80. Additionally, for CR3, we generated another library using TREU 667 cell line mRNA by selecting for transcripts of a larger size, instead of taking transcripts of all sizes (SDM79). This allowed us to enrich the library for transcripts that had initiated the editing process. Amplified cDNAs were then gel purified, barcoded and combined in equal molar amounts for sequencing. While the number of total reads obtained did vary by cell line and media used, surprisingly few transcript were fully edited (canonical AUG + ORF). For both RPS12 and CR3, the majority of reads (>80%) were completely unedited (Table 7). CR3, which has previously been shown to be preferentially edited in the BS stage, had the lowest percentage of fully edited transcripts, with only 0.1% – 0.2% translatable transcripts detected in both cell lines and under both growth conditions. In contrast, while RPS12 had similar levels of unedited transcripts, a larger percentage of translatable transcripts were found. For this essential transcript, the number of fully edited transcripts differed between the two different cell lines; 2.3% in TREU 667 and 0.9% in EATRO 164. Growth of the EATRO cells in low glucose media (SDM-80) did result in a substantial jump in the both the number of transcripts that initiated the editing process, and the number of fully edited transcripts (4.16%). This suggests that energy source may influence editing efficiency. While the predominance of completely unedited transcripts found for both CR3 and RPS12 was surprising, these numbers are in line with those found in other studies [131,148,149].

The ND7 5’ transcriptome analyses differed substantially from both RPS12 and CR3 in that the majority of these transcripts had initiated the editing process. The TREU cell line showed the highest editing efficiency with ~80% of transcripts having initiated editing and 9.7%

Table 7. Editing efficiencies of RPS12 (A), ND7 (B), and CR3 (C).

Transcript	Cell line and Media	Total # Reads	% unedited	% partial edited	% fully edited
RPS12	Treu 667, SDM79	787,584	89.8%	7.9%	2.3%
RPS12	Eatro 164, SDM79	846,549	92.6%	6.5%	0.9%
RPS12	Eatro 164, SDM80	1,381,092	81.3%	14.5%	4.2%
ND7 5’	Treu 667, SDM79	1,141,322	20.3%	70.0%	9.7%
ND7 5’	Eatro 164, SDM79	915,610	47.0%	52.8%	0.2%
ND7 5’	Eatro 164, SDM80	313,657	27.4%	72.1%	0.5%
CR3	Treu 667, SDM79	18,832	84.9%	15.0%	0.1%
CR3	Treu 667 enriched.	50,589	18.1%	73.1%	8.8%
CR3	Eatro 164, SDM79	348,210	93.2%	6.6%	0.2%
CR3	Eatro 164, SDM80	53,000	90.6%	9.3%	0.1%

of the transcripts fully edited and translatable. In contrast, in EATRO cells, only 53% of the transcripts had initiated editing, and a scant 0.2% had completed the editing process. As with RPS12, we did see an increase of efficiency in the cells grown in SDM80, with over 70% initiating editing. However, even with the large increase in initiation of the editing process, only a scant 0.5% of transcripts were fully edited (Table 7). The sharp drop in the ability to complete the editing process appears to be due to loss of an optimal gRNA for one region of this transcript (described below).

Editing Cascade and Reading Frame Analyses

In order to determine if the low editing efficiencies were due to any one step in the editing cascades, a full analysis of each editing step was done. For these analyses, we developed a pipeline that used our gRNA database to distinguish true alternative edits from both mis-edited and partially edited transcripts. This pipeline uses two programs, Segmentation of Kinetoplast Edited Transcripts to Characterize Editing Heterogeneity (SKETCH), and the gRNA database search program previously described [9]. The SKETCH program analyzes segments of transcripts that are defined by the relative range of coverage of each gRNA population used in

conventional editing patterns. Block sequences are compared to both the unedited sequence and the fully edited conventional sequence and then classified into unedited, fully edited and "unknown" blocks. Once the most abundant sequences of all segments are identified, transcripts containing each most abundant "unknown" sequences are used as queries against the gRNA database. If a gRNA is identified that can generate the edit, the sequence is considered a true alternative edit. If no plausible gRNA is identified the edit is considered a misedit or a junction, depending on the sequence and the status of other segments on that transcript with this sequence. By examining segments of transcripts independently, we were able to identify both branching and converging editing pathways.

RPS12 Analysis

As expected, the essential RPS12 showed the most robust editing path. In all three analyses, the majority of transcripts used the same series of 10 gRNA populations (A – J) (For full gRNA sequences and alignments see APPENDICES I and J). Use of the final gRNA population (gJ) in the cascade lead to only the RPS12 ORF, and we found no evidence of an alternative AUG or frameshift leading to utilization of the second ORF. We do note that there is a downstream start codon, that if translated, would be read in the alternative reading frame (APPENDIX J). While the editing cascades were relatively straight forward, we did see some minor deviations (Figure 14). Editing of block B could utilize a number of different gRNAs, including several that were used in one cell line only (dashed arrows). gRNAs B1 and B1* are variants of the same gRNA, with gB1* introducing a single amino acid (aa) change (V/Y) (Figure 15). While editing using the TREU specific $gB3^t$ and $gB4^t$ gRNAs lead to a distinct editing "dead end" (dead end = disruption of the next canonical anchor sequence, and no detection of any further editing), the

EATRO specific gB2^{e} did not disrupt the editing cascade. Use of this variant, however, did introduce a frameshift seven amino acids (aa) from the C-terminus (Figure 15). Because gB2^{e} did not disrupt editing, a significant percentage (5.3% in SDM79 and 7.6% in SDM80) of translatable RPS12 transcripts did contain the alternative C-terminus (J2 transcripts). This alternative C-terminus was previously reported in the 29-13 strain [148], however, it appears to be absent in the TREU cell line.

A drop in editing efficiency was seen at the D to E block transition due to the incorrect utilization of the gF^{p} guide RNA (D^{x}) that disrupted the editing cascade (Table 8). While mis-editing by gF^{p} was limited in the TREU 667 cell line (7.1% of D-block edited transcripts), it's use was much more prominent in the EATRO cell line (17.5%), leading to a significant drop in transcripts that could continue past D-block editing. Interestingly, growth in SDM80 lead to a significant increase in mis-editing by gF^{p}, with over 32.5% of trancripts using gF^{p} incorrectly, resulting in a significant portion of dead-end transcripts. The EATRO cell line had additional minor dead-end pathways at the D to E transition. Misediting by a ND7 gRNA (gE^{ep}) again disrupted any further editing, and mis-anchoring by the gE guide RNA (marked with box m) also led to the generation of an anchor sequence that could be used by a ND8 gRNA (gF^{ep}) disrupting any further editing. Interestingly, the editing efficiency did not drop as transcripts transitioned to the next block of editing (Table 8). In EATRO-SDM80 cells, the editing efficiency at level F is ~5.6%, and at level G it actually increases to 5.9%. Editing efficiency at the block level is

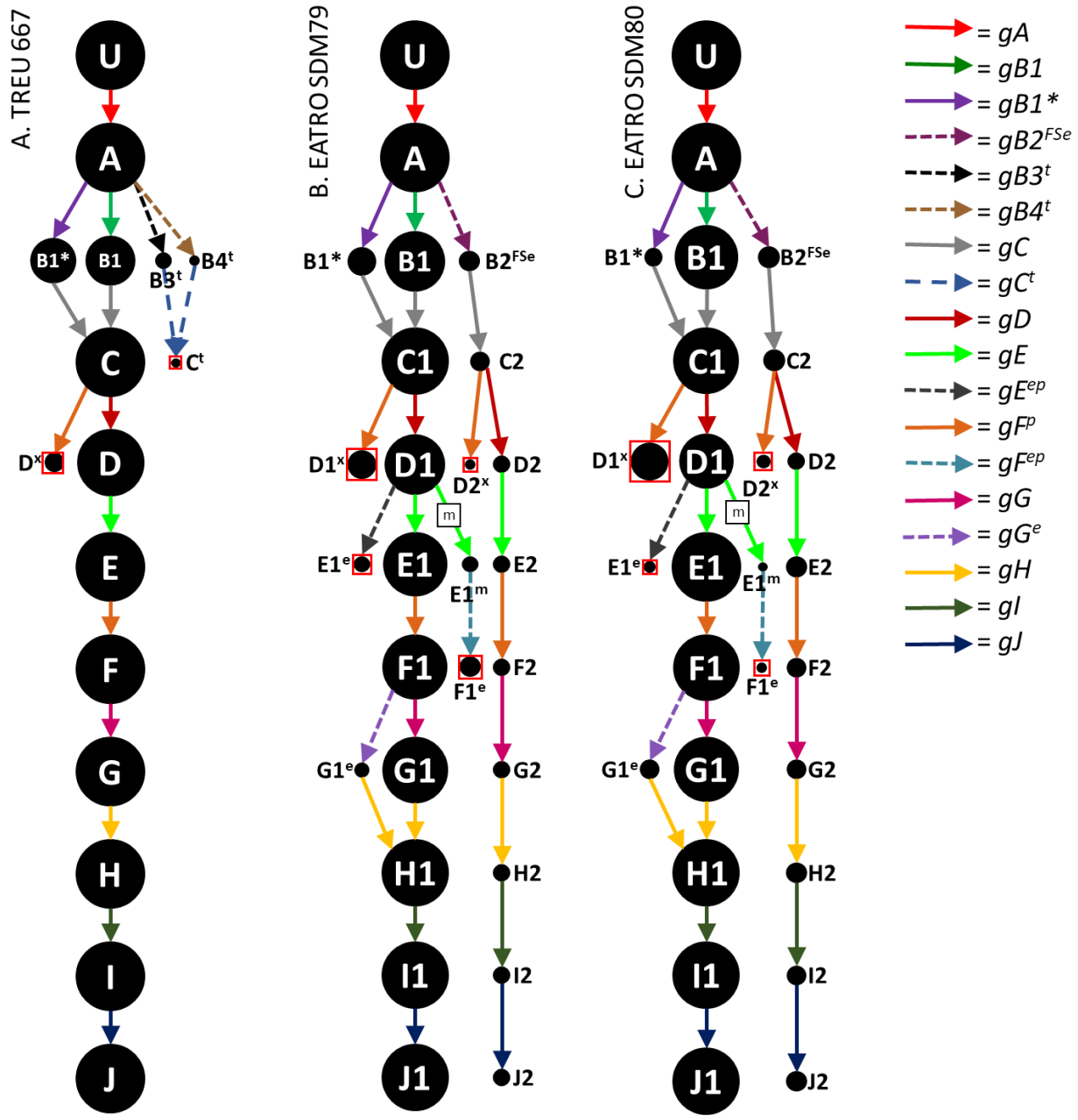


Figure 14. Observed RPS12 editing pathways in the TREU 667 cell line (A) and the EATRO 164 cell line grown in SDM79 (B) and SDM80 (C). U = unedited transcripts. Dot sizes are proportional to the percent of block level edited transcripts using the gRNA indicated. Colored arrows indicate the gRNA population used. Dashed arrows with closed heads represent gRNA populations used in only one cell line (superscript 'e' or 't'). gRNA names with superscript 'p' represent promiscuous gRNAs. Dots enclosed by a red box represent end point mRNAs with no AUG start codon. gF^{p} is a promiscuous gRNA that edits both in the D and F editing block of RPS12. Arrows with a boxed 'm' represent a gRNA that has mis-anchored.

calculated based on the number of transcripts that match any of the fully edited sequences in that block, regardless of the condition of earlier blocks. Analysis of editing intermediates suggest that this increase occurs due to the ability of the downstream gRNA (the gF^{p} population) to overwrite transcripts that have been previously edited through the G-level. Because of the overwriting gF^{p} population, mRNAs exist that are fully edited at the G editing block but are in a transition state in block F.

```
T. brucei B1          --MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWMLRCFFCC-FIGLVMSLFIIEGGGFVDLPGVKYYTRIVS---------
T. brucei B2FSe       --MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWMLRCFFCC-FIGLVMSLFIIEGGGFVDLPGYKILFTYCKLDLDIRYVF
T. vivax              --MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWLLRCFFCCVFFGLHLSLFIIEGGGFVDLPGIKYYTRMFIN--------
L. tarentolae         MRVLFLYGLCVRFLYFCLVLYLSPRLPSSGNRRCLYAICYMFNILWFFC-VFCCVCFL-NHLLFIVEGGGFIDLPGVKYFSRFFLNA-------
L. donovani           VRVLYLYGLCVRFLFFSLVLYLSPQLPSSGNRRCLYAISIMFNILWIFL-VFCCVFFV-VHLLFIVEGGGFIDLPGVKYFSRFFCKS-------
L. amazonensis        VRVLYLYGLCVRFLFLCLVLYLSPRLPSSGNRRCLYAISIMFNILWYFL-VFCCFVFV-IFQLFIVEGGGFIDLPGVKYFSRFCNVS-------
                        : :*** *:**: : *  *:**:*** **** ***:  ::*::* :  .***  :     ***:*****:**** *
```

Figure 15. Alignment of RPS12 proteins from *T. brucei*, *T. vivax*, *Leishmania tarentolae*, *Leishmania donovani*, and *Leishmania amazonensis*. Asterisks indicate identical residues, colons indicate conserved residues. Highlighted amino acids are changes introduced by alternative editing (S>P, gG^{e}, V>Y gB1*) Alignment was generated by Clustal Omega [116]. RPS12 signature sequence is shown in bold [150].

Table 8. Editing efficiency for each RPS12 gRNA population. Percentages were calculated based on the number of transcripts that had completed each editing level out of the total number of RPS12 transcripts.

Block	Percent complete editing of block		
	TREU 667 (SDM 79)	EATRO 164 (SDM 79)	EATRO 164 (SDM 80)
Initiated Editing	10.2	7.4	18.7
A	8.9	4.6	15.9
B	6.9	4.4	14.6
C	6.0	4.2	13.5
D	4.9	3.1	11.7
E	4.5	2.2	6.9
F	3.8	1.4	5.6
G	3.7	1.4	5.9
H	3.7	1.3	5.7
I	3.4	1.2	5.4
J	2.3	0.9	4.2

The only other minor variation was the use of the EATRO specific gG^{e} guide that occurs in a highly cytosine-rich region (Figure 16). Previous examinations of the gRNA coverage in this

region identified only rare gRNAs with multiple C:A basepairs, alignment mismatches and with gaps between adjacent gRNAs [9,119]. While this analysis did extend the identified gRNA population and eliminated the gap region, we did not identify either mRNA sequences or gRNAs that improved the alignment mismatches (Figure 16). The use of alternative base pairs is not unheard of. A study of *in vitro* deletions found that alternative base pairs such as C:A, C:U, and C:C were tolerated to varying extents [151]. Interestingly, this portion of RPS12 encodes the signature sequence, which is nearly universal [150]. Use of the gG^{e} variant gRNA results in a single point mutation, substituting a proline in place of a serine within this important sequence.

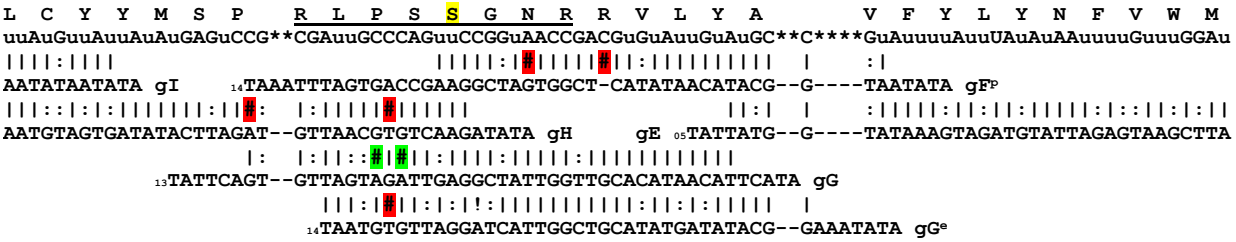


Figure 16. Regions with poor gRNA coverage and functionally conserved residues in RPS12. Functionally important aa residues are underlined [150]. Pipes (|) indicate Watson/Crick base pairs and colons (:) indicate G/U base pairs. Red highlighted hashtags (#) indicated gaps or mismatches, green highlighted # indicate C:A basepairs. The introduced substitution mutation introduced by use of the gG^{e} gRNA is highlighted in yellow (S>P).

ND7 5' analysis

Analyses of the ND7 5' targeted transcriptomes, indicate that full editing of the 5' domain requires five gRNA populations for both cell lines (Figure 17, APPENDIX K). Two variants of the terminal population (gE1 and gE2) were identified that resulted in different 5' terminal editing patterns (APPENDIX L). Translation of these editing patterns yields two different protein products in two different reading frames, one (RF1) encoding the canonical ND7 protein (E1) and the other (RF3) encoding a putative metabolite transporter (E2, see below) [141]. While transcripts for both open reading frames were found in both cell lines, there were notable

differences in the populations. The TREU 667 cell line had the highest editing efficiency with over 80% of the transcripts initiating the RNA editing process and ~9.7% of the transcripts fully edited through Block E. Use of the gE1 or gE2 gRNAs appeared to be equally efficient, resulting in nearly equal amounts of RF1 and RF3 fully edited transcripts. A small percentage of transcripts (4.9% of transcripts that completed block E editing) were observed that appeared to be mis-edited by a TREU specific gRNA (gE4^{t}), leading to a dead-end product (no ORF). In addition, gE4^{t} also appeared to be able to overwrite editing directed by gE2, to generate a small number of transcripts that could be translated in RF2 (pink E3^{t}).

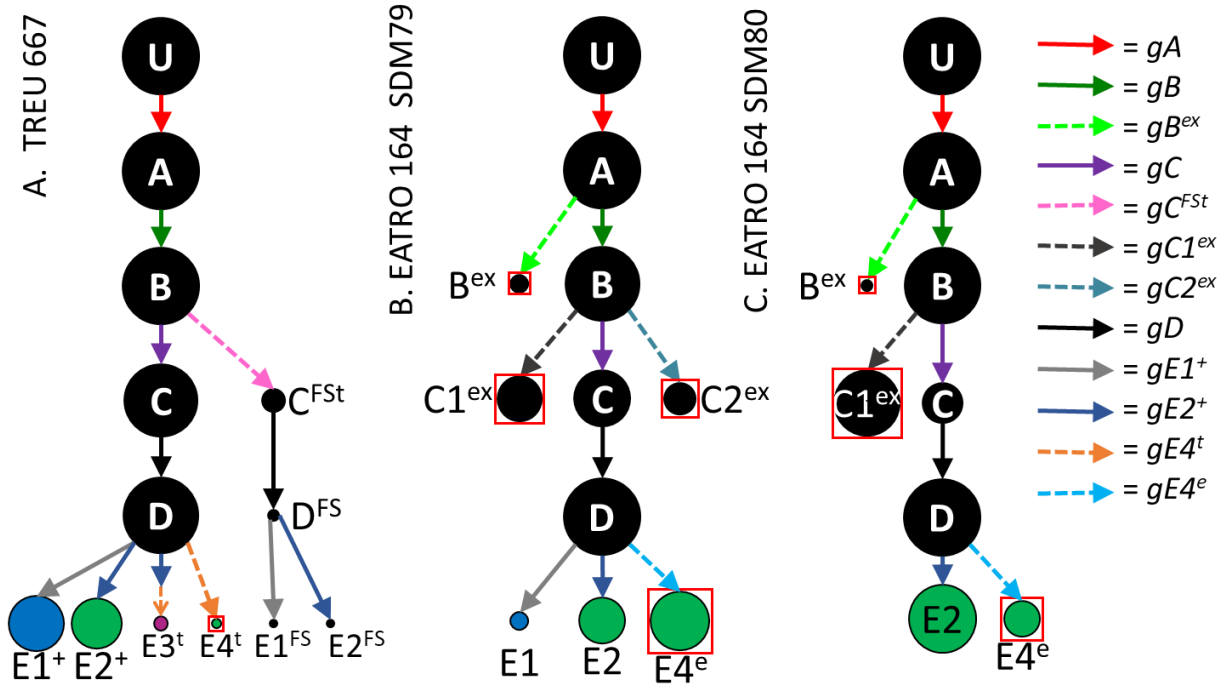


Figure 17. Observed ND7 5' editing pathways in the TREU 667 cell line (A) and the EATRO 164 cell line grown in SDM79 (B) and SDM80 (C). U = unedited transcripts. For arrow and gRNA naming descriptors see Figure 14. Dots enclosed by a red box represent end point mRNAs with no AUG start codon. + indicates that more than one mRNA form was condensed into this circle to simplify the figure (See APPENDIX M). Condensed forms encode largely the same amino acid sequence with only small variants. Terminal dots are colored blue for reading frame 1, magenta for reading frame 2, or green for reading frame 3. Boxed green dots have no functional start codon, but are translatable into reading frame 3 with the use of an alternative start codon (UUG).

While the number of "dead-end" pathways were very limited in the TREU cell line, use of the gC guide RNA population appeared to be very inefficient, resulting in a large drop in the percent of Block C-edited transcripts (25.8% drop, Table 9). A mutant gC gRNA (gC^{FSt}), did result in a small percentage of transcripts with a frameshift C-terminus. Interestingly, while 9.1% of C block transcripts used the gC^{FSt} gRNA, only 2.4% of the transcripts that have completed D-block editing come from this minor branch. This suggests that this alternative edit decreases the efficiency of use of the subsequent gRNAs. In contrast, full editing of the ND7 5' domain in the EATRO 164 cell line was very inefficient. While transcripts were able to initiate the editing process relatively efficiently (~50 – 70%, dependent on growth medium used), less than 1% of ND7 transcripts were fully edited at level E (Table 7). This appears to be due to the use of several EATRO specific gRNAs that disrupt further editing (Table 9). Again, the largest drop in editing efficiency occurred at the B to C-block transition. In addition, the EATRO specific use of gB^{ex}, $gC1^{ex}$ and $gC2^{ex}$ all disrupted the editing cascade (Table 9). This compounded the editing efficiency problem, with a majority of C-block edited transcripts (47% in SDM79 and 73.4% for SDM80), no longer editing competent. The 5' end of ND7 has multiple AUG sequences not created by the editing process. Translation predictions of these editing blocked transcripts ($C1^{ex}$, $C2^{ex}$) indicate that they do have ORFs that extend through the HR3 region. The protein product of B^{ex} transcripts is in the ARF, but is ten amino acids shorter, while the $C1^{ex}$ and $C2^{ex}$ products, which translate in the canonical ND7 reading frame, produce proteins that are both three amino acids shorter. Further drops in efficiency occurred due to an anchor mis-match (A:A) found in the gD guide RNA population (Figure 18, APPENDIX L). While the gD mutation is also observed in TREU, this cell line contains a sizable population of non-mutated gD guide

RNAs. Editing by the gE4[e] guide, results in a transcript with no in-frame AUG. However, translation of this transcript (E4[e]) in RF3 has no stop codons and we cannot rule out the possibility of a non-canonical START codon.

Table 9. Editing efficiency for each ND7 5'domain gRNA population. Percentages were calculated based on the number of transcripts that had completed each editing level out of the total number of ND7 transcripts.

Block	Percent complete editing of block		
	TREU 667 (SDM 79)	EATRO 164 (SDM 79)	EATRO 164 (SDM 80)
Initiated Editing	79.7	52.4	72.6
A	45.8	47.0	68.5
B	44.7	45.8	66.7
C	18.9	13.4	16.9
D	11.7	0.4	1.1
E	9.7	0.2	0.5

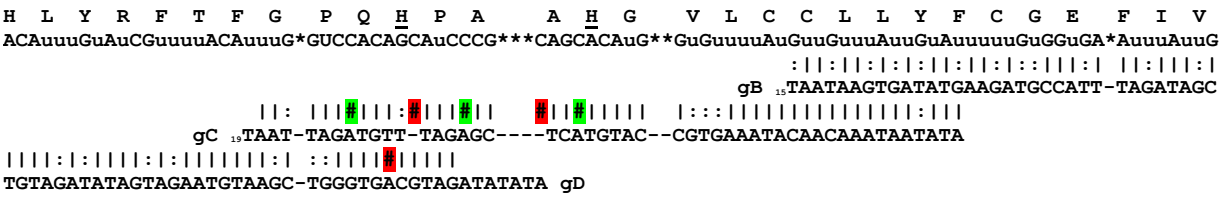


Figure 18. Regions with poor gRNA coverage and functionally conserved residues in ND7'5'. Functionally important residues are underlined [152]. Pipes indicate Watson/Crick base pairs and colons indicate G/U base pairs. Red highlighted hashtags (#) indicate gaps or mismatches, green highlighted # indicate C/A base pairs.

Similar to RPS12, ND7 5' has a cytosine-rich region with poor gRNA coverage (Figure 18). This cytosine-rich region contains two conserved residues involved in ND7 function and coincides with the C level of editing in the editing pathways where the largest drop in editing efficiency is observed (Table 9) [152]. The gC gRNA population is relatively rare (only 114 reads found in the TREU gRNA transcriptome, and 6185 reads in EATRO) and has 5 nt mismatches with the conventional ND7 sequence (including C:A basepairs; Figure 18).

CR3 Analysis

Previous work indicated that C-Rich region 3 is a putative Complex I member and that it is preferentially edited in the Bloodstream stage [47,120]. However, CR3 gRNAs are present in both life cycle gRNA transcriptomes, and PCR amplification of 5' edited transcripts were successfully cloned and sequenced in the TREU 667 procyclic cell line [9,119,141]. These studies indicated that multiple forms of the mRNA did exist that used different reading frames suggesting that CR3 is dual-coding and that it is selection of the terminal gRNA that determines which reading frame will be used [141]. In this study, we used primers flanking the editing domain in order to analyze the entire CR3 sequence. Interestingly, while 15% of the TREU CR3 transcripts had initiated the editing process, only 2.2% had completed editing by the initiating gA guide RNAs (Table 10). This suggests that the large drop in editing efficiency occurs due to incomplete editing by the block A guides. These gRNAs are fairly abundant, and we see no alignment issues, so it is unclear why editing of Block A is so inefficient (APPENDIX N).

Table 10. Editing efficiencies by block level of CR3. Transcripts whose gRNAs covered two blocks (DEs and FGs) were included in both blocks for this calculation.

Level	Percentage Complete			
	TREU 667	TREU 667 (Enriched)	EATRO 164 (SDM 79)	EATRO 164 (SDM 80)
Initiated editing	15.1	81.9	6.8	9.4
A	2.2	68.3	1.9	2.3
B	1.8	56.6	1.2	1.8
C	1.7	54.6	0.9	1.5
D	1.6	52.0	0.6	1.1
E	1.2	39.1	0.6	1.0
F	0.8	22.3	0.2	0.4
G/FG	0.4	8.8	0.2	0.3

While the percentage of fully edited transcripts was very low percentage (0.2 – 0.4%, Table 10), we were able to again identify the major 5' alternative editing patterns that direct translation to either the ORF or to the +1 Alternative Reading Frame (RF2). To increase the robustness of the analyses, we also generated a biased CR3 transcriptome, by size selecting for longer transcripts during the amplification process. Analyses of the TREU transcriptome indicates that the full CR3 editing pathway has multiple branches, resulting in a total of 12 major forms of fully edited CR3 (Figure 19). These 12 forms are comprised of three major 5' editing patterns, paired with any of four different 3' editing patterns. The two initiating gRNAs identified (gA1 and gA2), direct identical editing patterns except gA2 inserts an additional three U- 1 phenylalanine). The gB guide RNAs all anchor in different areas (Figure 20A) and do introduce substantial AA changes near the 3' end (Figure 20B). However, all gB guide RNAs generate the anchor binding site (ABS) that is recognized by gC, hence all 4 nodes merge to a common sequence guided by gC and gD (APPENDIX N).

The 5' end editing patterns begin to diverge after Block D editing. FG^{tx} transcripts are generated by the use of two subsequent gRNA populations, gE^{t} and gFG^{txp}. gFG^{txp} is a promiscuous gRNA (previously identified as a ND7 gRNA) that spans both the F and G editing blocks. These transcripts were more abundant than both G^{t} and FG^{t}, however final editing using this gRNA does not generate a AUG start codon. It has been proposed that trypanosomes can use UUG as an alternative start codon, thus we cannot rule out the possibility that FG^{tx} transcripts can be translated (Figure 20B). Analyses of intermediates suggest that the gE guide (red arrow) can in fact "overwrite" gE^{t}, indicating that a proportion of these may still be re-edited into other forms. Editing via the gE population required an additional gRNA to generate

the anchor for either gF^t or gFG^{tp}. Generation of G^t transcripts (canonical CR3) requires 2 additional gRNAs, while FG^t (+1 ORF) transcripts are generated by a single gRNA population (gFG^{tp}), another promiscuous gRNA (CR4).

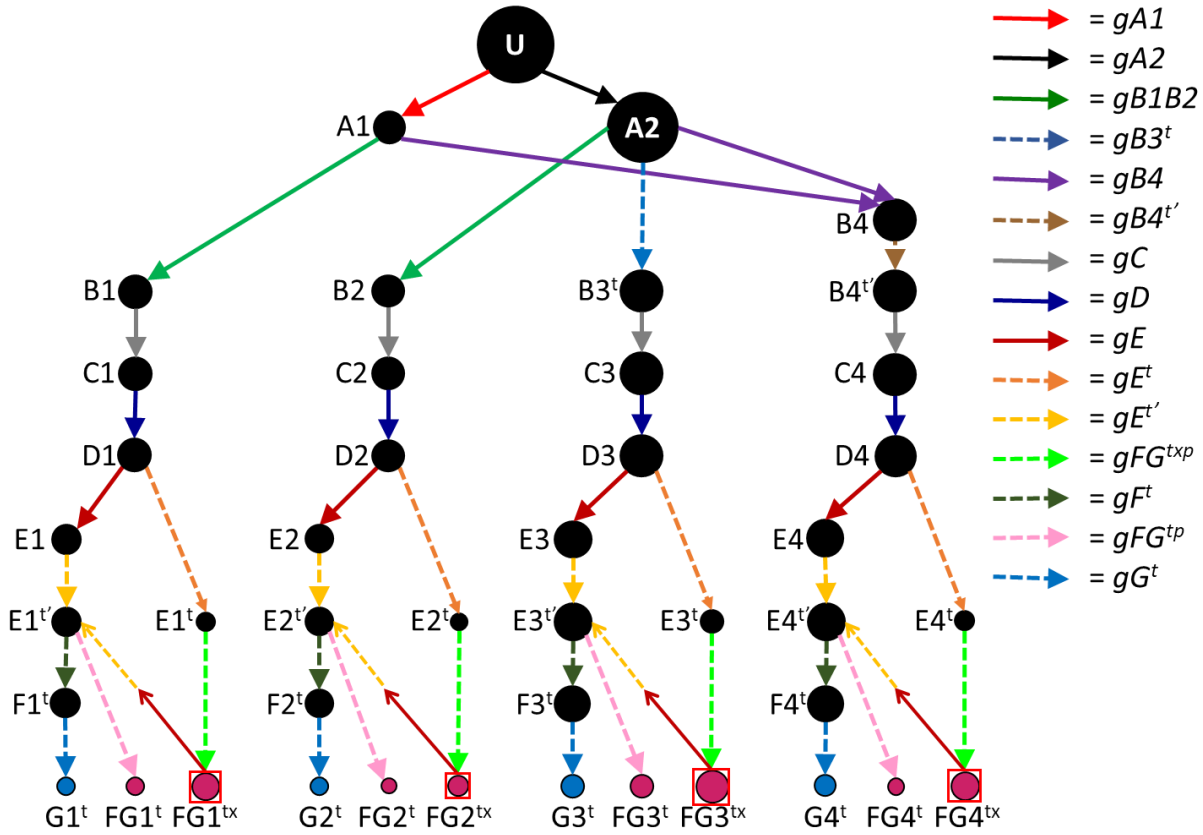


Figure 19. Observed CR3 editing pathways in the TREU 667 cell line. U = unedited transcripts. For arrow and gRNA naming descriptors see Figure 14. Dots enclosed by a red box represent end point mRNAs with no AUG start codon. Terminal dots are colored blue for reading frame 1 or magenta for reading frame 2. Boxed magenta dots have no functional start codon, but are translatable into reading frame 2 with the use of an alternative start codon (UUG).

Surprisingly, when we examined the editing pathways of CR3 in the EATRO 164 libraries, we discovered that while three of the four initial 3' editing patterns were found in this library, editing beyond those patterns was completely divergent (Figure 21, APPENDIX O). A completely different set of gRNAs were used to generate fully edited CR3 transcripts (APPENDIX

P). The divergent pathway did show some superficial similarities to the editing patterns observed in

Figure 20. Four different 3' end sequences found in the TREU 667 transcriptome for the CR3 transcript (A) and CR3 protein sequences (B). U-residues inserted by editing are indicated by lowercase; different sequences created by the different gRNAs are highlighted in RED. Thick underline sequence indicates the anchor binding site (ABS) for the initiating gRNAs (*gA1* and *gA2*). **Green** = ABS for *gB1B2*; **Blue** = ABS for *gB3*; **Purple** = ABS for *gB4*. Bolded amino acids show sequence variants and shaded sequence shows position of predicted transmembrane domains [134]

TREU 667 cells. While both the B1 and B2 transcripts were directly edited by gC^e, the B4 transcripts required an additional gRNA to generate the ABS recognized by gC^e. In EATRO cells, B4 transcripts could be edited by 3 different gRNAs ($gB5^e$, $gB6^e$ and $gB7^e$). While $gB7^e$ disrupted editing, both $gB5^e$ and $gB6^e$ generated the anchor that could be used by either gC or gC^e. Surprisingly, while the conventional CR3 gC guide RNA was clearly used by B4 transcripts, we saw no evidence of its use in the B1/B2 pathways. Transcripts using gC could be further extended by both gD and gE guide RNAs, however, no evidence of editing beyond the gE guides was observed. In contrast, use of the alternative gC^e guide RNA population, could be extended by a series of additional guide RNAs, generating transcripts with functional AUG start codons.

However, many of the gRNAs used were promiscuous, in that they had been previously identified as gRNAs of other transcripts. As with the TREU editing pathway, we observe transcripts capable of being translated in two reading frames with the FG^{e} mRNAs translating in RF1, and the FG^{e*} mRNAs translating in RF2 (Figure 21A, Figure 22). In addition, the G^{e} mRNAs, while not having a functional "AUG" do translate into RF2 if the first "UUG" is used. As with ND7, we observed a shift in editing pattern preference when the EATRO 164 cells were changed from SDM79 medium to SDM80. Interestingly, a new fully edited form of CR3 appeared in the EATRO164 SDM80 library only. The gRNA gC^{e80} is used in the EATRO SDM79 pathway, but editing appears to cease here. Cells grown in SDM80 continue this editing pathway with two additional gRNAs, gDE^{e80} and gFG^{e80} (Figure 21B). This mRNA is translatable, but produces a distinctly different and shorter protein product (Figure 21A). The protein products of the two different cell lines are highly dissimilar. Using bioinformatics tools to predict the secondary structure of these proteins, we find that the difference is most noticeable in the RF1s of the two cell lines (Figure 23). Interestingly, the RF2s have a very similar predicted secondary structure. This evidence suggests that the two different cell lines are able to use the CR3 transcript to create distinctly different protein products.

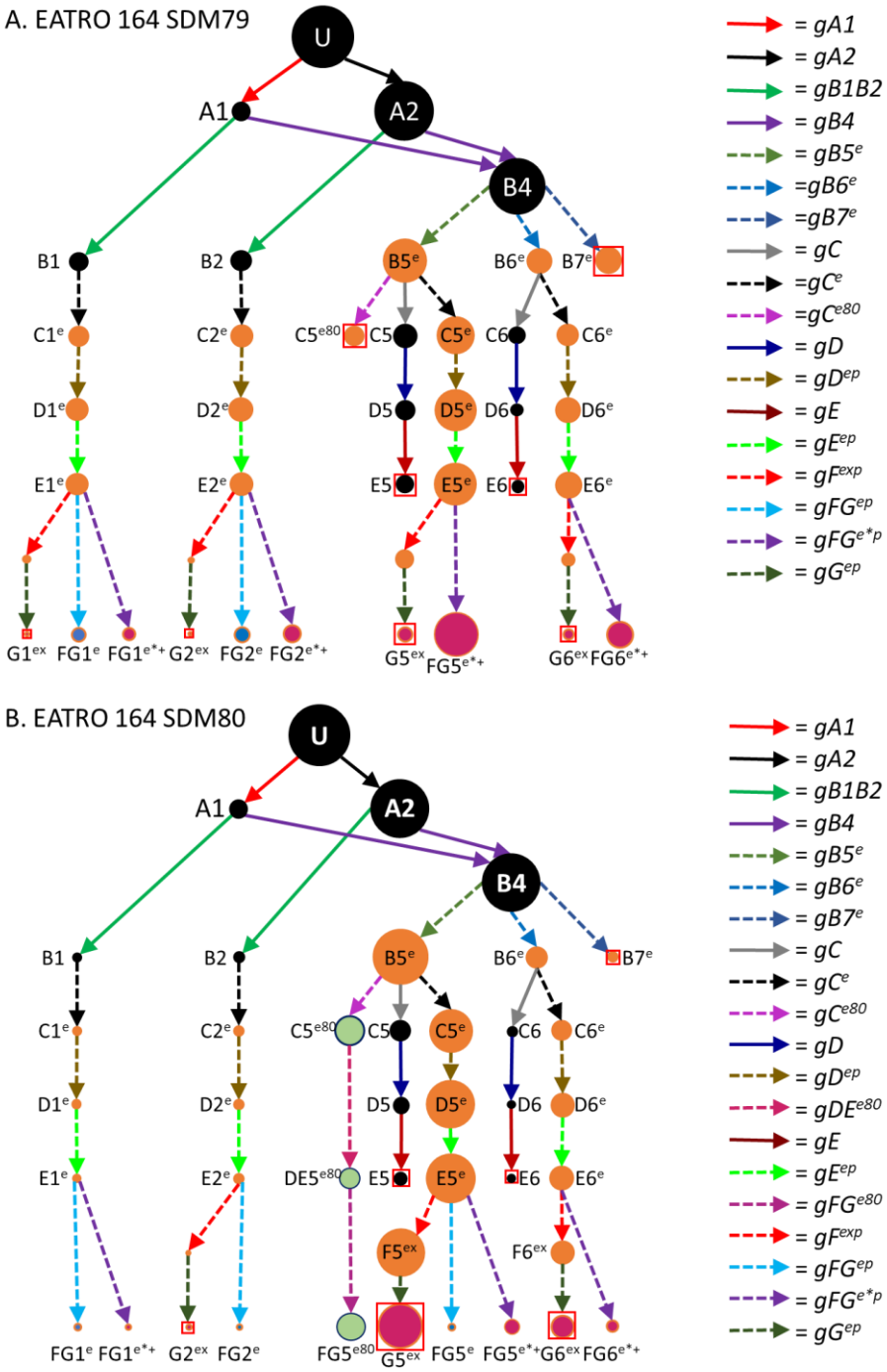


Figure 21. Observed CR3 editing pathways in the EATRO 164 cell line grown in SDM79 (A) and SDM80 (B). U = unedited transcripts. For arrow and gRNA naming descriptors see Figure 14. Dots enclosed by a red box represent end point mRNAs with no AUG start codon. Terminal dots are colored blue for reading frame 1 or magenta for reading frame 2. Boxed magenta dots have no functional start codon, but are translatable into reading frame 2 with the use of an alternative start codon (UUG). + indicates that more than one mRNA form was condensed into this circle to simplify the figure (See Figure 22). Condensed forms encode largely the same amino acid sequence with only small variants.

```
ORF
FG1e       MCMIYKYYHICVRWDFGDHCLFGCYELYFMFCYGYCFLFNMGLLLCF--ILQIFSVII-IIVYKFSLLD
FG2e       MCMIYKYYHICVRWDFGDHCLFGCYELYFMFCYGYCFLFNMGLLLCF--ILQIFSVIIFIIVYKFSLLD
FG5e       MCMIYKYYHICVRWDFGDHCLFGCYELYFMFCYGYCFLFNMGLLLCLYYLCIFIVVIIFIIVYKFSLLD

ARF
FG1e*v1    MCMIYKLTIVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLS-LLLYISFRY
FG2e*v1    MCMIYKLTIVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLSFLLLYISFRY
FG5e*v1    MCMIYKLTIVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYITYVFLLLLSFLLLYISFRY
FG6e*v1    MCMIYKLTIVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYIILCIFIVVIIFIIVYKFSLLD
FG1e*v2    MCMIYKLTYVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLS-LLLYISFRY
FG2e*v2    MCMIYKLTYVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLSFLLLYISFRY
FG5e*v2    MCMIYKLTYVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYITYVFLLLLSFLLLYISFRY
FG6e*v2    MCMIYKLTYVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYIILCIFIVVIIFIIVYKFSLLD
FG1e*v3    MCMIYKNIFVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLS-LLLYISFRY
FG2e*v3    MCMIYKNIFVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLSFLLLYISFRY
FG5e*v3    MCMIYKNIFVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYITYVFLLLLSFLLLYISFRY
FG6e*v3    MCMIYKNIFVLGGILVIIVYLVVMSCILCFVMVIVFYLIWVYCCVYIILCIFIVVIIFIIVYKFSLLD
G1ex          LLFGVLFLICFVYFIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLS-LLLYISFRY
G2ex          LLFGVLFLICFVYFIVYLVVMSCILCFVMVIVFYLIWVYCCVL-YYRFL-VLSFLLLYISFRY
G5ex          LLFGVLFLICFVYFIVYLVVMSCILCFVMVIVFYLIWVYCCVYITYVFLLLLSFLLLYISFRY
G6ex          LLFGVLFLICFVYFIVYLVVMSCILCFVMVIVFYLIWVYCCVYIILCIFIVVIIFIIVYKFSLLD

EATRO SDM 80 Only FGe80
MCMIYKNNGSCGFVGWFRLGYCYCECCSFCMIIL
```

Figure 22. Alignment of CR3 predicted protein variants from the EATRO 164 cell line. Bolded amino acids show sequence variants and shaded sequence shows position of predicted transmembrane domains [134].

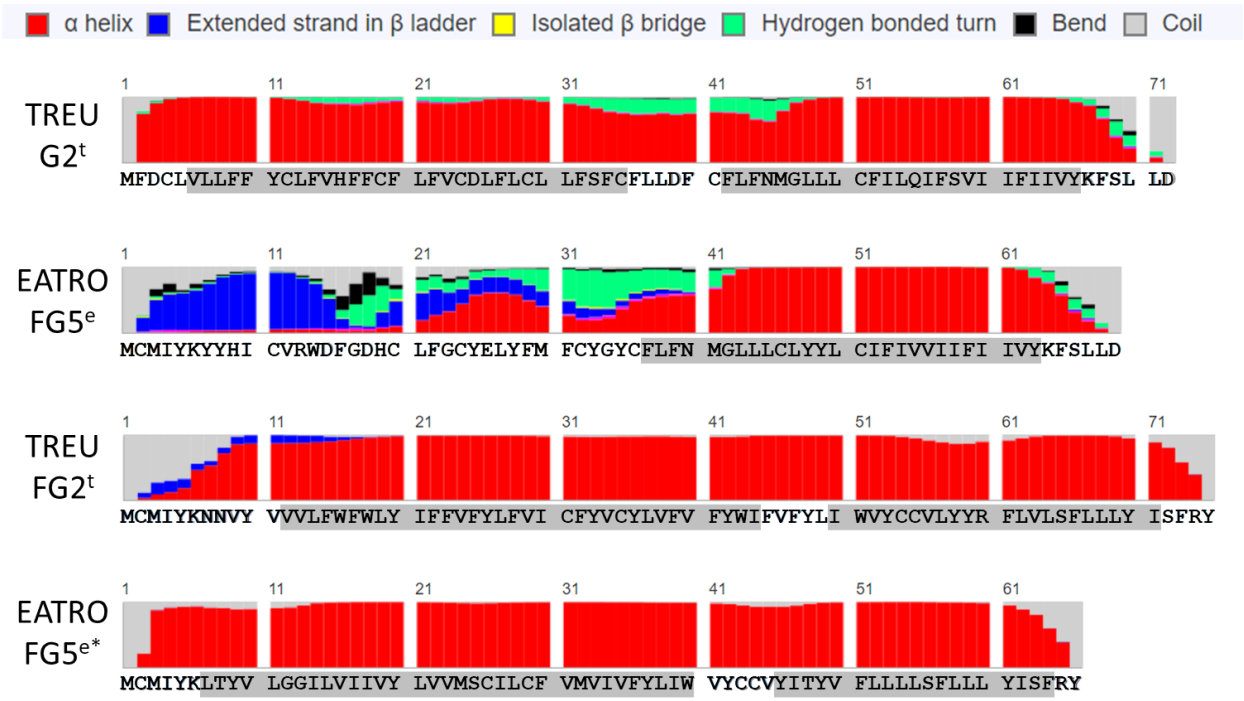


Figure 23. Predicted secondary structures of most abundant CR3 predicted proteins. Secondary structure predictions were generated by RaptorX [153–155]. Shaded regions indicate predicted transmembrane alpha helices predicted by Phobius [134].

Deletion of Encoded Uridines Directed by the Poly-U Tail

During the analyses of the mRNA deep sequencing data, we observed many partially edited transcripts where the deletion of encoded uridines appeared to occur early (prior to 3' insertion events). We hypothesize that the poly-U tail of preceding gRNAs may direct the removal of these encoded uridines. To examine this hypothesis, we examined partially edited RPS12 and ND7 5' transcripts with early deletion events, and determined their relative proximity to the preceding gRNA (Figure 24). (CR3 was excluded from this analysis due to the variability in gRNA location caused by the multiple branching editing pathways). These analyses revealed that for both RPS12 and ND7 5', encoded U sites that are near the poly-U tail of the preceding gRNA (U-Tail accessible sites) have a higher frequency of total deletions in the partially edited transcripts. This suggests that the proximity of the preceding gRNA's poly-U tail does impact deletion of encoded uridines and supports our hypothesis that poly-U tails can guide the deletion of encoded uridines.

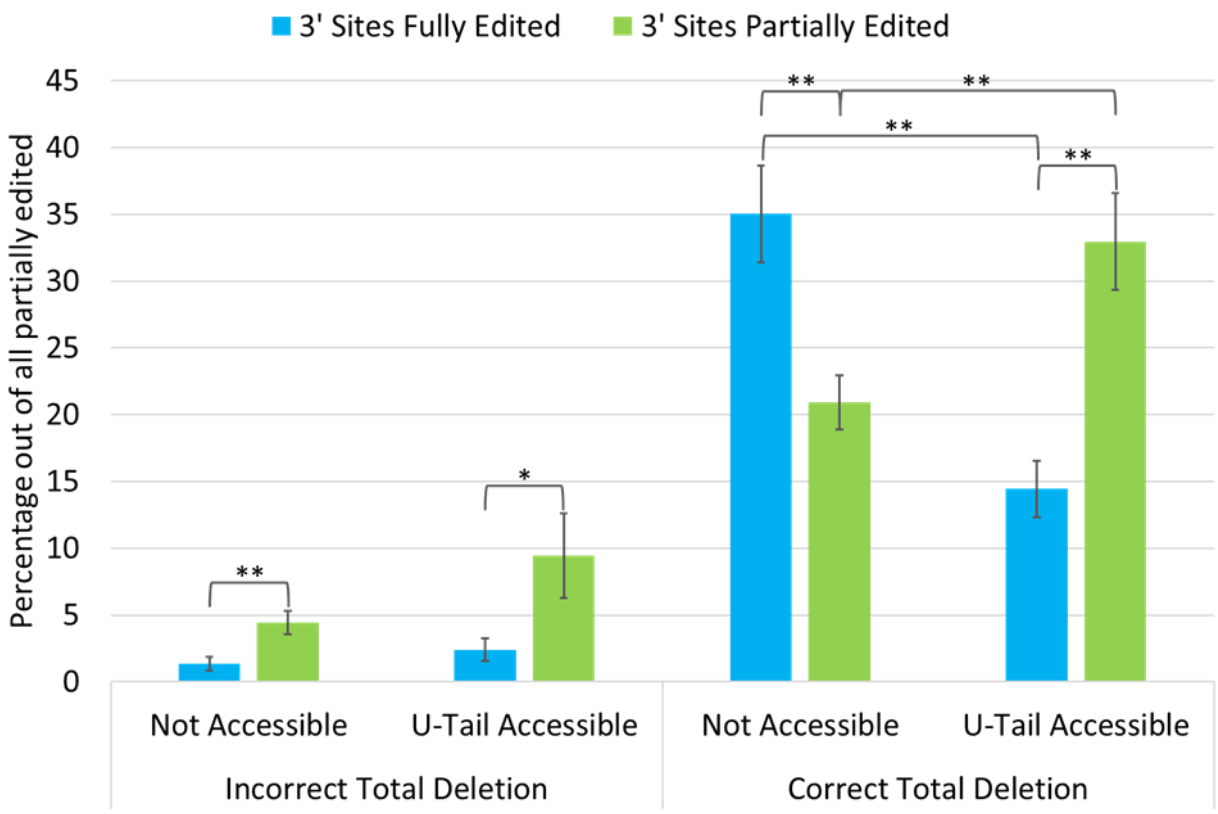


Figure 24. Frequencies of early total deletions of DNA encoded uridines in partially edited ND7 and RPS12 transcripts. Editing sites with DNA encoded Us were identified, and each encoded U site was then characterized based on the editing access of the preceding gRNAs' poly-U tail (U-Tail Accessible or Not Accessible) as well as whether the site in question was totally deleted in the final fully edited sequence (Correct Total Deletion or Incorrect Total Deletion). For each site, a window was defined consisting of the 6 sites upstream and 6 sites downstream of the site in question. Each editing sequence was then classified based on the states in the 3' end of the window as either matching the fully edited sequence or not. Error bars depict standard error. (*=p<0.05 **=p<0.01 unpaired T-test)

Discussion

In this work, we developed a new transcriptome analysis pipeline to fully characterize the editing pathways for three putative dual-coding genes, RPS12, ND75' and CR3. The pipeline uses a new program, SKETCH, in combination with our database search program [9]. Combining these two programs allowed us to separate true alternative edits from partially edited transcripts and allowed the precise mapping of the full progression of the editing process. This

characterization was done in two different cell lines (TREU 667 and EATRO 164) and under different energy conditions in order to determine the robustness of the editing process. Surprisingly, distinct differences in both editing progression as well as editing efficiency were observed in the two different cell lines. In addition, growth of parasites under different energy conditions also appeared to be able to influence the editing process. In both cell lines, the editing process appeared to be very inefficient, with most of the transcripts completely unedited. A comparison of the two cell lines grown in SDM79 did suggest that overall, the TREU 667 cells were more efficient in editing these three pan-edited transcripts. However, when the EATRO 164 cells were transferred from a high glucose medium (SDM79) to a glucose-restricted medium (SDM80), the number of transcripts that initiated the editing process more than doubled. For RPS12, the increase in editing initiation resulted in a 4-fold increase in the number of fully edited and translatable mRNAs.

Of the three transcripts characterized, the essential RPS12 showed the most robust editing progression. Editing of RPS12 is relatively linear, with only a few minor branching alternatives. For this mRNA, the first start codon found on the fully edited transcripts consistently translated into the canonical RPS12 open reading frame and we found no evidence of transcripts that access the alternative reading frame. The most prominent alternatively edited branch was only observed in the EATRO 164 cell line and causes a frame-shift that extends the reading frame at the 3' end (Figure 14 $B2^{e}$). Interestingly, this same alternative edit was previously described by Simpson et al. in the 29-13 strain, which shows that this edit is not an isolated occurrence in the EATRO 164 strain [148]. They observed the alternative at a low abundance, which agrees with our observations as well. Their analysis identified a large

amount of variance at the 5’ end, with 5.7% of transcripts being translatable in the canonical ORF. We did observe sloppy editing at the 5’ end, but found two primary forms 5’ end RPS12 editing. Interestingly, this data indicates that the 29-13 strain has a similar editing efficiency to the TREU 667 strain and EATRO 164 strain grown in SDM80.

In contrast to RPS12, we found distinct evidence that ND7 5’ is dual-coding. In both cell lines, alternative editing by different terminal gRNA variants resulted in transcripts with either RF1 (the canonical ND7) or RF3 (a putative metabolite transporter) linked to the first AUG [141]. Interestingly, ND7 5’ has also been sequenced in 29-13 cells [148]. While that study did not directly state evidence of dual-coding, they did indicate that a large proportion of the fully edited ND7 5’ transcripts had a single nucleotide difference in the 5’ UTR. This difference could very well be the same difference we observe in E2 transcripts that links an upstream AUG to the ARF. This suggests that the ability to access the two different reading frames is maintained across a number of different trypanosome cell lines.

While fully edited ND7 5’ transcripts were found in both cell lines, a major difference was observed in the efficiency of the editing process. In TREU 667 cells, over 79% of ND7 5’ transcripts had initiated the editing process and a full 9.7% are fully edited. In contrast, EATRO 164 cells grown under the same conditions (SDM79) had only 52.4% transcripts that initiated editing and a scant 0.2% fully edited. Growth of EATRO 164 cells in SDM80 did substantially increase the number of transcripts that had initiated RNA editing (72.6%), however, no corresponding increase in fully edited transcripts was observed. The major differences in editing efficiency appear to be due to both the use of alternative gRNAs that could disrupt the editing cascade and well as a gRNA mutation that affected the ability of the guide RNA to

efficiently anchor. Surprisingly, the gRNAs that disrupt editing in the EATRO cell line are also present in the TREU gRNA transcriptome. It is unclear why we see evidence of their use in only the EATRO cells. It may be that in the TREU cells, these gRNAs are more efficiently used in a different editing pathway. A full understanding of gRNA selection and use will require the characterization of the entire edited transcriptome. In addition to the large decrease in the efficiency of ND7 5' editing observed in the EATRO cells, we also saw a distinct shift in the number of fully edited transcripts that translate in RF3, the alternative open reading frame. This alternative protein has been previously predicted to be a metabolite transporter as it shares distant homology with a bacterial sugar transporter, SemiSWEET [141].

The most pronounced differences between the two cell lines was observed for the CR3 transcript. In both cell lines, CR3 utilizes a much more complicated editing pathway than either ND7 5' or RPS12 and the overall efficiency of the editing process is very low. Surprisingly, the number of CR3 transcripts that initiate RNA editing is comparable to the percentage observed for RPS12. However, editing by the initiating gRNA appears to be very inefficient. In TREU cells, while 15.1 % of the transcripts initiate editing, only 2.2% are fully edited through the first editing block. A similar drop is also observed in the EATRO cells. The identified gRNA population that initiates editing does not contain any mismatched base pairs and it is unclear why full editing by this gRNA is so inefficient. The canonical CR3 is a putative NADH Dehydrogenase complex I member (ND4L) [120]. Editing of Complex 1 members does appear to be developmentally regulated, with full editing only observed in the Bloodstream stage. [5,27–30] . It may be that editing is stalled right after initiation by a transcript specific mechanism. However, a small percentage of transcripts are still edited. Transcripts edited to the canonical

CR3 sequence were only observed in the TREU cell line. In this cell line, four different 3' editing pathways converge to an internal consensus sequence which then diverges again near the 5' end, generating a variety of different proteins in two different reading frames. In the EATRO cell line, editing initiates with the same 3' gRNAs, but diverges at the internal consensus sequence when they employ a different set of gRNAs for full editing. Because of how different their editing pathways are, the TREU and EATRO protein products cannot be directly compared. Searches were run on various different databases in order to determine the putative functions of the many CR3 proteins. Unfortunately, these searches yielded few significant results, with most proteins only sharing homology with the transmembrane domains of many different proteins. The very small percentage of CR3 transcripts that undergo full editing suggests that the protein products may not be made or utilized in this stage of the parasite life cycle. However, we hypothesize that the ability to alternatively edit transcripts may be an important evolutionary mechanism to maintain genetic plasticity.

The dual host life cycle of *T. brucei* leaves it vulnerable to genetic drift especially in regards to the mitochondrial ETC genes that are not always under selection. Previously, we proposed a mechanism that would contribute to the drift robustness of these mitochondrial genes. By overlapping ETC genes not under selection in the bloodstream stage with genes that are under selection during this stage of the life cycle, the accumulation of mutations can be prevented [141]. These overlapped genes share most gRNAs, and this strategy ensures that almost all of the genetic material is protected. We also hypothesize that the sequential nature of gRNA use and the sensitivity of the RNA editing process to both mRNA and gRNA mutations can also protect against genetic drift by increasing the deleterious effects of the mutations

(LaBar and Adami 2017). Increasing the lethality of mutations would insure that deleterious mutations are purged from the population during long periods of growth in the mammalian host. While the process of RNA editing may help weed out mutations by making them lethal, it would also prevent the population from generating beneficial mutations as well. This strategy leaves organisms no options for evolving. We suggest that alternative edits, such as those seen in the CR3 and others previously observed, editing pathways generate protein diversity without compromising the genetic information found within the genome [68].

Our analysis revealed a number of details about the larger mechanisms of gRNA selection in RNA editing. We found a surprising number of gRNAs with that had been identified to edit two different mRNA sequences. While some of these promiscuous gRNAs appear to be unproductive, generating dead end branches on their editing pathways, many appear to be productively used, as with the editing pathways of CR3 in the EATRO 164 cell line. The majority of these editing pathways are directed by promiscuous gRNAs and these pathways generate translatable transcripts. Interestingly, most of these gRNAs were identified to edit members of complex I, particularly ND8 and the 3' editing domain of ND7. It may be possible that this is another mechanism of increasing the drift robustness of *T. brucei* by giving these gRNAs multiple functions. Promiscuous gRNAs have been previously identified editing *L. tarentolae* RPS12 and ND3, however, they were not shown to be producing translatable transcripts [156,157].

In addition to this, we determined that RNA editing is not strictly sequential. While overall, editing proceeds from the 3' to 5' across of the editing domain, we have found evidence that shows that gRNAs can overwrite editing that has been previously generated.

These observations show that the RNA editing system can tolerate some amount of abnormal editing, despite the fragility of the system as a whole. Interestingly, another study that examined the use of alternative RPS12 gRNAs previously identified found that the three gRNAs in question were being utilized, and that despite not generating the conventional editing patterns, a very small number of transcripts returned to the canonical editing pattern after the alternatives [148]. This last observation may be another instance of gRNA overwriting. Non-sequential editing is also suggested by our proposal that the poly-U tail can be used to direct editing. We showed that deletions in partially edited transcripts were more common in regions close to the U tail of the preceding gRNA (Figure 24). Interestingly, in some cases, an upstream deletion site would be completely deleted while a downstream site would not be deleted. This phenomenon was also observed in another study, in an *in vitro* editing system [158].

In this analysis we determined the editing pathways of RPS12, ND7 5', and CR3 in EATRO 164 and TREU 667 cells, and we found evidence of dual-coding in ND7 and CR3. We also showed that editing patterns can vary quite significantly between cell lines and based on available energy sources. Using the gRNA transcriptomes to validate alternative edits was vital to completing this project. In light of the extreme variation we observed in the CR3 editing pathway, we believe that in order to fully understand the dynamics of the editing pathways as a whole, we need to sequence all of the edited mRNAs and gRNAs in multiple cell lines under multiple conditions.

Acknowledgements

We thank the Ken Stuart Lab for the trypanosome cell lines. We would also like to thank Hanyou Pan for his contributions to the data analysis of ND7 5'.

CHAPTER 5: CLUSTER CLASSIFICATION OF UNKNOWN GRNAS REVEALS THE ROBUSTNESS OF THE RNA EDITING SYSTEM

Abstract

The RNA editing system of *Trypanosoma brucei* uses small RNAs called guide RNAs to direct the insertion and deletion of mRNAs. Thousands of gRNAs are used in this system to render twelve mitochondrial mRNAs translatable. These gRNAs edit sequentially, with each gRNA generating the anchor binding sequence for the next gRNA. This means that the system is inherently fragile. gRNA transcriptomes have been generated and gRNAs were identified based on their complementarity to previously described edited mRNAs. This method was effective in identifying many gRNAs, but we found that many were still unidentified. To determine if these gRNAs were nonfunctional mutants or potentially had undescribed functions, we grouped all gRNAs into clusters, where each cluster had significant sequence conservation, using our new program ACORNS. This showed that most unidentified gRNAs were not related to any functionally known gRNAs and could be generating unobserved alternative editing patterns. Recently, the editing pathways of three genes, RPS12, ND7 5' and CR3 were described in detail. Each gene had branches in their editing pathways, but it was not always clear based on gRNA abundance data alone why one branch of the pathway was more expressed than another. Using the defined gRNA clusters, we screened all related members editing these three genes against their targets to determine what proportion of each family was able to productively edit, using another new program GUIDE. We found that most of these unidentified gRNAs were predicted to disrupt RNA editing. However, using this information in combination with the

mRNA data for these three genes, we found that these mutations are highly tolerated by the editing system. We also determined that gRNA abundance does not correlate with the mRNA editing preference. We also observed many gRNA populations of high abundance that were apparently not used to edit. In most cases these populations had no issues in their mRNA alignments but were seen to be only used in one cell line and not the other. Currently, the complete mechanism of gRNA selection is unknown.

Introduction

The RNA editing system of the trypanosomes is a unique and complex system. It requires the use of two genetic components, the protein coding genes whose transcripts require editing, and the small RNA genes encoding the guide RNAs (gRNAs) that direct the specific edits [7]. The genes that require editing are all mitochondrially encoded and are either associated with the electron transport chain or the mitochondrial ribosome. In this RNA editing process, mRNA transcripts can be dramatically altered by the insertion and deletion of uridines [4,6]. These editing events are catalyzed by a multi-subunit editosome complex [140]. Currently, over 47 proteins have been identified that are involved in the cleavage, uridylyl addition or deletion and subsequent re-ligation events that are required for every nucleotide change [140]. Formation of these specialized complexes as well as performing the hundreds of edits required is highly energetically expensive. The RNA editing process is also incredibly fragile to mutations, as it is sequentially dependent. RNA editing starts at the 3' end of the mRNA, and each gRNA generates the anchor for the next gRNA. A mutation in any gRNA could disrupt formation of the next anchor, halting the editing process and with it abort the expression of the protein.

It has been hypothesized that this energy intensive process evolved in response to the parasites complex life cycle [75,76,141]. A dixenous kinetoplastid, *Trypanosoma brucei*, undergoes substantial shifts in its energy sources over its life cycle, requiring extensive regulation of metabolic pathways. In the nutrient deprived environment of the insect, it relies primarily on metabolizing amino acids to drive the Krebs cycle and Oxidative Phosphorylation. Once transferred to the glucose rich bloodstream of its mammalian host however, it shifts from relying on mitochondrial respiration, and moves to using glycolysis alone. In addition, the exclusively bloodstream nature of *T. brucei* in the mammal host requires that they replicate continuously, escaping the host's immune response by periodically switching their surface glycoproteins [12]. These conditions should make the genes involved in mitochondrial respiration particularly susceptible to genetic drift. Recently, we showed that as many as six of the mitochondrial transcripts may be dual coding and that alternative editing near the 5'-end of the transcript can be used to access both open reading frames (ORFs). We hypothesized, that the alternative ORF may give the parasites a selective growth advantage in the mammalian host. This would significantly increase the deleterious effects of mutations and protect against genetic drift. The linking of an ETC gene with an essential gene has been previously shown for cytochrome oxidase III (COIII). For this transcript, alternative editing by one gRNA can link an ORF found in the unedited 5' sequence with the trans-membrane domains found in the fully edited carboxyl-end of the protein [67]. This alternative protein, AEP-1, is involved in mt DNA maintenance and appears to be essential during bloodstream growth [68]. The detection of such a large number of dual-coding genes (seven including COIII) in a genome that encodes only 17 proteins, suggests that this is an important mechanism to protect the integrity of the ETC

genes that are required only in the insect. Thus, the ability to overlay genetic information maybe a protective strategy made possible by RNA editing [99,141]. The changes in metabolism are reflected in other changes in the pattern of RNA editing found the different life cycle stages in *T. brucei* [5,24–30]. For instance, many of the transcripts encoding members of NADH dehydrogenase (Complex I) are only fully edited in the bloodstream stage. In contrast, cytochrome oxidase II (COII, Complex 3) and cytochrome B (Complex 3) are both preferentially edited in the insect stage. Recently, we even observed a shift in the editing pattern found in insect stage trypanosomes if the growth medium is switched from glucose rich to the glucose depleted. This suggests that editing can respond to its environment (Chapter 4). In addition, multiple studies have shown that different editing patterns can be observed for transcripts of the same mRNA, with edits being directed by multiple alternative gRNAs. The most dramatic example of this is seen in the putative NADH dehydrogenase subunit 4L, known as C-rich region 3 (CR3) [120]. For transcripts of this gene, two different sets of gRNAs were used in two different cell lines of *T. brucei*. Both editing pathways were complex and possessed multiple branches that were fully translatable. Curiously, the gRNAs required to generate the editing pathways of the two different cell lines were present in both cell lines, but apparently not selected for use for unknown reasons.

The gRNAs that direct the RNA editing process are stored on 1 kb circular DNA molecules called minicircles. These minicircles make up the bulk of the mitochondrial DNA of *T. brucei*, with up to 10,000 minicircles per cell, and there are an estimated ~1200 different sequence classes of minicircle at varying abundances [8,33]. gRNA transcriptomes have been generated for the insect and bloodstream stage trypanosomes [9,119]. Analyses of these

transcriptomes identified over 600 gRNA populations (gRNAs that edit the same region of the same gene) involved in editing the mitochondrial mRNA transcripts. These populations typically contained multiple different sequence classes of gRNAs (major classes). The sequence differences observed are all R to R or Y to Y changes, and since both A:U and G:U base pairs occur in the editing process, the multiple sequence classes can all guide the generation of the same mRNA sequence. These data also show extreme quantitative differences between the different gRNA populations, with population sizes ranging from <10 to >300,000 reads. We had postulated that regions with very low gRNA coverage are areas with mRNA sequence variations and that the editing of these regions may be directed by gRNAs not identified due to internal mismatches with conventional sequence. In addition, we also identified some abundant, mutated gRNAs that could potentially introduce frameshifts or create sequence that disrupts the upstream anchor binding site. This was surprising, as we had hypothesized that these mutations would be selected against due to the fragile nature of the editing process. Because of the high stringency of our initial screen, only gRNAs with mismatch mutations biased towards the 3' or 5' ends of the gRNAs were initially identified and it was unclear how prevalent mutated gRNAs that could disrupt editing were. Analyses of our gRNA libraries did indicate that they contain millions of "unidentified" reads that had key guide RNA characteristics (characteristic transcription start sites, U-tail and ~ length).

In this manuscript, we describe the development of two new pipelines that allows us to begin to characterize these "unidentified" transcripts in order to determine what impact they might have on RNA editing efficiency. ACORNS (Assemble Clusters Of Related Nucleotide Sequences), allows us to identify and classify gRNA sequences. A second program, GUIDE (gRNA

Uridine Insertion/Deletion Editor), simulates the RNA editing process and allows us to predict the effect of the gRNA mutation on the sequencing process. Using the CR3, RPS12 and ND7 5' mRNA transcriptome data, we identified the gRNA populations that can edit or disrupt editing of these genes. We found a surprisingly high toleration of mismatches and gaps in gRNA/mRNA alignments, as well as many instances where the most abundant gRNAs present were not those preferentially used. A deeper look at the gRNA transcriptomes also revealed the presence of surprisingly many unidentified and functionally unknown gRNAs, suggesting that more work is needed to discover their roles in the RNA editing system.

Materials and Methods

Cluster analysis of related gRNAs by ACORNS

In order to determine the relationships of previously identified and unknown gRNAs, a new program was created called ACORNS (Assemble Clusters of Related Nucleotide Sequences). This program functions to identify putative gRNAs, determine relationships between gRNAs, and group related gRNAs into clusters.

Identification of putative gRNAs. As described previously, identical sequences were collapsed and sequences without four consecutive Ts were filtered out, indicating the lack of a poly-U tail [9]. To perform this analysis, a new program was generated, called Assemble Clusters Of Related Nucleotide Sequences (ACORNS). ACORNS first filtered out all putative gRNAs, based on two criteria: having 40 nucleotides prior to the start of the poly-U tail, and having a transcription start site that either matches one of the top twenty most common six nucleotide gRNA transcription start sites, or is one mutation away from one of these sites [9,119]. Maxicircle edited and unedited sequences as well as ribosomal RNAs were also filtered

from the sequence file. Identical sequences were collapsed, but retained their overall total read abundance. Additionally, if sequences were identical with the exception of the of the start position of the poly-U tail, or 5' end location of the gRNA, they were still consolidated, keep the sequence of the most abundant transcript.

Identification of gRNA families. A pair of related gRNAs are defined as two gRNAs that only differ by a single substitution, insertion or deletion mutation. Once putative gRNAs were identified, ACORNS aligned each gRNA with every other gRNA to determine which gRNAs were related. Alignments were scored using the python-levenshtein package [159]. In order to prevent errors due to 5' exonuclease activity or difference in poly-U site, overhanging nucleotides on either end of the alignment were not counted as mismatches.

ACORNS then grouped gRNAs into families based on their relationships. Each family was grouped together by starting with the most abundant transcript, and including any other gRNAs related to that transcript, as well as any gRNAs related to those gRNAs, and so on. Once gRNAs were grouped into families, sequences were trimmed to the same length and any new identical sequences were collapsed.

Visualization of these related gRNA clusters was performed by a subprogram, and clusters were visualized in two different ways, with color coding based on gRNA identity or gRNA abundance. Identity was defined as the gene known to be edited by previously identified gRNAs, with other all other gRNAs labeled as "unknown". Color coding by abundance was based on a log scale, and each scale created for each cluster, with the least abundant gRNA being colored purple and the most abundant gRNA being colored red, and all other gRNAs being scaled accordingly.

Prediction of RNA editing by GUIDE

A reference library of all edited forms previously described of RPS12, ND7 5', and CR3 were collected and annotated with the editing states in each transcript (Chapter 4). For this analysis we generated a new piece of software known as the gRNA uridine insertion/deletion editor (GUIDE). GUIDE uses the outputs of ACORNS, and for each gene, GUIDE pulls the families of gRNAs that have members that had been previously identified to edit the given gene. Editing of each member of each family was then simulated. For each family, the template generated by the previously identified members of that family were used. If gRNAs generating different forms of the same mRNA were in the same family, all appropriate templates were tested on each family member, and the predicted edit with the longest alignment with the gRNA was saved.

In order to predict the edits for each gRNA, GUIDE determines the most likely anchor binding location on the template. Anchor sequences were required to use Watson-Crick base pairs only and be located in the first twenty nucleotides (nt) of the gRNA. The first twenty nt of each were scanned along the template, and the longest consecutive stretch of Watson-Crick base pairs was identified. Additionally, a second anchor was identified, using a weighted score, where each G:C base pair was worth two points and A:U base pairs were worth one point. Editing from both anchors were tested and the anchor that generated the longest edited sequence was saved.

The rules of editing used by the program are the standard uridine insertion/deletion rules observed in the editing system. If a nucleotide from the gRNA is aligned to a nucleotide from the mRNA that is "illegal" (anything other than Watson-Crick or G:U base pairs), a uridine

will either be inserted or deleted to resolve the mismatch. If this cannot resolve the mismatch, editing ends. Once edits were predicted, they were classified based on their effect on the editing process as a whole and their effect on the predicted protein.

Results

In our characterization of the EATRO 164 gRNA transcriptome, we identified over 3 million reads and ~64,000 unique gRNA sequences capable of generating conventional editing patterns [9]. These gRNAs were identified by finding the longest common substring to conventionally edited mRNAs and retaining only those scoring 45 or more (Watson-Crick base pairs = 2; G:U base pairs = 1). In these studies, we found that lowering the stringency and/or allowing for mismatches lead to the identification of thousands of gRNAs with characteristics suggesting that they were misaligned. While this initial analysis did identify an almost full cohort of guide RNAs, there were still millions of reads in the transcriptome that could not be classified. To determine how many of these reads were unidentified gRNAs for possible alternative sequence or mutated conventional gRNAs, we developed a new pipeline. ACORNS (Assemble Clusters Of Related Nucleotide Sequences) has a number of features that allows it to identify and classify putative guide RNAs:

1. It filters out all contaminating nuclear and maxicircle sequences.
2. It Identifies putative gRNAs based on three criteria: presence of a U-tail (defined as 4 consecutive U-residues), length (40 nts prior to last stretch of U-residues) and transcription start site (based on sequences defined in the previous analyses) [9,119].

3. It identifies related gRNAs by scoring the best alignment of each gRNA against all other putative gRNAs in the library. gRNAs with a single mismatch or gap are then classified as related.

4. It clusters related gRNAs by starting with the most abundant transcript, grouping all relatives of that transcript into the cluster, then adding all relatives of any relative to the cluster.

5. Using a subprogram, the clusters of related guide RNAs can be visualized, revealing the relationships between the previous identified gRNAs and unknown gRNAs.

ACORNS analyses were done for both the TREU 667 and EATRO 164 procyclic gRNA transcriptomes, leading to the identification of 1256 clusters in TREU and 1168 clusters in EATRO 164 (Tables 11 and 12). This allowed us to identify all of the gRNAs that were distinctly related to a conventional gRNA but had undergone a mutational event. In addition, we also identified a number of clusters that had no previously identified members. The clusters identified varied greatly in size, with the largest cluster containing over 3400 sequence members with a total of 1,377,190 transcript reads. These large clusters (1000+ members) were actually quite rare, with only 10 and 4 clusters of this size identified in the TREU and EATRO gRNA libraries, respectively (Table 13). Interestingly, the majority of clusters were quite small, containing fewer than 25 sequence members and an average transcript number of approximately 250 reads (Table 13). The majority of the small clusters were “unidentified”, in that they contained no previously identified gRNA member (Table 13). An increase in the size of the cluster, also increased the probability that they contained a previously identified gRNA and that cluster members could then be identified.

Table 11. Summary of ACORNS Results. gRNAs are classified as previously identified, related to a previously identified gRNA based on sequence similarity or unrelated to any previously identified gRNAs.

	TREU 667 Procyclic	EATRO 164 Procyclic
Initial Reads	15,251,292	11,387,683
Final Reads after ACORNS Step 2.	11,199,364	9,049,005
Previously Identified Reads	5,121,216	4,413,142
Previously Unidentified tagged as Related	2,636,714	2,072,099
Previously Unidentified tagged as Unrelated	3,441,434	2,563,764

Table 12. Cluster summary. Clusters were defined as a group of 10 or more related gRNAs. Cluster characteristics describe the percent of the cluster members that had been previously identified.

Cluster Characteristics	TREU 667 Procyclic		EATRO 164 Procyclic	
	# of Clusters	Reads	# of Clusters	Reads
≥95% Previously Identified	238	4,232,188	191	3,281,190
>0% Previously Identified	267	3,488,118	215	3,180,019
0% Previously Identified	751	3,289,359	762	2,401,148
Unclustered	NA	189,699	NA	186,648

Table 13. Cluster size summary. Cluster size is determined by the number of unique sequence classes in a cluster. % Unidentified clusters is the percentage of all clusters in each bin that are completely unidentified.

	# of clusters identified		**Total # Reads**		**% Unidentified clusters**	
Cluster Size	**TREU 667**	**EATRO 164**	**TREU 667**	**EATRO 164**	**TREU 667**	**EATRO 164**
10 - 24	518	503	137,413	134,707	**68%**	**77%**
25 - 49	291	265	240,804	211,154	**63%**	**64%**
50 - 99	185	184	427,029	425,535	**55%**	**61%**
100 - 249	173	139	1,575,518	1,398,488	**45%**	**54%**
250 - 499	60	47	2,167,031	2,153,210	**45%**	**26%**
500 - 999	19	26	2,354,552	3,379,901	**21%**	**27%**
1000 +	10	4	4,107,318	1,159,362	**30%**	**25%**

Figure 25 shows an example of three different clusters assembled by the ACORNS program. The visualization program represents each individual gRNA sequence as a dot, with lines connecting dots representing relationships between the individual sequences. Each member is numbered in order of abundance of transcript reads (0 = most abundant). gRNAs can be further characterized within the cluster by using color to identify either transcript specific gRNAs (Figure 25 A, C and E), or to indicate Log transcript abundance (Figure 25 B, D and F). The

first cluster shown illustrates a relatively small cluster with 104 different gRNA sequence members (Figure 25 A and B). In this visualization, it is clear that most of the members were previously identified (red dots), including the most abundant member (dot 0). Figure 25B shows the transcript abundance of each cluster member. This cluster is very characteristic of most of the clusters we identified, in that it has a central very abundant gRNA (red dot in center, >5,000 reads) with most other cluster members identified having very few reads (purple = 1 read in the transcriptome). 98.1% of the transcripts in this cluster were previously identified. Figure 25 C and D illustrate a large cluster with 1801 sequence members. In contrast to the first cluster, most of the identified sequence members were previously unidentified (gray dots), a few of the members however, had been previously tagged as RPS12 specific gRNAs (red dots). The previously identified RPS12 gRNAs were tagged as editing the Block H region of RPS12. The originally identified transcripts were rare, with only 255 reads found in the EATRO 164 library. The ACORNS analysis allowed us to identify an additional 276,358 reads that cover this region. However, most of these newly identified gRNAs, including the most abundant sequence class, have mutations that affect their ability to correctly edit the mRNA transcript.

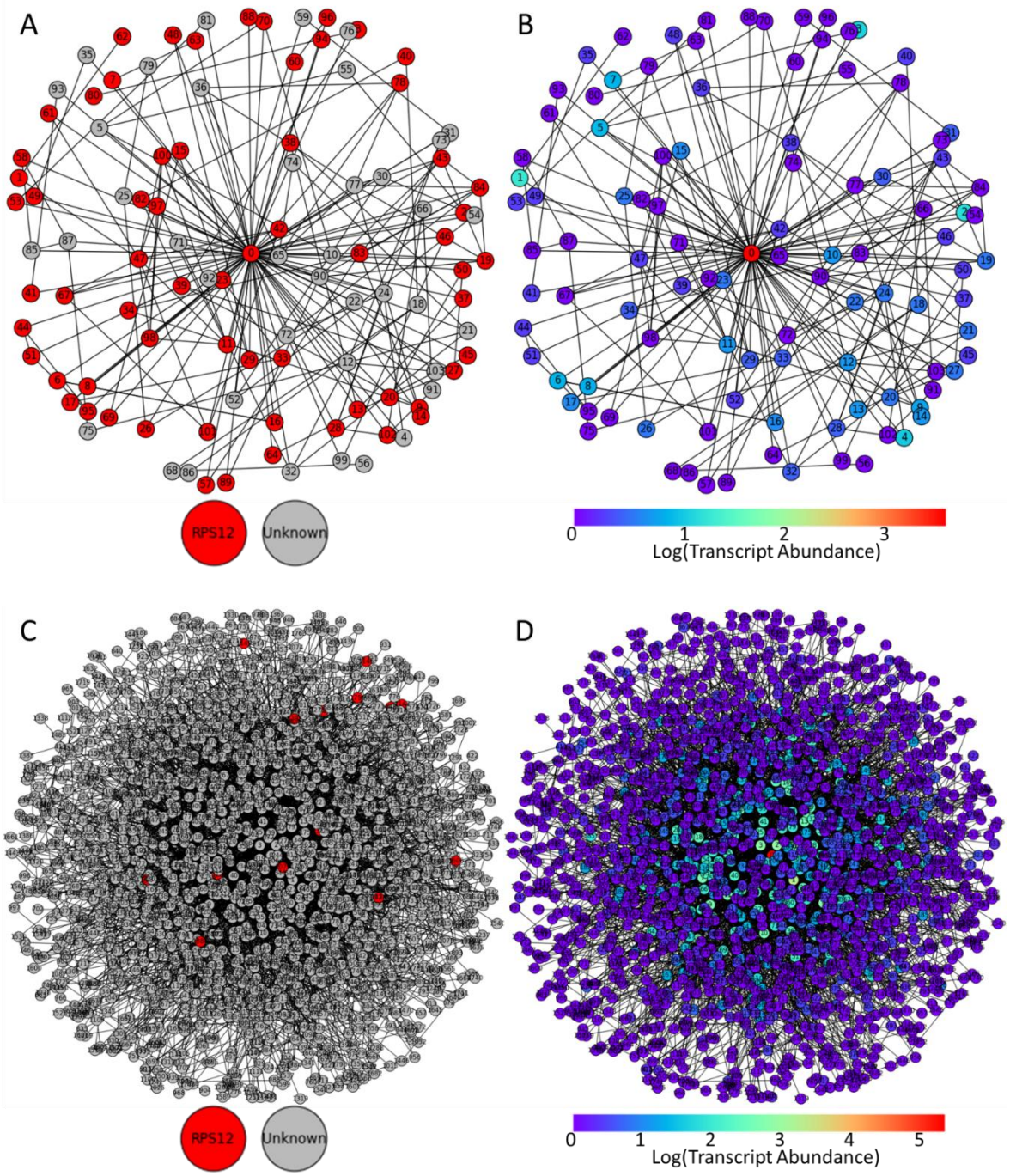
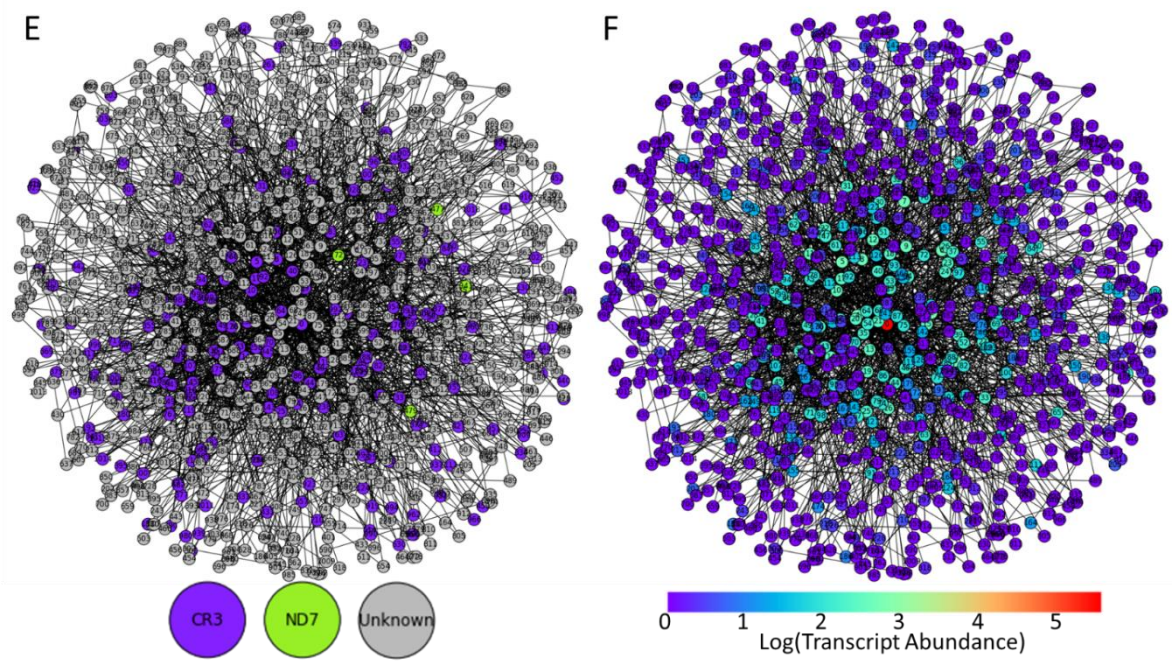


Figure 25. Example clusters of related gRNAs generated by ACORNS from the EATRO 164 PC gRNA transcriptome. Each dot represents an individual gRNA sequence and lines connecting dots represent relationships between gRNAs. Each cluster is shown with two different color schemes; editing gene identity (A, C and E) and gRNA read abundance (B, D and F). The first two clusters (A,B and C,D) both contain RPS12 identified gRNAs (red dots) and previously unidentified gRNAs (gray dots). The third cluster (E and F) contains a majority of previously unidentified gRNAs (grey), but also contains CR3 identified gRNAs (purple) and a few gRNAs that had been tagged for a dead-end ND7 5' alternative editing pattern (green).

Figure 25 (cont'd).



The final cluster shown also illustrates a large cluster with 1039 members. (Figure 25 E and F). For this cluster, the most abundant member (member 0) contained over >400,000 reads and had been previously identified as an alternative CR3 gRNA (gC^{e80}). The ACORN cluster analysis allowed us to identify and additional 829 sequences and over 12,700 reads as being related to this alternative gRNA. Interestingly, in the visualization of this cluster, we noted that 4 of the members were previously tagged as involved in the generation of a disruptive (dead-end) alternative edit of ND7 5'. An examination of these transcripts indicate that they are in fact a specific mutant subclass of the gRNA cluster that are now capable of anchoring and creating a misedit in ND7. This cluster was the only example of a cluster containing related gRNAs with different targets (distinct from promiscuous gRNAs that all have multiple targets), and could be an example of how alternative editing originates.

These data analyses suggest that both procyclic libraries have large numbers of unidentified gRNAs of unknown function. It may be that these gRNAs are directing alternative editing events that have not yet been characterized. The full mRNA transcriptome has not been sequenced and the limited amount of mRNA sequence available does suggest an abundance of alternative editing (Chapters 3 and 4) [9,67,68,99,141,148,160]. Surprisingly, both libraries also contained large numbers of mutated conventional gRNAs. This was unexpected, due to the fragile nature of the RNA editing process. We had hypothesized, that the sequential nature of the RNA editing process, would decrease the tolerance for mutations in both the gRNA and the mRNA genes (Chapter 4) [9]. In order to determine how these mutated gRNAs might influence the RNA editing process, we developed a second program (GUIDE, gRNA Uridine Insertion/Deletion Editor) that simulates the RNA editing process. This program takes the fully edited mRNA templates for the identified gRNAs in each cluster, finds the best anchor for each gRNA and then simulates editing based on the conventional base-pairing rules (A:U and G:U pairing both allowed). For each gRNA, the anchor length, length of complementarity and the number of sites showing non-conventional editing are determined. This allows the classification of the gRNA into different bins: 1) low quality anchor; 2) does not fully edit; 3) conventionally edits and 4) alternatively edits. Guide RNAs that generate alternatively edited sequence were then further classified as 1) Disruptive edits (does not generate the anchor for the next gRNA); 2) Frameshift edits (does not disrupt editing but generates a frameshift); 3) missense editing (does not disrupt editing but generates a missense mutation). We note that in the GUIDE program, we initially classified C:A base pairs as disruptive (stopping the editing process). However, because C:A base pairs have been previously identified within known gRNA

alignments, we did sort gRNAs containing a C:A base pair into their own bin [9,119]. These analyses were done for all gRNA clusters identified for RPS12, ND7 5' and CR3. These 3 genes were chosen because the mRNA transcriptomes and full editing pathways had been previously characterized (Chapter 4).

Of the unidentified gRNAs that are related to previously identified gRNAs, 942,397 and 958,140 reads were related to RPS12, ND7 5' or CR3 editing gRNAs in TREU and EATRO cells respectively. Of these newly identified gRNAs, 92.6% in TREU and 95.5% in EATRO were predicted to be incapable of fully editing, to fail to create anchor for the next gRNA or to cause a frameshift. This suggests that a large number of gRNAs could be disruptive, however, because of the large differences in population sizes for the different guide RNAs, it is important to evaluate the percent of disruptive editing at the population level. This data could indicate how tolerant the RNA editing process is to mutations in the gRNA population.

RPS12

Editing of the essential RPS12 transcript was relatively straight forward with a limited number of alternative edits (Chapter 4, Figure 26). The main editing pathway involved 10 gRNA populations (gA – gJ). Analyses of these populations still show a large variation in the abundance of the different populations even after the cluster analyses (Figure 26, Table 14). Surprisingly, we saw no correlation between the abundance of the gRNA populations and RNA editing efficiency. For example, in the TREU cell line, the B1 and B1* mRNA transcripts are equally abundant (Figure 26A). However, the gB1* gRNA is almost 50-fold more abundant than gB1 (Figure 26B and Table 14). In contrast, in the EATRO cell line, the B1 mRNA transcript is significantly more abundant (5x), while the gB1 and gB1* guide RNAs are approximately equal

in abundance. Interestingly, the only gRNA identified that can generate the B1* edits requires a gap in the alignment with the mRNA to create the correct sequence. Our GUIDE program bins this gRNA into the “disrupts editing” bin (brown) and it is unclear why this gap is tolerated. We do note that there are three G:C base pairs surrounding the gap that would help stabilize this alignment. We hypothesize however, that the gap in the alignment does in fact affect the ability of this gRNA to efficiently edit. In the TREU cells, the abundance of gB1* may contribute to its editing efficiency compensating for the gap and increasing the amount of B1* mRNA observed. This pattern appears to be repeated in EATRO, where the B1 mRNA is almost five times as abundant as the B1* mRNA, but the gRNAs are nearly equally abundant. Interestingly, in both transcriptomes the gRNAs responsible for generating the B1 edits both require C:A base pairs for the gRNAs to function correctly. This was the first of many instances observed where editing is impossible without the use of C:A base pairs.

A significant number of the gRNA populations required for the editing of RPS12 did have large numbers of mutated gRNAs that were predicted to stop (mismatch cannot be resolved by the insertion or deletion of a U-residue) or disrupt editing (does not generate the correct anchor sequence for the next gRNA). For example, while the gC populations had a small number of gRNAs with perfect alignment to the canonical sequence (1.6% in TREU and 5.9% of the gC populations in EATRO), most of these gRNAs contain an illegal base pair (G:A or G:G). Similarly, more than 70% of the gD population in the TREU cell line have a U:U mis-match in the middle of the alignment and both the gG and gH populations have a majority of gRNAs that contain a single mismatch in the best alignment with their editing block.(Figures 27 and 28). Surprisingly, when the GUIDE program simulated the edits generated by the mutants of the gD

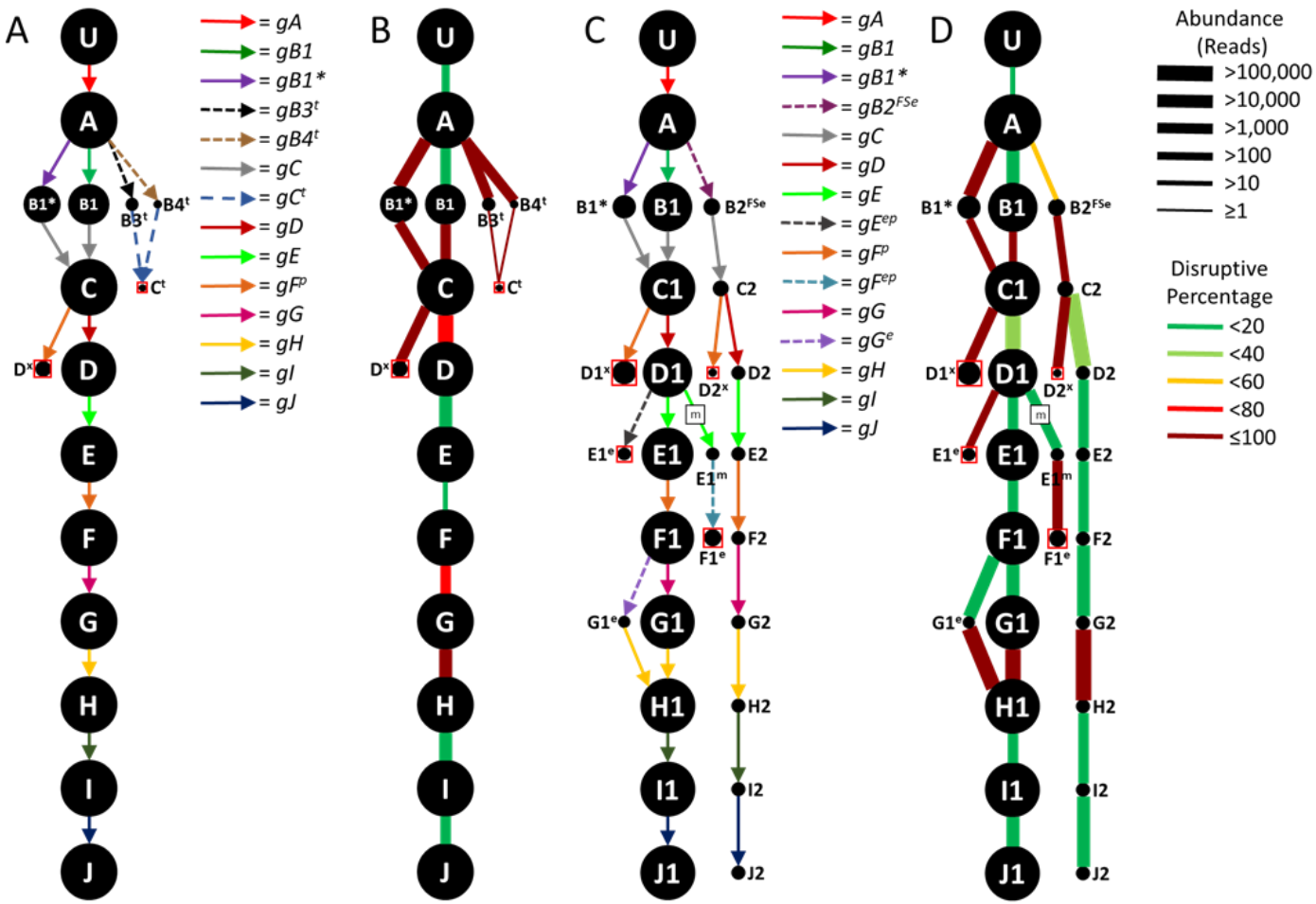


Figure 26. Observed RPS12 editing pathways in the TREU 667 cell line (A and B) the EATRO cell line (C and D). U = unedited transcripts. Dot sizes are proportional to the percent of block level edited transcripts using the gRNA indicated. Colored arrows indicate the gRNA population used (A and C). Dashed arrows represent gRNA populations used in only one cell line (superscript 'e' or 't'). gRNA names with superscript 'p' represent promiscuous gRNAs. Dots enclosed by a red box represent end point mRNAs with no AUG start codon. Lines connecting dots (B and D) indicate gRNA population size and functionality. Disruptive gRNAs were considered to be those that were predicted by GUIDE to be unable to complete editing (excluding those that only required a C:A base pair to finish editing), unable to generate the anchor for the next productive gRNA, or generated a frameshift mutation.

Table 14. gRNA population analyses for RPS12.

Editing Block	Population	TREU 667 gRNA Reads	TREU 667 mRNA Reads	EATRO 164 gRNA Reads	EATRO 164 mRNA Reads
A	gA	416	69,923	95	38,946
B	gB1	1,901	27,805	19,878	28,494
	gB1*	84,459	23,318	17,975	5,880
	$gB3^{t}$	1,692	2,463	0	0
	$gB4^{t}$	2,784	946	0	0
	$gB2^{FSe}$	1	0	15	2,820
C	gC	3,081	46,436	304	35,255
	gC^{t}	6	577	0	0
D	gD	387,246	35,563	460,928	20,067
	gF^{p} (Dx edit)	2,877	2,724	2,371	4,617
E	gE	11,593	35,046	5,269	17,723
	gE^{ep}	1,162	0	178	879
F	gF^{p}	2,940	29,838	2,371	10,818
	gF^{ep}	0	0	1,843	955
G	gG	4,505	28,973	19,028	11,163
	gG^{e}	105,217	0	99,153	468
H	gH	41,824	29,331	287,393	10,843
I	gI	14,251	26,417	6,136	9,747
J	gJ+	5,573	17,870	18,094	7,665

population, it predicted that the mutants would generate an alternative sequence that lacks the anchor binding site for the gE population (Figures 27 and 28). However, this alternative sequence was not identified in the RPS12 mRNAs of the TREU or EATRO cells (Chapter 4). This suggests that the alignment error is tolerated in the generation of the canonical sequence. For most of the mismatches detected, the error in the alignments are all immediately flanked by multiple G:C base pairs. It may be that the presence of multiple stable G:C pairs allows the editing machinery to tolerate these mismatches. The EATRO gH gRNAs are the exception. This population contains a majority of transcripts with a single point mutation near the end of the gRNA that should prevent it from generating the full anchor sequence for gI. There are no

stabilizing G:C pairs that flank this mismatch. The other notable difference between the TREU and EATRO editing patterns were found in an alternative edit (G^{e}) observed in the EATRO cell line only. We note that the gG^{e} guide RNA is very abundant in both transcriptomes (>100,000 reads in TREU cells and >90,000 reads in EATRO cells). However, we found evidence of its use in only the EATRO cell line. In summary, we identified multiple populations of gRNAs that were predicted to be nonfunctional or disruptive in the RPS12 editing pathway. However, these predictions are contradicted by the observed RPS12 mRNA data, suggesting that many of these mismatches that we had previously considered to render gRNAs nonfunctional or disruptive are tolerated by the RNA editing system.

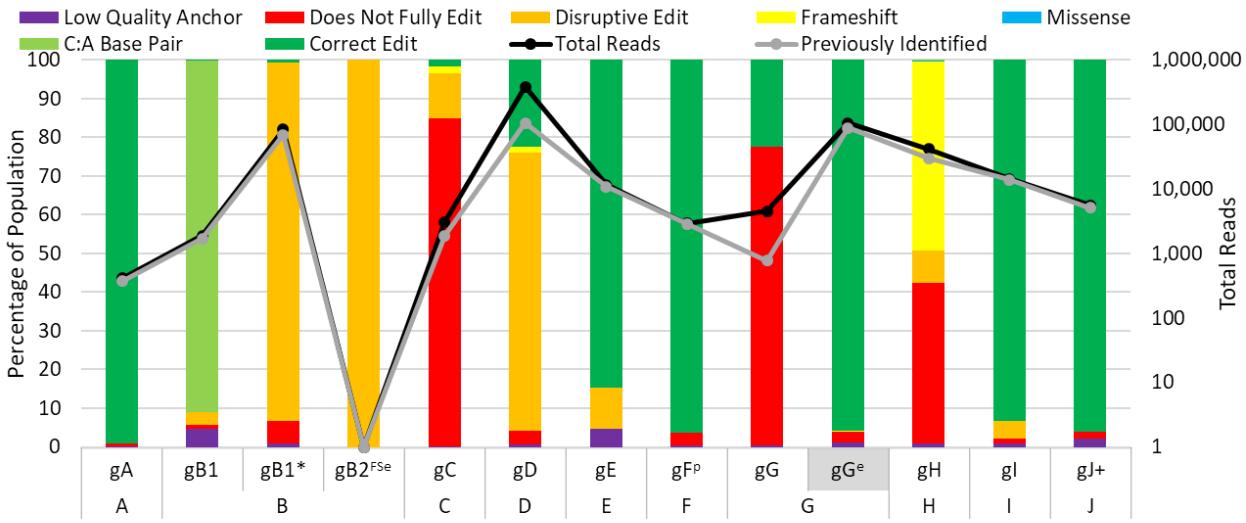


Figure 27. Analysis of functionality and abundance of productive gRNAs populations that edit RPS12 in TREU 667 cells. The functionality of each subpopulation is shown as a bar, with percentage shown on the left y-axis, and subpopulation abundance before and after identification of gRNA relatives is shown on the right y-axis. gRNAs labeled as 'Disruptive Edit' failed to generate the anchor for the subsequent gRNA, and gRNAs labeled as 'C:A base pair' required a C:A base pair to be tolerated for the editing to be completed correctly. gRNAs with shaded names were found in the TREU 667 gRNA transcriptome but were not used in editing the TREU 667 RPS12 mRNAs, despite being used in EATRO 164 cells.

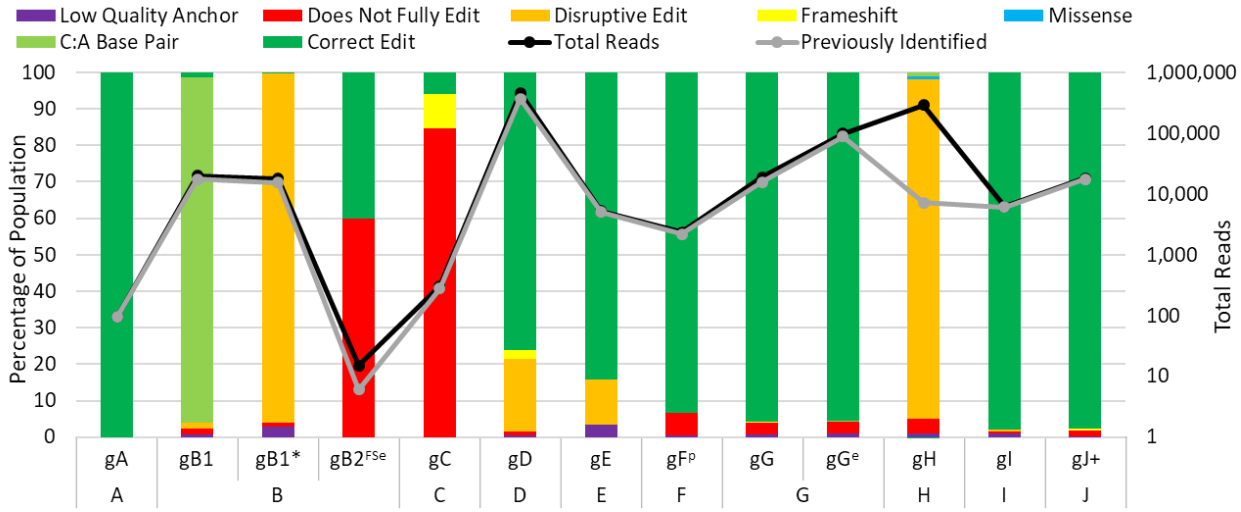


Figure 28. Analysis of functionality and abundance of productive gRNAs that edit RPS12 in EATRO 164 cells. For description of axes and gRNA functionality labels, please see Figure 27.

ND7

As with RPS12, the ACORNS program identified gRNAs related to those previously identified to generate the ND7 5' editing pathways (Figure 29). The ND7 5' editing domain contains five editing block levels, and the editing pathway is relatively straight forward until the final editing block, where alternative editing generates transcripts translatable in multiple reading frames (Chapter 4). In full editing of this transcript, two of the 5 gRNA populations appear to be problematic; gC and gD (Figure 29). In both cell lines, the gC population requires both C:A base pairs as well as multiple mismatches or gaps in the best alignments with the canonical sequence (Figure 30). These multiple alignment errors do not appear to be well tolerated by the editing system, as a severe decrease in editing efficiency was observed in both cell lines from the B to C block level (25.8% in TREU cells and 32.4% in EATRO cells) (Chapter 4). For the ND7 5' gD guide RNAs, both cell lines have a majority of gRNA transcripts with a base pair mismatch in their anchor. In the EATRO cells, almost all gRNAs in the gD population either

require a C:A or A:A base pair in the anchor (Figure 31). While these mismatches are surrounded by G:C base pairs, editing efficiency does appear to suffer with a 13.0% drop in editing efficiency and only 0.4% of transcripts completing the D block level (Table 15). In contrast, in TREU cells, while most gRNAs have a A:A mismatch in the anchor, ~31% of the population has the ability to form a conventional Watson-Crick anchor (Figure 30). The drop in editing efficiency is less (7.2%) than observed in the EATRO cells, and a full 11.7% of transcripts complete D block level editing. In these two problematic gRNA populations of ND7 5' editing, we begin to see what the limits of the RNA editing system are. We find that multiple mismatches as well as mismatches in the anchor binding region of a gRNA appear to have severe impacts on editing efficiency. Surprisingly, however, these aberrant gRNAs do not appear to halt editing altogether.

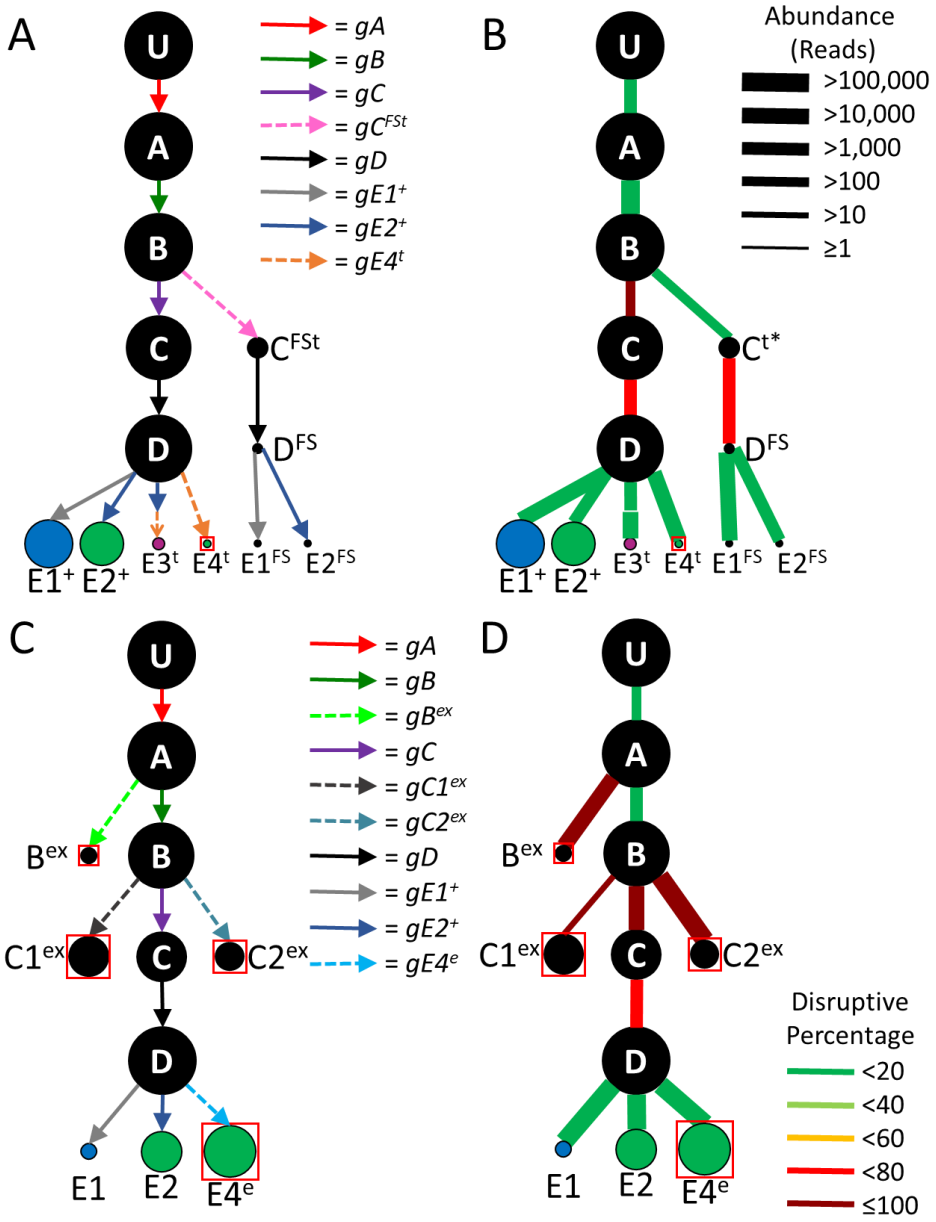


Figure 29. Observed ND7 5' editing pathways in TREU 667 cell line (A and B) and the EATRO 164 cell line (C and D). For descriptions of dots and arrows, please see Figure 26. Dashed arrows with open heads represent a hypothetical rewrite. + indicates that more than one mRNA form was condensed into this circle to simplify the figure . Condensed forms encode largely the same amino acid sequence with only small variants. Terminal dots are colored blue for reading frame 1, magenta for reading frame 2, or green for reading frame 3. Boxed green dots have no functional start codon but are translatable into reading frame 3 with the use of an alternative start codon (UUG). Lines connecting dots (B and D) indicate gRNA population size and functionality. Disruptive gRNAs were considered to be those that were predicted by GUIDE to be unable to complete editing (excluding those that only required a C:A base pair to finish editing), unable to generate the anchor for the next productive gRNA, or generated a frameshift mutation.

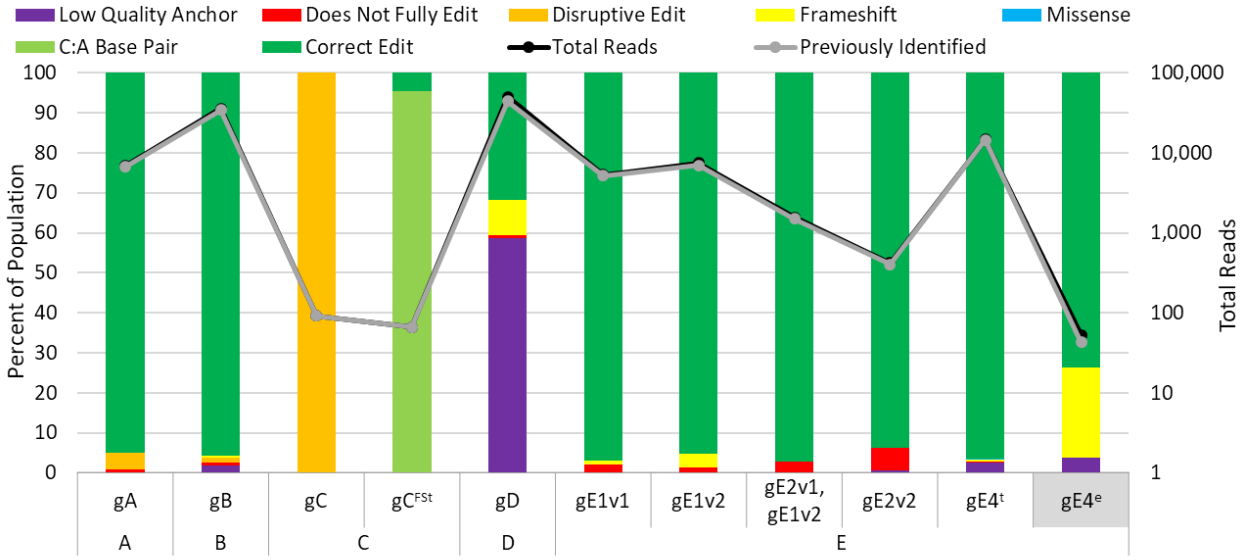


Figure 30. Analysis of functionality and abundance of productive gRNAs that edit that edit ND7 5' in TREU 667 cells. For description of axes and gRNA functionality labels, please see Figure 27. gRNAs with shaded names were found in the TREU 667 gRNA transcriptome but were not found to be utilized in editing the TREU 667 ND7 5' mRNAs, despite being utilized in EATRO 164 cells. The population labeled gE2v1,gE1v2 contained members that generated both sequence patterns and could not be separated.

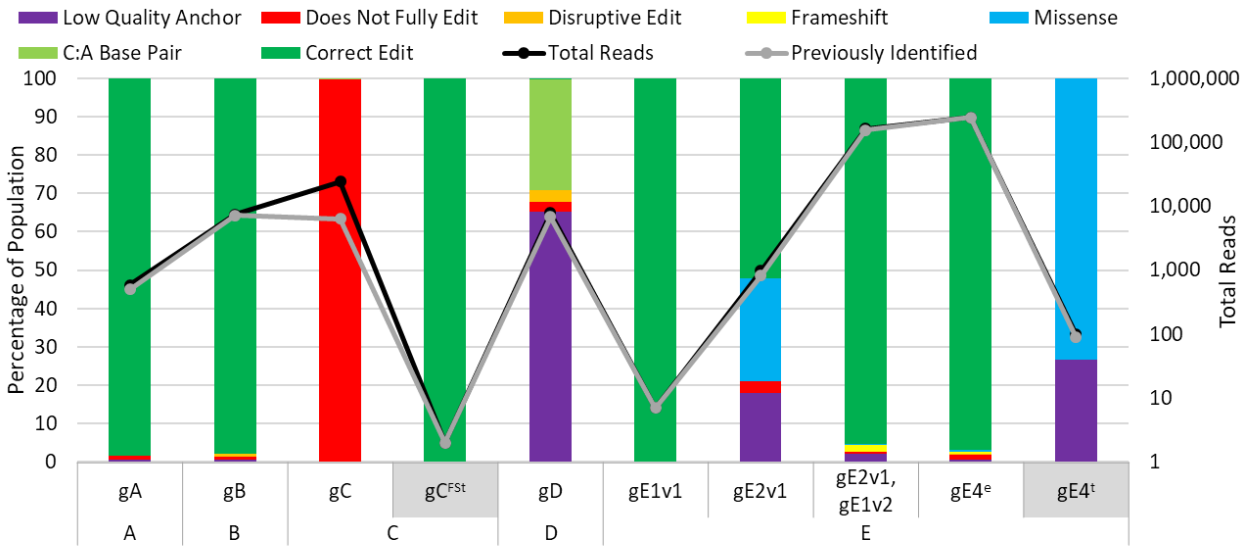


Figure 31. Analysis of functionality and abundance of productive gRNAs that edit ND7 5' in EATRO 164 cells. For description of axes and gRNA functionality labels, please see Figure 27. gRNAs with shaded names were found in the EATRO 164 gRNA transcriptome but were not found to be utilized in editing the EATRO 164 ND7 5' mRNAs, despite being utilized in TREU 667 cells. The population labeled gE2v1,gE1v2 contained members that generated both sequence patterns and could not be separated.

Table 15. gRNA population analysis for ND7 5'

Editing Block	Population	TREU 667 gRNA Reads	TREU 667 mRNA Reads	EATRO 164 gRNA Reads	EATRO 164 mRNA Reads
A	gA	6,819	523,128	589	430,266
B	gB	35,890	510,653	7,478	395,621
	gB^{ex}	113	0	21,775	23,440
C	gC	92	196,504	24,332	67,080
	gC^{FSt}	66	19,717	2	0
	$gC1^{ex}$	161	0	94	38,479
	$gC2^{ex}$	538,052	0	434,196	21,471
D	gD	49,883	133,558	7,781	3,552
E	gE1+/gE2+	14,804	106,379	163,481	942
	$gE4^{t}$	14,772	5,651	98	0
	$gE4^{e}$	53	0	246,486	1,301

CR3

The editing pathways of CR3 are significantly different from those of RPS12 and ND7 5' (Figure 32, Chapter 4). The most obvious difference is the fact that the pathways are highly branched, generating multiple distinctly different mRNA products. The other key difference is that most of the editing pathways defined in the TREU and EATRO cells are not shared. The two cell lines appear to use almost completely different sets of gRNAs to edit CR3, with the exception of some edits at the very 3' end of the transcript. What was most surprising about these data was the fact that most of the gRNAs necessary to generate both the TREU and EATRO CR3 editing pathways were found in both gRNA transcriptomes in similar relative abundances (Table 16, Chapter 4). In addition, similar to the RPS12 and ND7 5' data, we saw very little correlation between the abundance of the gRNA population and the corresponding abundance of the mRNA transcript generated. For example, in the EATRO cell line, the gB1B2 population (38,663 reads) was significantly more abundant than the gB4 guide RNA population (100 reads). Nevertheless, the number of mRNAs with the B4 editing pattern was 5-fold higher

than the B1B2 mRNAs. The EATRO B4 mRNAs can be further edited by three different gRNAs, gB5^{e}, gB6^{e} and gB7^{e}. Again, while gB6^{e} is the most abundant of the three, the predominant mRNA found is the B5^{e} transcript (Table 16). In addition, while the gB5^{e} guide RNA is also abundant in TREU cells, we found no corresponding B5^{e} mRNAs in this cell line. The most significant divergence of the EATRO CR3 editing pattern occurs at the B to C editing block transition. In TREU, all of the different 3' editing patterns converge and are further edited by the gC guide RNA population. In EATRO cells, the gC population is much less abundant (1590 reads instead of >17,000). Correspondingly, very few mRNA transcripts were observed that used a gC guide. Instead, the bulk of the B-level transcripts were further edited by gC^{e} (a rare gRNA with only 43 reads detected). The C-block gRNAs also contained the most abundant gRNA detected for CR3 editing, gC^{e80}. This gRNA has close to 500,000 reads in both cell line transcriptomes. It can anchor to all of the B block mRNAs in both the TREU and EATRO pathways but was only observed editing the B5e template in the EATRO cells. Despite the overwhelming abundance of the gC^{e80} guide RNA population, the corresponding mRNA had the smallest number of detected reads.

Surprisingly, there were very few CR3 gRNA populations in either cell line that had sequence members that could disrupt editing (Figures 33 and 34). In the TREU cell line, the FG^{tx} 5' end pattern of mRNA editing is generated by a promiscuous gRNA, gFG^{txp} (originally identified as a CR4 gRNA). This gRNA population is predicted to cause a frameshift just upstream of the anchor region of gFG^{txp}. However, this frameshift is not observed in the mRNA population, and it appears that this gap is tolerated like those observed in RPS12.

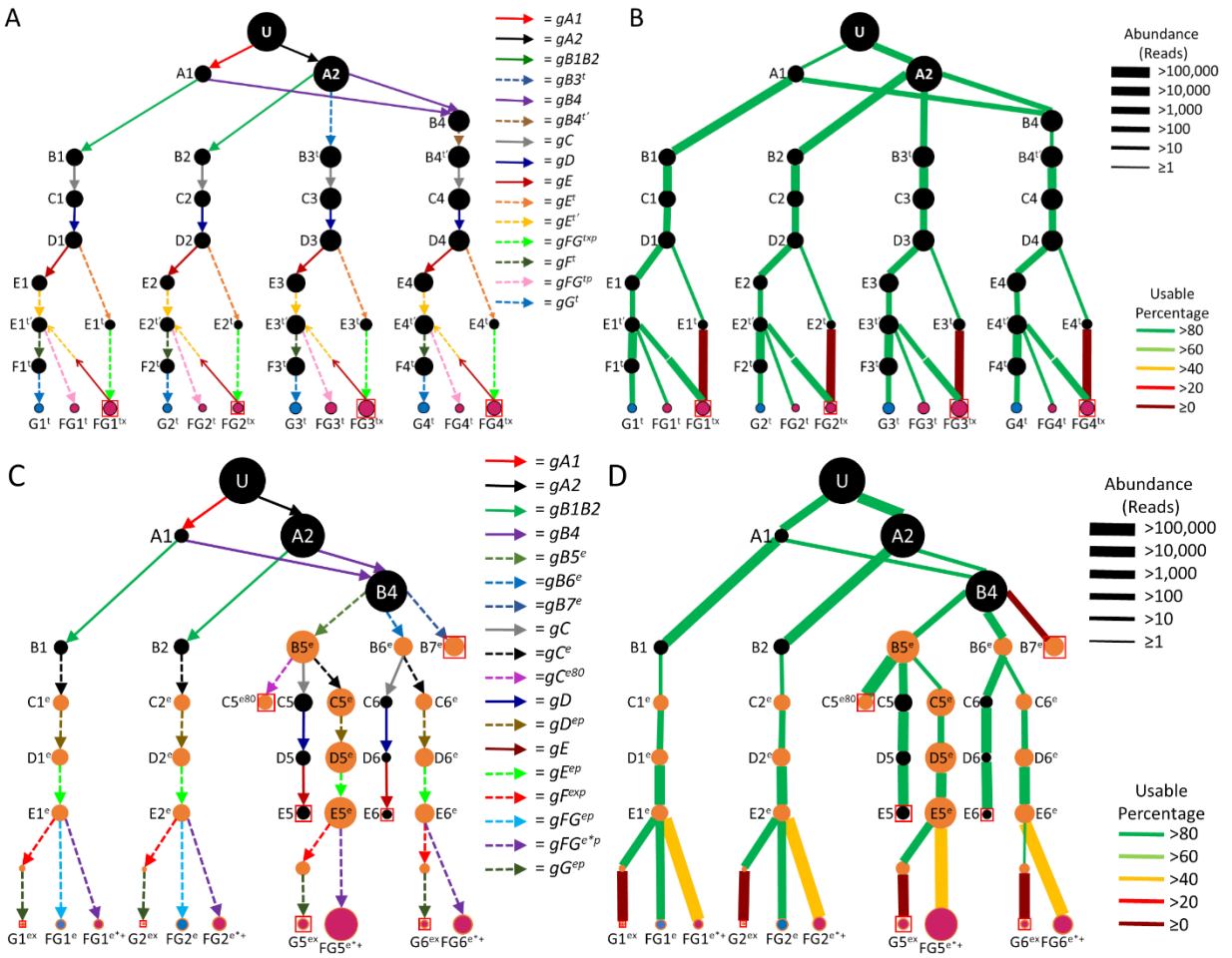


Figure 32. Observed CR3 editing pathways in TREU 667 (A and B) and EATRO 164 (C and D) cell lines. For descriptions of dots and arrows, please see Figure 26. Dashed arrows with open heads represent a hypothetical rewrite. + indicates that more than one mRNA form was condensed into this circle to simplify the figure. Condensed forms encode largely the same amino acid sequence with only small variants. Terminal dots are colored blue for reading frame 1, or magenta for reading frame 2. Boxed magenta dots have no functional start codon but are translatable into reading frame 2 with the use of an alternative start codon (UUG). Lines connecting dots (B and D) indicate gRNA population size and functionality. Usable gRNAs were considered to be those that correctly edit, utilize C:A base pairs to edit, or generate only small missense mutations.

Table 16. CR3 gRNA population analysis

Editing Block	Population	TREU 667 gRNA Reads	TREU 667 mRNA Reads	EATRO 164 gRNA Reads	EATRO 164 mRNA Reads
A	gA1	100	5,907	2,368	534
	gA2	1,729	28,654	19,947	5,934
B	gB1B2	13,278	10,776	38,663	757
	$gB3^{t}$	1,598	9,158	0	0
	gB4	945	9,536	100	3,720
	$gB4^{t'}$	12	8,692	0	0
	$gB5^{e}$	3,800	0	803	2,025
	$gB6^{e}$	221	0	2,359	689
	$gB7^{e}$	1,376	0	369	744
C	gC	17,045	27,646	1,590	616
	gC^{e}	37	0	43	2,079
	gC^{e80}	538,230	0	434,196	270
D	gD	1,649	26,287	30,125	280
	gD^{ep}	24	0	118	1,981
DE	gDE^{e80}	301	0	120	0
E	gE	9,135	15,016	15,242	204
	gEt'	567	14,687	582	0
	gE^{t}	34	5,099	0	0
	gE^{ep}	14,062	0	30,982	1,864
F	gF^{t}	34,666	8,073	189,828	0
	gF^{exp}	0	0	1,265	112
FG	gFG^{tp}	14	1,035	90	0
	gFG^{txp}	49,892	2,172	3,220	0
	gFG^{ep}	219	0	1,912	79
	gFG^{e*p}	69,138	0	104,747	691
	gFG^{e80}	255	0	589	0
G	gG^{t}	466	1,231	21,788	0
	gG^{ep}	38,001	0	17,318	68

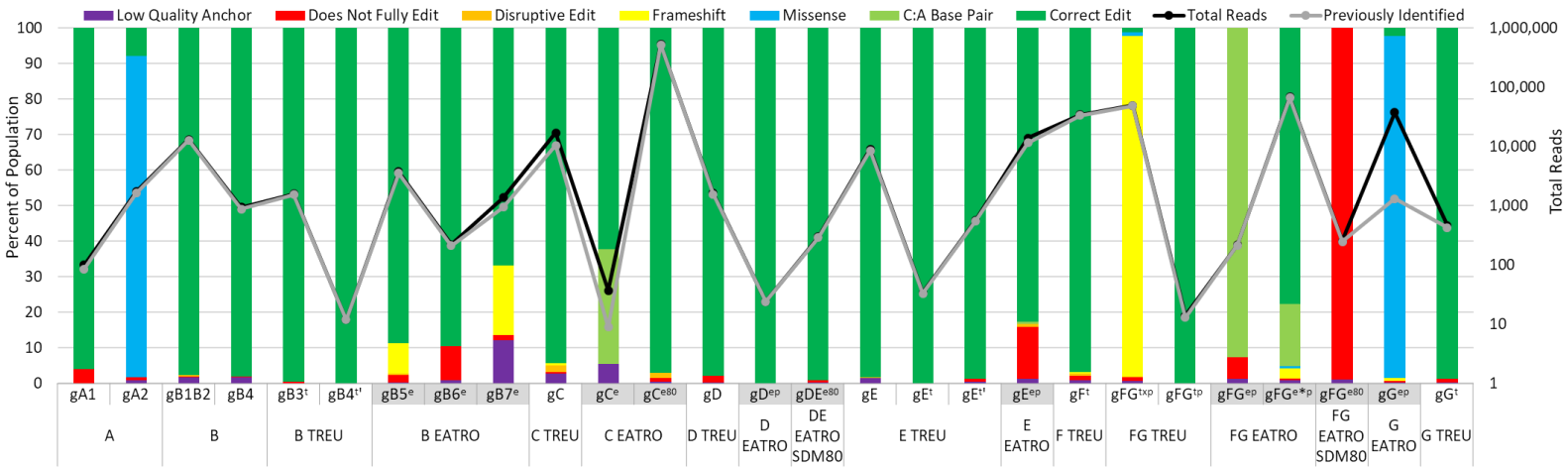


Figure 33. Analysis of functionality and abundance of gRNA subpopulations that edit CR3 in TREU 667 cells. For description of axes and gRNA functionality labels, please see Figure 27. gRNAs with shaded names were found in the TREU 667 gRNA transcriptome but were not found to be utilized in editing the TREU 667 CR3 mRNAs, despite being utilized in EATRO 164 cells.

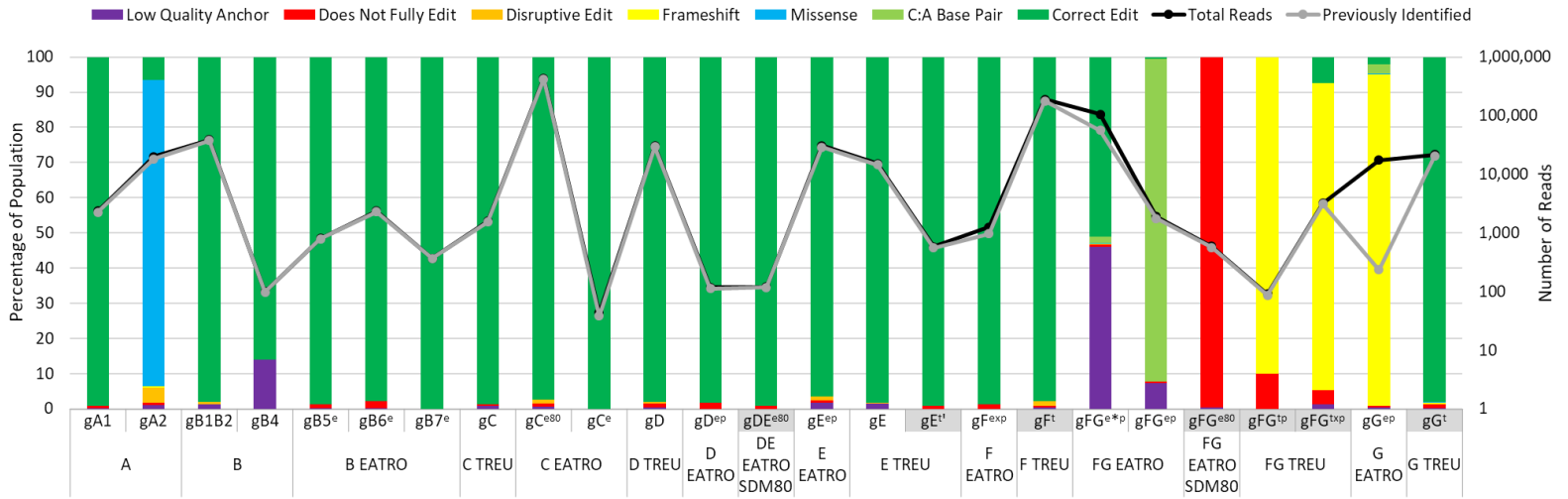


Figure 34. Analysis of functionality and abundance of gRNA subpopulations that edit CR3 in EATRO 164 cells. For description of axes and gRNA functionality labels, please see Figure 27. gRNAs with shaded names were found in the EATRO 164 gRNA transcriptome but were not found to be utilized in editing the EATRO 164 CR3 mRNAs, despite being utilized in TREU 667 cells.

In the EATRO cell line, only two terminal gRNA populations have predicted editing problems. About half of the gFG^{e*p} subpopulation has a low-quality anchor (<5 nt), however the subpopulation does have >53,000 reads that have higher quality anchors, so this mutation may be tolerated. The gG^{ep} population is also interesting, in that it is predicted to generate a frameshift when editing was simulated by our GUIDE program (Figure 34). This frameshift is not observed in the mRNA transcriptome data however because editing appears to stop before reaching the last two sites (Chapter 4).

Discussion

ACORNS and GUIDE are two new tools that can aide in the understanding of the complex dynamics of the kinetoplastid RNA editing system. The ability of ACORNS to cluster related gRNAs proved to be a powerful mechanism for the identification of gRNAs with mutations that disrupt their alignment to fully edited sequence. These analyses allowed us to identify and characterize nearly 2 million additional gRNA transcripts from our transcriptome libraries. In addition, these analyses identified a large cohort of gRNA clusters that are not involved in directing the sequence changes associated with the known canonical transcripts. More than 25% of the gRNA transcripts found in the EATRO and TREU gRNA transcriptomes have no known functional relatives. This strongly suggests that the coding capacity of the mitochondrial genome is much larger than previously thought.

Full characterization of the known gRNA population was also informative. In our initial gRNA characterization study, we had noted the extreme population differences found between the different identified gRNAs [9]. We had hypothesized that the low copy number gRNAs were an artifact of the high stringency of our initial screen. The cluster analyses suggest that extreme

population size differences do exist between the different gRNAs. In both the TREU and EATRO transcriptomes, ~30 clusters were identified with over 500 sequence members each. These clusters accounted for the bulk of the gRNA transcript reads (6,461,870 reads in TREU and 4,539,263 reads in EATRO). In contrast, over 500 clusters were found that contained fewer than 25 sequence members. These tiny clusters account for only ~135,000 transcript reads. The very large numbers of low copy number gRNAs are suggestive of high plasticity in the gRNA encoding minicircles. Studies in *Leishmania* have suggested that minicircle sequence class frequencies are extremely variable [63,78]. How these huge differences in gRNA abundance influences gRNA selection and use is unclear. In only ~50% of the editing branch points characterized for RPS12, ND7 5' and CR3, did the abundance of the gRNAs involved somewhat correlate with the preferred editing path. However, even when gRNA abundance did align with mRNA abundance, they were often not proportional. In addition, block editing by an abundant gRNA is often followed by editing using a rare gRNA, with no equivalent drop in editing efficiency. In addition, multiple instances were observed, where highly abundant gRNAs did not appear to be used in one cell line. In one example, the gRNAs responsible for generating the $B5^e$, $B6^e$ and $B7^e$ mRNA forms of RPS12 are present in both cell lines, but only apparently act in the EATRO cell line, despite being all more than ten times more abundant than their only competitor. It may be that protein factors play a predominant role in gRNA selection. One study examined the role of the RNA editing mediator complex (REMC), which is heterogenous and consists of one primary subunit TbRGG2 that formed associations with either MRB8170 or MRB8180 [161]. They showed that depletion of MRB8180 caused global effects on RNA editing, but depletion of MRB8170 had transcript specific effects, substantially increasing the amount of

pre-edited RPS12 transcripts, but not significantly affecting the amount of pre-edited ND7 5' transcripts. This specificity is intriguing, and it is possible that the REMC or other protein factors are involved in gRNA selection.

The identification of all relatives of conventional gRNAs, including those that had undergone a mutational event, also allowed us to more carefully characterize the effect of mutational noise on the RNA editing system. These analyses indicate that the RNA editing system can tolerate more mutational noise than we had originally hypothesized. Surprisingly, a large number of the gRNA populations we characterized had base pair mismatches with their best aligned mRNA transcripts. In this study, we saw that while most of the mismatches were specifically C:A base pairs, almost every other mismatch was also observed. Even gaps in the alignment of the gRNA to the mRNA appeared to be tolerated. In almost all cases however, the mismatch base pair and/or gap appeared to be stabilized by multiple flanking G:C base pairs. While we cannot rule out the possibility that rare, perfect match gRNAs do exist for these regions, it may be that the incompletely base paired interaction is the most stable structure possible, hence is generated by that gRNA [162]. While a minimum number of non-paired nucleotides does appear to be tolerated within the guiding region, mismatches or non-Watson-Crick base pairs within the anchor region do not appear to be tolerated, greatly affecting the efficiency of the RNA editing process [119].

In a large number of the gRNA populations containing alignment mismatches, our GUIDE program predicted that the mismatch would drive the generation of an alternative edit. However, these alternative edits were not observed in the mRNA transcriptome data. These data contradict one of the most prominent models of RNA editing progression, known as the

"mismatch recognition" model [33]. In this model, when the gRNA/mRNA duplex initially forms, the editosome proceeds to edit beginning at the first mismatch site closest to the anchor binding region. Once this mismatch is resolved either by the insertion or deletion of a uridine, the next mismatch site is edited, and no sites further will be edited until the sites nearest the anchor binding region are resolved. When GUIDE predicts editing patterns, it follows this model, predicting each editing site, moving from the anchor binding region towards the poly-U tail. If editing did follow this strict mismatch recognition model, the alternative sequences predicted by GUIDE should have been observed in the mRNA transcriptome data. An alternative model suggests that RNA editing occurs via a more "dynamic interaction" [162]. This model proposes that when a gRNA/mRNA duplex forms, the editosome targets regions of the duplex with low thermodynamic stability and edits those regions. As editing progresses, the duplex realigns, changing the targets of the editing system. These cycles of progressive realignment proceed until the gRNA/mRNA duplex reaches maximum stability. In this way, RNA editing does not necessarily proceed in a strict 3' to 5' directional manner. This model suggests that mismatches and gaps in alignments can be tolerated because they do not significantly impact the stability of the final gRNA/mRNA duplex. Supporting this, are the frequent observations of neighboring G:C base pairs, which would substantially enhance the stability of these mismatches. The "dynamic interaction" model is further supported based on the existence of "junction regions" in partially edited mRNAs [161–165,156,166,148,160]. These regions adjoin the unedited and fully edited regions of a partially edited mRNA, but do not match either the unedited or fully edited sequence. Junction regions vary significantly across partially edited transcripts, possessing no consensus sequence. These regions can vary in size

and depletion of different protein factors affects their occurrence and length, but the presence of junction regions remains ubiquitous across all partially edited mRNAs [161–165,156,166,148,160]. The mismatch recognition model reconciles the presence of junction regions as areas of mis-editing (utilization of the wrong gRNA or a misaligned gRNA), hence all junction regions would have a gRNA capable of generating the sequence. In our characterization of the editing pathways of RPS12, ND7 5' and CR3, highly abundant mRNA sequences were screened against the gRNA transcriptomes at low stringency levels in order to identify true alternative edits. While we were able to identify and number of gRNAs that could direct alternative edits, a large number of "junction sequences" were identified that do not match any gRNA in our databases. If the "dynamic interaction" model is correct, these multiple variable junction sequences could be generated by the same gRNA during the editing process. Alternative base pairs have also been shown to be tolerated to different extents in an *in vitro* gRNA directed deletion assay [151]. In this study, substitutions were made of the nucleotides immediately upstream of a deletion site. This study found that when the base pair upstream of the deletion site was C:A, C:U or a C:C, the site was still found deleted in the mRNAs. Deletions were not observed when the base pair was a G:A or G:G.

Another facet of gRNA utilization is the existence of promiscuous gRNAs, gRNAs editing more than one target. In this analysis we showed many populations that edit more than one gene, and one population editing the same gene in two different locations (gF^p of RPS12). Interestingly, most of the promiscuous populations were found in the CR3 data set, and these promiscuous gRNAs were the only productive promiscuous gRNAs identified. No promiscuous populations were found editing ND7 5', and the only promiscuous gRNAs found to edit RPS12

lead to editing pathway dead ends. Predictions made by GUIDE indicate that many of the promiscuous gRNAs generating the CR3 EATRO specific pathways should also be able to productively edit in TREU. Because these are promiscuous gRNA, it is possible that their availability is impacted by their use in editing other transcripts. A full mRNA transcriptome would allow a full analysis of the global impacts of editing on any particular gRNA cluster.

This study revealed the surprising amount of noise and errors that are tolerated in the RNA editing system of *Trypanosoma brucei*. Some questions still remain, such as the functions of the large proportion of unknown gRNAs, how gRNAs are selected, and what is the true extent of gRNA promiscuity. In order to answer these questions, we believe that a full deep sequence of all edited mRNAs paired with gRNA transcriptomes of multiple cell lines would shed more light on this complicated situation.

Acknowledgements

We would like to thank the Ken Stuart Lab for the trypanosome cell lines and Chris Adami for his assistance in the conceptualization of this work.

CHAPTER 6: SUMMARY AND DISCUSSION

Introduction

Trypanosoma brucei is one of the few organisms that utilizes the kinetoplastid RNA editing system. This system seems unnecessarily complex, using two genetic components to generate one fully functional product. Moreover, this system is prone to malfunction; each gRNA generates the anchoring region for the next gRNA, making the system sequentially dependent. This makes the mutation or loss of any gRNA along the editing pathway extremely detrimental, especially considering that two of the twelve edited genes are essential [17,100]. This problem is made worse by the fact that some gRNAs are incredibly rare, and during replication, the 5,000-10,000 minicircles encoding the gRNAs are divided asymmetrically, making minicircle loss not only possible, but routine [8,167]. This system should not work, but it does. Kinetoplastids are some of the most successful parasites on Earth, infecting insects, plants, mammals, fish, birds, and reptiles [168].

The study of this editing system inevitably brings up questions of how this system evolved and how it continues to be maintained. The concept of drift robustness begins to explain this surprising amount of fragile complexity [80]. Drift robustness is a form of genetic robustness that allows an organism to be protected from extreme events of genetic drift by making mutations either neutral or lethal. *T. brucei* frequently undergoes population bottlenecks throughout its life cycle and as its mitochondria is completely asexual, it seems particularly prone to genetic drift [22]. The fragility of the RNA editing system seems to coincide remarkably well with the idea of drift robustness. By rendering mutations or loss of

minicircles lethal, this would prevent accumulation of slightly deleterious mutants in the population. But such a system should suffer from another disadvantage. In a system where mutations are lethal or neutral, how can the organisms continue to evolve?

In the first examination of the gRNA transcriptome, many gRNAs were identified that were capable of generating alternative edits [9]. These edits ranged from having no effect on the protein sequence to causing a frameshift and altering a large portion of the protein. These findings sparked this project, which sought to understand the impact of alternative editing on genetic integrity, developmental regulation, protein diversity and editing efficiency. These goals were accomplished through the generation of the bloodstream gRNA transcriptome and comparative analysis of it with the insect stage gRNA transcriptome, analysis of dual-coding genes that utilize alternative edits to access multiple reading frames, the generation of libraries of putative dual-coding mRNAs at different states of editing, and analysis of gRNA population diversity and the impact of that diversity on the editing system as a whole.

Summary of Chapter 2

This chapter characterized the gRNA transcriptome of EATRO 164 bloodstream stage *Trypanosoma brucei*, and compared it to the gRNA transcriptome of procyclic stage trypanosomes. As with the procyclic gRNA transcriptome, conventionally accepted fully edited mRNA sequences were used to identify the gRNAs, and a comparison of the two life cycle transcriptomes show a 3.5:1 ratio of procyclic to bloodstream gRNA reads. This ratio varies significantly by gene and by gRNA populations within genes. The variation in the abundance of the initiating gRNAs for each gene, however, displays a trend that correlates with the developmental pattern of edited gene expression. Surprisingly, there were very few gRNAs

found in both transcriptomes, but there were many gRNAs that appeared to be related between transcriptomes. Comparing these related major classes from each transcriptome revealed a median value of ten single nucleotide variations per major class. Nucleotide variations were much less likely to occur in the consecutive Watson-Crick anchor region, indicating a very strong bias against G:U base pairs in this region. In spite of the variation we saw between related gRNAs, we did find several conserved gRNA characteristics, such as transcription start site sequence, length of complementarity, and non-base pairing nucleotides at the 5' and 3' ends of gRNAs. Overall, gRNA coverage of edited mRNAs as well as overlap between adjacent gRNAs was lower in the bloodstream gRNA transcriptome than in the procyclic gRNA transcriptome.

This work indicates that gRNAs are expressed during both life cycle stages, and that the differences in the extent of editing previously reported for different mRNA transcripts are not due to the presence or absence of gRNAs. However, the abundance of the initiating gRNAs may be important in the developmental regulation of RNA editing.

Summary of Chapter 3

In this work, we show that many of the mitochondrially edited mRNAs in *T. brucei* can alter the choice of open reading frame by alternative editing of the 5' end. Dual-coding genes have specific mutational biases, such as an increase of the ratio of nonsynonymous to synonymous mutations, and an increase in mutational frequency overall. Analyses of mutational bias of all mitochondrial genes indicate that six of the pan-edited genes may be dual-coding. These analyses include measuring the conservation of editing patterns between *T. brucei* and *T. vivax* transcripts and frequencies of different types of mutations. These data were

used in a principal component analysis, which showed a distinct difference between alternative reading frames of dual-coding genes and single-coding genes. Discovery of alternative gRNAs reveal that RNA editing can allow access to both reading frames. We predicted the functions of two of these alternative reading frames as small metabolite transmembrane transporters. We hypothesize that dual-coding genes can protect genetic information by overlapping genes that are under selection at different portions of the life cycle.

Summary of Chapter 4

In this study, we analyzed the editing patterns of three putative dual-coding genes, ribosomal protein S12, the 5' editing domain of NADH dehydrogenase subunit 7, and C-rich region 3, and constructed detailed editing pathway maps using mRNA and gRNA transcriptome data. While editing of RPS12 showed only transcripts that produce the canonical RPS12 protein, we did observe a second downstream start codon capable of producing the alternative protein, if selected by the ribosome. In ND7 5' and CR3, we found evidence that both of these transcripts are edited to express protein products in more than one reading frame. Moreover, we found that CR3 has a very complex set of highly branched editing pathways that vary significantly between cell lines, with a different set of gRNAs being used in each cell line, despite both sets of gRNAs being present in both cell lines. We also found that changing the energy source available to cells also alters the editing preferences of both CR3 and ND7 5'. In addition to this, we found evidence that the poly-U tail that is added post transcriptionally to gRNAs may also be used in editing. These findings suggest that these reading frames can be alternatively selected based on the current environment, and that alternative editing may be a way for the trypanosomes to continue evolving this rigid editing system.

Summary of Chapter 5

In the analysis of the gRNA transcriptomes, gRNAs were identified based on complementarity to edited mRNAs. While millions of gRNAs were found using this method, we discovered that many gRNAs were still left unidentified. This high proportion of unidentified gRNAs was shocking, as we had predicted that the RNA editing system should be intolerant of mutations, and the presence of so many unidentified gRNAs meant that many editing pathways had not been characterized, or many mutated gRNAs were also present. To determine the identity and function of these gRNAs, two new programs were created, ACORNS and GUIDE. The first program functions to group related gRNAs into clusters, where each cluster had significant sequence conservation. This analysis showed that more than half of all unidentified gRNAs were not related to any functionally known gRNAs and could be generating uncharacterized alternative editing patterns.

However, there were still many gRNAs that were related to previously identified gRNAs. These gRNAs could be capable of disrupting the editing process or generating small alternative edits that are tolerated. In order to investigate this, our second program, GUIDE examined the gRNAs responsible for generating the editing pathways of RPS12, ND7 5' and CR3, as well as their previously unidentified relatives. Using the defined gRNA clusters, GUIDE screened all members editing these three genes against their targets to determine what proportion of each family was able to productively edit. The initial analyses of these previously unidentified gRNAs revealed than nearly all were predicted to disrupt the editing system, but in our examination of the mRNA data, we found this not to be the case.

By combining the analyses of the GUIDE program and the mRNA transcriptome data, we learned more about the robustness of the RNA editing system. In examining the RPS12 gRNAs we found that single mismatches or gaps were highly tolerated in gRNA alignments, with only very small drops in editing efficiency. In the examination of the ND7 5' gRNAs, however, we identified the limit of this tolerance. Significant drops in editing efficiency were observed when gRNAs either possessed multiple mismatches or gaps, or possessed mismatches that disrupted the anchor binding region. Finally, in our analysis of the CR3 gRNAs, we observed many gRNA populations of high abundance that were apparently not used to edit. We found that RNA editing preference does not correlate positively or negatively with gRNA abundance, and in most cases, the apparently unused gRNA populations had no issues in their mRNA alignments. We predict that these gRNAs may be in use elsewhere in the editing system, and to fully understand this system, we recommend that a full mRNA transcriptome be analyzed.

Genetic Integrity

The introduction and maintenance of kinetoplastid RNA editing

As previously mentioned, the sheer size and complexity of the RNA editing system has left many in search of the answers to how it evolved and how it has been maintained. Many hypotheses have been proposed, such as it being a relic of the old RNA world, being a product of constructive neutral evolution, or that the system co-evolved with G-quadruplex structures that served to protect the genetic information [69–74]. As for how it has been maintained, one prominent theory is that RNA editing is advantageous because it is a mechanism by which an organism can fragment and scatter essential genetic information throughout a genome [75,76]. Because kinetoplast DNA is less stable than chromosomal DNA, and minicircles are frequently

lost due to asymmetric division, this hypothesis suggests that scattering essential guide RNA genes throughout the DNA network would prevent fast growing deletion mutants from outcompeting more metabolically versatile parasites during growth in the mammalian host [76,77].

We propose that the RNA editing system, and its inherent fragility operate as a system to weed out deleterious mutations by making them lethal, as a form of drift robustness. Drift robustness is not adaptive, however, and prevents the population from generating beneficial mutations as well. This strategy leaves organisms no options for evolving. We suggest that alternative edits, such as those seen in CR3 and others previously observed, editing pathways generate this evolution without compromising the rigid conservation of other genes such as the essential RPS12 [68].

Supporting this is the fact that, of the kinetoplastids, *Trypanosoma* have some of the harshest life cycles in terms of maintaining genetic integrity, with the electron transport genes under very relaxed selection in the glucose rich bloodstream stage, and very strict selection imposed in the insect stage. In conjunction with this, we observe that *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Trypanosoma vivax* all maintain more genes that are edited and more extensive editing than their other kinetoplastid counterparts with milder life cycles [4,16,19–21,52–56,58,59]. For example, *Phytomonas serpens* infects important crops and is transmitted by sap feeding bugs. These parasites have glucose readily available in both life cycle stages, and are unique in that they lack a fully functional respiratory electron transport chain, and this species is also missing two edited genes entirely, and pan-edits six genes and partially edits three genes, compared to the nine pan-edited and three partially edited genes in the

Trypanosoma spp. [64,65]. For *Leishmania spp.*, all life cycle stages possess an active Krebs cycle and ETC linked to the generation of ATP, but these cells are never restricted from access to glucose [61,62,139]. *Leishmania* do not pan-edit ATPase 6 or ND7, but only partially edit them [63]. These observations suggest that RNA editing provides a larger advantage to organisms with a more complex life cycle.

Dual-coding and dual-function genes

One oversite that has not been accounted for in the hypotheses that attempt to explain the advantage of kinetoplastid RNA editing is how genetic material that is not under selection is maintained. There has been considerable debate on the necessity of Complex I subunits for either stage of the trypanosome life cycle. Studies using RNAi and knockout cell lines of nuclear-encoded members of Complex I have shown that the complex is unnecessary for survival in either life cycle stage, and in this work we were unable to find complete gRNA coverage of the edited ND subunits in the bloodstream stage, despite the ND subunits generally being more fully edited or only fully edited in the bloodstream stage [5,26–29,111,112]. However, the nuclear encoded Complex I member genes are maintained [42], and the vast majority of the gRNAs required to edit mitochondrially encoded ND subunits were found in both life cycle stages. This evidence together suggests that the Complex I proteins are vulnerable to genetic drift but have somehow been maintained.

In this work, we propose that by overlapping Complex I genes not under strict selection with genes that are under selection, the accumulation of mutations can be prevented. Because these overlapped genes share most gRNAs, and alternative edits only occur in the terminal gRNAs, this strategy ensures that almost all of the genetic material is protected. The genes

predicted to be dual-coding based on our mutational bias analysis include almost all of the edited ND transcripts, excluding only ND8. In the examination of procyclic ND7 5' and CR3 (putative ND4L) we found that these two genes are alternatively edited to produce more than one protein product [120]. This evidence shows that not only are dual-coding genes being utilized in *T. brucei*, but also that RNA editing is facilitating the use of these dual-coding genes, providing the world with a concrete advantage of utilizing this type of RNA editing.

Based on limited sequence homology, we hypothesize that the trypanosome mitochondrial alternative reading frames (ARFs) encode small metabolite transporters that provide a distinct growth advantage to bloodstream form parasites. These proteins would function differently from all other edited proteins, which do not function as transporters. While it has been previously suggested that both alternative editing and dual-function proteins are important mechanisms for expanding the functional diversity of proteins found in trypanosomes, a duplication event could easily alleviate the evolutionary constraints imposed by dual-coding genes [67,97–99]. We maintain that in salivarian trypanosomes, these genes must provide the additional benefit of protecting genetic information in order to continue to be overlapped. Protection of the mitochondrial genome during growth in the mammal would increase the capacity for successful transfer to an insect vector and maximize the parasites long-term survival and spread.

Analyses of other trypanosomes do show that some of the ARFs have intriguing homology to the ARFs identified in *T. brucei* and *T. vivax*. However, most of the ARFs are punctuated with stop codons. It is possible that these genes have since lost their function and are no longer required to be dual-coding due to the reduced selective pressures endured during

their life cycles, or it is possible that these stop codons are removed by alternative editing events.

In addition to the dual-coding genes we have described, another dual-coding gene, COIII, is accessed through alternative editing, by connecting an unedited 5' reading frame with an edited 3' reading frame, and the alternative protein produced, AEP-1, has been shown to be essential [67,68]. In addition, there are known Krebs cycle proteins (α-ketoglutarate dehydrogenase E2 and α-ketoglutarate decarboxylase) that have two functions, rendering them protected while they are not under selection as well [97,98]. These data show that trypanosomes are utilizing dual-coding and dual-function genes to protect genetic integrity.

Another set of dual-function genes appear in the gRNAs, as promiscuous gRNAs. The editing pathways of CR3 are littered with gRNAs identified to edit other genes, such as ND8 and the 3' editing domain of ND7, both of which were not predicted to be dual-coding, and are under less protection. We believe that this is yet another mechanism of increasing the drift robustness of *T. brucei* through the use of alternative RNA editing.

Developmental Regulation

In the examination of the bloodstream gRNA transcriptome, we found the abundance of the initiating gRNAs in the procyclic and bloodstream gRNA transcriptome is correlated with the developmental editing patterns of the genes they edit. However, we cannot rule out the possibility that not all of the populations of initiating gRNAs were identified. We identified alternative initiating gRNAs for CR3, and without deep sequencing all pan-edited mRNAs, we can't know that others don't exist.

It has been previously reported that gRNA presence did not correlate with developmental RNA editing patterns in *T. brucei* [50,51]. This, however, was reported on the observation of a very limited number of gRNAs, but we found that for the most part, this held true, with gRNA populations having similar relative abundances across both transcriptomes, with, of course, the exception of the initiating gRNAs.

In our examination of the editing pathways of RPS12, ND7 5', and CR3, we found many editing patterns that were only observed or were more prominent in only one cell line. This was most prevalent in CR3, where the editing pathway of the TREU cells is almost completely different from that of the EATRO cells. Curiously, the gRNAs required for both pathways are present in the transcriptomes of both cell lines in relatively equal abundances. Additionally, another exclusive editing pathway was discovered when the EATRO cells were moved into glucose depleted medium. This pathway produced a transcript that would make a unique protein, totally different from all other CR3 protein products. These gRNAs appear functional in both cell lines but are perhaps being used to edit a different gene when they are not observed to be editing CR3. Indeed, many of the EATRO specific CR3 gRNAs are promiscuous gRNAs known to edit other genes.

To better understand the complexities of gRNA selection, we examined the editing branch points of the RPS12, ND7 5' and CR3 pathways. Examination of editing branch points revealed that the abundance of the gRNAs involved did not correlate with the observed editing preference. This work has raised many questions about how gRNAs are selected and editing pattern preferences are exerted, and more study is needed to understand this system.

Protein Diversity

In the examination of the RPS12 mRNAs, we found one major alternative editing event. This event was observed in the EATRO 164 cell line only, and causes a frame shift that extends the reading frame at the 3' end by nine amino acids. This same alternative edit was previously described in the 29-13 strain as well, which shows that this edit is not an isolated occurrence in the EATRO 164 strain [148].Frameshifting gRNAs were also identified to edit ATPase 6 in both bloodstream and procyclic EATRO 164 cells. The predicted frameshifts also occur close to the 3' end of the transcript and alters the C terminus of the protein. As the frameshifts occur downstream of the highly conserved amino acid region involved in proton translocation, it may be that this is also tolerated [31].

Edits of the ND7 5' mRNAs generate transcripts that can translate into two reading frames in TREU 667 and EATRO 164 cells. Interestingly, this gene was also sequenced in 29-13 cells [148]. That study indicated that a large proportion of the fully edited ND7 5' transcripts had a single nucleotide difference in the 5' UTR. As the upstream start codon that allows the ARF to be translated is in what is known as the 5' UTR, this difference could be the same alternative edit that generates the ARF transcripts. This suggests that this alternative editing may be widespread.

Like ND7 5', we also detected evidence of dual-coding in CR3. The alternative edits in both cell lines produce multiple variations of the CR3 proteins. While the editing efficiencies of CR3 are very low, possibly preventing the generation of these CR3 proteins, we propose that this mechanism of branched editing pathways is a safe way for the trypanosomes to introduce variation into the rigid system and continue to evolve.

In addition to the high level of alternative editing observed in CR3, in our re-examination of the procyclic EATRO 164 and TREU 667 gRNA transcriptomes, we identified many gRNAs in the existing transcriptomes still have no known function. These gRNAs may be generating alternative editing events, thus further increasing the protein diversity of *T. brucei*.

Editing Efficiency

The editing efficiencies of all three genes we examined were surprisingly low. The efficiencies of TREU cell line mRNAs were all less than ten percent, while the EATRO 164 cells grown in SDM79 were all less than one percent. With CR3 and RPS12, more than 80% of mRNAs were completely unedited, while those numbers were much lower in ND7 5'. There are many factors we observed that had the potential to affect the overall editing efficiency.

Mutations and non-canonical base pairs

Because gRNAs utilize both canonical (Watson-Crick) as well as G:U base-pairing to direct the change in sequence, most transition mutations in the gRNA, would not lead to changes in the mRNA sequence and would not be selected against [33]. In our observations of the gRNA transcriptome, we found a very strong bias against A to G transitions in the anchor regions of the gRNAs, suggesting that G:U base-pairing is not well tolerated in this region. However, we also found many populations of gRNAs where non-canonical base pairs appeared to be tolerated, even in essential genes, ATPase 6 and RPS12 [17,100,107,110]. In the bloodstream database, a gRNA population with a C:U base pair must be tolerated to be able to complete the editing pathway of ATPase 6, and in RPS12, we observed a gRNA population that requires toleration of a gap in the alignment to be capable of fully editing.

In addition to this, RPS12 and ND7 5' possess to highly cytosine rich regions that typically have poor gRNA coverage. These regions both encode conserved amino acids vital to the functions of the proteins. When we deep sequenced these regions of the mRNAs, we had hoped to find alternative sequences that allowed us to identify a more abundant population to carry out these edits, but we only found the previously described editing patterns. Using these sequences and performing a search of the gRNA databases at a low stringency yielded populations with imperfect anchors, requiring noncanonical base pairs to be tolerated. In RPS12, this did not appear to affect editing efficiency, but in ND7 5', the effects were severe. However, the gRNA population identified to edit this region of ND7 5' had multiple mismatches and gaps in both cell lines, whereas the RPS12 populations did not.

These examples are far from isolated incidents. Many other populations were identified where noncanonical base pairs were required to generate fully edited mRNA sequences. In most cases, these alternative base pairs do not seem to affect the editing efficiency. The most prominent mismatch was the C:A base pair, but almost every other mismatch was observed. Even gaps in the alignment of the gRNA to the mRNA seem to be tolerated. Of note, we did observe that most mismatches were flanked by nearby G:C base pairs that may have assisted in stabilizing the alignments. The use of alternative base pairs has been previously shown to be tolerated at different extents, with the use of C:A, C:C and C:U base pairs not completely disrupting editing [151]. This evidence suggests that the RNA editing system will tolerate some amount of illegal base pairs, there is a limit to what will be tolerated.

The use of non-canonical base pairs in RNA editing support the model of editing known as the "dynamic interaction" model [162]. In this model, editing of an mRNA in a gRNA duplex

does not proceed in a site by site fashion, but instead, the editosome targets regions of low stability first, and continues to edit and re-edit the mRNA until a thermodynamically stable gRNA/mRNA duplex is achieved. In this model, non-canonical base pairs may be tolerated if they do not significantly disrupt the stability of the duplex.

Overwriting

In our exploration of partially edited mRNAs, we found that RNA editing is not always strictly sequential. While overall, editing proceeds from the 3' to 5' across of the editing domain, we have found evidence that shows that gRNAs can overwrite editing that has been previously generated. Another study that examined the use of alternative RPS12 gRNAs previously identified found that the three gRNAs in question were being utilized, and that despite not generating the conventional editing patterns, a very small number of transcripts returned to the canonical editing pattern after the alternatives [148]. This last observation may be another instance of gRNA overwriting. This is another source of noise, lowering overall editing efficiency.

Future Work

In this work, we proposed the hypothesis that *T. brucei* utilizes dual-coding genes to protect vulnerable genetic material from genetic drift and accesses the overlapped reading frames through alternative RNA editing. In order to test this hypothesis, we should confirm the presence of the alternative protein products *in vivo* and knock down these products to determine if do provide a benefit to the cells. Knocking down edited transcripts has proved difficult, but not impossible. One study was able to genetically engineer an artificial site-specific RNA endonuclease to target the ATPase 6 edited mRNA [110]. This engineering

requires an eight-nucleotide specific target sequence, and in order to use this technique, this would require finding a target sequence specific to the alternatively edited mRNAs only, and leave intact the mRNAs expressing the canonical ETC proteins, which could easily prove difficult as we have found that these sequences can be quite similar.

Another approach to tackle this problem would be to sequence the dual-coding transcripts from cells grown under a variety of energetic conditions. By varying the conditions and identifying which conditions the alternative proteins were most prominently expressed in, this could help in elucidating the functions of these proteins. Once a baseline for each set of conditions was determined, the artificial site-specific RNA endonucleases could be engineered to target the unedited transcripts for each dual-coding gene and observe the effects on cell growth in the varied conditions. This could give insight into the importance of the two different overlapped proteins at different points in the trypanosome life cycle.

Another direction to pursue is understanding the mechanisms of gRNA selection. We observed many promiscuous gRNAs, and it is possible that even more promiscuity exists in the system. To better understand this, it would be necessary to deep sequence all of the edited mRNAs and have paired gRNA transcriptome data. Ideally, this would be done in multiple cell lines, and under various energetic conditions, as we saw significant variation in our two cell lines and media conditions. This would allow us to see the full picture of alternative mRNA editing and gRNA usage. Then we could begin to determine how much gRNA promiscuity plays a role in gRNA selection, or if other factors are in play.

Conclusion

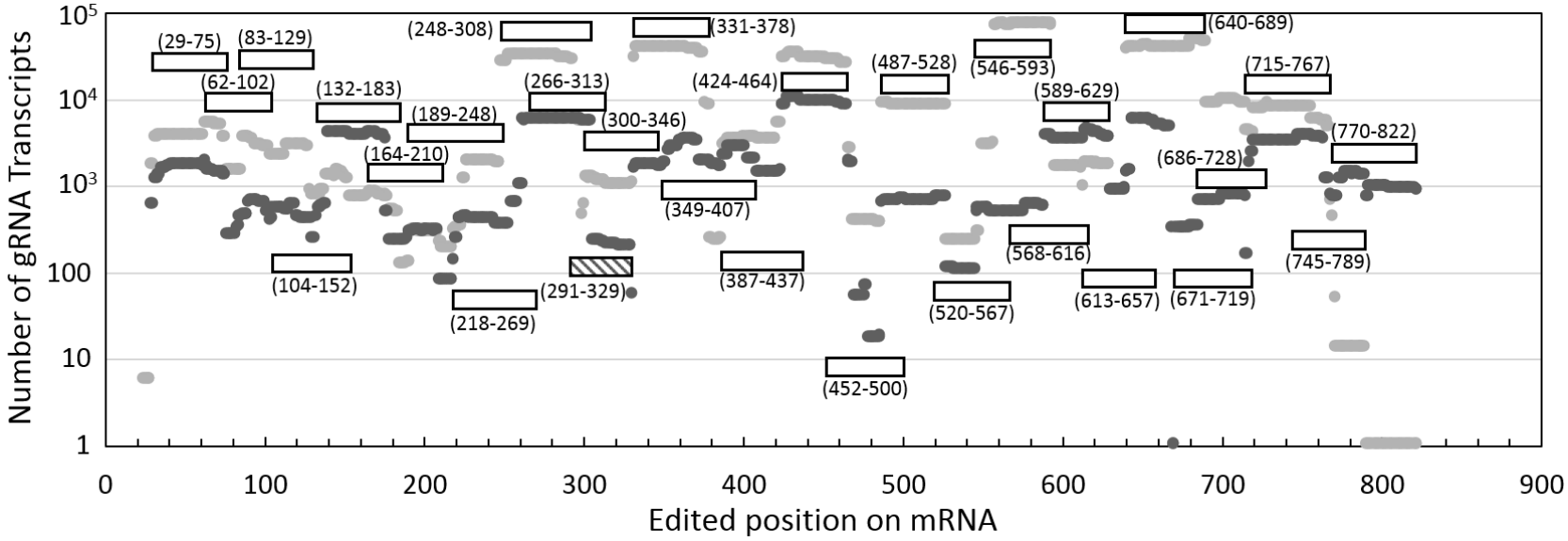
This work found that alternative editing is pervasive in *Trypanosoma brucei*, and brought to light the use of overlapped reading frames, providing another strong reason for the utility and maintenance of RNA editing. This work also showed that the RNA editing system is surprisingly robust and tolerant of mutational noise.

APPENDICES

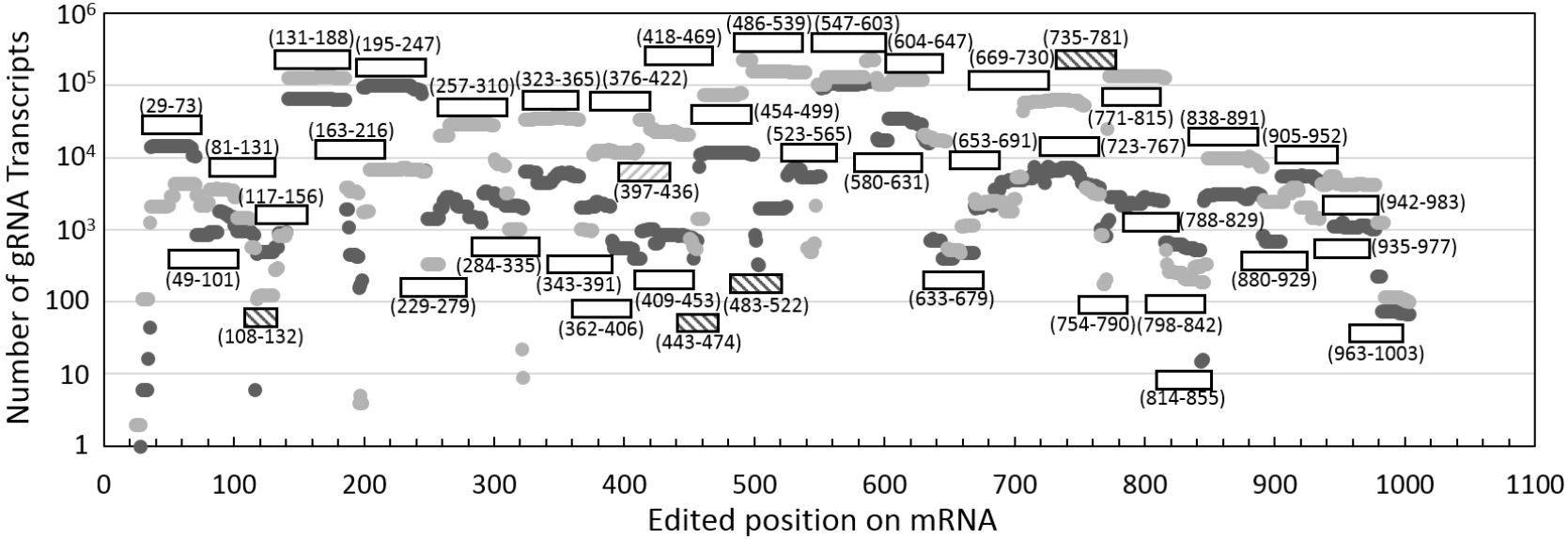
APPENDIX A. Quantification of the number of identified bloodstream and procyclic gRNA transcripts that cover a respective nucleotide in the fully edited mRNA.

Bloodstream gRNAs are shown in dark gray and procyclic gRNAs are shown in light gray. Nucleotides and deletion sites were both numbered as edited positions in the mRNA transcripts starting from the 5' end (+1 =0). Boxes indicate the positions of identified populations of gRNAs (coverage ranges shown in parenthesis). Boxes with dark gray or light gray diagonal stripes indicate populations identified only in the bloodstream or procyclic transcriptomes respectively. A. ATPase subunit 6; B. Cytochrome oxidase III; C. C-rich region 3; D. C-rich region 4; E. NADH dehydrogenase subunit 3; F. NADH dehydrogenase subunit 7; G. NADH dehydrogenase subunit 8; H. NADH dehydrogenase subunit 9; I. Ribosomal Protein S12. All individual data points were designated with solid circles. Close overlapping of individual data points generate the observed solid lines.

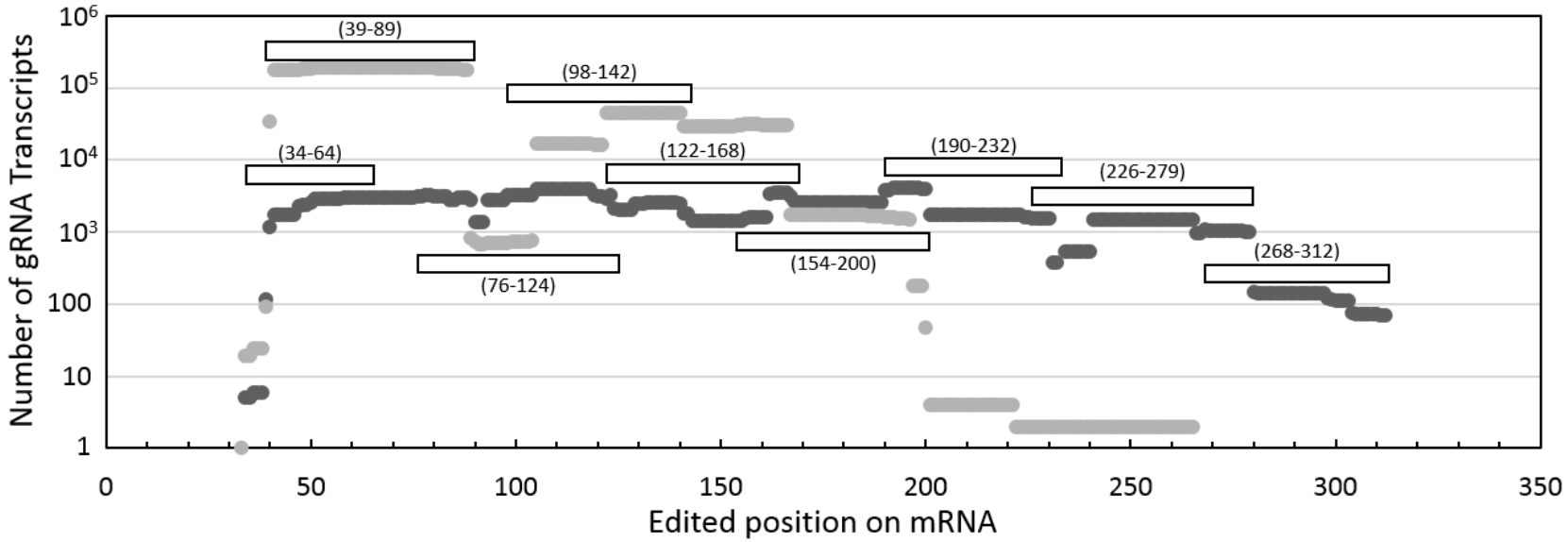
A. ATPase subunit 6



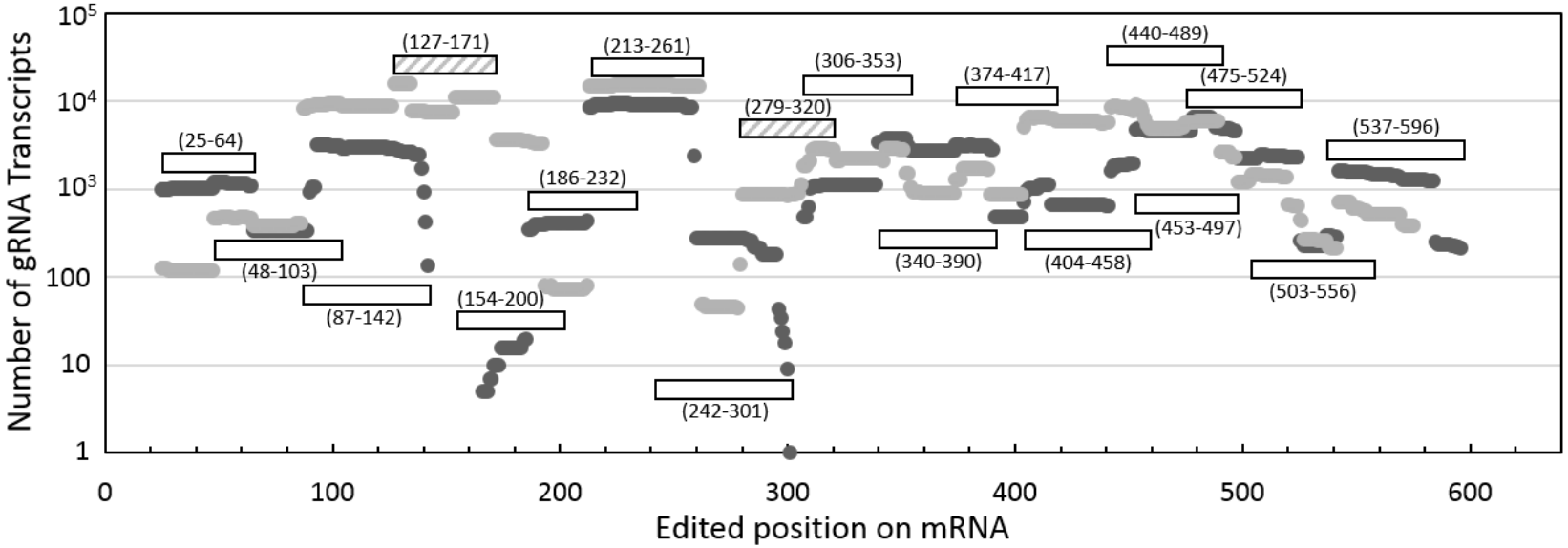
B. Cytochrome oxidase III



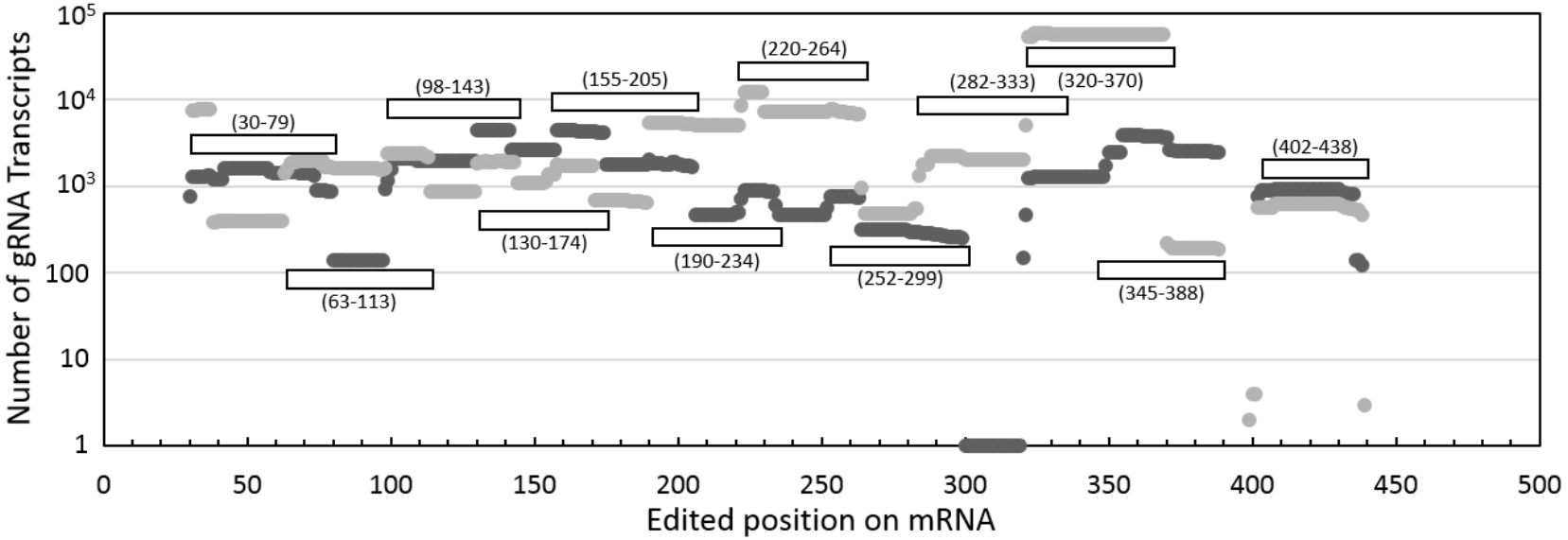
C. C-rich region 3



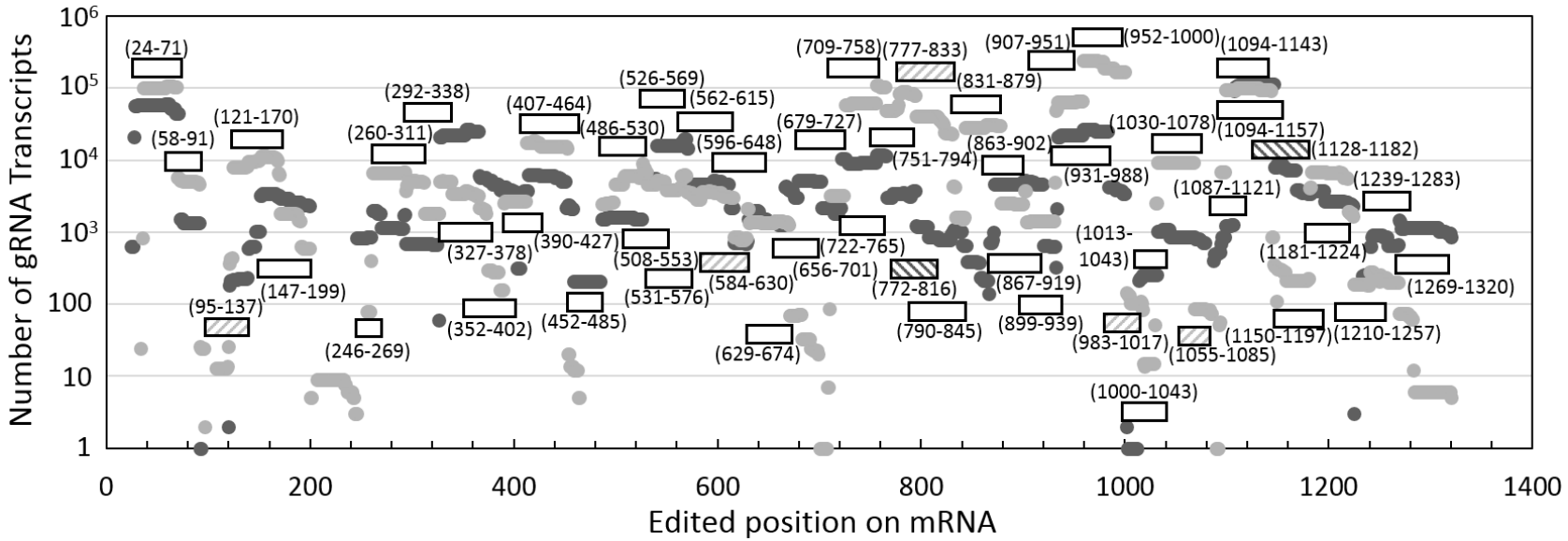
D. C-rich region 4



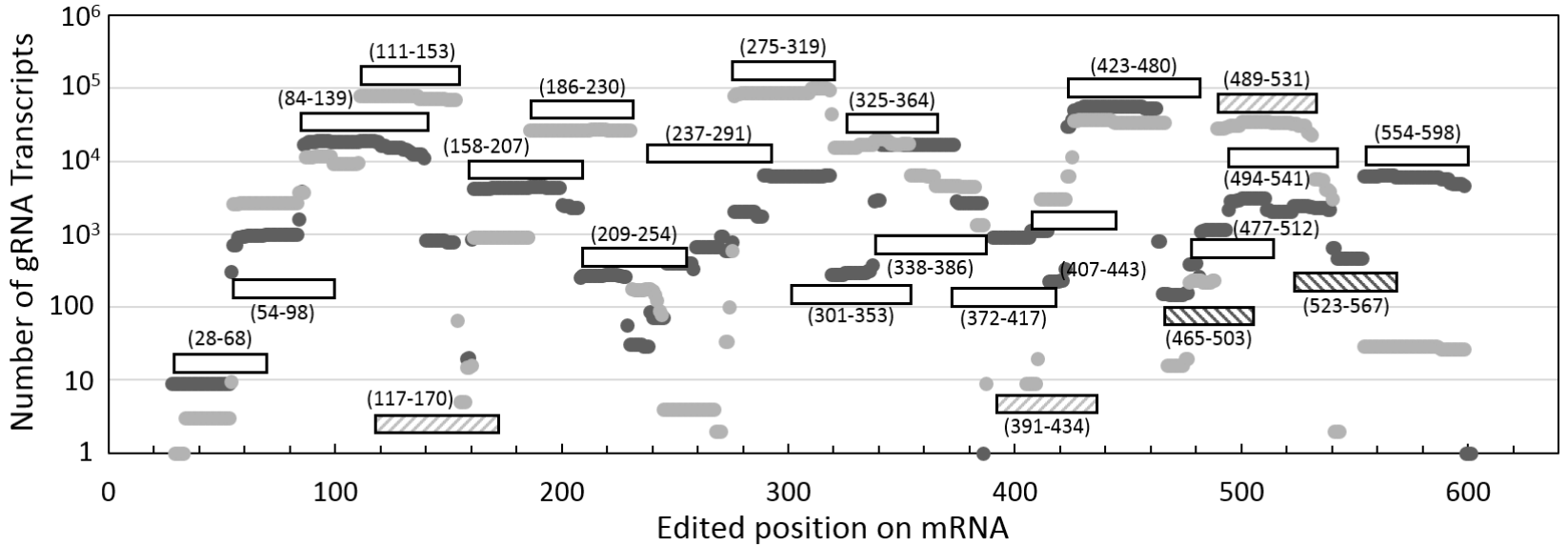
E. NADH dehydrogenase subunit 3



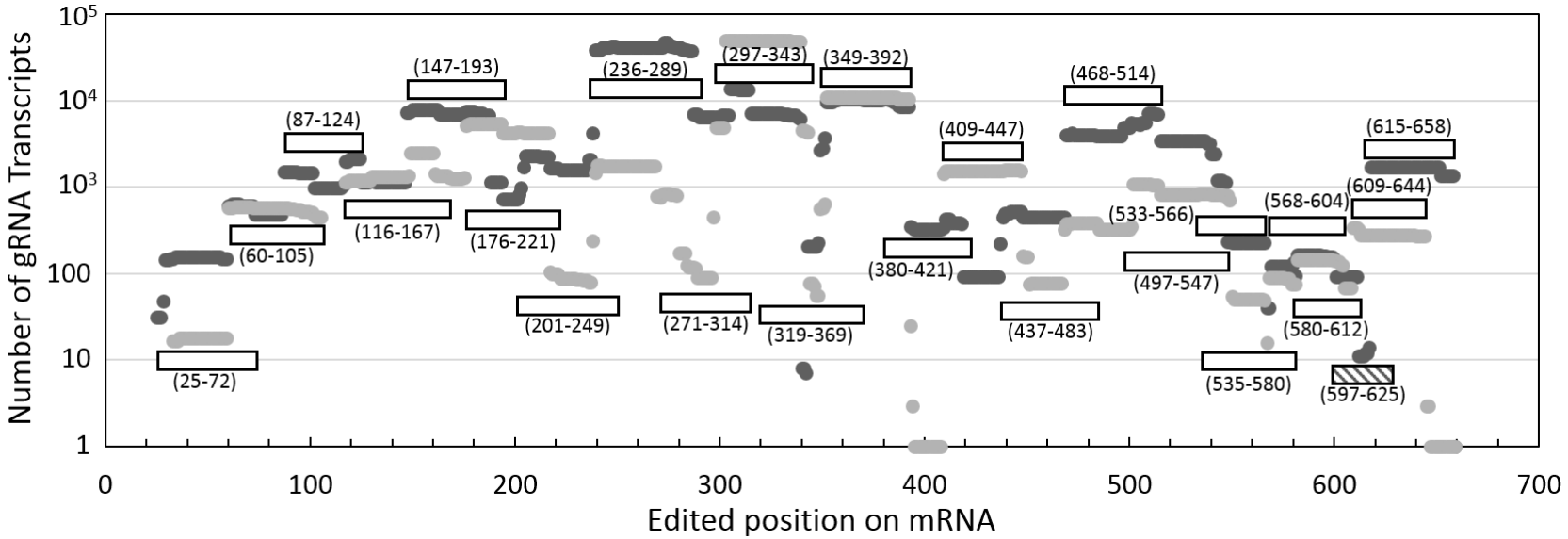
F. NADH dehydrogenase subunit 7



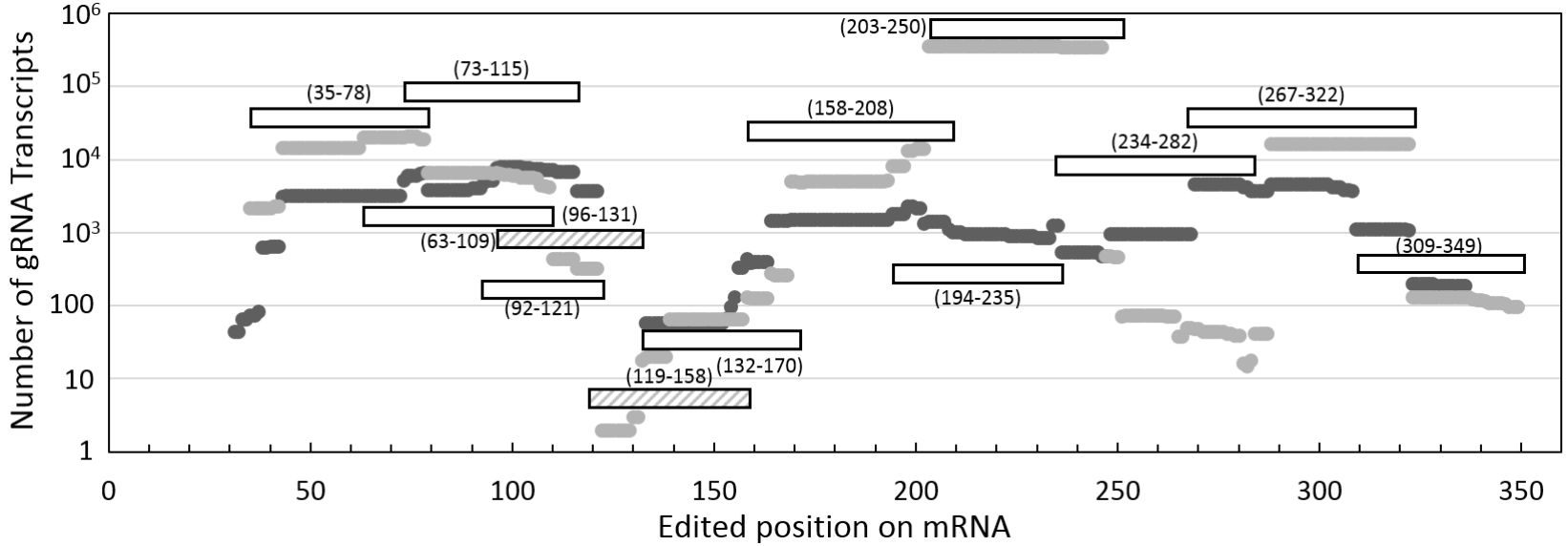
G. NADH dehydrogenase subunit 8



H. NADH dehydrogenase subunit 9



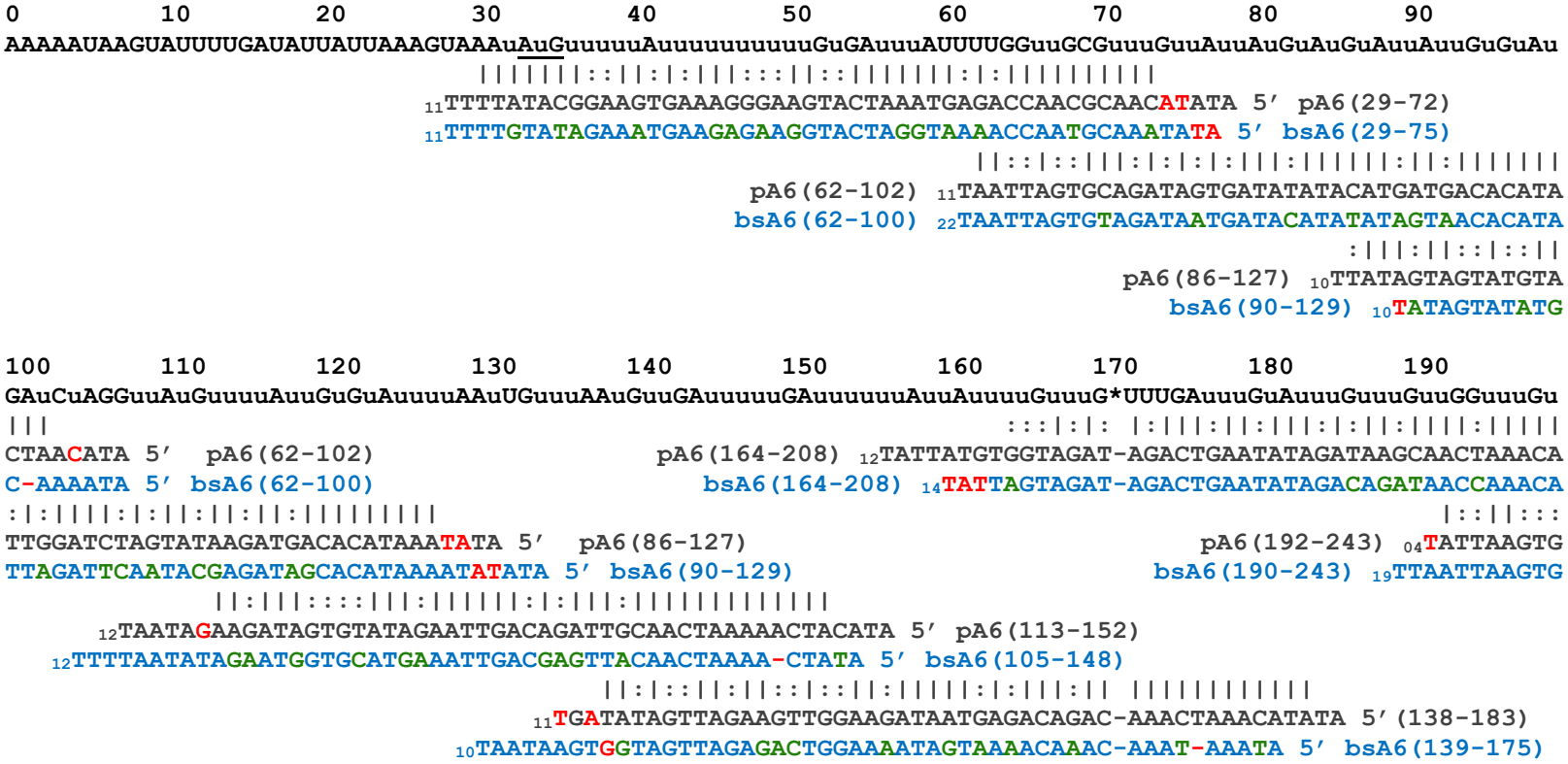
I. Ribosomal Protein S12

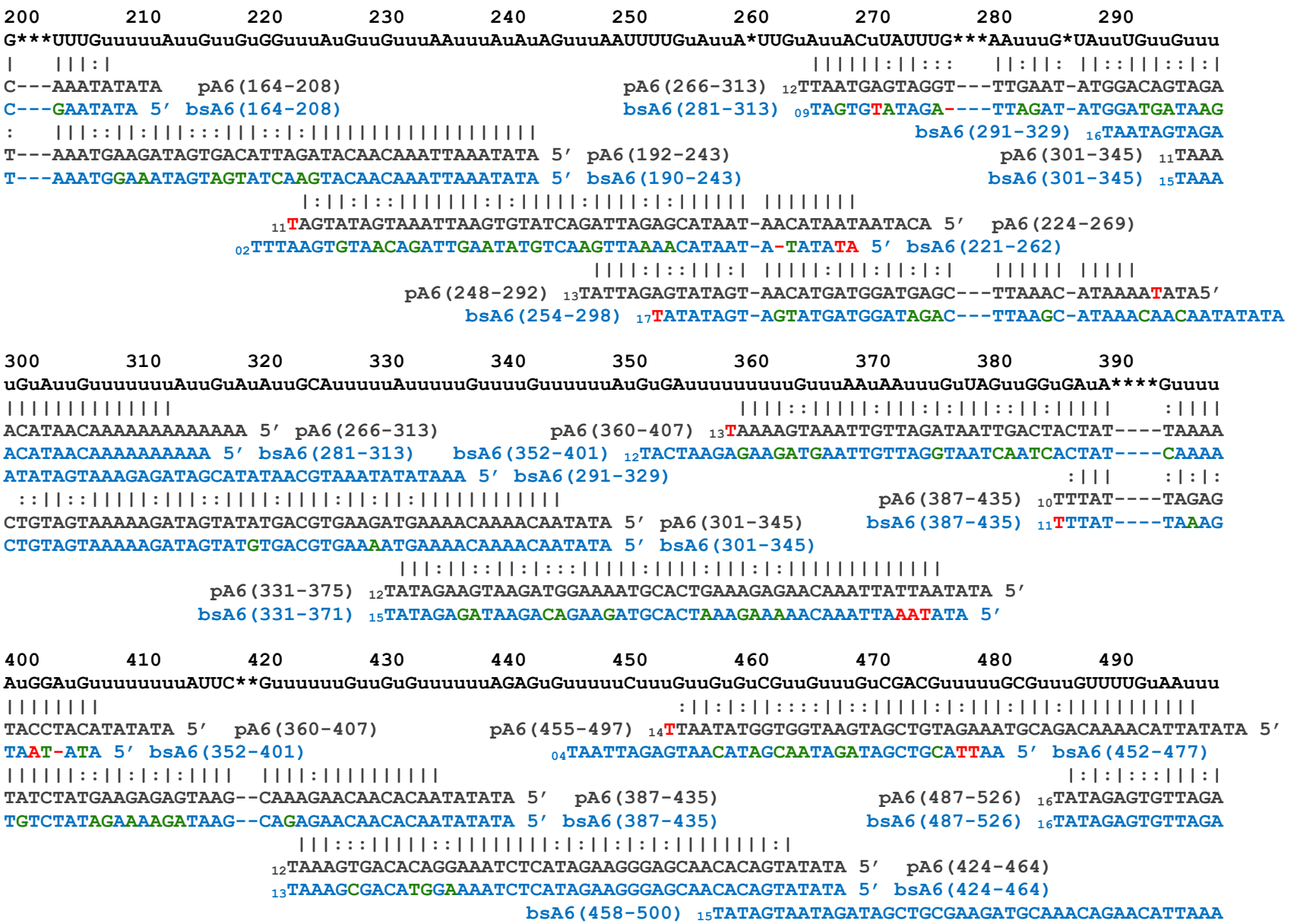


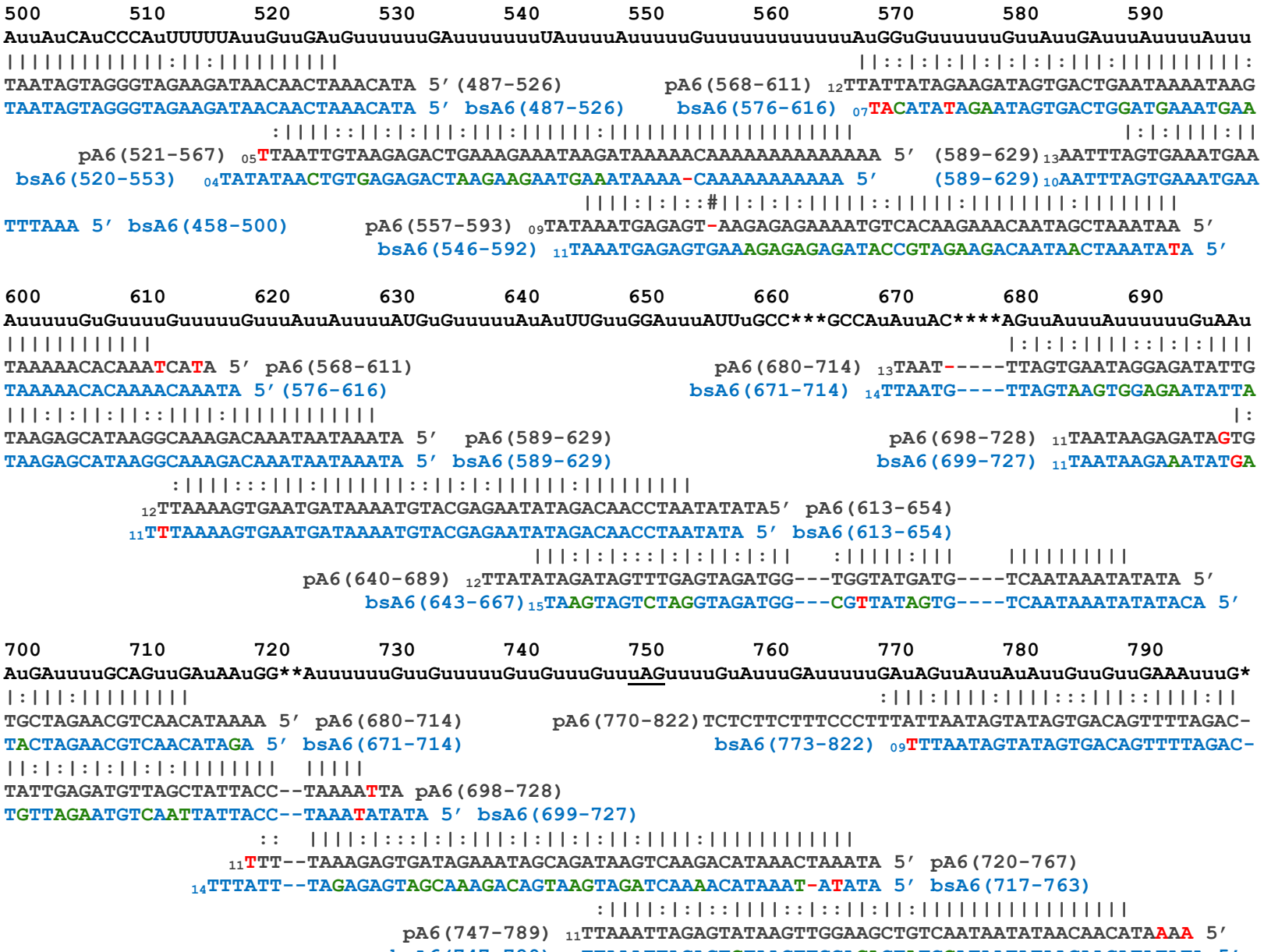
APPENDIX B. Alignment of the mitochondrial fully edited mRNAs and the most abundant gRNAs required for full coverage identified in the bloodstream (blue) and procyclic (gray) life cycle stages.

Conservative mutations between gRNAs are shown in green and mutations that disrupt alignment are shown in red. Lowercase u's indicate uridylates added by editing, asterisks indicate encoded uridylates deleted during editing. Nucleotides and deletion sites in the fully edited mRNA were numbered starting from the 5' end (+1=0). Watson-Crick (|) and G:U (:) base pairs are indicated. Mismatches are indicated by the number sign (#). A) ATPase 6; B) Cytochrome Oxidase III; C) C-Rich Region 3; D) C-Rich Region 4; E) Cytochrome b; F) Maxicircle Unidentified Reading Frame II (Murf II); G) NADH Dehydrogenase Subunit 3; H) NADH Dehydrogenase Subunit 7; I) NADH Dehydrogenase Subunit 8; J) NADH Dehydrogenase Subunit 9; K) Ribosomal Protein S12.

A) ATPase subunit 6




200 210 220 230 240 250 260 270 280 290
G***UUUGuuuuuuAuuGuuGuGGuuuAuGuuGuuuAAuuuAuAuAGuuuAAUUUUGuAuuA*UUGuAuuACuUAUUUG***AAuuuG*UAuuUGuuGuuu
C---AAATATATA pA6(164-208)
C---GAATATA 5′ bsA6(164-208)
T---AAATGAAGATAGTGACATTAGATACAACAAATTAAATATA 5′ pA6(192-243)
T---AAATGGAAATAGTAGTATCAAGTACAACAAATTAAATATA 5′ bsA6(190-243)
pA6(266-313) 12TTAATGAGTAGGT---TTGAAT-ATGGACAGTAGA
bsA6(281-313) 09TAGTGTATAGA----TTAGAT-ATGGATGATAAG
bsA6(291-329) 16TAATAGTAGA
pA6(301-345) 11TAAA
bsA6(301-345) 15TAAA
11TAGTATAGTAAATTAAGTGTATCAGATTAGAGCATAAT-AACATAATAATACA 5′ pA6(224-269)
02TTTAAGTGTAACAGATTGAATATGTCAAGTTAAAACATAAT-A-TATATA 5′ bsA6(221-262)
pA6(248-292) 13TATTAGAGTATAGT-AACATGATGGATGAGC---TTAAAC-ATAAAATATA5′
bsA6(254-298) 17TATATAGT-AGTATGATGGATAGAC---TTAAGC-ATAAACAACAATATATA
300 310 320 330 340 350 360 370 380 390
uGuAuuGuuuuuuuAuuGuAuAuuGCAuuuuuAuuuuuGuuuuGuuuuuuAuGuGAuuuuuuuuuGuuuAAuAAuuuGuUAGuuGGuGAuA****Guuuu
ACATAACAAAAAAAAAAAAA 5′ pA6(266-313)
ACATAACAAAAAAAAAA 5′ bsA6(281-313)
ATATAGTAAAGAGATAGCATATAACGTAAATATATAAA 5′ bsA6(291-329)
CTGTAGTAAAAAGATAGTATATGACGTGAAGATGAAAACAAAACAATATA 5′ pA6(301-345)
CTGTAGTAAAAAGATAGTATGTGACGTGAAAATGAAAACAAAACAATATA 5′ bsA6(301-345)
pA6(331-375) 12TATAGAAGTAAGATGGAAAATGCACTGAAAGAGAACAAATTATTAATATA 5′
bsA6(331-371) 15TATAGAGATAAGACAGAAGATGCACTAAAGAAAAACAAATTAAATATA 5′
pA6(360-407) 13TAAAAGTAAATTGTTAGATAATTGACTACTAT----TAAAA
bsA6(352-401) 12TACTAAGAGAAGATGAATTGTTAGGTAATCAATCACTAT----CAAAA
pA6(387-435) 10TTTAT----TAGAG
bsA6(387-435) 11TTTAT----TAAAG
400 410 420 430 440 450 460 470 480 490
AuGGAuGuuuuuuuuAUUC**GuuuuuuGuuGuGuuuuuuAGAGuGuuuuuCuuuGuuGuGuCGuuGuuuGuCGACGuuuuuGCGuuuGUUUUGuAAuuu
TACCTACATATATA 5′ pA6(360-407)
TAAT-ATA 5′ bsA6(352-401)
TATCTATGAAGAGAGTAAG--CAAAGAACAACACAATATATA 5′ pA6(387-435)
TGTCTATAGAAAAGATAAG--CAGAGAACAACACAATATATA 5′ bsA6(387-435)
pA6(455-497) 14TTAATATGGTGGTAAGTAGCTGTAGAAATGCAGACAAAACATTATATA 5′
04TAATTAGAGTAACATAGCAATAGATAGCTGCATTAA 5′ bsA6(452-477)
pA6(487-526) 16TATAGAGTGTTAGA
bsA6(487-526) 16TATAGAGTGTTAGA
12TAAAGTGACACAGGAAATCTCATAGAAGGGAGCAACACAGTATATA 5′ pA6(424-464)
13TAAAGCGACATGGAAAATCTCATAGAAGGGAGCAACACAGTATATA 5′ bsA6(424-464)
bsA6(458-500) 15TATAGTAATAGATAGCTGCGAAGATGCAAACAGAACATTAAA

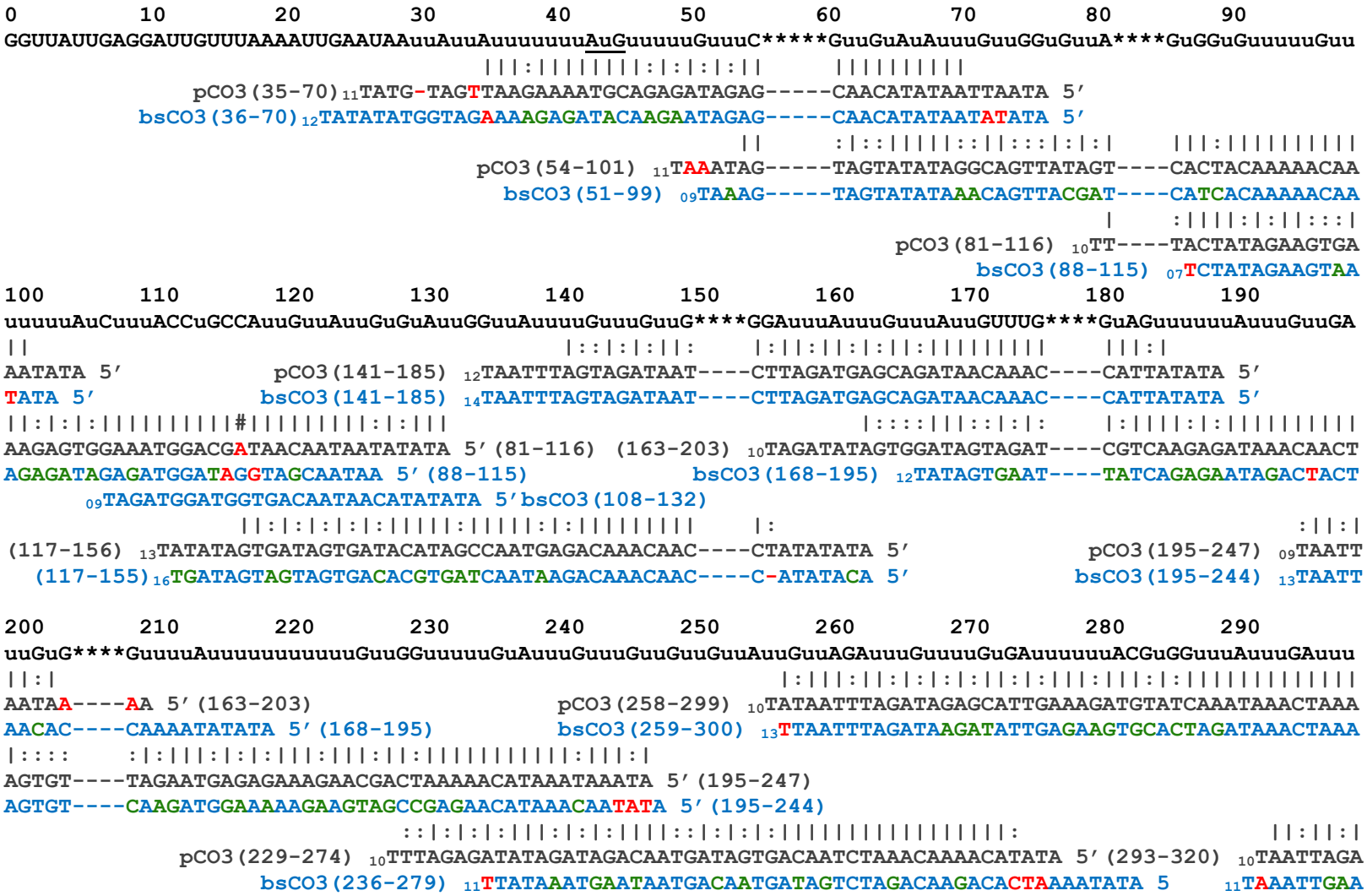

500 510 520 530 540 550 560 570 580 590
AuuAuCAuCCCAuUUUUUAuuGuuGAuGuuuuuuGAuuuuuuuuUAuuuuAuuuuuGuuuuuuuuuuuuuAuGGuGuuuuuuGuuAuuGAuuuAuuuuAuuu
TAATAGTAGGGTAGAAGATAACAACTAAACATA 5′(487-526)
pA6(568-611) 12TTATTATAGAAGATAGTGACTGAATAAAATAAG
TAATAGTAGGGTAGAAGATAACAACTAAACATA 5′ bsA6(487-526)
bsA6(576-616) 07TACATATAGAATAGTGACTGGATGAAATGAA
pA6(521-567) 05TTAATTGTAAGAGACTGAAAGAAATAAGATAAAAACAAAAAAAAAAAAAA 5′ (589-629)13AATTTAGTGAAATGAA
bsA6(520-553) 04TATATAACTGTGAGAGACTAAGAAGAATGAAATAAAA-CAAAAAAAAAAA 5′
(589-629)10AATTTAGTGAAATGAA
TTTAAA 5′ bsA6(458-500)
pA6(557-593) 09TATAAATGAGAGT-AAGAGAGAAAATGTCACAAGAAACAATAGCTAAATAA 5′
bsA6(546-592) 11TAAATGAGAGTGAAAGAGAGAGATACCGTAGAAGACAATAACTAAATATA 5′
600 610 620 630 640 650 660 670 680 690
AuuuuuGuGuuuuGuuuuuGuuuAuuAuuuuAUGuGuuuuuAuAuUUGuuGGAuuuAUUuGCC***GCCAuAuuAC****AGuuAuuuAuuuuuuGuAAu
TAAAAACACAAATCATA 5′ pA6(568-611)
pA6(680-714) 13TAAT-----TTAGTGAATAGGAGATATTG
TAAAAACACAAAACAAATA 5′(576-616)
bsA6(671-714) 14TTAATG----TTAGTAAGTGGAGAATATTA
TAAGAGCATAAGGCAAAGACAAATAATAAATA 5′ pA6(589-629)
pA6(698-728) 11TAATAAGAGATAGTG
TAAGAGCATAAGGCAAAGACAAATAATAAATA 5′ bsA6(589-629)
bsA6(699-727) 11TAATAAGAAATATGA
12TTAAAAGTGAATGATAAAATGTACGAGAATATAGACAACCTAATATATA5′ pA6(613-654)
11TTTAAAAGTGAATGATAAAATGTACGAGAATATAGACAACCTAATATA 5′ bsA6(613-654)
pA6(640-689) 12TTATATAGATAGTTTGAGTAGATGG---TGGTATGATG----TCAATAAATATATA 5′
bsA6(643-667)15TAAGTAGTCTAGGTAGATGG---CGTTATAGTG----TCAATAAATATATACA 5′
700 710 720 730 740 750 760 770 780 790
AuGAuuuuGCAGuuGAuAAuGG**AuuuuuuGuuGuuuuuGuuGuuuGuuuAGuuuuuGuAuuuGAuuuuuuGAuAGuuAuuAuAuuGuuGuuGAAAuuuG*
TGCTAGAACGTCAACATAAAA 5′ pA6(680-714)
pA6(770-822)TCTCTTCTTTCCCTTTATTAATAGTATAGTGACAGTTTTAGAC-
TACTAGAACGTCAACATAGA 5′ bsA6(671-714)
bsA6(773-822) 09TTTAATAGTATAGTGACAGTTTTAGAC-
TATTGAGATGTTAGCTATTACC--TAAAATTA pA6(698-728)
TGTTAGAATGTCAATTATTACC--TAAATATATA 5′ bsA6(699-727)
11TTT--TAAAGAGTGATAGAAATAGCAGATAAGTCAAGACATAAACTAAATA 5′ pA6(720-767)
14TTTATT--TAGAGAGTAGCAAAGACAGTAAGTAGATCAAAACATAAAT-ATATA 5′ bsA6(717-763)
pA6(747-789) 11TTAAATTAGAGTATAAGTTGGAAGCTGTCAATAATATAACAACATAAAA 5′
bsA6(747-789) 11TTAAATTAGAGTGTAAGTTGGAGACTATCGATAATATAACAACATATATA 5′

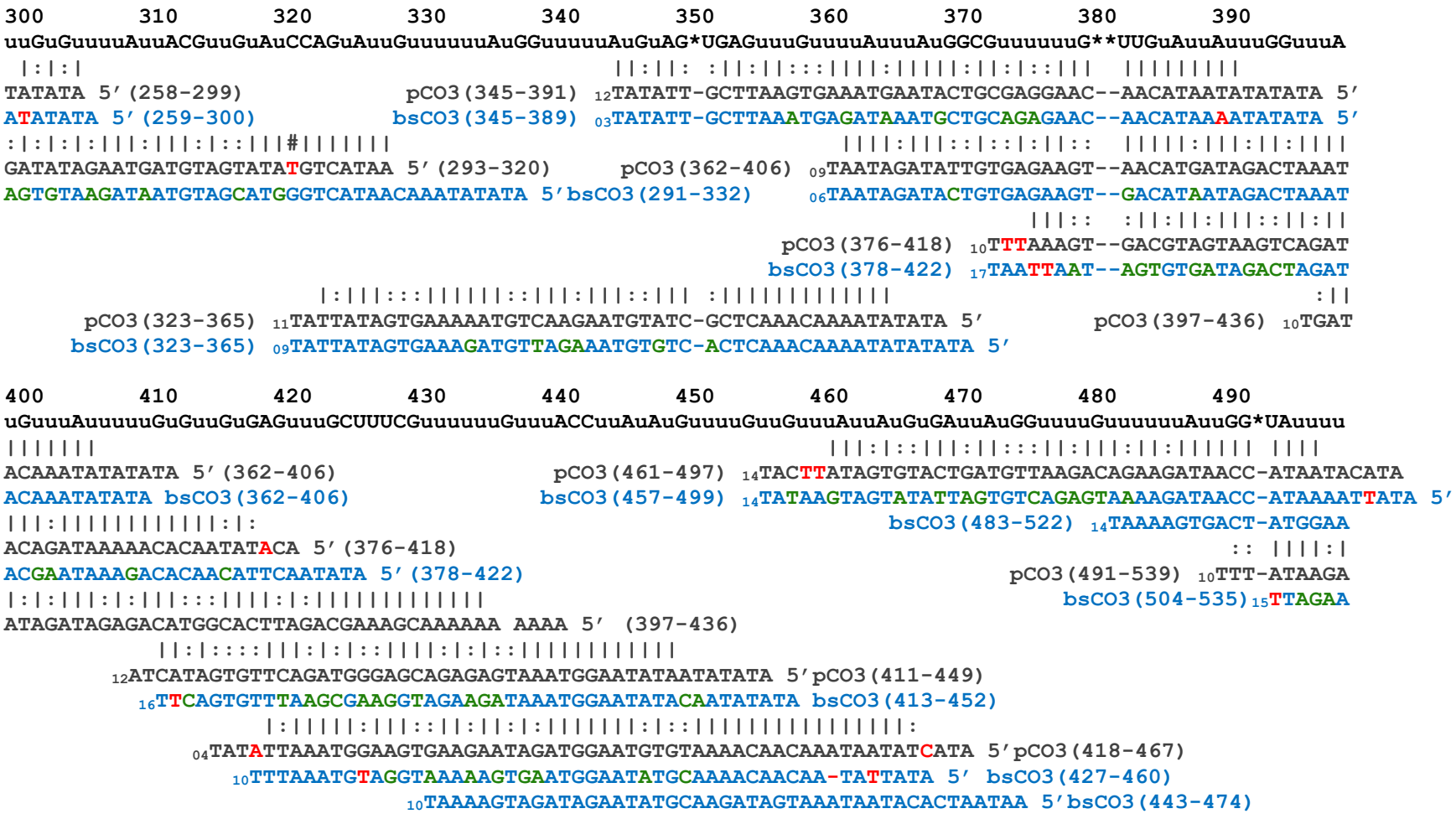
```
800       810       820       830       840
*GuuUGuuA**UUGGAGUUAUAGAAUAAGAUCAAAUAAGUUAAUAAUA_
 :||:||||  |:||||||||||
-TAAGCAAT--AGCCTCAATATCAGG 5′
-TAAGCAAT--AGCCTCAATATCATATA 5′
```

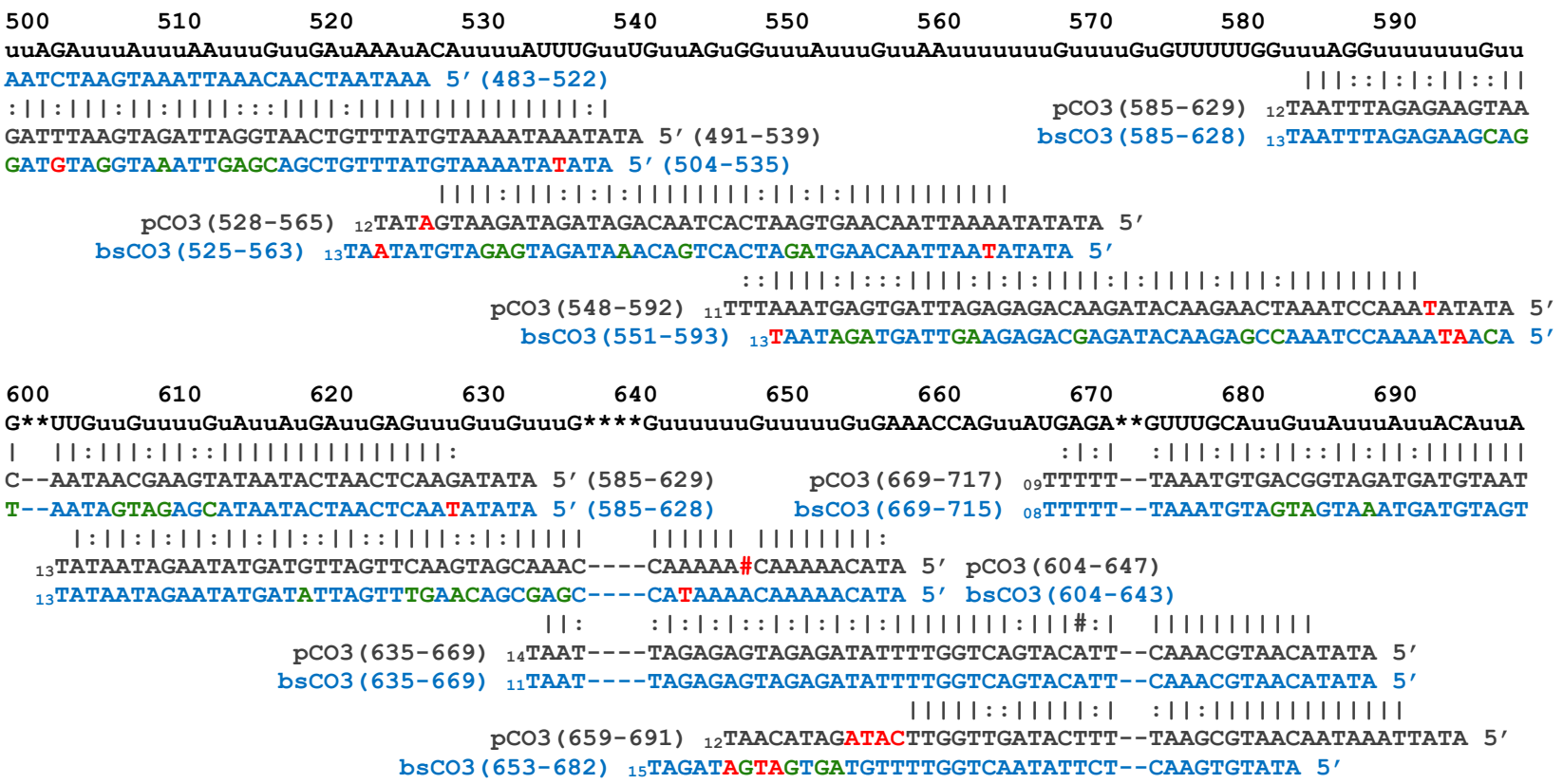
Alternate initiating gRNA (procyclic transcriptome only)

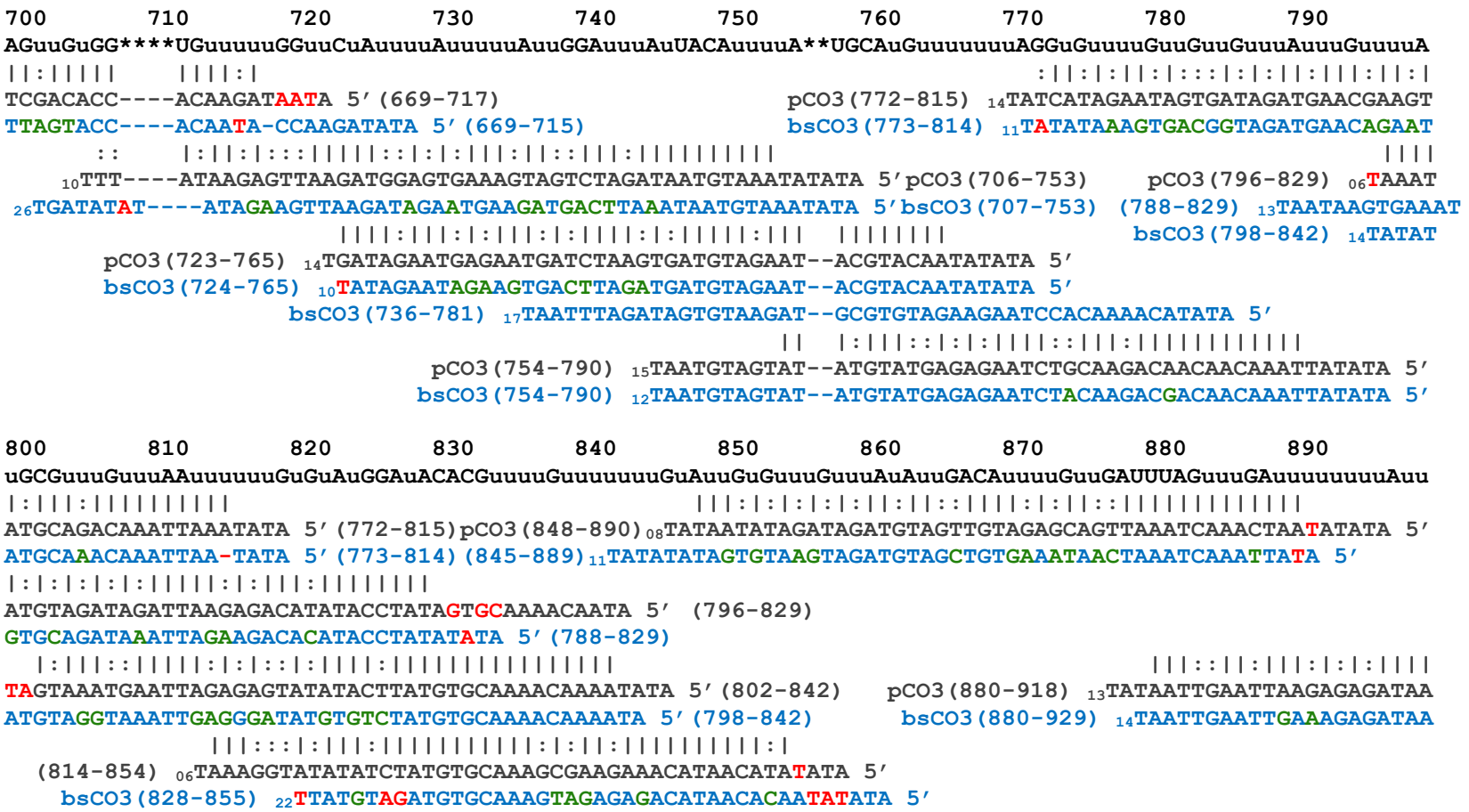
```
750       760       770       780       790       800       810       820       830
uAGuuuuGuAuuuGAuuuuuGAuAGuuAuuAuAuuGuuG*uGAAA*uuG**GuuuUGuuA**UUGGAGUUAUAGAAUAAGAUCAAAU
                        :||||:|||:::::|| |:||| |||  :|||:||:|  ||||||||||||
         pA6(774-822) *14TAATAGTATGGTGAC-ATTTT-GAC--TAAAGCAGT--AACCTCAATATCATA 5′
```

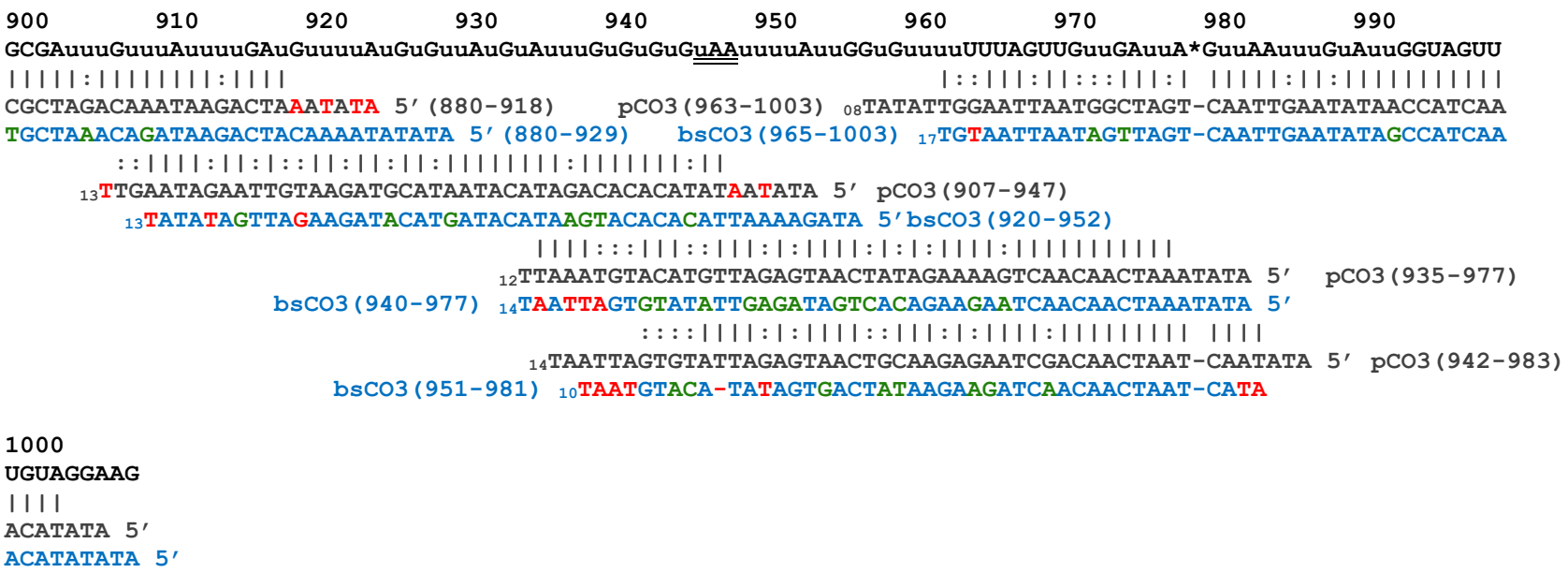
B) Cytochrome Oxidase III



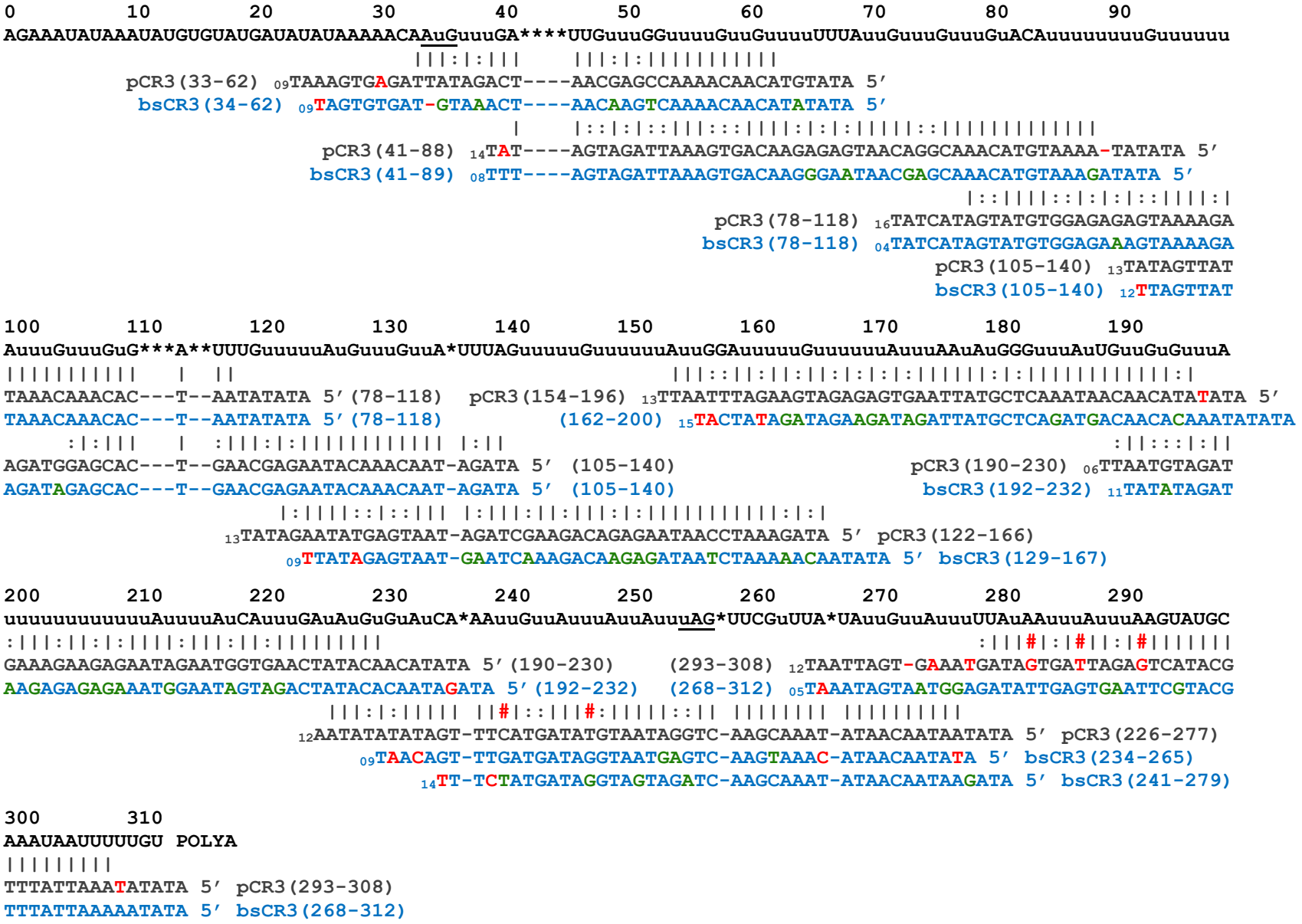

300 310 320 330 340 350 360 370 380 390
uuGuGuuuuAuuACGuuGuAuCCAGuAuuGuuuuuuAuGGuuuuuAuGuAG*UGAGuuuGuuuuAuuuAuGGCGuuuuuuG**UUGuAuuAuuuGGuuuA
TATATA 5′(258-299)
ATATATA 5′(259-300)
GATATAGAATGATGTAGTATATGTCATAA 5′(293-320)
AGTGTAAGATAATGTAGCATGGGTCATAACAAATATATA 5′bsCO3(291-332)
pCO3(345-391) 12TATATT-GCTTAAGTGAAATGAATACTGCGAGGAAC--AACATAATATATATA 5′
bsCO3(345-389) 03TATATT-GCTTAAATGAGATAAATGCTGCAGAGAAC--AACATAAAATATATA 5′
pCO3(362-406) 09TAATAGATATTGTGAGAAGT--AACATGATAGACTAAAT
06TAATAGATACTGTGAGAAGT--GACATAATAGACTAAAT
pCO3(376-418) 10TTTAAAGT--GACGTAGTAAGTCAGAT
bsCO3(378-422) 17TAATTAAT--AGTGTGATAGACTAGAT
pCO3(323-365) 11TATTATAGTGAAAAATGTCAAGAATGTATC-GCTCAAACAAAATATATA 5′
bsCO3(323-365) 09TATTATAGTGAAAGATGTTAGAAATGTGTC-ACTCAAACAAAATATATATA 5′
pCO3(397-436) 10TGAT
400 410 420 430 440 450 460 470 480 490
uGuuuAuuuuuGuGuuGuGAGuuuGCUUUCGuuuuuuGuuuACCuuAuAuGuuuuGuuGuuuAuuAuGuGAuuAuGGuuuuGuuuuuuuAuuGG*UAuuuu
ACAAATATATATA 5′(362-406)
ACAAATATATA bsCO3(362-406)
ACAGATAAAAACACAATATACA 5′(376-418)
ACGAATAAAGACACAACATTCAATATA 5′(378-422)
ATAGATAGAGACATGGCACTTAGACGAAAGCAAAAAA AAAA 5′ (397-436)
pCO3(461-497) 14TACTTATAGTGTACTGATGTTAAGACAGAAGATAACC-ATAATACATA
bsCO3(457-499) 14TATAAGTAGTATATTAGTGTCAGAGTAAAAGATAACC-ATAAAATTATA 5′
bsCO3(483-522) 14TAAAAGTGACT-ATGGAA
pCO3(491-539) 10TTT-ATAAGA
bsCO3(504-535)15TTAGAA
12ATCATAGTGTTCAGATGGGAGCAGAGAGTAAATGGAATATAATATATA 5′pCO3(411-449)
16TTCAGTGTTTAAGCGAAGGTAGAAGATAAATGGAATATACAATATATA bsCO3(413-452)
04TATATTAAATGGAAGTGAAGAATAGATGGAATGTGTAAAACAACAAATAATATCATA 5′pCO3(418-467)
10TTTAAATGTAGGTAAAAAGTGAATGGAATATGCAAAACAACAA-TATTATA 5′ bsCO3(427-460)
10TAAAAGTAGATAGAATATGCAAGATAGTAAATAATACACTAATAA 5′bsCO3(443-474)


500 510 520 530 540 550 560 570 580 590
uuAGAuuuAuuuAAuuuGuuGAuAAAuACAuuuuAUUUGuuUGuuAGuGGuuuAuuuGuuAAuuuuuuuGuuuuGuGUUUUUGGuuuAGGuuuuuuuGuu
AATCTAAGTAAATTAAACAACTAATAAA 5′(483-522)
pCO3(585-629) 12TAATTTAGAGAAGTAA
GATTTAAGTAGATTAGGTAACTGTTTATGTAAAATAAATATA 5′(491-539)
bsCO3(585-628) 13TAATTTAGAGAAGCAG
GATGTAGGTAAATTGAGCAGCTGTTTATGTAAAATATATA 5′(504-535)
pCO3(528-565) 12TATAGTAAGATAGATAGACAATCACTAAGTGAACAATTAAAATATATA 5′
bsCO3(525-563) 13TAATATGTAGAGTAGATAAACAGTCACTAGATGAACAATTAATATATA 5′
pCO3(548-592) 11TTTAAATGAGTGATTAGAGAGACAAGATACAAGAACTAAATCCAAATATATA 5′
bsCO3(551-593) 13TAATAGATGATTGAAGAGACGAGATACAAGAGCCAAATCCAAAATAACA 5′
600 610 620 630 640 650 660 670 680 690
G**UUGuuGuuuuGuAuuAuGAuuGAGuuuGuuGuuuG****GuuuuuuGuuuuuuGuGAAACCAGuuAUGAGA**GUUUGCAuuGuuAuuuAuuACAuuA
C--AATAACGAAGTATAATACTAACTCAAGATATA 5′(585-629)
pCO3(669-717) 09TTTTT--TAAATGTGACGGTAGATGATGTAAT
T--AATAGTAGAGCATAATACTAACTCAATATATA 5′(585-628)
bsCO3(669-715) 08TTTTT--TAAATGTAGTAGTAAATGATGTAGT
13TATAATAGAATATGATGTTAGTTCAAGTAGCAAAC----CAAAAA#CAAAAACATA 5′ pCO3(604-647)
13TATAATAGAATATGATATTAGTTTGAACAGCGAGC----CATAAAACAAAAACATA 5′ bsCO3(604-643)
pCO3(635-669) 14TAAT----TAGAGAGTAGAGATATTTTGGTCAGTACATT--CAAACGTAACATATA 5′
bsCO3(635-669) 11TAAT----TAGAGAGTAGAGATATTTTGGTCAGTACATT--CAAACGTAACATATA 5′
pCO3(659-691) 12TAACATAGATACTTGGTTGATACTTT--TAAGCGTAACAATAAATTATA 5′
bsCO3(653-682) 15TAGATAGTAGTGATGTTTTGGTCAATATTCT--CAAGTGTATA 5′

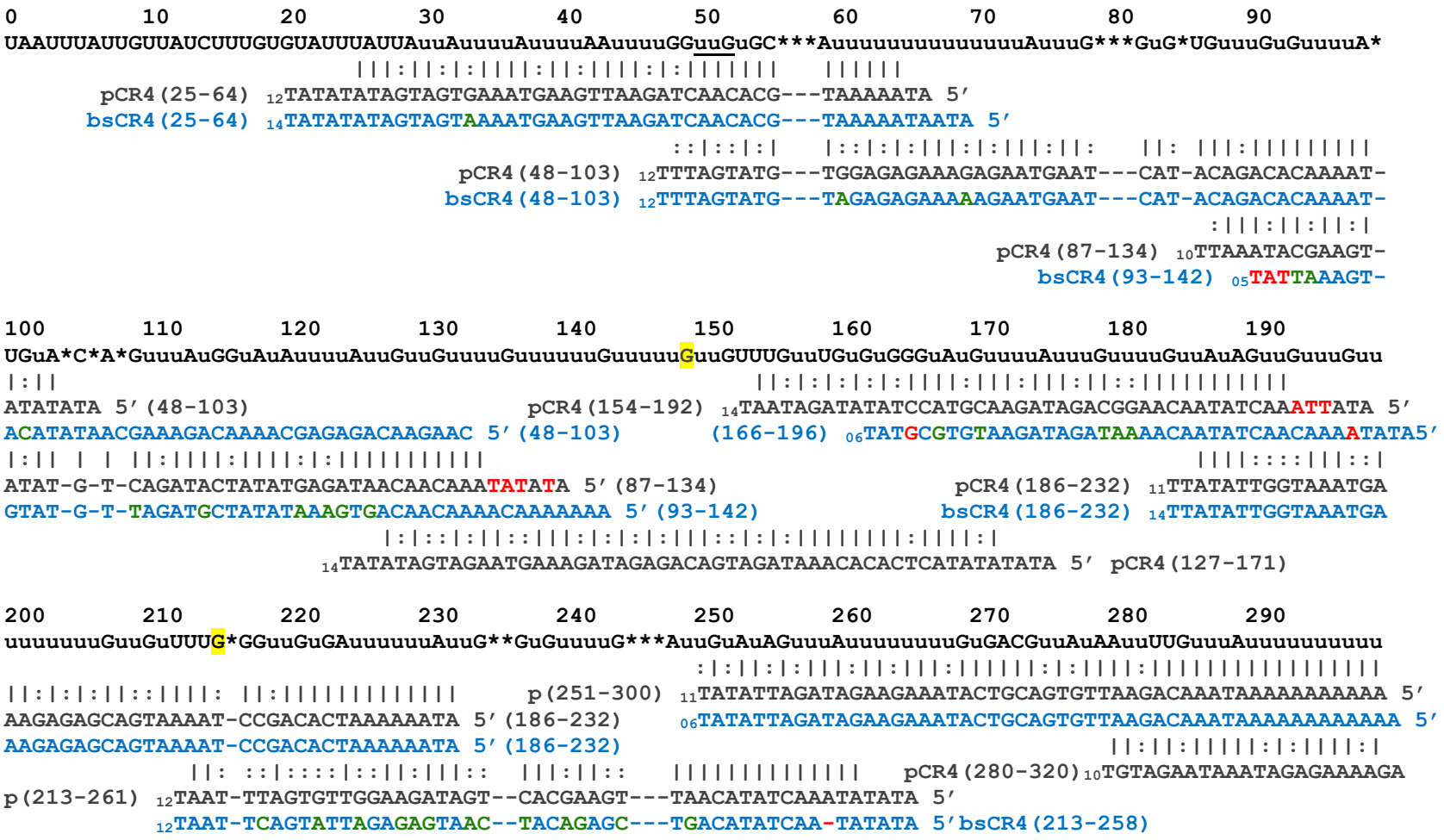

700 710 720 730 740 750 760 770 780 790
AGuuGuGG****UGuuuuuGGuuCuAuuuuAuuuuuAuuGGAuuuAuUACAuuuuA**UGCAuGuuuuuuuuAGGuGuuuuGuuGuuGuuuAuuuGuuuuA
TCGACACC----ACAAGATAATA 5′(669-717)
TTAGTACC----ACAATA-CCAAGATATA 5′(669-715)
pCO3(772-815) 14TATCATAGAATAGTGATAGATGAACGAAGT
bsCO3(773-814) 11TATATAAAGTGACGGTAGATGAACAGAAT
10TTT----ATAAGAGTTAAGATGGAGTGAAAGTAGTCTAGATAATGTAAATATATA 5′pCO3(706-753)
pCO3(796-829) 06TAAAT
26TGATATAT----ATAGAAGTTAAGATAGAATGAAGATGACTTAAATAATGTAAATATA 5′bsCO3(707-753)
(788-829) 13TAATAAGTGAAAT
bsCO3(798-842) 14TATAT
pCO3(723-765) 14TGATAGAATGAGAATGATCTAAGTGATGTAGAAT--ACGTACAATATATA 5′
bsCO3(724-765) 10TATAGAATAGAAGTGACTTAGATGATGTAGAAT--ACGTACAATATATA 5′
bsCO3(736-781) 17TAATTTAGATAGTGTAAGAT--GCGTGTAGAAGAATCCACAAAACATATA 5′
pCO3(754-790) 15TAATGTAGTAT--ATGTATGAGAGAATCTGCAAGACAACAACAAATTATATA 5′
bsCO3(754-790) 12TAATGTAGTAT--ATGTATGAGAGAATCTACAAGACGACAACAAATTATATA 5′
800 810 820 830 840 850 860 870 880 890
uGCGuuuGuuuAAuuuuuuuGuGuAuGGAuACACGuuuuGuuuuuuuuGuAuuGuGuuuGuuuAuAuuGACAuuuuGuuGAUUUAGuuuGAuuuuuuuuuAuu
ATGCAGACAAATTAAATATA 5′(772-815)pCO3(848-890)08TATAATATAGATAGATGTAGTTGTAGAGCAGTTAAATCAAACTAATATATA 5′
ATGCAAACAAATTAA-TATA 5′(773-814)(845-889)11TATATATAGTGTAAGTAGATGTAGCTGTGAAATAACTAAATCAAATTATA 5′
ATGTAGATAGATTAAGAGACATATACCTATAGTGCAAAACAATA 5′ (796-829)
GTGCAGATAAATTAGAAGACACATACCTATATATA 5′(788-829)
TAGTAAATGAATTAGAGAGTATATACTTATGTGCAAAACAAAATATA 5′(802-842)
pCO3(880-918) 13TATAATTGAATTAAGAGAGATAA
ATGTAGGTAAATTGAGGGATATGTGTCTATGTGCAAAACAAAATA 5′(798-842)
bsCO3(880-929) 14TAATTGAATTGAAAGAGATAA
(814-854) 06TAAAGGTATATATCTATGTGCAAAGCGAAGAAACATAACATATATA 5′
bsCO3(828-855) 22TTATGTAGATGTGCAAAGTAGAGAGACATAACACAATATATA 5′

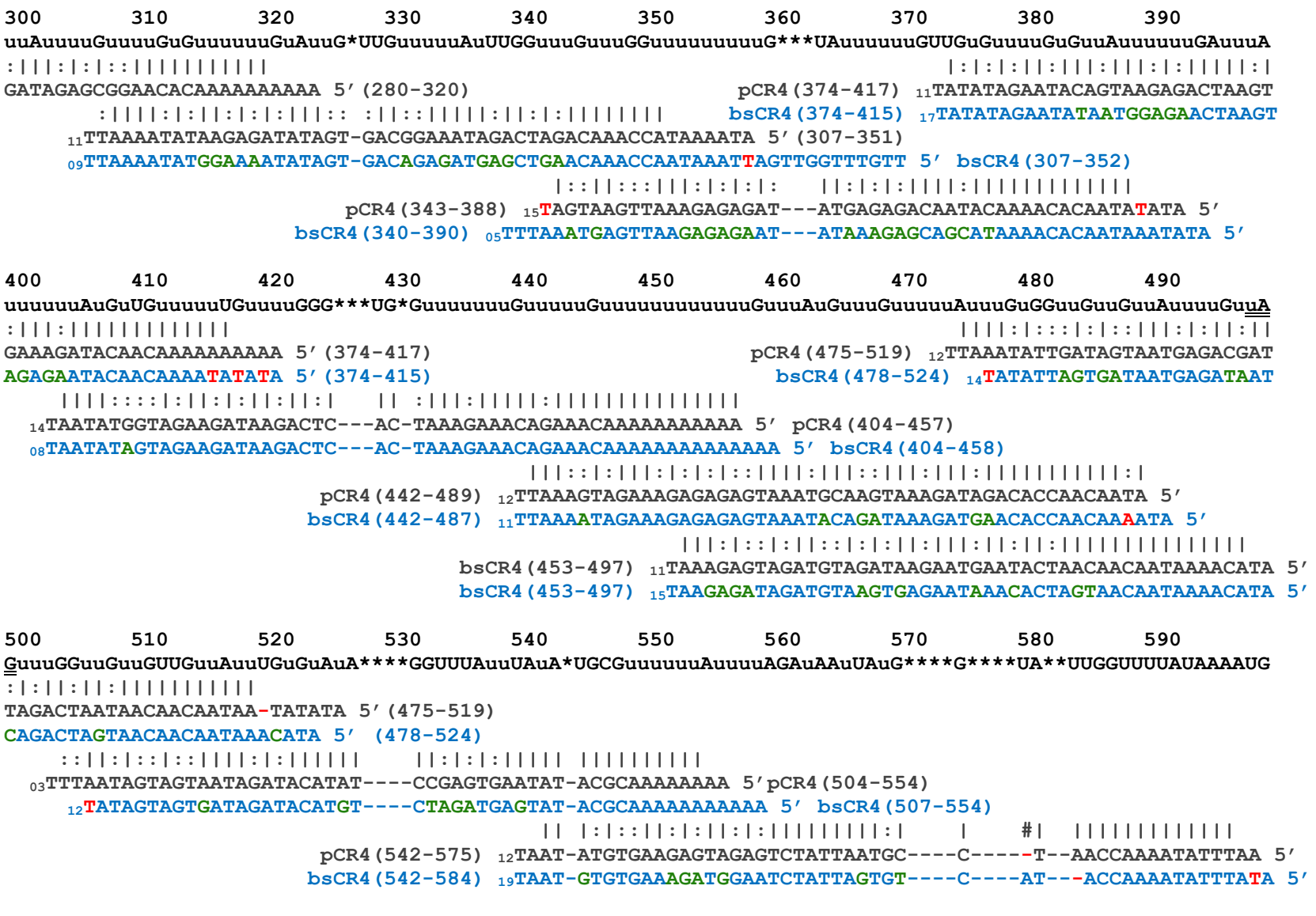

900 910 920 930 940 950 960 970 980 990
GCGAuuuGuuuAuuuuGAuGuuuuAuGuGuuAuGuAuuuGuGuGuuAAuuuuAuuGGuGuuuuUUUAGUUGuuGAuuA*GuuAAuuuGuAuuGGUAGUU
CGCTAGACAAATAAGACTAAATATA 5′(880-918) pCO3(963-1003) 08TATATTGGAATTAATGGCTAGT-CAATTGAATATAACCATCAA
TGCTAAACAGATAAGACTACAAAATATATA 5′(880-929) bsCO3(965-1003) 17TGTAATTAATAGTTAGT-CAATTGAATATAGCCATCAA
13TTGAATAGAATTGTAAGATGCATAATACATAGACACACATATAATATA 5′ pCO3(907-947)
13TATATAGTTAGAAGATACATGATACATAAGTACACACATTAAAAGATA 5′bsCO3(920-952)
12TTAAATGTACATGTTAGAGTAACTATAGAAAAGTCAACAACTAAATATA 5′ pCO3(935-977)
bsCO3(940-977) 14TAATTAGTGTATATTGAGATAGTCACAGAAGAATCAACAACTAAATATA 5′
14TAATTAGTGTATTAGAGTAACTGCAAGAGAATCGACAACTAAT-CAATATA 5′ pCO3(942-983)
bsCO3(951-981) 10TAATGTACA-TATAGTGACTATAAGAAGATCAACAACTAAT-CATA
1000
UGUAGGAAG
ACATATA 5′
ACATATATA 5′

C) C-rich region 3

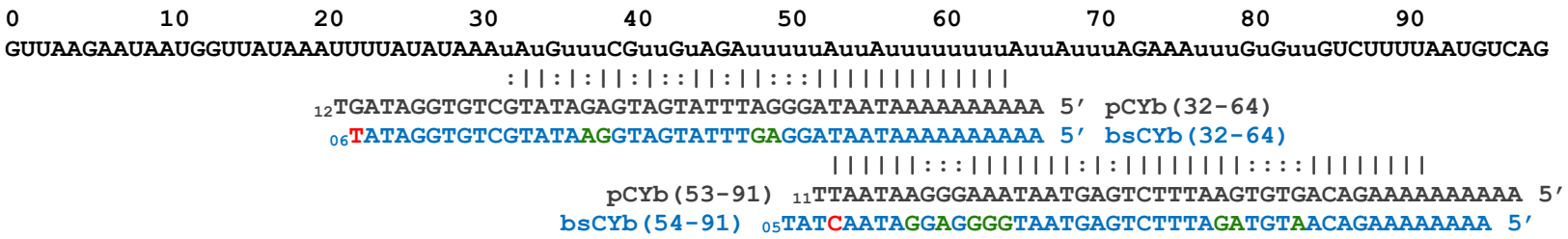


D) C-rich region 4. Resequencing of the mRNA indicated that there were 2 errors in the original sequence (yellow highlights).

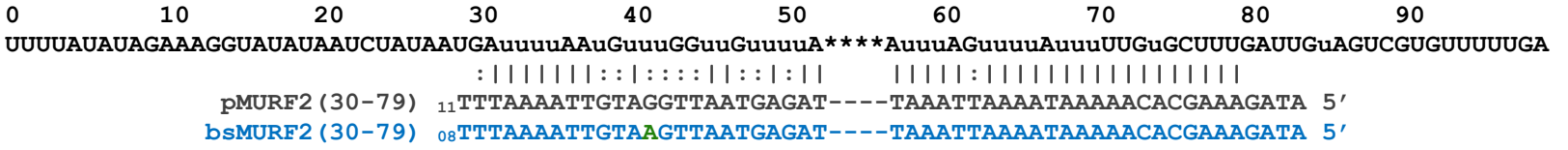



300 310 320 330 340 350 360 370 380 390
uuAuuuuGuuuuGuGuuuuuuuGuAuuG*UUGuuuuuuAuUUGGuuuGuuuGGuuuuuuuuuuG***UAuuuuuuuGUUGuGuuuuGuGuuAuuuuuuuGAuuuA
GATAGAGCGGAACACAAAAAAAAAA 5′(280-320)
pCR4(374-417) 11TATATAGAATACAGTAAGAGACTAAGT
bsCR4(374-415) 17TATATAGAATATAATGGAGAACTAAGT
11TTAAAATATAAGAGATATAGT-GACGGAAATAGACTAGACAAACCATAAAATA 5′(307-351)
09TTAAAATATGGAAAATATAGT-GACAGAGATGAGCTGAACAAACCAATAAATTAGTTGGTTTGTT 5′ bsCR4(307-352)
pCR4(343-388) 15TAGTAAGTTAAAGAGAGAT---ATGAGAGACAATACAAAACACAATATATA 5′
bsCR4(340-390) 05TTTAAATGAGTTAAGAGAGAAT---ATAAAGAGCAGCATAAAACACAATAAATATA 5′
400 410 420 430 440 450 460 470 480 490
uuuuuuAuGuUGuuuuuUGuuuuGGG***UG*GuuuuuuuuGuuuuuGuuuuuuuuuuuuuuGuuuAuGuuuGuuuuuAuuuGuGGuuGuuGuuAuuuuGuuA
GAAAGATACAACAAAAAAAAAA 5′(374-417)
pCR4(475-519) 12TTAAATATTGATAGTAATGAGACGAT
AGAGAATACAACAAAATATATA 5′(374-415)
bsCR4(478-524) 14TATATTAGTGATAATGAGATAAT
14TAATATGGTAGAAGATAAGACTC---AC-TAAAGAAACAGAAACAAAAAAAAAAA 5′ pCR4(404-457)
08TAATATAGTAGAAGATAAGACTC---AC-TAAAGAAACAGAAACAAAAAAAAAAAAA 5′ bsCR4(404-458)
pCR4(442-489) 12TTAAAGTAGAAAGAGAGAGTAAATGCAAGTAAAGATAGACACCAACAATA 5′
bsCR4(442-487) 11TTAAAATAGAAAGAGAGAGTAAATACAGATAAAGATGAACACCAACAAAATA 5′
bsCR4(453-497) 11TAAAGAGTAGATGTAGATAAGAATGAATACTAACAACAATAAAACATA 5′
bsCR4(453-497) 15TAAGAGATAGATGTAAGTGAGAATAAACACTAGTAACAATAAAACATA 5′
500 510 520 530 540 550 560 570 580 590
GuuuGGuuGuuGUUGuuAuuUGuGuAuA****GGUUUAuuUAuA*UGCGuuuuuuAuuuuAGAuAAuUAuG****G****UA**UUGGUUUUAUAAAAUG
TAGACTAATAACAACAATAA-TATATA 5′(475-519)
CAGACTAGTAACAACAATAAACATA 5′ (478-524)
03TTTAATAGTAGTAATAGATACATAT----CCGAGTGAATAT-ACGCAAAAAAAA 5′pCR4(504-554)
12TATAGTAGTGATAGATACATGT----CTAGATGAGTAT-ACGCAAAAAAAAAAA 5′ bsCR4(507-554)
pCR4(542-575) 12TAAT-ATGTGAAGAGTAGAGTCTATTAATGC----C-----T--AACCAAAATATTTAA 5′
bsCR4(542-584) 19TAAT-GTGTGAAAGATGGAATCTATTAGTGT----C----AT--ACCAAAATATTTATA 5′
600
UUUUUUCU polyA

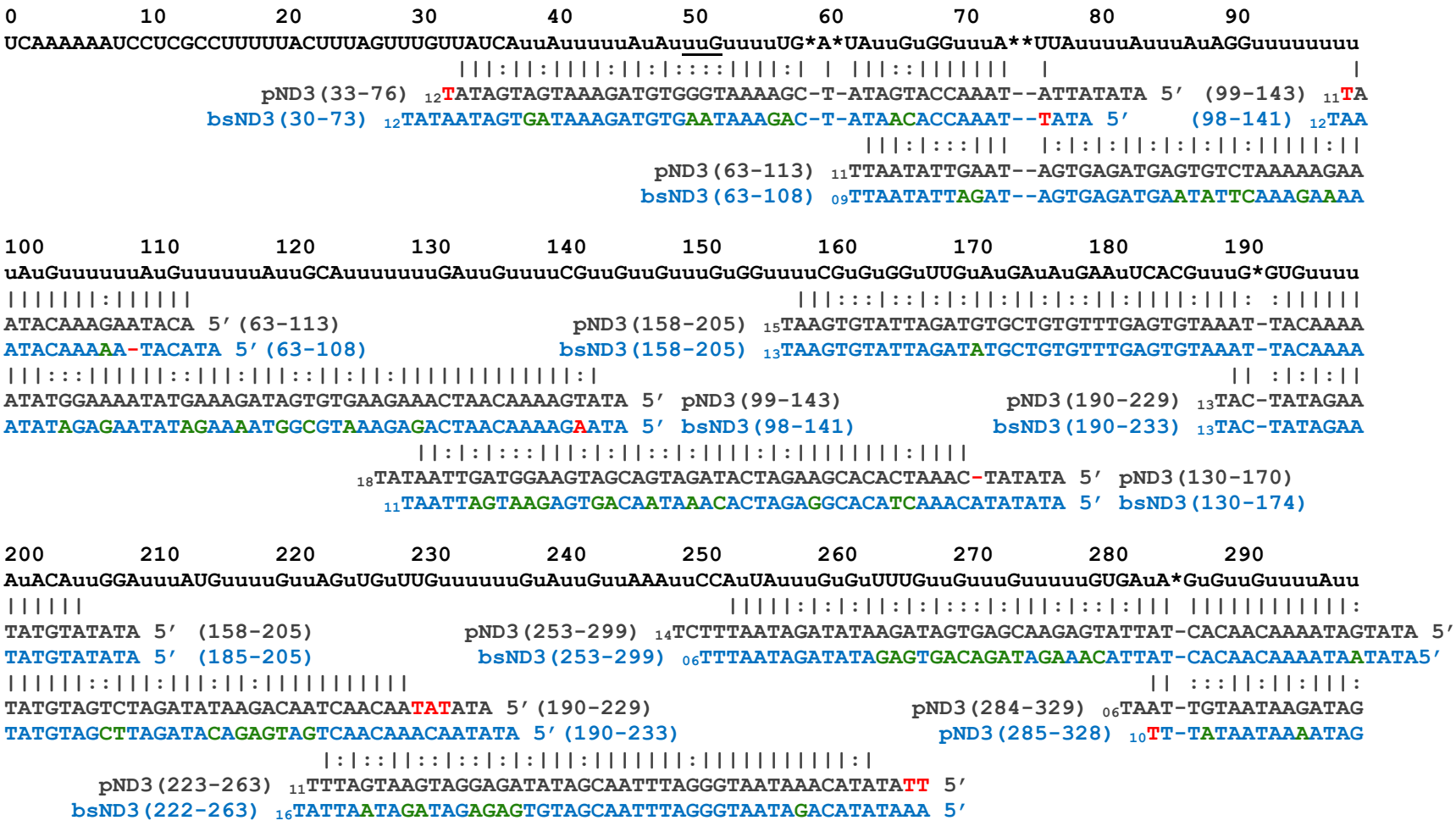
E) Cytochrome b



F) Maxicircle unidentified reading frame II

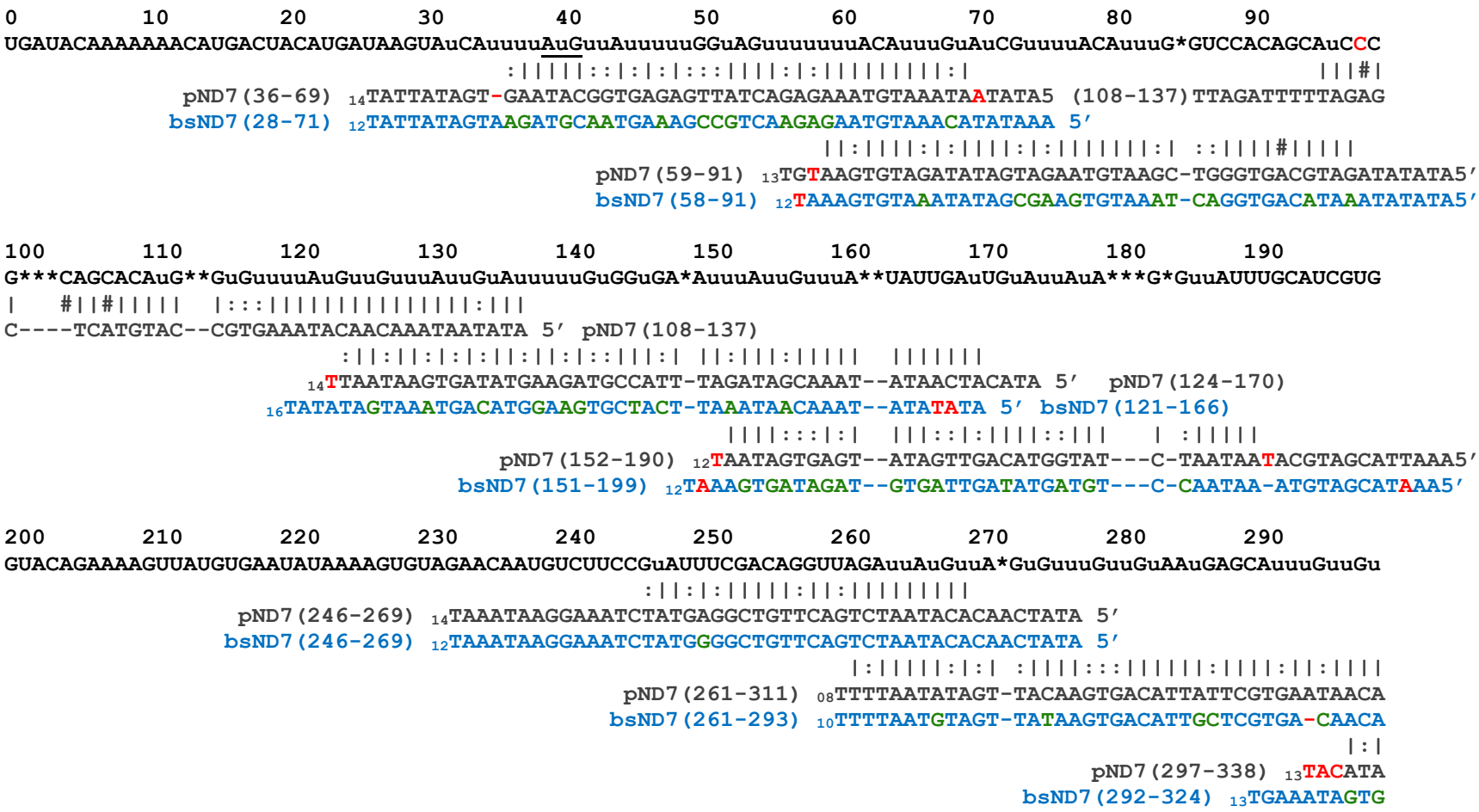


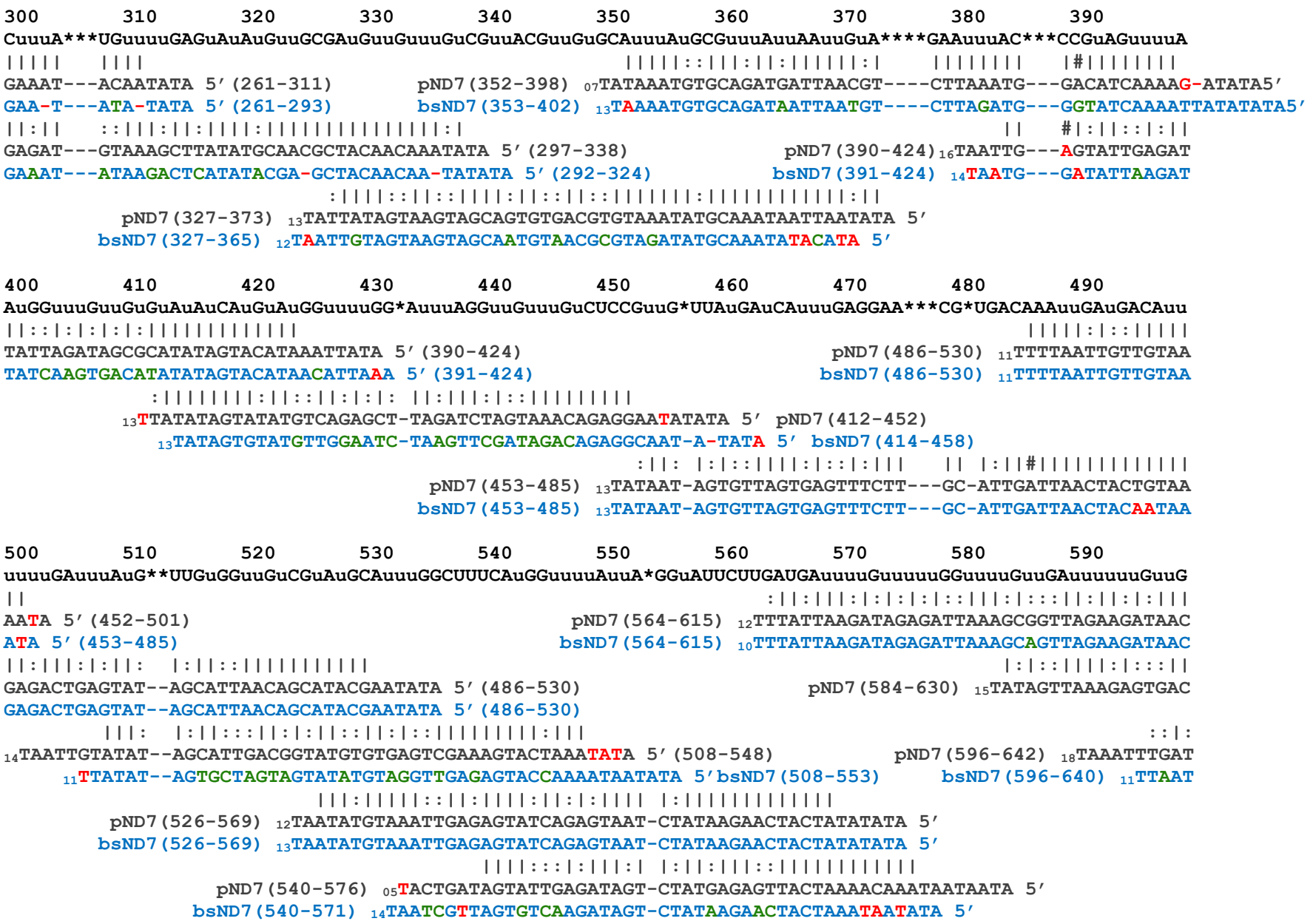
G) NADH Dehydrogenase subunit 3

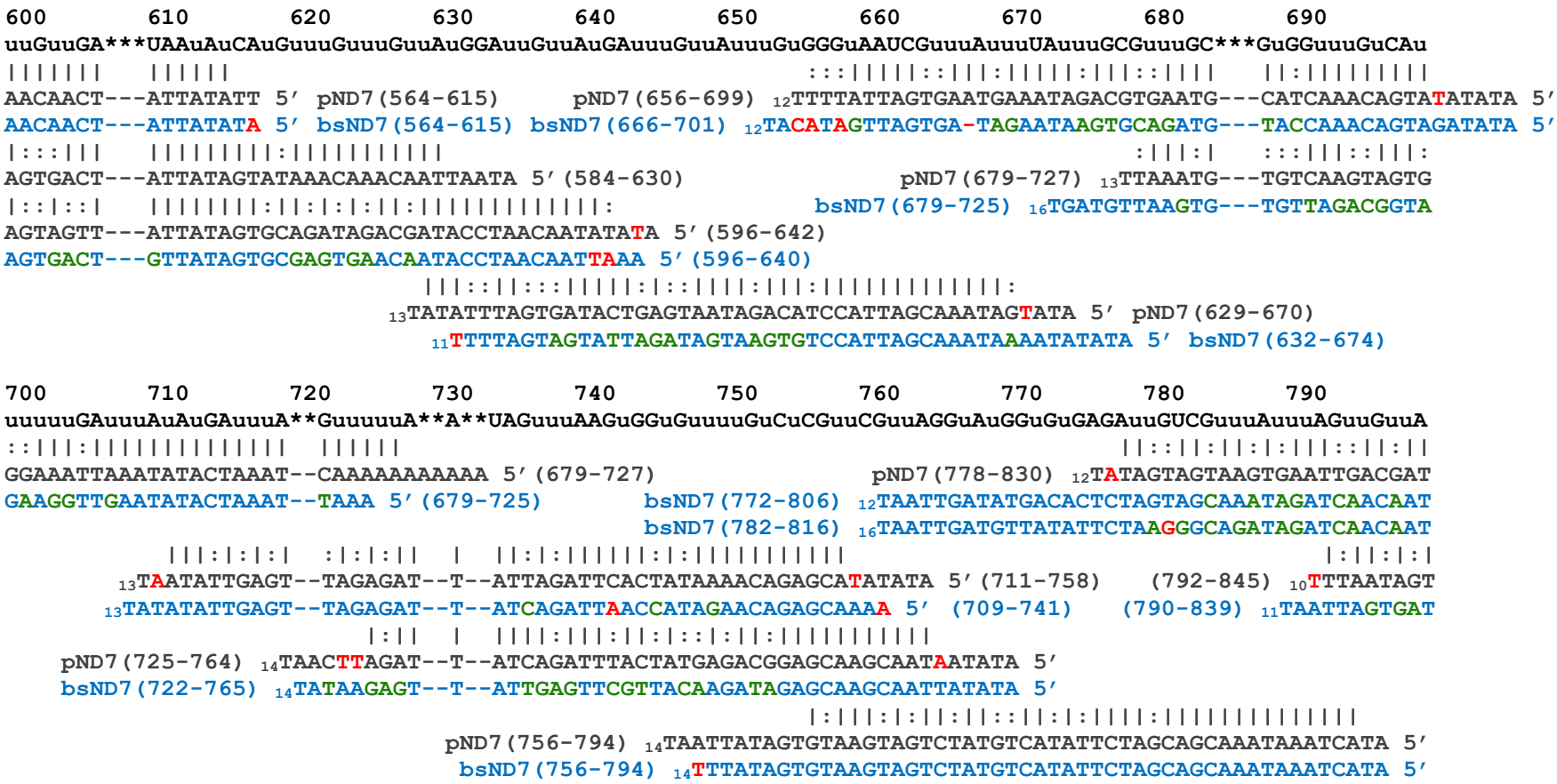


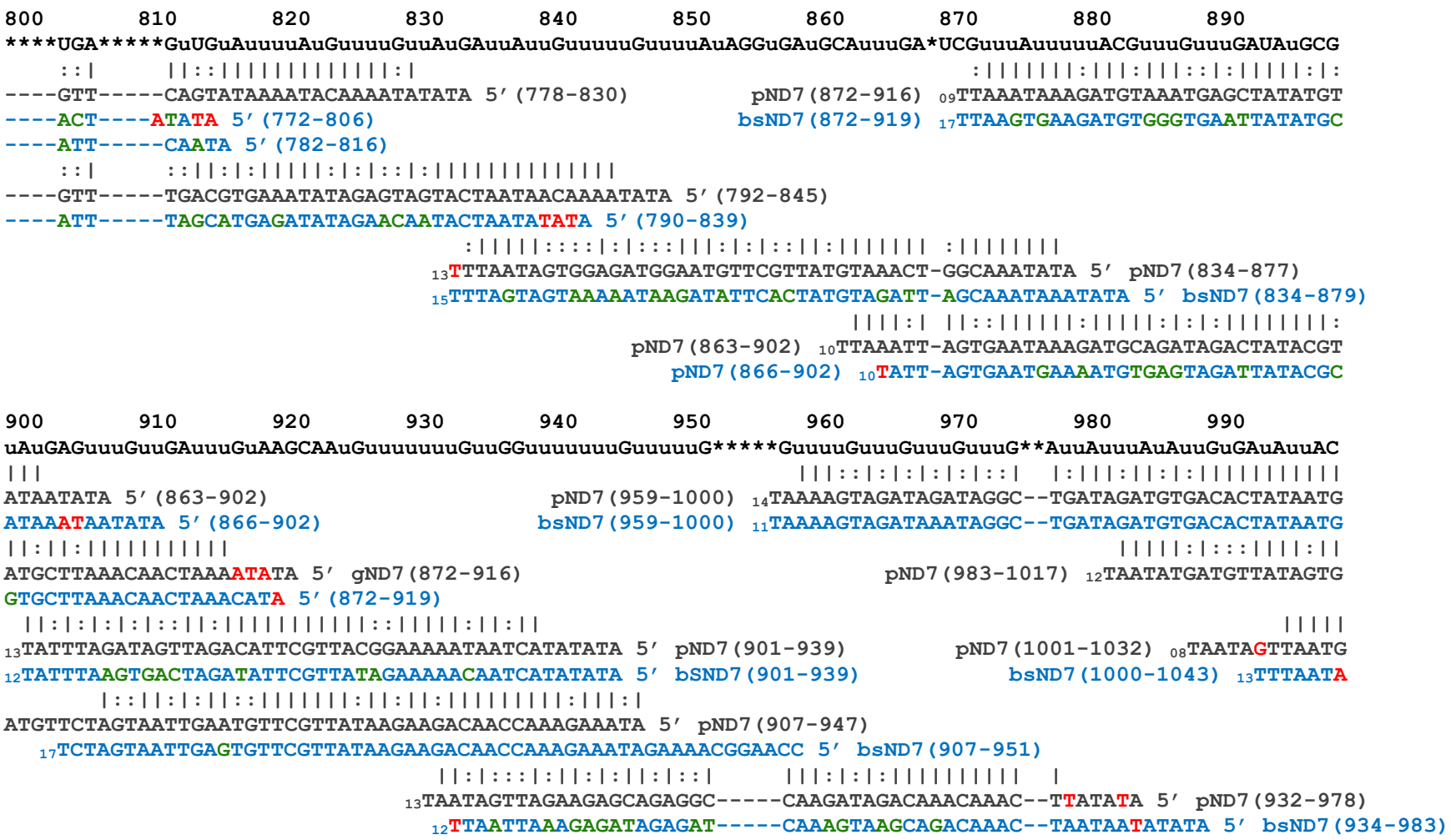

300 310 320 330 340 350 360 370 380 390
uuuGuuAUGGuuuuuuGuUUUUGuGGuuuuuGuuuuuuGuuGuAuGuAuAG****GAuuUGuGuGGuAuuuuuGGGAUCAC*GuAuAUUUGUGUGGUGUA
pND3(402-438) 13TATATTGC
AGATAATATCGAAAGATAAAAACACCAAAATATA 5′(284-329)
bsND3(402-435) 08TGC
AGATAATACCGGAAGATAGAAACACCAAA-TA 5′(285-328)
pND3(322-369) 13TTATTAAGAGTAGAAGGTAGCATGTATATT----CTAAGTACACCATAA-TATATA 5′
bsND3(322-370) 12TTATTAAGAGTAGAAGGTAGCATGTATATT----CTAAGTACACCATAAATA 5′
pND3(347-388) 18TATTTTATC----TTAAGTGTATCATAGAGACTTTAGTG-CATATAACAAATGTATA 5′
bsND3(355-388) 11TTT---ATTAAGTGTACTGTGAAAGCTTTGGTG-TATATAACAAACAAATATA 5′
400 410 420 430 440 450 460
AUUUUAuuuuGuuuAuGA**UGuuuUUUGUUGUAUUAUACAUAUUAUAUUAAUAAAUAUAUAAAA
TTAAGTAGAGTGAATACT--GTAAAAGACAACATAATATTATA 5′ (402-438)
TTAAATGAGATAAATGCT--GCAGAGAACAACATAAAAT-ATATA 5′ (402-435)

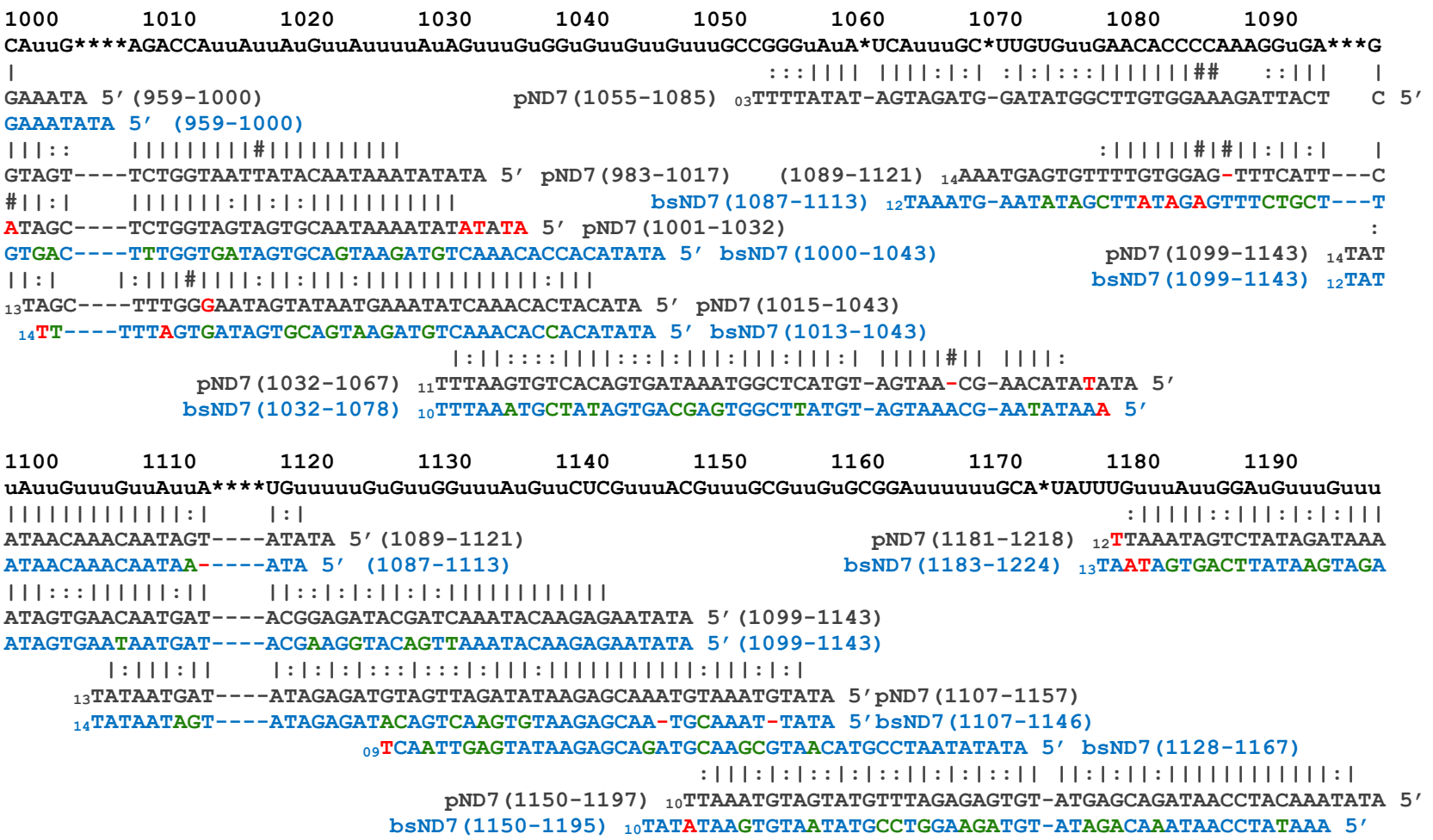
H) NADH Dehydrogenase subunit 7

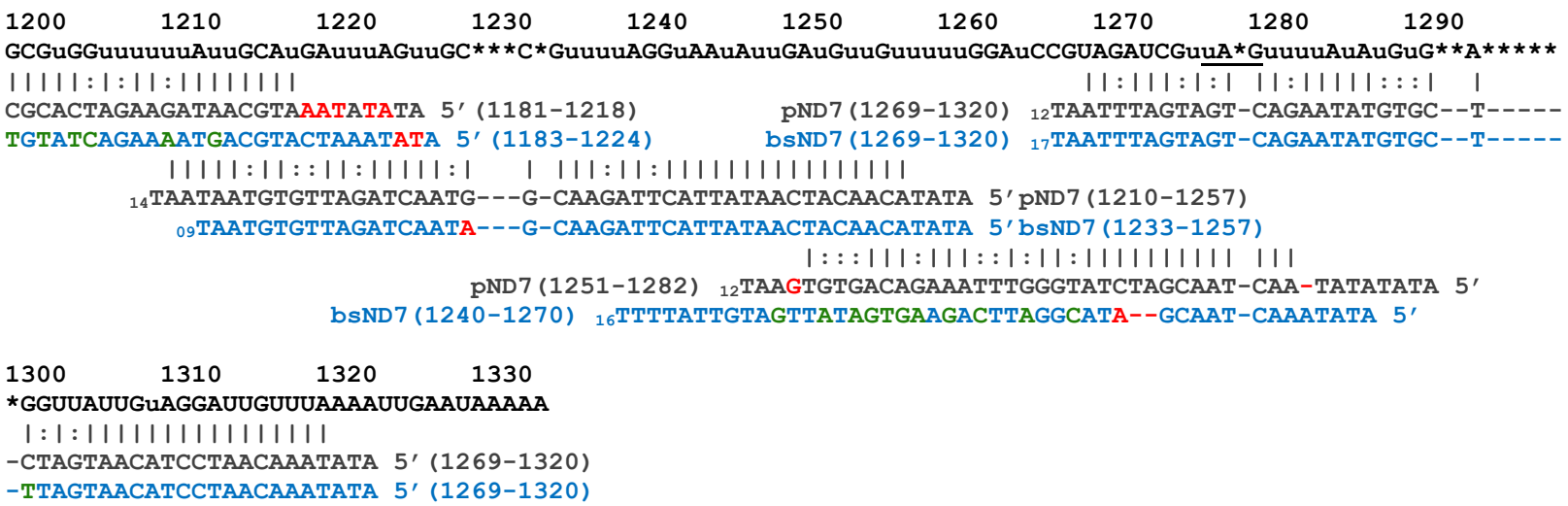



300 310 320 330 340 350 360 370 380 390
CuuuA***UGuuuuGAGuAuAuGuuGCGAuGuuGuuuGuCGuuACGuuGuGCAuuuAuGCGuuuAuuAAuuGuA****GAAuuuAC***CCGuAGuuuuA
GAAAT---ACAATATA 5′(261-311)
pND7(352-398) 07TATAAATGTGCAGATGATTAACGT----CTTAAATG---GACATCAAAAG-ATATA5′
GAA-T---ATA-TATA 5′(261-293)
bsND7(353-402) 13TAAAATGTGCAGATAATTAATGT----CTTAGATG---GGTATCAAAATTATATATA5′
GAGAT---GTAAAGCTTATATGCAACGCTACAACAAATATA 5′(297-338)
pND7(390-424)16TAATTG---AGTATTGAGAT
GAAAT---ATAAGACTCATATACGA-GCTACAACAA-TATATA 5′(292-324)
bsND7(391-424) 14TAATG---GATATTAAGAT
pND7(327-373) 13TATTATAGTAAGTAGCAGTGTGACGTGTAAATATGCAAATAATTAATATA 5′
bsND7(327-365) 12TAATTGTAGTAAGTAGCAATGTAACGCGTAGATATGCAAATATACATA 5′
400 410 420 430 440 450 460 470 480 490
AuGGuuuGuuGuGuAuAuCAuGuAuGGuuuuGG*AuuuAGGuuGuuuGuCUCCGuuG*UUAuGAuCAuuuGAGGAA***CG*UGACAAAuuGAuGACAuu
TATTAGATAGCGCATATAGTACATAAATTATA 5′(390-424)
pND7(486-530) 11TTTTAATTGTTGTAA
TATCAAGTGACATATATAGTACATAACATTAAA 5′(391-424)
bsND7(486-530) 11TTTTAATTGTTGTAA
13TTATATAGTATATGTCAGAGCT-TAGATCTAGTAAACAGAGGAATATATA 5′ pND7(412-452)
13TATAGTGTATGTTGGAATC-TAAGTTCGATAGACAGAGGCAAT-A-TATA 5′ bsND7(414-458)
pND7(453-485) 13TATAAT-AGTGTTAGTGAGTTTCTT---GC-ATTGATTAACTACTGTAA
bsND7(453-485) 13TATAAT-AGTGTTAGTGAGTTTCTT---GC-ATTGATTAACTACAATAA
500 510 520 530 540 550 560 570 580 590
uuuuGAuuuAuG**UUGuGGuuGuCGuAuGCAuuuGGCUUUCAuGGuuuuAuuA*GGuAUUCUUGAUGAuuuuGuuuuuGGuuuuGuuGAuuuuuuGuuG
AATA 5′(452-501)
pND7(564-615) 12TTTATTAAGATAGAGATTAAAGCGGTTAGAAGATAAC
ATA 5′(453-485)
bsND7(564-615) 10TTTATTAAGATAGAGATTAAAGCAGTTAGAAGATAAC
GAGACTGAGTAT--AGCATTAACAGCATACGAATATA 5′(486-530)
pND7(584-630) 15TATAGTTAAAGAGTGAC
GAGACTGAGTAT--AGCATTAACAGCATACGAATATA 5′(486-530)
14TAATTGTATAT--AGCATTGACGGTATGTGTGAGTCGAAAGTACTAAATATA 5′(508-548)
pND7(596-642) 18TAAATTTGAT
11TTATAT--AGTGCTAGTAGTATATGTAGGTTGAGAGTACCAAAATAATATA 5′bsND7(508-553)
bsND7(596-640) 11TTAAT
pND7(526-569) 12TAATATGTAAATTGAGAGTATCAGAGTAAT-CTATAAGAACTACTATATATA 5′
bsND7(526-569) 13TAATATGTAAATTGAGAGTATCAGAGTAAT-CTATAAGAACTACTATATATA 5′
pND7(540-576) 05TACTGATAGTATTGAGATAGT-CTATGAGAGTTACTAAAACAAATAATAATA 5′
bsND7(540-571) 14TAATCGTTAGTGTCAAGATAGT-CTATAAGAACTACTAAATAATATA 5′

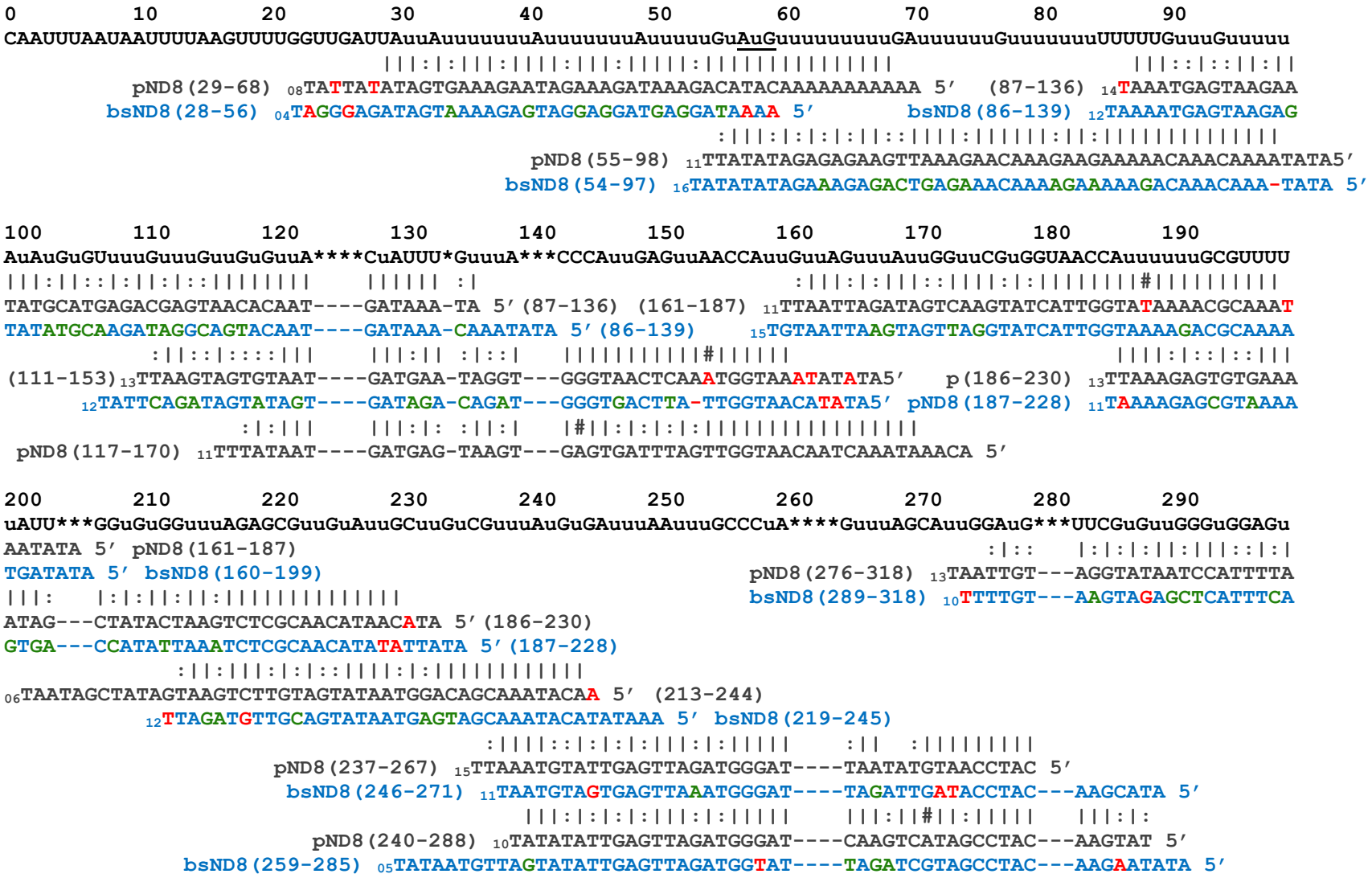

600 610 620 630 640 650 660 670 680 690
uuGuuGA***UAAuAuCAuGuuuGuuuGuuAuGGAuuGuuAuGAuuuGuuAuuuGuGGGuAAUCGuuuAuuuUAuuuGCGuuuGC***GuGGuuuGuCAu
AACAACT---ATTATATT 5′ pND7(564-615) pND7(656-699) 12TTTTATTAGTGAATGAAATAGACGTGAATG---CATCAAACAGTATATATA 5′
AACAACT---ATTATATA 5′ bsND7(564-615) bsND7(666-701) 12TACATAGTTAGTGA-TAGAATAAGTGCAGATG---TACCAAACAGTAGATATA 5′
AGTGACT---ATTATAGTATAAACAAACAATTAATA 5′(584-630) pND7(679-727) 13TTAAATG---TGTCAAGTAGTG
bsND7(679-725) 16TGATGTTAAGTG---TGTTAGACGGTA
AGTAGTT---ATTATAGTGCAGATAGACGATACCTAACAATATATA 5′(596-642)
AGTGACT---GTTATAGTGCGAGTGAACAATACCTAACAATTAAA 5′(596-640)
13TATATTTAGTGATACTGAGTAATAGACATCCATTAGCAAATAGTATA 5′ pND7(629-670)
11TTTTAGTAGTATTAGATAGTAAGTGTCCATTAGCAAATAAAATATATA 5′ bsND7(632-674)
700 710 720 730 740 750 760 770 780 790
uuuuuGAuuuAuAuGAuuuA**GuuuuuA**A**UAGuuuAAGuGGuGuuuuGuCuCGuuCGuuAGGuAuGGuGuGAGAuuGUCGuuuAuuuAGuuGuuA
GGAAATTAAATATACTAAAT--CAAAAAAAAAAA 5′(679-727) pND7(778-830) 12TATAGTAGTAAGTGAATTGACGAT
GAAGGTTGAATATACTAAAT--TAAA 5′(679-725) bsND7(772-806) 12TAATTGATATGACACTCTAGTAGCAAATAGATCAACAAT
bsND7(782-816) 16TAATTGATGTTATATTCTAAGGGCAGATAGATCAACAAT
13TAATATTGAGT--TAGAGAT--T--ATTAGATTCACTATAAAACAGAGCATATATA 5′(711-758) (792-845) 10TTTAATAGT
13TATATATTGAGT--TAGAGAT--T--ATCAGATTAACCATAGAACAGAGCAAAA 5′ (709-741) (790-839) 11TAATTAGTGAT
pND7(725-764) 14TAACTTAGAT--T--ATCAGATTTACTATGAGACGGAGCAAGCAATAATATA 5′
bsND7(722-765) 14TATAAGAGT--T--ATTGAGTTCGTTACAAGATAGAGCAAGCAATTATATA 5′
pND7(756-794) 14TAATTATAGTGTAAGTAGTCTATGTCATATTCTAGCAGCAAATAAATCATA 5′
bsND7(756-794) 14TTTATAGTGTAAGTAGTCTATGTCATATTCTAGCAGCAAATAAATCATA 5′

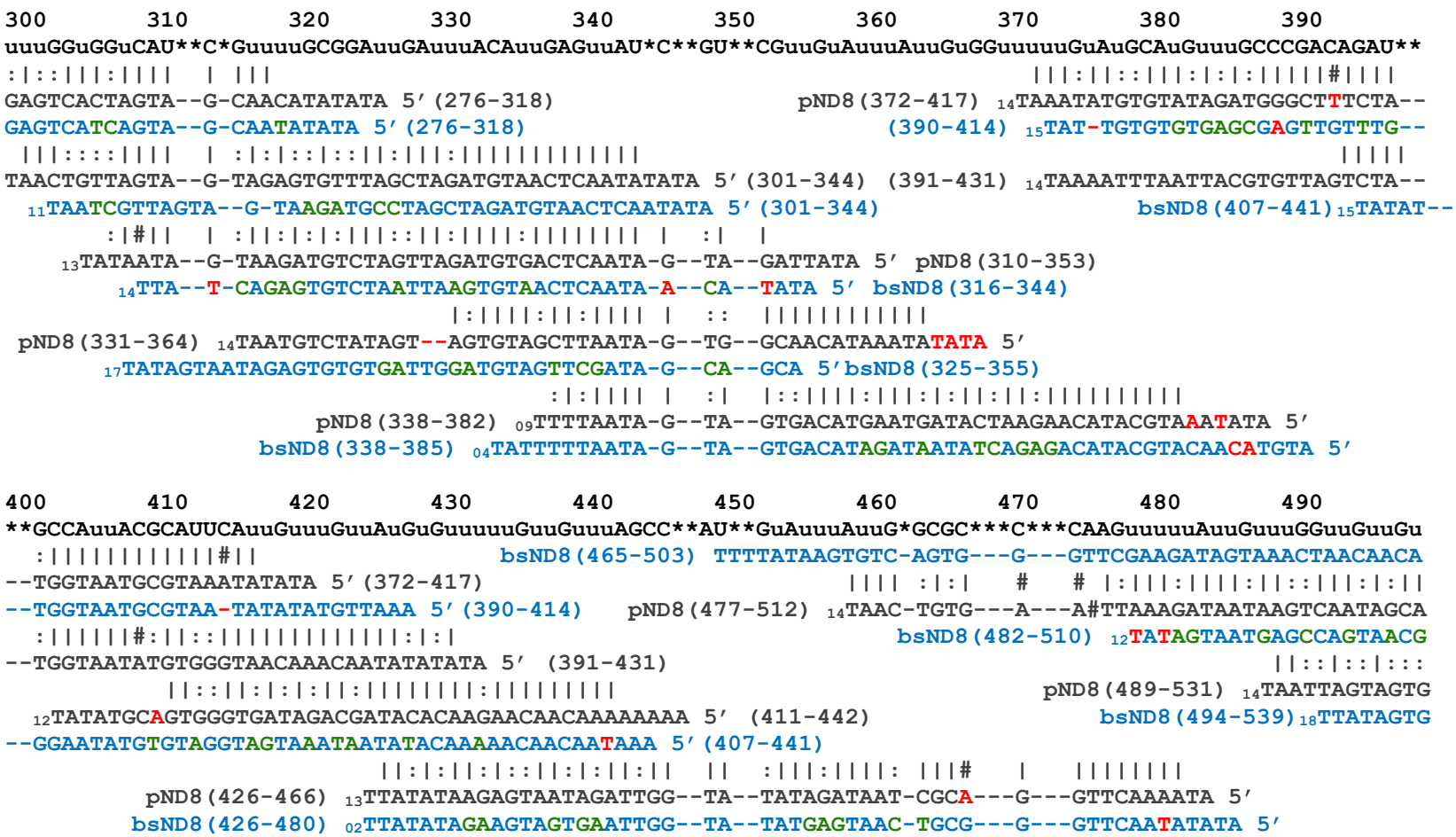

800 810 820 830 840 850 860 870 880 890
****UGA*****GuUGuAuuuuAuGuuuuGuuAuGAuuAuuGuuuuuGuuuuAuAGGuGAuGCAuuuGA*UCGuuuAuuuuuACGuuuGuuuGAUAuGCG
----GTT-----CAGTATAAAATACAAAATATATA 5′(778-830)
----ACT----ATATA 5′(772-806)
----ATT-----CAATA 5′(782-816)
----GTT-----TGACGTGAAATATAGAGTAGTACTAATAACAAAATATA 5′(792-845)
----ATT-----TAGCATGAGATATAGAACAATACTAATATATA 5′(790-839)
pND7(872-916) 09TTAAATAAAGATGTAAATGAGCTATATGT
bsND7(872-919) 17TTAAGTGAAGATGTGGGTGAATTATATGC
13TTTAATAGTGGAGATGGAATGTTCGTTATGTAAACT-GGCAAATATA 5′ pND7(834-877)
15TTTAGTAGTAAAAATAAGATATTCACTATGTAGATT-AGCAAATAAATATA 5′ bsND7(834-879)
pND7(863-902) 10TTAAATT-AGTGAATAAAGATGCAGATAGACTATACGT
pND7(866-902) 10TATT-AGTGAATGAAAATGTGAGTAGATTATACGC
900 910 920 930 940 950 960 970 980 990
uAuGAGuuuGuuGAuuuGuAAGCAAuGuuuuuuuGuuGGuuuuuuuGuuuuuG*****GuuuuGuuuGuuuGuuuG**AuuAuuuAuAuuGuGAuAuuAC
ATAATATA 5′(863-902)
ATAAATAATATA 5′(866-902)
ATGCTTAAACAACTAAAATATA 5′ gND7(872-916)
GTGCTTAAACAACTAAACATA 5′(872-919)
pND7(959-1000) 14TAAAAGTAGATAGATAGGC--TGATAGATGTGACACTATAATG
bsND7(959-1000) 11TAAAAGTAGATAAATAGGC--TGATAGATGTGACACTATAATG
pND7(983-1017) 12TAATATGATGTTATAGTG
13TATTTAGATAGTTAGACATTCGTTACGGAAAAATAATCATATATA 5′ pND7(901-939)
12TATTTAAGTGACTAGATATTCGTTATAGAAAAACAATCATATATA 5′ bSND7(901-939)
ATGTTCTAGTAATTGAATGTTCGTTATAAGAAGACAACCAAAGAAATA 5′ pND7(907-947)
17TCTAGTAATTGAGTGTTCGTTATAAGAAGACAACCAAAGAAATAGAAAACGGAACC 5′ bsND7(907-951)
pND7(1001-1032) 08TAATAGTTAATG
bsND7(1000-1043) 13TTTAATA
13TAATAGTTAGAAGAGCAGAGGC-----CAAGATAGACAAACAAAC--TTATATA 5′ pND7(932-978)
12TTAATTAAAGAGATAGAGAT-----CAAAGTAAGCAGACAAAC--TAATAATATATA 5′ bsND7(934-983)

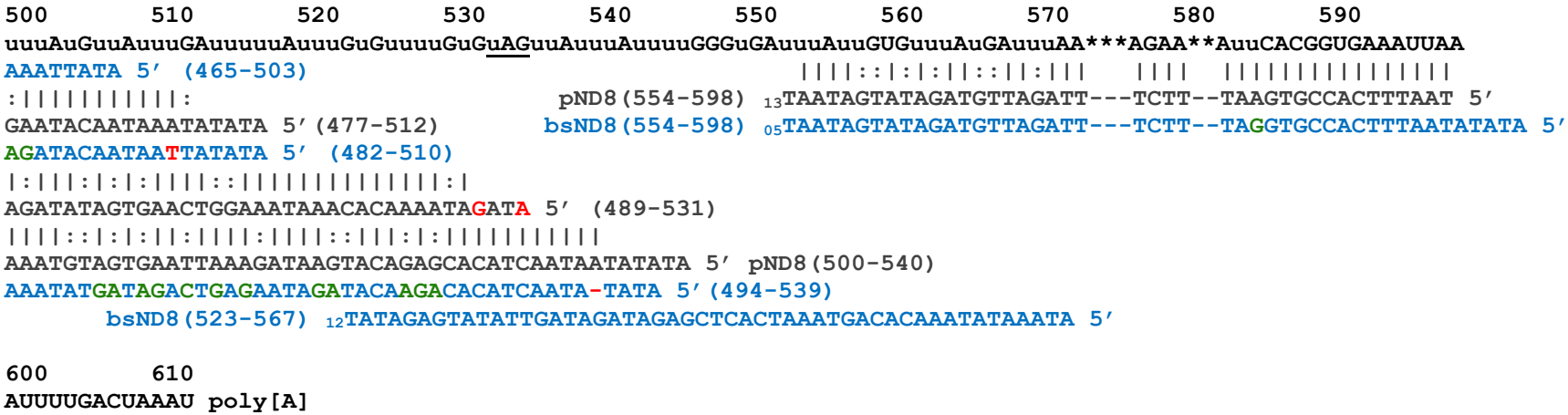

1000 1010 1020 1030 1040 1050 1060 1070 1080 1090
CAuuG****AGACCAuuAuuAuGuuAuuuuAuAGuuuGuGGuGuuGuuGuuuGCCGGGuAuA*UCAuuuGC*UUGUGuuGAACACCCCAAAGGuGA***G
GAAATA 5′(959-1000)
pND7(1055-1085) 03TTTTATAT-AGTAGATG-GATATGGCTTGTGGAAAGATTACT C 5′
GAAATATA 5′ (959-1000)
GTAGT----TCTGGTAATTATACAATAAATATATA 5′ pND7(983-1017)
(1089-1121) 14AAATGAGTGTTTTGTGGAG-TTTCATT---C
bsND7(1087-1113) 12TAAATG-AATATAGCTTATAGAGTTTCTGCT---T
ATAGC----TCTGGTAGTAGTGCAATAAAATATATATA 5′ pND7(1001-1032)
GTGAC----TTTGGTGATAGTGCAGTAAGATGTCAAACACCACATATA 5′ bsND7(1000-1043)
pND7(1099-1143) 14TAT
bsND7(1099-1143) 12TAT
13TAGC----TTTGGGAATAGTATAATGAAATATCAAACACTACATA 5′ pND7(1015-1043)
14TT----TTTAGTGATAGTGCAGTAAGATGTCAAACACCACATATA 5′ bsND7(1013-1043)
pND7(1032-1067) 11TTTAAGTGTCACAGTGATAAATGGCTCATGT-AGTAA-CG-AACATATATA 5′
bsND7(1032-1078) 10TTTAAATGCTATAGTGACGAGTGGCTTATGT-AGTAAACG-AATATAAA 5′
1100 1110 1120 1130 1140 1150 1160 1170 1180 1190
uAuuGuuuGuuAuuA****UGuuuuuGuGuuGGuuuAuGuuCUCGuuuACGuuuGCGuuGuGCGGAuuuuuuGCA*UAUUUGuuuAuuGGAuGuuuGuuu
ATAACAAACAATAGT----ATATA 5′(1089-1121)
pND7(1181-1218) 12TTAAATAGTCTATAGATAAA
ATAACAAACAATAA-----ATA 5′ (1087-1113)
bsND7(1183-1224) 13TAATAGTGACTTATAAGTAGA
ATAGTGAACAATGAT----ACGGAGATACGATCAAATACAAGAGAATATA 5′(1099-1143)
ATAGTGAATAATGAT----ACGAAGGTACAGTTAAATACAAGAGAATATA 5′(1099-1143)
13TATAATGAT----ATAGAGATGTAGTTAGATATAAGAGCAAATGTAAATGTATA 5′pND7(1107-1157)
14TATAATAGT----ATAGAGATACAGTCAAGTGTAAGAGCAA-TGCAAAT-TATA 5′bsND7(1107-1146)
09TCAATTGAGTATAAGAGCAGATGCAAGCGTAACATGCCTAATATATA 5′ bsND7(1128-1167)
pND7(1150-1197) 10TTAAATGTAGTATGTTTAGAGAGTGT-ATGAGCAGATAACCTACAAATATA 5′
bsND7(1150-1195) 10TATATAAGTGTAATATGCCTGGAAGATGT-ATAGACAAATAACCTATAAA 5′


1200 1210 1220 1230 1240 1250 1260 1270 1280 1290
GCGuGGuuuuuuAuuGCAuGAuuuAGuuGC***C*GuuuuAGGuAAuAuuGAuGuuGuuuuuGGAuCCGUAGAUCGuuA*GuuuuAuAuGuG**A*****
CGCACTAGAAGATAACGTAAATATATA 5′(1181-1218)
pND7(1269-1320) 12TAATTTAGTAGT-CAGAATATGTGC--T-----
TGTATCAGAAAATGACGTACTAAATATA 5′(1183-1224)
bsND7(1269-1320) 17TAATTTAGTAGT-CAGAATATGTGC--T-----
14TAATAATGTGTTAGATCAATG---G-CAAGATTCATTATAACTACAACATATA 5′pND7(1210-1257)
09TAATGTGTTAGATCAATA---G-CAAGATTCATTATAACTACAACATATA 5′bsND7(1233-1257)
pND7(1251-1282) 12TAAGTGTGACAGAAATTTGGGTATCTAGCAAT-CAA-TATATATA 5′
bsND7(1240-1270) 16TTTTATTGTAGTTATAGTGAAGACTTAGGCATA--GCAAT-CAAATATA 5′
1300 1310 1320 1330
*GGUUAUUGuAGGAUUGUUUAAAAUUGAAUAAAAA
-CTAGTAACATCCTAACAAATATA 5′(1269-1320)
-TTAGTAACATCCTAACAAATATA 5′(1269-1320)

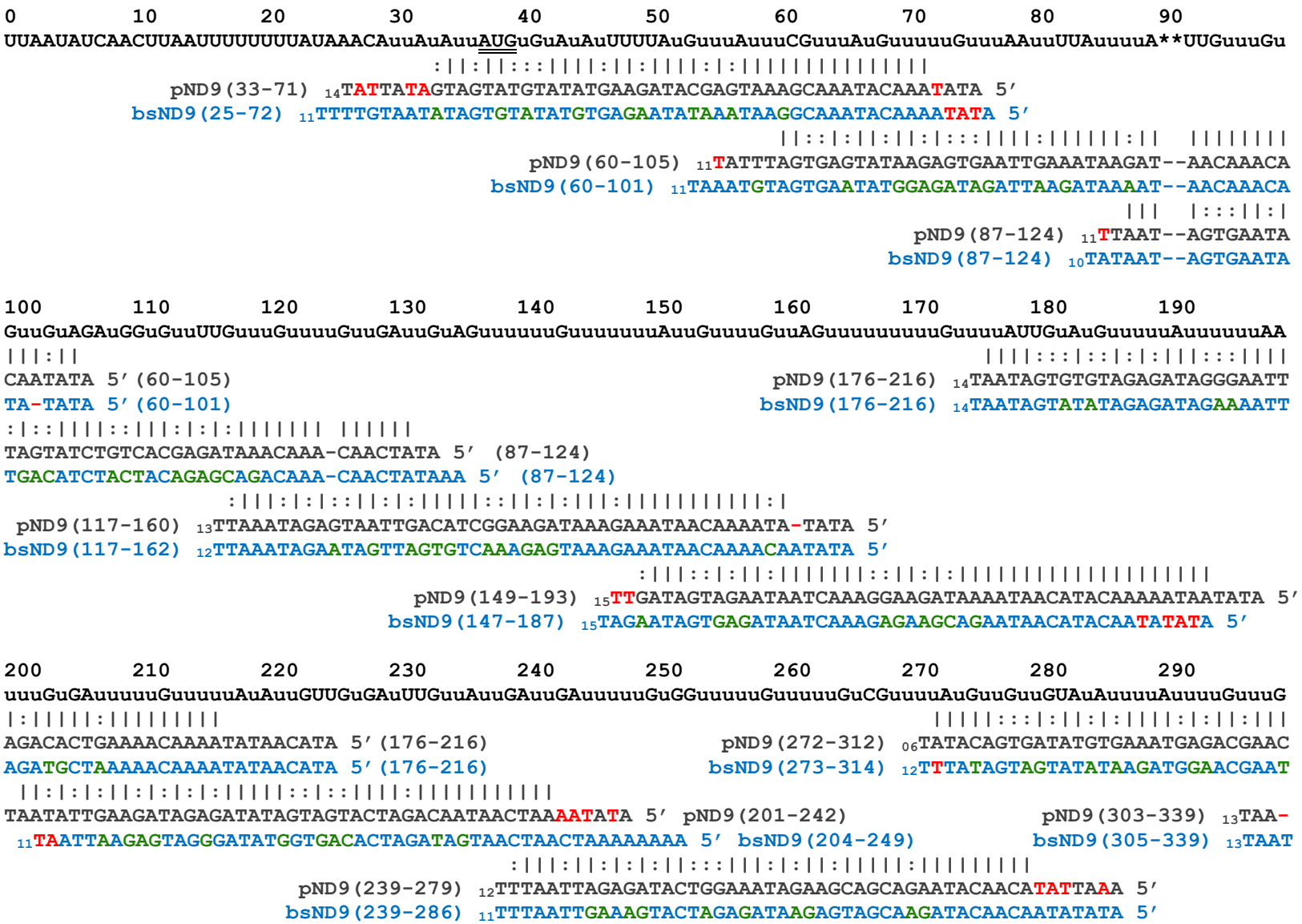
I) NADH Dehydrogenase subunit 8

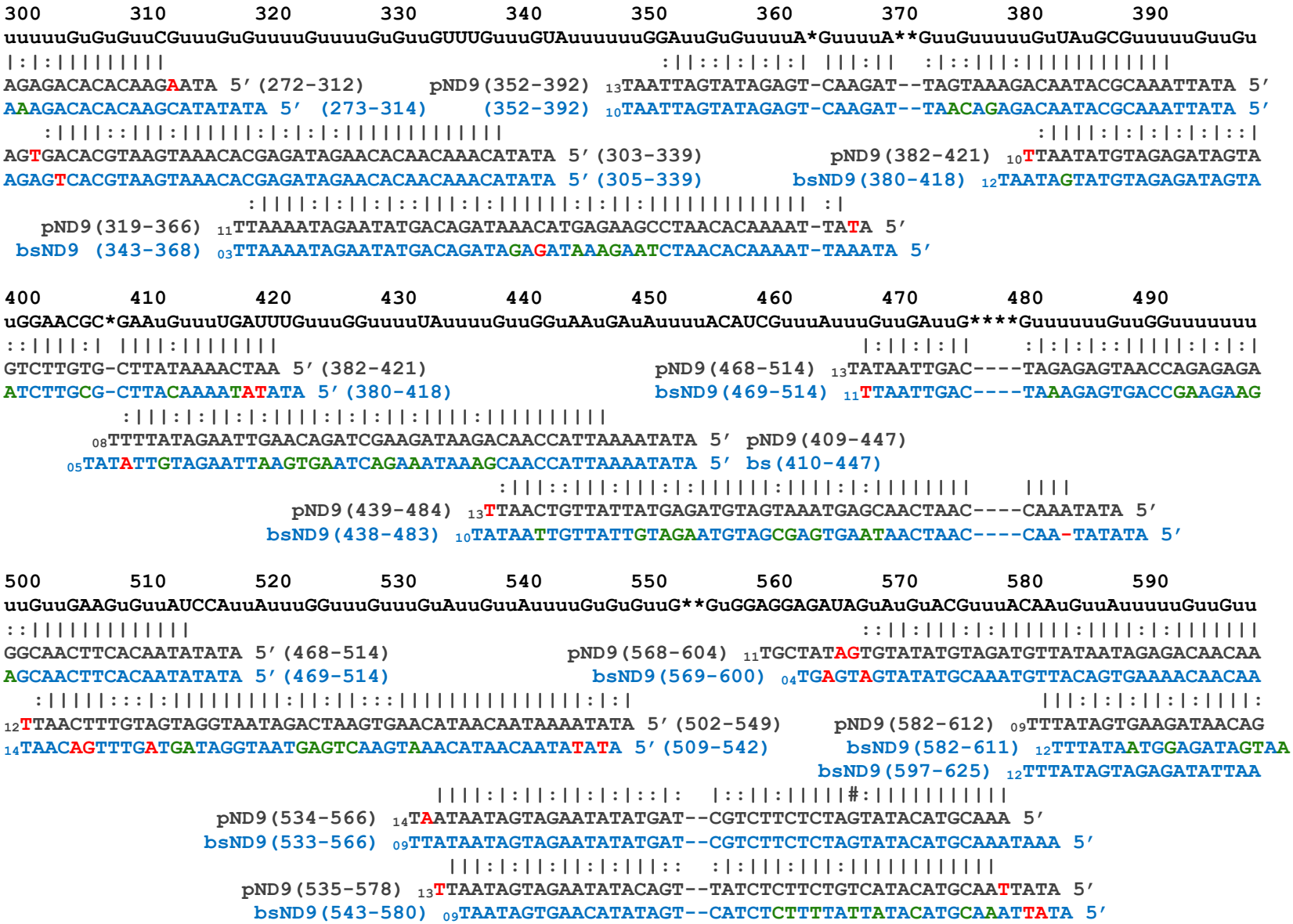


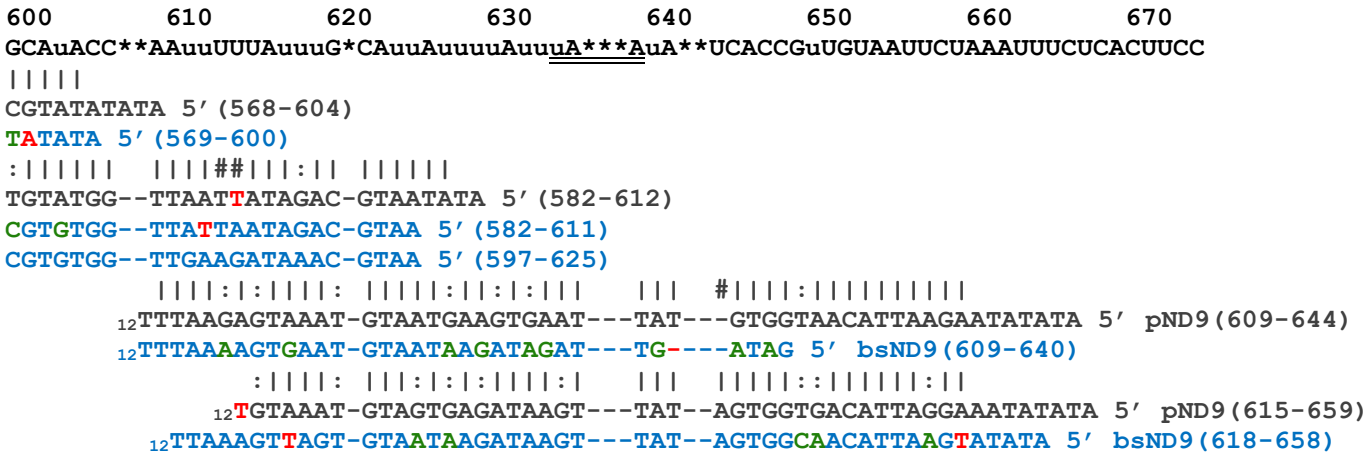

300 310 320 330 340 350 360 370 380 390
uuuGGuGGuCAU**C*GuuuuGCGGAuuGAuuuACAuuGAGuuAU*C**GU**CGuuGuAuuuAuuGuGGuuuuuGuAuGCAuGuuuGCCCGACAGAU**
GAGTCACTAGTA--G-CAACATATATA 5′(276-318)
GAGTCATCAGTA--G-CAATATATA 5′(276-318)
TAACTGTTAGTA--G-TAGAGTGTTTAGCTAGATGTAACTCAATATATA 5′(301-344)
11TAATCGTTAGTA--G-TAAGATGCCTAGCTAGATGTAACTCAATATA 5′(301-344)
13TATAATA--G-TAAGATGTCTAGTTAGATGTGACTCAATA-G--TA--GATTATA 5′
14TTA--T-CAGAGTGTCTAATTAAGTGTAACTCAATA-A--CA--TATA 5′
pND8(331-364) 14TAATGTCTATAGT--AGTGTAGCTTAATA-G--TG--GCAACATAAATATATA 5′
17TATAGTAATAGAGTGTGTGATTGGATGTAGTTCGATA-G--CA--GCA 5′
pND8(338-382) 09TTTTAATA-G--TA--GTGACATGAATGATACTAAGAACATACGTAAATATA 5′
bsND8(338-385) 04TATTTTTAATA-G--TA--GTGACATAGATAATATCAGAGACATACGTACAACATGTA 5′
400 410 420 430 440 450 460 470 480 490
GCCAuuACGCAUUCAuuGuuuGuuAuGuGuuuuuGuuGuuuAGCCAU**GuAuuuAuuG*GCGC***C***CAAGuuuuuAuuGuuuGGuuGuuGu
bsND8(465-503) TTTTATAAGTGTC-AGTG---G---GTTCGAAGATAGTAAACTAACAACA
--TGGTAATGCGTAAATATATA 5′(372-417)
--TGGTAATGCGTAA-TATATATGTTAAA 5′(390-414)
pND8(477-512) 14TAAC-TGTG---A---A#TTAAAGATAATAAGTCAATAGCA
bsND8(482-510) 12TATAGTAATGAGCCAGTAACG
--TGGTAATATGTGGGTAACAAACAATATATATA 5′ (391-431)
pND8(489-531) 14TAATTAGTAGTG
12TATATGCAGTGGGTGATAGACGATACACAAGAACAACAAAAAAAA 5′ (411-442)
bsND8(494-539)18TTATAGTG
--GGAATATGTGTAGGTAGTAAATAATATACAAAACAACAATAAA 5′(407-441)
pND8(426-466) 13TTATATAAGAGTAATAGATTGG--TA--TATAGATAAT-CGCA---G---GTTCAAAATA 5′
bsND8(426-480) 02TTATATAGAAGTAGTGAATTGG--TA--TATGAGTAAC-TGCG---G---GTTCAATATATA 5′


500 510 520 530 540 550 560 570 580 590
uuuAuGuuAuuuGAuuuuuAuuuGuGuuuuGuGuAGuuAuuuAuuuuGGGuGAuuuAuuGUGuuuAuGAuuuAA***AGAA**AuuCACGGUGAAAUUAA
AAATTATA 5′ (465-503)
pND8(554-598) 13TAATAGTATAGATGTTAGATT---TCTT--TAAGTGCCACTTTAAT 5′
GAATACAATAAATATATA 5′(477-512)
bsND8(554-598) 05TAATAGTATAGATGTTAGATT---TCTT--TAGGTGCCACTTTAATATATA 5′
AGATACAATAATTATATA 5′ (482-510)
AGATATAGTGAACTGGAAATAAACACAAAATAGATA 5′ (489-531)
AAATGTAGTGAATTAAAGATAAGTACAGAGCACATCAATAATATATA 5′ pND8(500-540)
AAATATGATAGACTGAGAATAGATACAAGACACATCAATA-TATA 5′(494-539)
bsND8(523-567) 12TATAGAGTATATTGATAGATAGAGCTCACTAAATGACACAAATATAAATA 5′
600 610
AUUUUGACUAAAU poly[A]

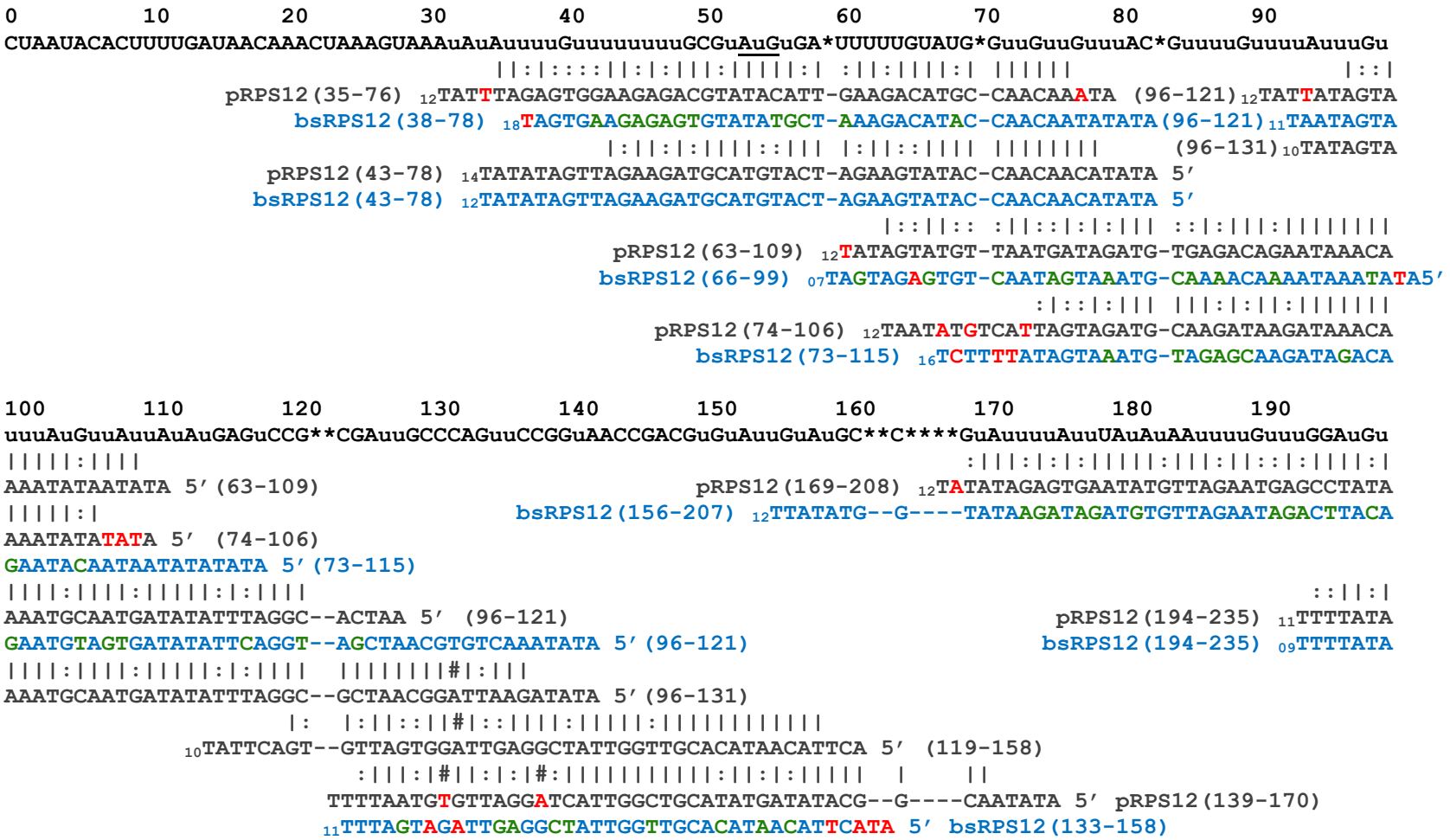
J) NADH Dehydrogenase subunit 9




300 310 320 330 340 350 360 370 380 390
AGAGACACACAAGAATA 5′(272-312) pND9(352-392) 13TAATTAGTATAGAGT-CAAGAT--TAGTAAAGACAATACGCAAATTATA 5′
AAAGACACACAAGCATATATA 5′ (273-314) (352-392) 10TAATTAGTATAGAGT-CAAGAT--TAACAGAGACAATACGCAAATTATA 5′
AGTGACACGTAAGTAAACACGAGATAGAACACAACAAACATATA 5′(303-339) pND9(382-421) 10TTAATATGTAGAGATAGTA
AGAGTCACGTAAGTAAACACGAGATAGAACACAACAAACATATA 5′(305-339) bsND9(380-418) 12TAATAGTATGTAGAGATAGTA
pND9(319-366) 11TTAAAATAGAATATGACAGATAAACATGAGAAGCCTAACACAAAAT-TATA 5′
bsND9 (343-368) 03TTAAAATAGAATATGACAGATAGAGATAAAGAATCTAACACAAAAT-TAAATA 5′
400 410 420 430 440 450 460 470 480 490
GTCTTGTG-CTTATAAAACTAA 5′(382-421) pND9(468-514) 13TATAATTGAC----TAGAGAGTAACCAGAGAGA
ATCTTGCG-CTTACAAAATATATA 5′(380-418) bsND9(469-514) 11TTAATTGAC----TAAAGAGTGACCGAAGAAG
08TTTTATAGAATTGAACAGATCGAAGATAAGACAACCATTAAAATATA 5′ pND9(409-447)
05TATATTGTAGAATTAAGTGAATCAGAAATAAAGCAACCATTAAAATATA 5′ bs(410-447)
pND9(439-484) 13TTAACTGTTATTATGAGATGTAGTAAATGAGCAACTAAC----CAAATATA 5′
bsND9(438-483) 10TATAATTGTTATTGTAGAATGTAGCGAGTGAATAACTAAC----CAA-TATATA 5′
500 510 520 530 540 550 560 570 580 590
GGCAACTTCACAATATATA 5′(468-514) pND9(568-604) 11TGCTATAGTGTATATGTAGATGTTATAATAGAGACAACAA
AGCAACTTCACAATATATA 5′(469-514) bsND9(569-600) 04TGAGTAGTATATGCAAATGTTACAGTGAAAACAACAA
12TTAACTTTGTAGTAGGTAATAGACTAAGTGAACATAACAATAAAATATA 5′(502-549) pND9(582-612) 09TTTATAGTGAAGATAACAG
14TAACAGTTTGATGATAGGTAATGAGTCAAGTAAACATAACAATATATA 5′(509-542) bsND9(582-611) 12TTTATAATGGAGATAGTAA
bsND9(597-625) 12TTTATAGTAGAGATATTAA
pND9(534-566) 14TAATAATAGTAGAATATATGAT--CGTCTTCTCTAGTATACATGCAAA 5′
bsND9(533-566) 09TTATAATAGTAGAATATATGAT--CGTCTTCTCTAGTATACATGCAAATAAA 5′
pND9(535-578) 13TTAATAGTAGAATATACAGT--TATCTCTTCTGTCATACATGCAATTATA 5′
bsND9(543-580) 09TAATAGTGAACATATAGT--CATCTCTTTTATTATACATGCAAATTATA 5′


600 610 620 630 640 650 660 670
GCAuACC**AAuuUUUAuuuG*CAuuAuuuuAuuuA***AuA**UCACCGuUGUAAUUCUAAAUUUCUCACUUCC
CGTATATATA 5′(568-604)
TATATA 5′(569-600)
TGTATGG--TTAATTATAGAC-GTAATATA 5′(582-612)
CGTGTGG--TTATTAATAGAC-GTAA 5′(582-611)
CGTGTGG--TTGAAGATAAAC-GTAA 5′(597-625)
12TTTAAGAGTAAAT-GTAATGAAGTGAAT---TAT---GTGGTAACATTAAGAATATATA 5′ pND9(609-644)
12TTTAAAAGTGAAT-GTAATAAGATAGAT---TG----ATAG 5′ bsND9(609-640)
12TGTAAAT-GTAGTGAGATAAGT---TAT--AGTGGTGACATTAGGAAATATATA 5′ pND9(615-659)
12TTAAAGTTAGT-GTAATAAGATAAGT---TAT--AGTGGCAACATTAAGTATATA 5′ bsND9(618-658)

K) Ribosomal Protein S12




200 210 220 230 240 250 260 270 280 290
uGCGuuGuuuuuuuuGuuGuuuuAuuGGuuuAGuuAuG**UCAuuAuuuAuuAuAGA***GGGUGGuGGuuuuGuuGAuuuACCC***G****GuG*UAA
ACGCAACAATAAATA 5′ (169-208)
pRPS12(267-322) 07TTTAAAGTGACTAGATAGG---T----CAT-ATT
ATGCAACATATA 5′(156-207)
bsRPS12(288-322) 12TTTAAAGTAACTAGATGGA---T----CAT-ATT
bsRPS12(269-308) 10TTAGTACAAGAGCAGTTAAATGGG---C----TAC-ATT
14TATAT--AGTAGTGAATGATGTCT---TTTACCGTCAAAACAACTAGAATATA 5′pRPS12(234-280)
15TATAT--AGTAGTGAGTGATATCT---TTTACCGTCAAAACAACTAGAATATA 5′bsRPS12(234-280)
ATGTAGTAGAGAAGATGACGAAATAGTCAAATCAAT-C--AGTAATATATA 5′(194-235)
ATGTAGCAGAGAAGATGACGAAATAGTCAAATCAAT-C--AGTAATATATA 5′(198-235)
14TATAATAGAGAGAGTAACGGAGTAATCGAGTCAATGC--AGTAATATATATA 5′ pRPS12(203-246)
07TATAATAGAGAGAGTAATGGGATAGTTAAATCAATAC--AGTAATTAAAATA 5′ bsRPS12(203-245)
300 310 320 330 340 350
AGuAuuAuACA*CG**UAuuGuAAGuuAGA*UUUAGAuAUAAGAUAUGUUUUU[AAUA]POLYA
TTATAGTATGT-GC--ATAACATATA 5′ (267-322)
TCATAATATATA 5′(269-308)
TTATAATGTGT-GC--ATAACATATA 5′bsRPS12(288-322)
16TAAGT-GT--ATAATGTTCAATCT-AAGTCTATATTCTATACAATATAAA 5′ pRPS12(309-349)
24TAAGT-GT--GTAATGTTCAATCT-AAATCT-TAGTCTATACAAAATAAA 5′bsRPS12(309-336)

APPENDIX C. All gRNA major classes pulled for ATPase 6 in the EATRO 164 procyclic (shaded gray) and bloodstream (white) transcriptomes.

Populations of gRNAs are bordered boxes. A) ATPase 6; B) Cytochrome Oxidase III; C) C-Rich Region 3; D) C- Rich Region 4; E) Cytochrome b; F) Maxicircle Unidentified Reading Frame II (Murf II); G) NADH Dehydrogenase Subunit 3; H) NADH Dehydrogenase Subunit 7; I) NADH Dehydrogenase Subunit 8; J) NADH Dehydrogenase Subunit 9; K) Ribosomal Protein S12.

A) ATPase 6

5'	3'	Reads	ATPase 6 gRNA Sequences
31	75	2,044	AT ATAAACGTAACTGAAATGAATCACGAGAGAAAGATAAAGATATAT AT_{12}
29	72	1,630	ATATAC AACGCAACCAGAGTAAATCATGAAGGGAAAGTGAAGGCATATTT T_{11}
31	75	143	AT ATAAACGTAACTGAAATGAATCGCGAGAGAAAGATAAAGATATAT AT_{21}
31	75	489	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAAAGATATGT AT_{09}
29	75	395	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAAAGATATGTTT T_{11}
35	75	224	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAAAGAT T_{13}
29	62	203	ATATACAACGCAACC AGATAAATCATGAAGAGAAAGTGAAGGTATATTT T_{09}
33	75	144	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAAAGATAT T_{12}*
31	62	95	AACGCAACC AGATAAATCATGAAGAGAAAGTGAAGGTATAT AT_{14}
39	75	82	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAA T_{16}
37	75	40	AT ATAAACGTAACCAAAATGGATCATGGAAGAGAAGTAAAG TCAACTTAAT$_{08}$*
62	102	1,435	ATACA ATCATACACAGTAGTACATATATAGTGATAGACGTGATTAA T_{11}
62	100	154	ATAAAA CATACACAATGATATATACATAGTAATAGATGTGATTAA T_{23}
62	102	92	ATATAA ATCATACACGATAATATATGCGTAGTAACAGATGTGATTAA T_{18}
74	102	10	ATATAT ATCATACACGATAATGCATATGTAGTAAC T_{13}
64	100	9	ATAAAA CATACACAATGATATATACATAGTAATAGATGTGATT T_{14}*
86	127	2,158	ATAT AAATACACAGTAGAATATGATCTAGGTTATGTATGATGATAT T_{11}
90	129	177	ATATA TAAAATACACGATAGAGCATAACTTAGATTGTATATGATA T_{20}
84	118	85	ATAATAATACAC ATAGAACATGACCTAGATTGTACATAGTGATATAT T_{12}
82	118	38	ATATAATAATACAC ATAGAACATGACCTAGATTGTACATAGTGATATATAT T_{13}
86	118	28	ATATAATAATACAC ATAGAACATGACCTAGATTGTACATAGTGATAT T_{15}
83	118	22	ATAATAATACAC ATAGAACATGACCTAGATTGTACATAGTGATATATA AT_{18}
91	129	21	ATATA TAAAATACACGATAGAGCATAACTTAGATTGTATATGAT T_{18}
113	152	743	ATAC ATCAAAAATCAACGTTAGACAGTTAAGATATGTGATAGAA GATAAT$_{12}$
105	152	54	AC ATCAAAAATCGACATTAGATAATTGAGGTATGTGATAGAGTATAATTT T_{11}
105	148	128	ATATC AAAATCAACATTGAGCAGTTAAAGTACGTGGTAAGATATAATTT T_{12}
116	152	84	AC ATCAAAAATCAACATTGAGCAATTGAGGTACATGATA TGATATAAT$_{09}$
104	148	16	ATATC AAAATCAACATTGAGCAGTTAAAGTACGTGGTAAGATATAATTTA T_{07}*
138	183	430	AT ATACAAATCAAACAGACAGAGTAATAGAAGGTTGAAGATTGATAT AGT$_{11}$
144	177	210	ATATC ATCAAACAAACAGAATAATAGAGAATCAGAGGT GAATGTTAAGT$_{15}$
139	175	3,468	ATAAA TAAACAAACAAAATGATAAAAGGTCAGAGATTGATG GTGAATAAT$_{08}$*
132	177	205	ATAT ATCAAACAAACAAAGTAATAGAAAGTCAGAGATTGATGTTAAATA T_{10}
164	208	54	ATAT ATAAACACAAATCAACGAATAGATATAAGTCAGATAGATGG TGTATTAT$_{12}$*
176	210	36	AT AAACAAACACAAATCAGTAGACGAGTACAAGT GAGATGGACGTATAGAT$_{07}$
165	208	24	ATAAT ACAAACACAAACTGATAGACGAATACGAGTTAGATGGACG TAT$_{06}$
158	208	14	ATAT ACAAACACAAACTGACGAATAGATACAGATTAAGTGAATGAAATAAT T_{11}
164	208	218	AT ATAAGCACAAACCAATAGACAGATATAAGTCAGATAGATGA TTAT$_{14}$
192	243	172	ATATAAATTAAACAACATAGATTACAGTGATAGAAGTAAATGTGAATTA T_{04}
218	248	147	ATC AGACTATGTGAGTTAGATGACGTGAATTATA CTGTATAT$_{12}$
190	243	52	ATATAAATTAAACAACATGAACTATGATGATAAAGGTAAATGTGAATTAAT T_{19}
189	243	10	ATATAAATTAAACAACATGAACTATGATGATAAAGGTAAATGTGAATTAATG TACT*
207	243	7	ATATAAATTAAACAACATGAACTATGATGATAAAGGT T_{10}
193	246	5	*ATTGTATAAATTAAACAACATGAACTATGATGATAAAGGTAAATGTGAATT T_{03}
224	269	864	ACATAA TAATACAATAATACGAGATTAGACTATGTGAATTAAATGATATGA T_{11}
226	269	808	ACATAA TAATACAATAATACGAGATTAGACTATGTGAATTAAATGATAT T_{13}
221	262	149	ATATAT ATAATACAAAATTGAACTGTATAAGTTAGACAATGTGAATT TT
219	262	105	ATATAT ATAATACAAAATTGAACTGTATAAGTTAGACAATGTGAATTAT T_{14}
218	262	57	ATATAT ATAATACAAAATTGAACTGTATAAGTTAGACAATGTGAATTATA T_{21}
221	267	32	AC ATACAATAATACAGAATTAAACTGTGTAAGTTAGATAGTGTAAATT T_{10}*

5'	3'	Reads	ATPase 6 gRNA Sequences cont.
248	292	25,157	ATATA AAATACAAATTCGAGTAGGTAGTACAATGATATGAGATTA T_{13}
253	298	5,405	ATATAT AACAACAAATATAGATTCAAGTAAGTGATGTAGTAATATGA T_{11}
255	292	244	ATATA AAATACAAATTCGAGTAGGTAGTACAATGATAT TATTATTAAT$_{15}$
252	292	134	ATATA AAATACAAATTCGAGTAGGTAGTACAATGATATAGA TTATTAAT$_{07}$
262	304	4,881	ATAT ATACAAAACAACAGATATAGATTCGGATAGGTAATATGA GATCT$_{13}$
259	304	395	AT ATATAAAACAACAGATATGAATTCAAGTGAGTGATACAGTA TAT$_{15}$*
254	298	290	ATATAT AACAACAAATACGAATTCAGATAGGTAGTATGATGATATA T_{15}
266	313	586	AAAAAA AAAAAAACAATACAAGATGACAGGTATAAGTTTGGATGAGTAAT T_{12}
281	313	14	AAA AAAAAAACAATACAGAATAGTAGGTATAGATT AGATATGTGAT$_{09}$*
284	310	13	ATAT AGAACAATACAAAATAACGAGTACAG T_{08}
283	310	2	TAT AGAACAATACAAAATAACGAGTACAGG ATAAGTGATAT$_{08}$
281	312	2	AT AGAAAACAATACAGAATAGTAGGTATAGATT AGATATGTGAT$_{19}$
291	329	133	AAATATAT AAATGCAATATACGATAGAGAAATGATATAAGATGATAA T_{16}*
293	329	12	AAATATAT AAATGCAATATACGATAGAGAAATGATATAAGATGAT T_{14}
301	345	647	ATAT AACAAAACAAAAGTAGAAGTGCAGTATATGATAGAAAAATGATGT CAAAT$_{11}$
300	346	263	ATAT AAACAAAACAGAAATAGAAATGCAATATACGATAAGAAAATGGTATA T_{12}
301	335	125	ATAT ACAAAACAT AAATAAAAGTGCAGTATATGATAAAGAGATAATAT T_{11}
301	345	33	ATAT AACAAAACAAAAGTAAAAGTGCAGTGTATGATAGAAAAATGATGT CAAAT$_{15}$
331	375	24,736	ATAT AATTATTAAACAAGAGAAAGTCACGTAAAAGGTAGAATGAAGATA T_{12}
332	378	8,776	AT ATAAATTATTAAACAGAAAGAGATCATGTAGAAAGTGAGATAGAAAT T_{14}
331	371	3,561	ATATAA ATTAAACAAAAAGAAATCACGTAGAAGACAGAATAGAGATA T_{13}
331	375	712	ATAT AATTATTAAACAAGAGAAAGTCACGTAAAAAGTAGAATGAAGATA TTAT$_{05}$
332	378	387	AT ATAAATTATTAAACAGAAAAGAGTCATATAGAAAATAAGATAGAAAT T_{12}
331	374	302	ATAT ATTATTAAACAAAGAGAAATCATATAAGAGACAGAATGAGAATA T_{15}
332	378	144	AT ATAAATTATTAAACAGAAAGAGATCATGTAGAAAGTGAGATAAAAAT TTT
331	371	1,547	ATATAA ATTAAACAAAAAGAAATCACGTAGAAGACAGAATAGAGATA T_{15}*
360	407	181	ATATAT ACATCCATAAAATTATCATCAGTTAATAGATTGTTAAATGAAAA TTTT
349	389	41	ATATAA ATCACCAACTAATAAGTTATTGAATGAGAGAAAGTTATATA T_{12}
352	401	705	ATATA ATAAAACTATCACTAACTAATGGATTGTTAAGTAGAAGAGAATCAT T_{11}
360	407	429	ATATAT ACATCCATAAAATTATCATCGGTTAATAGATTGTTAAATGAAAA T_{11}
354	401	160	ATATA ATAAAACTATCACTAACTAATGGATTGTTAAGTAGAAGAGAATC CT$_{07}$
362	407	143	ATATAT ACATCCATAAAATTATCATCGGTTAATAGATTGTTAAATGAA T_{12}
349	389	104	ATATAA ATCACCAACTAATAAGTTATTGAATGAGAGAAAGTTATATA T_{09}
361	407	74	ATATAT ACATCCATAAAATTATCATCGGTTAATAGATTGTTAAATGAAA TTTTTTTTTTTTTTT
387	435	1,428	ATATAT AACACAACAAGAAACGAATGAGAGAAGTATCTATGAGATTATT T_{14}*
387	435	1,049	ATATAT AACACAACAAGAGACGAATAGAAAAGATATCTGTGAAATTATT T_{13}
387	437	934	ATATAT AAAACACAATAGAAAACGGATAAGAGAGATATTCATAGAGTTATT T_{13}
387	435	664	ATATAT AACACAACAAGAGACGAATAGAAAAGATATCTGTGAAATTATT T_{12}
390	435	635	ATATAT AACACAACAAGAGACGAATAGAAAAGATATCTGTGAAATT T_{11}
398	435	28	ATATAT AACACAACAAGAGACGAATAGAAAAGATATCTGTGA T_{05}
397	435	23	ATATAT AACACAACAAGAGACGAATAGAAAAGATATCTGTGAA T_{13}
424	464	25,624	ATAT ATGACACAACGAGGGAAGATACTCTAAAGGACACAGTGAAA T_{12}
427	467	2,307	ATAT ATAACGACACAATAGAGAAAGATGCTCTGAGAGATGTAATA T_{13}
421	460	1,864	ATAAAT TACAACAAAGAAAGATACTCTAGAAAGCACAGTGAGAAAT T_{16}
424	457	368	AAATTAACGACA AACAAAGAGAAATACTCTGAGAAATATGATGAAA T_{12}
424	464	6,879	ATAT ATGACACAACGAGGGAAGATACTCTAAAAGGTACAGCGAAA T_{13}*
427	468	1,781	ATAT AACAACGATACGACAGAGAAAGATATTCTAAGAGATATGACA T_{13}*
424	457	232	AAATTAACGACA AACAAAGAGAAATACTCTGAGAAATATGATGAAA T_{12}
5'	3'	Reads	ATPase 6 gRNA Sequences cont.
455	491	368	ATATATAATTAC AAACAAACGCAGAGATGTCGGTAAATAATGATATAAT T_{11}
455	497	22	ATAT ATTACAAAACAGACGTAAAGATGTCGATGAATGGTGGTATAAT T_{14}
452	477	54	AATT ACGTCGATAGATAACGATACAATGAG ATTAATTTT
476	500	14	AAATT TAAATTACAAGACAAACGTAGAAGC T_{24}
458	500	1	AAATT TAAATTACAAGACAAACGTAGAAGCGTCGATAGATAATGATAT$_{15}$
487	526	8,723	ATACAA ATCAACAATAGAAGATGGGATGATAATAGATTGTGAGATA T_{17}
487	528	1	ATAC ACATCAACAATAGAAGATGGGATGATAATAGATTGTGAGATA T_{27}
487	526	635	ATACAA ATCAACAATAGAAGATGGGATGATAATAGATTGTGAGATA T_{16}
521	567	232	AA AAAAAAAAAAAACAAAAATAGAATAAAGAAAGTCAGAGAATGTTAAT T_{05}
520	553	85	AAAAAAAAAAAC AAAATAAAGTAAGAAGAATCAGAGAGTGTCAATA TATTTT
521	553	11	AAAAAAAAAAC AAAATAAAGTAAGAAGAATCAGAGAGTGTCAAT T_{12}

5'	3'	Reads	ATPase 6 gRNA Sequences cont.
557	593	69,619	AATAAATCGATAACAAAGAACACTGTAAAAGAGAGAA TGAGAGTAAATAT$_{09}$
549	593	2,587	AC AATAAATCAATAACAGAGAATATCATAGAGAGGAAAGATAGAAAT T$_{12}$*
549	592	181	ATAT ATAAATCAATGACAAGAAGCACTGTAGAAAAAGAGAGTGAAAAT T$_{13}$
546	592	313	AT ATAAATCAATAACAGAAGATGCCATAGAGAGAGAAAGTGAGAGTAAA T$_{11}$
546	593	76	AT AATAAATCAATAACAAAGAACATTGTAAAAGAGAAAAGTGAGAATAAA T$_{13}$
549	592	30	ATAT ATAAATCAATGACAAGAAGCACTGTAGAAAAAGAGAGTGAAAAT T$_{13}$
568	611	670	ATACT AAACACAAAAATGAATAAAATAAGTCAGTGATAGAAGATATTAT T$_{12}$
576	616	115	AT AAACAAAACACAAAAATAAGTAAAGTAGGTCAGTGATAAGA TATACAT$_{07}$*
589	629	854	AT AAATAATAAACAGAAACGGAATACGAGAATAAGTAAAGTGA TTTAAT$_{13}$
589	629	3,292	AT AAATAATAAACAGAAACGGAATACGAGAATAAGTAAAGTGA TTTAAT$_{10}$
613	654	618	ATATAT AATCCAACAGATATAAGAGCATGTAAAATAGTAAGTGAAAAT T$_{12}$
613	657	183	AT ATAAATCCAACAAGTATAAGAACATATAGAATAGTAGGTGAAAAT T$_{12}$
613	654	459	ATAT AATCCAACAGATATAAGAGCATGTAAAATAGTAAGTGAAAAT T$_{11}$
613	657	399	ATAT ATAAATCCAACAAGTATGAAGACACGTAAAATAGTAAATGAAAAT T$_{14}$*
640	689	39,063	ATAT ATAAATAACTGTAGTATGGTGGTAGATGAGTTTGATAGATATA T$_{12}$
647	689	678	ATAT ATAAATAACTGTAGTATGGTGGTAGATGAGTTTGAT T$_{11}$
640	689	131	ATAT ATAAATAACTGTAGTATGGCGGTAGATGAGTTTGATAGATATA T$_{11}$
654	689	234	ATAT ATAAATAACTGTAGTATGGTGGTAGATGA TTTTGATAGATATAT$_{12}$
643	667	4,401	ACATATATAAATAACTGTGATATTGCGGTAGATGGATCTGATGAAT T$_{14}$†
640	668	250	ATATAATAAATAACTATAATAAGGTGGTAAGTGAGTTCAGTGAATATA T$_{14}$†
640	662	165	AAAACTGTAATATGGAGGTAAGTGAATTTGATAGATGTA TTAATTTT†
680	714	7,581	AAAATA CAACTGCAAGATCGTGTTATAGAGGATAAGTGATT TAAT$_{13}$
680	719	1,291	ATATAA ATTATCAACTGTGAGATTATATTACAAGGAATAAGTGATT T$_{13}$
672	716	119	ACACA ATCAACTGCAGAATTATATTACAGAGAGTGAGTAATTGTAA AAT$_{12}$
680	714	105	AAATA CAACTGCAAGATCGTGTTGTAGAGGATAAGTGATT TAAT$_{11}$
671	714	319	AGATA CAACTGCAAGATCATATTATAAGAGGTGAATGATTGTAAT $_{14}$T*
685	714	309	ATATA TAACTGCAGAATCATATTATAAGGGATGAA CGATTGT$_{13}$
698	728	740	ATT AAAATCCATTATCGATTGTAGAGTTATGT GATAGAGAATAAT$_{11}$
686	728	12	ATATATT AAAATCCATTATCGATTGTAGAGTTATGTTATAGAGAATAA TAT$_{21}$
699	727	88	AAATCCATTATCAGTTGCGAGATTGTA GTATAAAGAATAAT$_{14}$
699	727	24	ATATAT AAATCCATTATTAACTGTAAGATTGTA GTATAGT$_{17}$
720	767	4,588	AT AAATCAAATACAGAACTGAATAGACGATAAAGATAGTGAGAAATTT T$_{11}$
715	755	2,272	ATATATAT AAACTAAACAAATAGCAGAGACAGTGAGAGATTCGTTAT AAT$_{13}$
728	765	920	AT ATCAAATACAAAACTGAGCAGATGACAGAGATAGTAAA TGATTTAT$_{12}$
720	767	165	AT AAATCAAATACAGAACTAGATGAACAATAGAGATAGTGAGAAATTT T$_{12}$
717	763	1,613	ATATA TAAATACAAAACTAGATGAATGACAGAAACGATGAGAGATTTATT T$_{14}$*
718	763	488	ATA TAAATACAAAACTAGATGAATGACAGAAACGATGAGAGATTTAT AAT$_{17}$
720	763	452	ATATA TAAATACAAAACTAGATGAATGACAGAAACGATGAGAGATTT T$_{12}$
720	767	269	AT AAATCAAATACAGAACTAGATAGACGATAGAGATAGTGAGAAATTC TTTT*
720	767	177	AT AAATCAAATACAGAACTAGATAGACGATAGAGATAGTGAGAAATTT T$_{17}$
747	789	13	ATAAAT ACAACAATATAATAACTGTCGAAGGTTGAATATGAGATTAAAT T$_{11}$
747	789	587	ATATAT ACAACAATATAATAGCTATCAGAGGTTGAATGTGAGATTAAAT T$_{11}$
745	789	91	ATATAT ACAACAATATAATAGCTATCAGAGGTTGAATGTGAGATTAAATGA T$_{11}$
790	822	8,663	ATA CTATAACTCCAATGACGAAATCAGTTTTA CAGTGATATGATAATT T$_{12}$
770	822	1	GGA CTATAACTCCGATAACGAATCAGATTTTGACAGTGATATGATAATTATT*
773	822	428	ATATA CTATAACTCCGATAACGAATCAGATTTTGACAGTGATATGATAATT T$_{09}$
774	822	223	ATATA CTATAACTCCGATAACGAATCAGATTTTGACAGTGATATGATAAT
777	822	195	ATATA CTATAACTCCGATAACGAATCAGATTTTGACAGTGATATGAT T$_{14}$

B) Cytochrome Oxidase III

5'	3'	Reads	COIII gRNA Sequences
35	70	1,179	ATAATT AATATACAACGAGATAGAGACGTAAAAGAAT TGATGTAT$_{12}$
36	73	826	AT ATAAATATACAACGAGATGAAGGCATAGAGAAA AGATGGTATATAAT$_{14}$
29	70	112	ATATAC AATATACAACGGAATGAGAATATAAGAAAGTGATGATA TTAT$_{11}$
36	70	14,200	ATATAT AATATACAACGAGATAAGAACATAGAGAAA AGATGGTATATAT$_{13}$
54	101	1,386	ATAT AAAACAAAAACATCACTGATATTGACGGATATATGATGA TAAAT$_{12}$
51	99	721	ATATAT AACAAAAACACTACTAGCGTTGACAGATATATGATGAAAT T$_{12}$
51	99	364	ATAT AACAAAAACACTACTAGCATTGACAAATATATGATGAAAT T$_{13}$*
52	101	229	AT AAAACAAAAACACTGCTAATATCGACGAATATATGATGGAA AAT$_{14}$*
50	92	122	ATATAAAAT ACACCACTGATATCAACGAGTATATGATGAGATA T$_{14}$
49	95	78	ATATAT AAAACACCACTGACATCGATAAGTATATAGTGAAGTGA TTAAT$_{15}$
81	112	834	GTA GAGTGAAGATAGAGAAATAAAGATATCGTT T$_{13}$
81	116	550	ATATATAATAACAATA GCAGGTAAAGGTGAGAAAGTGAAGATATCATT T$_{10}$
81	131	1	TACATAATAACAGTGGCGGGTAGAGATAGAAGAATAAAGATACTATT T$_{08}$
88	115	443	AACGATGGA TAGGTAGAGATAGAGAAATGAAGATATT TTAT$_{05}$*
88	115	117	AATAACAATGGA TAGGTAGAGATAAAGAAATGAAGATATC T$_{06}$
88	115	96	AATAACGATGGA TAGGTAGAGATAGAGAAATGAAGATATC T$_{07}$
81	112	64	ATACATAACAGTGGCAGA GTGAAGATAGAGAAATAAAGATATCATT T$_{11}$*
108	132	6	AT ATATACAATAACAGTGGTAGGTAGA T$_{09}$
117	156	104	ATATATA TCCAACAAACAGAGTAACCGATACATAGTGATAGTG ATAT$_{13}$
117	155	465	ACATATA CCAACAAACAGAATAACTAGTGCACAGTGATGATG ATAGT$_{16}$
118	150	36	ATATA CAACAAACAAAATAATCGATGCACAGTGATAGT AGTAGT$_{13}$
141	185	126,513	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGATGA TTTAAT$_{13}$
141	185	761	ATAT ATTACCAAACAATAGATGAGTAGATTCTAATAGATGA TTTAATTAAGTTTT*
134	188	542	ATAT AAAACTACCAAACAGTAAATAGATAAGTTCTAATAAGTGAGATAATT T$_{11}$
146	185	350	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATA TATGATTTAAT$_{12}$
142	185	199	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGATA TTTAATTAT$_{05}$
141	185	183	ATAT ATTACCAAACAATAGACAAGTAGATTCTAATAGATGA TTTAAT$_{13}$
145	185	172	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAG CTGATTTAAT$_{13}$
143	185	116	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGAT CATTTAATTAT$_{05}$
131	188	109	ATAT AAAACTACCAAACAGTAAATAGATAAGTTCTAATAAGTGAGATAATTAAT TGTTAT$_{16}$
147	185	106	ATAT ATTACCAAACAATAGACGAGTAGATTCTAAT CGATGATTTAATTAAT$_{17}$
141	185	99	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAAATGA TTTAATTAAT$_{08}$
141	185	62,901	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGATGA TTTAAT$_{14}$
141	187	810	ATAT AAACTACCAAATAATGAACAGATAAATTTCAGTGAGTGA TTTAAT$_{13}$*
143	185	685	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGAT T$_{13}$
141	185	497	AT ACTACCAAACGATAAGCAGATAAGTCTCAGTGAATGA TGTAACT$_{15}$
134	188	332	ATAT AAAACTACCAAACAGTAAATAGATAAGTTCTAATAAGTGAGATAATT T$_{10}$
141	188	308	ATAT AAAACTACCAAACGATAGACGAATAAGTTCTGATAAGTGA TATAT$_{12}$
142	185	140	ATAT ATTACCAAACAATAGACGAGTAGATTCTAATAGATG T$_{15}$
163	203	2,487	AAA ATAATCAACAAATAGAGAACTGCTAGATGATAGGTGA TATAGAT$_{13}$
163	211	861	ATATT AACCACAATCAATAAGTAAGAGACTACTAGATGATAGATAA T$_{14}$
168	195	232	ATATATAAAACCACAATCAT CAGATAAGAGACTATTAAGTGATA T$_{12}$*†
185	216	138	ATATAT AATAAAACCACAATTAGCAAGTAAGAAG GTATCAGATGATAATTAT$_{06}$†
204	247	5,008	ATATAT ACAAACAAATACAGAGATCGACGAGAAAGAAAGTGAGATT TAT$_{12}$
195	247	688	ATAAATAAATACAAAAATCAGCAAGAAAGAGAGTAAGATTGTGATTAAT T$_{08}$
199	243	462	ATAT ACAAATACAAAAATCGATAGAAAAGAAAGTGAGATCATGATT TAT$_{12}$
195	247	457	ATAAATAAATACAAAAATCGACAGAGAGAAAAGTAGGATTGTGATTAAT T$_{12}$
195	244	83,016	ATAT AACAAATACAAGAGCCGATGAAGAAAAAGGTAGAACTGTGATTAAT T$_{12}$
199	243	9,810	ATATAT ACAAATACAGAAACTGACGAAAGAGAGAATGAAGTTATGAT CT$_{17}$
202	247	2,126	ATATAT ACAAACAAATATGAGAACTAACAAGAGAGAAAGTGAGATTAT T$_{12}$*
201	247	1,676	ATATAT ACAAACAAATATGAGAACTAACAAGAGAGAAAGTGAGATTATA T$_{17}$
204	247	1,379	ATATAT ACAAACAAATATGAGAACTAACAAGAGAGAAAGTGAGATT T$_{13}$
199	244	363	ATATAT AACAAATACAGAAGCCAACGAGAGAAGGAATAAGATTGTAAT T$_{10}$
202	243	269	ATATAT ACAAATACAGAAACTGACGAAAGAGAGAATGAAGTTAT T$_{14}$
204	244	149	ATAT AACAAATACAAGAGCCGATGAAGAAAAAGGTAGAACT TTGATTAAT$_{12}$
195	243	130	ATAT ATAAATACAAAAACTAACGAAAGAAAAGATGGAACTGTGGTTAAT T$_{12}$
204	243	126	ATAT ACAAATACAGAAACTGACGAAAGAGAGAATGAAGTT T$_{19}$
202	244	102	ATAT AACAAATACAAGAGCCGATGAAGAAAAAGGTAGAACTGT TATTAT$_{14}$

5'	3'	Reads	COIII gRNA Sequences cont.
229	274	289	ATA TACAAAACAAATCTAACAGTGATAGTAACAGATAGATATAGAGATT T_{10}
236	279	575	ATAT AAAATCACAGAACAGATCTGATAGTAACAGTAATAAGTAAATAT T_{11}
229	270	559	ATATAT AAACAAATCTAATGATAACGATGACGGATAGATATAGAGATT TAT_{16}
238	268	120	ATATAAACCACAAT ACAGATCTGACAGTAATGATGATAGGTAAAT T_{05}
238	279	37	ATAT AAAATCACAGAACAGATCTGATAGTAACAGTAATAAGTAAAT $TTCT_{14}$
234	264	33	ACAAAACAT ATCTAGCAGTAACAGTGACGAATAGATACAA T_{07}
234	279	29	ATAT AAAATCACAGAACAGATCTGATAGTAACAGTAATAAGTAAATATAA TTTT
258	299	18,740	ATATAT AAATCAAATAAACTATGTAGAAAGTTACGAGATAGATTTAATA T_{10}
265	308	4,452	AAA AAAACACAAAAATCAAGTGAACTATGTAGAGGATTGTAAGATAA T_{11}
265	310	4,418	ATAAAACACAAAAATCAAGTGAACTATGTAGAGGATTGTAAGATAA T_{13}
258	300	814	ATATAT AAAATCAAATAAATTACGTAGAGAGTTACAGAATAAGTTTAAT T_{10}
267	306	435	ATATAT AACACAAAAATCAGATAGACTATGTAGAAGATTGTGAAAT T_{11}
261	299	181	ATATAT AAATCAAATAAACTATGTAGAAAGTTACGAGATAGATTT T_{08}
258	300	347	ATATAT AAAATCAAATAGATCACGTGAAGAGTTATAGAATAGATTTAAT T_{14}
261	307	328	ATAC AAACACAGAAATCAGATAGATCACGTAGAGAGTTATAAGATAAATTT T_{08}
262	300	244	ATATATATT AAAATCAGATAAGCCACGTAGAAGATTGTAAAGTGAATT AT_{12}
261	300	195	ATATATATT AAAATCAGATAAGCCACGTAGAAGATTGTAAAGTGAATTT T_{09}
258	289	182	AATCACGTGAAAGATCGTAGAATGAGTTTAAT T_{13}
263	307	71	ATAC AAACACAGAAATCAGATAGATCACGTAGAGAGTTATAAGATAAAT AT_{08}
257	300	39	ATATAT AAAATCAAATAGATCACGTGAAGAGTTATAGAATAGATTTAATA T_{12}
293	320	1,024	AATACTGT ATATGATGTAGTAAGATATAGAGATTAA T_{10}
284	321	21	ATATAAA GATACAACGTAATAAGGCATAGAAGTTAAGTGAATTAT TGT_{12}
293	335	1	AAAAAACAATACTGGATATGATGTAGTAAGATATAGAGATTAA $TAACT_{06}$
291	332	1,923	ATATAT AAACAATACTGGGTACGATGTAATAGAATGTGAAAGTTAAAT T_{14}
293	321	151	AATACTGA GATACGACGTGATAAGATATAGAAGTTAA T_{11}
323	365	33,179	ATAT ATAAAACAAACTCGCTATGTAAGAACTGTAAAAAGTGATATT AT_{12}
330	365	856	ATAT ATAAAACAAACTCGCTATGTAAGAACTGTAAAAAA GTGATATT T_{09}
323	365	301	ATAT ATAAAACAAACTCGCTATGTAAGAACTGTAAAAAGCGATATT AT_{14}
323	365	4,275	ATATAT ATAAAACAAACTCACTGTGTAAAGATTGTAGAAAGTGATATT AT_{24}
323	360	139	ATACAT ATAAACTCACTGCATAAGAATCATAGAGAGTGATATT AT_{11}*
345	391	522	ATATAT ATAATACAACAAGGAGCGTCATAAGTAAAGTGAATTCGTTATAT T_{12}
345	391	203	ATATT ATAATACAACAGAAAATGTCATAAGTGAGATGAATTCGTTATAT T_{08}
345	389	490	ATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAATTCGTTATAT TTT
349	389	461	ATATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAATTCGTT T_{06}
347	389	365	ATATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAATTCGTTAT T_{11}
343	389	262	ATATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAATTCGTTATATAA T_{15}
357	391	106	ATATT ATAATACAACAGAAAATGTCATAAGTGAGATGA TTCGTTAT$_{06}$
353	389	79	ATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAATTT TTT
352	388	48	ACAAAT ATACAACAAAAGATGCCGTAGATAAGATAGATTTG GTATATTTT
354	390	41	ATATAT TAATACAACAGAAGACGCTATAAGTGAGATAGATT GATTATAT$_{27}$
362	406	101	ATATAT ATAAACATAAATCAGATAGTACAATGAAGAGTGTTATAGATAA T_{09}
362	406	139	ATAT ATAAACATAAATCAGATAATACAGTGAAGAGTGTCATAGATAA T_{06}
376	418	9,942	ACA TATAACACAAAAATAGACATAGACTGAATGATGCAGTGAAA T_{13}
378	422	1,634	ATAT AACTTACAACACAGAGATAGACATAGATCAGATAATGTGATAA T_{13}
384	418	185	ACA TATAACACAAAAATAGACATAGACTGAATGATGCA TTGAAAT$_{11}$
376	418	169	ACA TATAACACAAAAATAGACATAGACTGAATGATGCAGTAAAA TTTTAACT$_{06}$*
378	422	375	ATAT AACTTACAACACAGAAATAAGCATAGATCAGATAGTGTGATAA TTAAT$_{17}$
397	426	15	AAAACGAT AGCAAATTCATGACGTGAAAATAGATGTAA T_{11}
397	436	10	AAAA AAAAAACGAAAGCAGATTCACGGTACAGAGATAGATATAG T_{10}
411	449	20,772	ATATATA ATATAAGGTAAATGAGAGACGAGGGTAGACTTGTGATAC TAT_{12}
409	438	1,420	ATAATAAGGTAT ACAGAGAACGGAAGCAGACTTATGATATAA T_{12}
413	449	224	ATATATA ATATAAGGTAAATGAGAGACGAGGGTAGACTTGTGAT T_{10}
410	449	125	ATATA ATATAAGGTAAATGAGAGACGAGGGTAGACTTGTGATATA T_{09}
413	452	545	ATATAT AACATATAAGGTAAATAGAAGATGGAAGCGAATTTGTGAC T_{16}
418	453	60	ATATATATAAACAAC AGACATATAAGGTAAGTAAGAGATGAAGGTAAATTT T_{09}
413	452	20	ATATAT AACATATAAGGTAAATAGAAGATGGAAGCGAATTTGTGAT TCT_{05}
418	467	451	ATAC TATAATAAACAACAAAATGTGTAAGGTAGATAAGAAGTGAAGGTAAATT ATATTTT
437	469	238	A TACATAATAAACAATGAGATATATAAGGTGAAT CGAAAGTGAAATAT$_{12}$
427	460	218	ATATTAT AACAACAAAACGTATAAGGTAAGTGAAAAATGGA TGTAAAT$_{12}$
443	474	37	A ATAATCACATAATAAATGATAGAACGTATAAG ATAGATGAAAAT$_{10}$
5'	3'	Reads	COIII gRNA Sequences cont.

461	497	66,677	ATACAT AATACCAATAGAAGACAGAATTGTAGTCATGTGATA TTCAT$_{14}$
461	497	5,941	ATACAT AATACCAATAGAAGACAGAATCGTAGTCATGTGATA TTCAT$_{13}$
456	499	936	ATAT AAAATACCAATAAAGAACAGAATTATAGTTGTATGATAGATAA AT$_{12}$
462	497	371	ATACAT AATACCAATAGAAGACAGAATTGTAGTCATGTGAT T$_{11}$
461	497	139	ATACAT AATACCAATAGAAGACAAAATTGTAGTCATGTGATA TTCATAT$_{11}$
457	499	6,621	ATATT AAAATACCAATAGAAAATGAGACTGTGATTATATGATGAATA T$_{14}$*
458	499	3,602	ATATT AAAATACCAATAGAAAATGAGACTGTGATTATATGATGAAT T$_{14}$
454	499	295	ATAT AAAATACCAATAAAGAACAGAATTATAGTTACATGATAGATAATA T$_{10}$
462	499	185	ATATT AAAATACCAATAGAAAATGAGACTGTGATTATATGAT T$_{15}$
460	501	185	AAA AAAAAATACCAGTAGAAGATAAGACCATAATCATGTGATAA TTTT
461	499	107	ATATT AAAATACCAATAGAAAATGAGACTGTGATTATATGATG T$_{14}$
483	522	288	AAATA ATCAACAAATTAAATGAATCTAAAAGGTATCAGTGAAAA T$_{14}$
491	539	145,031	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAATATT T$_{10}$
488	539	4,441	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAATATTAAT TTCT$_{08}$
490	539	1,630	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAATATTA T$_{15}$
498	539	1,118	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAA TATTAATAG TTTT
495	539	1,052	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAAT T$_{14}$*
491	539	849	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAGTTTAGAGAATATT T$_{10}$
487	539	847	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAATATTAATA T$_{14}$
486	539	653	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAGAATATTAATAG TTTT
504	535	599	ATAT ATAAAATGTATTTGTCGACGAGTTAAATGGAT GTAGAAGAT$_{12}$
491	539	482	ATA TAAATAAAATGTATTTGTCAATGGATTAGATAAATTTAGAGAATATT T$_{11}$
502	539	448	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTT TAGAATAT$_{12}$
491	539	353	ATA TAAATAAAATGTATTTGTCAATGGATTAGATGAATTTAGAAAATATT T$_{09}$
504	535	1,604	ATAT ATAAAATGTATTTGTCGACGAGTTAAATGGAT GTAGAAGAT$_{15}$
528	565	188	ATATAT AAAATTAACAAGTGAATCACTAACAGATAGATAGAATG ATAT$_{12}$*
528	564	181	ATATT AAATTAACAGATAAGCCACTGACAAATAGATAGAGTG ATAT$_{12}$
524	564	77	ATATAT AAATTAACAAATAGACTACTAATAAGTGAGTAAGATGTATT AATTTATATATTTT*
525	563	3,592	ATATAT AATTAACAAGTAGATCACTGACAAATAGATGAGATGTAT AAT$_{13}$*
528	565	640	ATATAT AAAATTAACAGATAGATCATTAACGAGTAGATAAAGTG ATAT$_{11}$
523	563	336	ATATAT AATTAACAAGTAGATCACTGACAAATAGATGAGATGTATTT T$_{08}$
526	564	420	ATATT AAATTAACAAATAAACTATTAATGGATGAGTGAGATGTA ATTAT$_{15}$
527	563	144	ATATAT AATTAACAAGTAGATCACTGACAAATAGATGAGATGT T$_{16}$*
548	592	99,540	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAGTAAATT T$_{12}$
555	594	27,720	ACAT AAAAACCTAAACTGAGAATACGAGACAAAGAAATTAGTGA TTAAAT$_{12}$
547	592	1,442	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAGTAAATTA AT$_{11}$
551	592	599	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAGTAA T$_{13}$
554	592	504	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAG AAAAT$_{13}$
558	592	309	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAG GGAGTAAAT$_{14}$
548	592	238	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGCGAGTAAATT T$_{09}$
550	592	176	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAGTAAA AT$_{11}$
545	592	150	ATATAT AAACCTAAATCAAGAACATAGAACAGAGAGATTAGTGAGTAAATTATT T$_{10}$
558	594	147	ACAT AAAAACCTAAACTGAGAATACGAGACAAAGAAATTAG GGATTAAAT$_{11}$
558	593	128	ACATT AAAACCTAAACTGAGAATACAAGACAGAGAGATTAA GGATTAAAT$_{13}$
548	592	114	ATATAT AAACCTAAATCAAGAACATAAAACAGAGAGATTAGTGAGTAAATT T$_{10}$
551	593	84,835	ACAAT AAAACCTAAACCGAGAACATAGAGCAGAGAAGTTAGTAGATAA T$_{13}$
555	593	6,897	ACATT AAAACCTAAACTGAGAATACAAGACAGAGAGATTAATGA TTAAAT$_{14}$
552	593	4,207	ACAAT AAAACCTAAACCGAGAACATAGAGCAGAGAAGTTAGTAGATA T$_{13}$
556	593	1,803	ATATAT AAAACCTAAATCAGAGACGCAGAATAGAGAGATTGATA TAT$_{10}$*
553	593	1,137	ACAAT AAAACCTAAACCGAGAACATAGAGCAGAGAAGTTAGTAGAT T$_{12}$
551	596	722	A AAAAAAACCTAAACCGAGAACATAGAGCAGAGAAGTTAGTAGATAA T$_{15}$
557	593	486	ACATT AAAACCTAAACTGAGAATACAAGACAGAGAGATTAAT T$_{12}$
551	603	456	ACAAAAAAACCTAAACCGAGAACATAGAGCAGAGAAGTTAGTAGATAA T$_{05}$
555	593	206	ATAT AAAACCTAAACCAAAGATATGAGACAGAGAGATTAGTGA TATGT$_{13}$

5'	3'	Reads	COIII gRNA Sequences cont.
585	629	85,916	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGAGATTTAA T$_{12}$

5'	3'	Reads	COIII gRNA Sequences cont.
587	629	9,106	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGAGATTT T_{11}
586	629	1,991	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGAGATTTA T_{12}
585	631	1,447	ATAT ATAAACTCAATCATAGTATAAGATGACGACAATGAGAAGATTTAA T_{13}
585	629	276	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAAAGATTTAA T_{12}
592	629	269	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGA TTTAAT$_{14}$
591	629	263	ATATA GAACTCAATCATAATATGAAGCAATAACAATAAGAGA TTTAAT$_{07}$
588	629	221	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGAGATT AAT$_{13}$
594	629	150	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAA TAGATTTAA T_{06}
580	622	145	ATATAT ATCATAATACAAGGCAATGACGACGAGAAGATTTAGATTAA T_{11}
590	629	131	ATATA GAACTCAATCATAATATGAAGCAATAACAATGAAGAGA ATTAATTAT$_{06}$
603	634	17,313	AT ATAACAAACTTAATCGTAATATGAAACAACGA GAATGAGAAAAT$_{13}$
585	628	8,519	ATATAT AACTCAATCATAATACGAGATGATAATGACGAAGAGATTTAA T_{13}
580	624	3,564	ATATA TAATCATAATACAGAACGATGGCAGTGAAGAGATTTAGATTAA T_{12}
587	630	2,850	ATATATAT CAAACTCAATTGTAGTACGAGACAATAATGATGAGAAGATTT T_{10}
587	628	628	ATATAT AACTCAATCATAATACGAGATGATAATGACGAAGAGATTT T_{08}
585	630	582	ATATATAT CAAACTCAATTGTAGTACGAGACAATAATGATGAGAAGATTTAA T_{11}
586	628	231	ATATAT AACTCAATCATAATACGAGATGATAATGACGAAGAGATTTA T_{11}
587	631	135	ATATAT ACAAACTCAGTCATAGTATAAGACAGTGATAATGAGAGAATTT T_{12}
591	630	112	ATATATAT CAAACTCAATTGTAGTACGAGACAATAATGATGAGAAG T_{13}
604	647	10,315	ATACAAAAAC AAAAACCAAACGATGAACTTGATTGTAGTATAAGATAATA T_{13}
605	647	7,606	ATACAAAAAC AAAAACCAAACGATGAACTTGATTGTAGTATAAGATAAT T_{13}
604	643	250	ATACAAAAACAAAAT ACCGAGCGACAGATTTGATTGTAGTATAAGATAATA T_{11}
607	647	637	ATACAAAAAC AAAAACCAAACGATGAACTTGATTGTAGTATAAGATA T_{21}
604	643	173	ATACAAAAACAAAAT ACCGAGCGACAAGTTTGATTATAGTATAAGATAATA T_{13}*
605	643	150	ATACAAAAACAAAAT ACCGAGCGACAAGTTTGATTATAGTATAAGATAAT T_{15}*
635	669	482	ATATACAATGCAAACTTA CATGACTGGTTTTATAGAGATGAGAGATTAA T_{14}
635	669	268	ATATACAATGCAAACTTA CATGACTGGTTTTATAGAGATGAGAGATTAA T_{15}
635	676	60	ATATACAATGT ACTCTCATAATTGGTTTCATAGAGATAGAAGATTAA T_{15}
637	669	42	ATATACAATGCAAACTTA CATGACTGGTTTTATAGAGATGAGAGATT T_{08}*
633	679	25	ATATA CAAACTCTTATAACTGGTTTTACGAGAATGAGAAATTAAAT T_{15}
659	691	649	ATAT TAAATAACAATGCGAATTTTCATAGTTGGTT CATAGATACAAT$_{12}$
653	682	39	AT ATGTGAACTCTTATAACTGGTTTTGTAG TGATGATAGAT$_{15}$
669	717	758	ATAAT AGAACACCACAGCTTAATGTAGTAGATGGCAGTGTAAATTTTT T_{10}
669	722	374	ATAT AATCAAAAACACCGTAACTTGATGTAGTAGATAGTAGTGTAAATTTTT T_{07}
669	706	354	ATATAGAACCAAAAACAG TGCAATTTAGTGTGATAGATGATAGTGTAAATTTTT T_{07}*
669	715	160	ATATAGAACCAAT AACATCGCGACTTAGTGTGATAAGTAATAGTGTAAATTTTT T_{07}
684	722	1,329	ATATAT AACCAAAAACACTGCGATTTGATGTAATAAGTGAC TGTGTAAAT$_{11}$
689	726	1,181	ATAT ATAGAACCAAGAACACCATAGTTTGATGTGATAG TGATAGTGTAAAT$_{14}$*
669	715	911	ATATAGAACCAT AACACCATGATTTGATGTAGTAAATGATGATGTAAATTTTT T_{09}
669	722	738	ATAT AACCAAAAACACTGTAACTTGATGTAGTAGATAGTAGTGTAAATTTTT T_{08}*
695	730	129	ATAT TAAAATAGAACTAAAGACACTGTAACTTAGTG AGTAAATGATATTAAT$_{10}$*
675	715	107	ATATAGAACCAT AACACCATGATTTGATGTAGTAAATGATGATGTAAAT AT$_{08}$
677	715	102	ATATAGAACCAT AACACCATGATTTGATGTAGTAAATGATGATGTAA T_{11}
675	717	90	ATAAATT AAAACACCATAATTTGATGTGATAAGTAATGATGTAAAT AT$_{17}$

5'	3'	Reads	COIII gRNA Sequences cont.
706	753	31,331	ATATAT AAATGTAATAGATCTGATGAAAGTGAGGTAGAATTGAGAATATT T_{10}
707	753	8,037	ATATAT AAATGTAATAGATCTGATGAAAGTGAGGTAGAATTGAGAATAT AT$_{14}$*

706	753	6,744	ATATAT AAATGTAATAGATCCAATGAAGGTAAGATAGAACTGAGAATATT T_{09}
707	753	3,791	ATATAT AAATGTAATAGATCCAATGAAGGTAAGATAGAACTGAGAATAT AATTAT$_{07}$
701	748	2,726	ATA TAATAAATCCAATGAAGATAAAGTAGAGTCAGAGATATTATGAT AT_{13}
706	753	1,214	ATATAT AAATGTAATAGATTCAATGAAGGTAAGATAGAACTGAGAATATT T_{09}
707	753	1,124	ATATAT AAATGTAATAGATTCAATGAAGGTAAGATAGAACTGAGAATAT AAT_{14}
699	748	977	ATATA TAATAAATCCAATGAAGATAAAGTAGAGTCAGAGATATTATGATTT T_{10}
707	752	848	ATATAT AATGTAATAAATCTAATAGAGATAAGATAGAACTGAGGATAT AT_{12}
713	753	595	ATATAT AAATGTAATAGATCTGATGAAAGTGAGGTAGAATTGAGAAT T_{09}*
707	753	206	ATATAT AAATGTAATAGATCTGATAAAAGTGAGGTAGAATTGAGAATAT AT_{13}
715	753	182	ATATAT AAATGTAATAGATCTGATGAAAGTGAGGTAGAATTGAGA $TTTAT_{08}$
713	753	154	ATATAT AAATGTAATAGATCCAATGAAGGTAAGATAGAACTGAGAAT T_{14}
719	753	131	ATATAT AAATGTAATAGATCTGATGAAAGTGAGGTAGAATT TAGAATATAT$_{08}$
715	750	1,457	ATATAT TGTAATAAATCCGATAGAAGTAAGATAGAACTGAGA TAT_{15}
714	748	688	ATATA TAATAAATCCAATAAGAATAAGATGGAACTGAAGA GATTATGAT$_{11}$
719	746	304	AT ATAAATCCAATAAAAGTGAAATAGAATC TGAGTTTT
707	753	416	ATAT AAATGTAATAAATTCAGTAGAAGTAAGATAGAATTGAAGATAT ATATAGT$_{26}$
723	765	2,739	ATATAT AACATGCATAAGATGTAGTGAATCTAGTAAGAGTAAGATAG T_{14}
728	767	1,193	ATATAT AAAACATGCATAGAGTGTAGTAAGTTCAGTGAAAGTGA TATAGT$_{09}$
724	765	2,604	ATATAT AACATGCATAAGATGTAGTAGATTCAGTGAAGATAAGATA T_{14}
725	765	456	ATATAT AACATGCATAAGATGTAGTAGATTCAGTGAAGATAAGAT T_{10}
728	767	284	ATATAT AAAACATGCATAGAGTGTAGTAAGTTCAGTGAAAGTGA TATAGTTTT
736	781	495	ATAT ACAAAACACCTAAGAAGATGTGCGTAGAATGTGATAGATTTAAT T_{16}
737	781	159	ATAT ACAAAACACCTAAGAAGATGTGCGTAGAATGTGATAGATTTAA AT_{15}
739	781	40	ATAT ACAAAACACCTAAGAAGATGTGCGTAGAATGTGATAGATTT T_{23}
735	779	19	AAAACACCTAAGAAGATGTGCGTAGAATGTGATAGATTTAATA T_{06}
750	778	19	AAACACCTAGAGAAACGTGCATGAGAT TGTGATAGATTAAT$_{07}$
754	790	153	ATATAT TAAACAACAACAGAACGTCTAAGAGAGTATGTATA TGATGTAAT$_{15}$
754	790	89	ATATAT TAAACAACAGCAGAACATCTAAGAGAGTATGTATA TGATGTAAT$_{12}$
772	815	111,116	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAAGATACT AT_{14}
771	815	25,628	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAAGATACTT T_{10}
775	815	323	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAAGAT T_{14}
778	815	315	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAA T_{12}
772	815	281	ATAT AAATTAAACAAACGTATGAAGCAAGTAGATAGTGATAAGATACT AT_{11}
772	815	230	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGCGATAAGATACT AT_{15}
777	815	188	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAAG TTACTAT$_{05}$
771	814	138	ACATAT AATTAAACAAACGTATAGAGCAAGTAAGTAGCAGTGAAATATTT T_{11}
773	815	136	ATAT AAATTAAACAGACGTATGAAGCAAGTAGATAGTGATAAGATAC AAT_{12}
773	814	1,459	ATAT AATTAAACAAACGTATAAGACAAGTAGATGGCAGTGAAATAT AT_{11}
771	814	527	ATAT AATTAAACAAACGTATAAGACAAGTAGATGGCAGTGAAATATTT T_{13}
796	829	24	ATAACAAAACGTGA TATCCATATACAGAGAATTAGATAGATGTATAAA T_{06}
788	829	5	ATATA TATCCATACACAGAAAATTAGATGAACGTGTAAAATGAGTAA $TTAT_{13}$
788	829	92	ATATA TATCCATACACAGAAGATTAAATAGACGTGTAAAGTGAATAA T_{15}
802	842	116	ATAT AAAACAAAACGTGTATTCATATATGAGAGATTAAGTAAATG AT_{15}
798	842	386	AT AAAACAAAACGTGTATCTGTGTATAGGGAGTTAAATGGATGTATA TAT_{14}
799	842	156	ATAT AAAACAAAACGTGTATCTGTGTATAGGGAGTTAAATGGATGTAT T_{20}
814	854	66	ATAT ATACAATACAAAGAAGCGAAACGTGTATCTATATATGGAAA T_{06}
828	855	5	ATATAT AACACAATACAGAGAGATGAAACGTGTA GATGTAT$_{23}$
829	855	1	ATATAT AACACAATACAGAGAGATGAAACGTGT T_{22}
830	854	1	AT ACATAATACAAAGAGACAGAACGTG ATATCT$_{16}$
848	890	9,675	ATATAT AATCAAACTAAATTGACGAGATGTTGATGTAGATAGATATAAT AT_{08}
846	891	138	ATAT AAATCAAACTAAATTAGCAAGGTGTCAGTGTAAATGAGCATGATAT T_{13}
838	882	109	ATACAA TAAATCAACAAGATGTCGATATAGATGGATATGATATGAGAGAAT T_{15}
845	889	2,508	AT ATTAAACTAAATCAATAAAGTGTCGATGTAGATGAATGTGATATA TAT_{11}
846	889	202	AT ATTAAACTAAATCAATAAAGTGTCGATGTAGATGAATGTGATAT T_{11}
846	891	127	ATAT AAATCAAACTAAATTAGCAAGGTGTCAGTGTAAATGAGCATGATAT TAT_{13}

5'	3'	Reads	COIII gRNA Sequences cont.
880	918	1,822	ATATAA ATCAGAATAAACAGATCGCAATAGAGAGAATTAAGTTAA TAT_{14}
882	927	600	AA ATATAAAACATCAAGATAAATGGATTGTGATAGAGAAAGTTAAATT T_{11}
880	929	222	ATATATAAAACATCAGAATAGACAAATCGTAATAGAGAAAGTTAAGTTAA T_{14}

880	921	209	ATAT AATATCAAAATAAACAGATCGTAGTAAAAGAAGTTAGATTAA T_{12}
882	929	159	ATATATAAAACATCAGAATAGACAAATCGTAATAGAGAAAGTTAAGTT T_{13}*
881	929	28	ATATATAAAACATCAGAATAGACAAATCGTAATAGAGAAAGTTAAGTTA TTTT
907	947	787	ATATAA TATACACACAGATACATAATACGTAGAATGTTAAGATAAGT T_{16}
909	950	126	ATATA ATTACACACACAGATACGTGATATATAGAATGTTAAGGTAA $TATAAT_{10}$
913	946	558	ATAC ATACACACAAATATATAACATATAGAGCATTGAG TTAGATAAT $_{15}$
905	944	3,472	ATATAA ACACACAAATATATGGCATATAGAGCATTGAAGTAGATAA T_{13}*
905	944	1,321	ATATAA ACACACAAATATATGACATATAGAGCATTGAAGTAGATAA T_{15}
920	952	162	ATAG AAAATTACACACATGAATACATAGTACATAGAA GATTGATATAT $_{13}$
935	977	2,920	ATATA AATCAACAACTGAAAAGATATCAATGAGATTGTACATGTAAAT T_{15}
940	977	522	ATATA AATCAACAACTAAGAAGACACTGATAGAGTTATATGTG $ATTAAT_{14}$*
935	977	354	ATATA AATCAACAACTGAAAAGATATCAATGAGATTGTACATGTAAAT T_{08}
939	977	28	ATATA AATCAACAACTGAAAAGATATCAATGAGATTGTACATGT T_{09}
942	983	1,160	ATAT AACTAATCAACAGCTAAGAGAACGTCAATGAGATTATGTG $ATTAAT_{14}$
951	981	127	AT ACTAATCAACAACTAGAAGAATATCAGTGA $TATACATGTAAT_{10}$
951	981	25	AT ACTAATCGACAACTAGAGGGACATCAGTGA $TTTATACGTAT_{15}$
963	1003	111	ATA TACAAACTACCAATATAAGTTAACTGATCGGTAATTAAGG TTATAT $_{15}$
965	1003	68	ATATA TACAAACTACCGATATAAGTTAACTGATTGATAATTAA $TGTTCT_{14}$

C) C-Rich Region 3

5'	3'	Reads	CR3 gRNA Sequences
34	62	18	ATATGT ACAACAAAACCGAGCAATCAGATAT AGAGTGAAAT$_{09}$
34	62	3	ATATAT ACAACAAAACTGAACAATCAAATGT AGTGTGAT$_{09}$
36	64	1	ATCGCAAGGTCGT GGACAACAAAACTGAACAATCAAAT T$_{11}$
34	62	1	ATATAT ACGACAAAACTGAACAATCAAATGT AGTGTGT$_{08}$*
41	88	140,541	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGAT AT$_{14}$
40	88	34,162	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGATT T$_{12}$
48	87	3,434	ATATAATT AAATGTACAGACAAATGATAGAGAGACGATGAGATTAAGT TATAT$_{12}$
51	86	3,016	AAAAATT AATGTACAAATAAACGATAGAGAGACAGTGAGATTA TGAT$_{13}$
51	88	2,553	ATAT AAAATGTACAGACGAGCAGTGAAGAGACAGTGAGATTA TACAT$_{11}$
47	88	1,270	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATG T$_{11}$
48	88	902	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGAT T$_{14}$
41	88	693	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTAAAATTAGATGAT AT$_{12}$
52	88	648	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATT T$_{09}$
51	88	632	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTA TATGATAATATTTT
46	88	468	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGA AATAT$_{14}$
40	77	213	ATATAT AAAATGTACATACGAACGATAAAAGGGCAGTGAAATTAGATAATT TT
41	88	212	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGCGAAATTAGATGAT ATCT$_{07}$
50	88	182	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAG TTGTTAATAT$_{11}$
41	88	181	ATATAT AAAATGTACAAACGGACAATGAGAAAACAGTGAAATTAGATGAT ATCT$_{16}$
49	88	177	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGA AGATATAT$_{11}$
55	88	170	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAA T$_{19}$
41	88	155	ATATAT AAAATGTACAAACGAACAATGAGAGAACAGTGAAATTAGATGAT AT$_{13}$
40	88	151	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTAAAATTAGATGATT T$_{12}$
41	88	135	ATATAT AAAATGTACAAATGGACAATGAGAGAACAGTGAAATTAGATGAT AATAT$_{07}$
56	88	133	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGA T$_{12}$
41	88	106	ATATAT AAAATGTACAAACAGACAATGAGAGAACAGTGAAATTAGATGAT AATAT$_{08}$
40	89	598	ATAT AGAAATGTACAAACGAGCAATAAGGGAACAGTGAAATTAGATGATT TCT$_{07}$
41	89	526	ATAT AGAAATGTACAAACGAGCAATAAGGGAACAGTGAAATTAGATGAT AAT$_{10}$*
40	77	437	ATATATAAAATGTACAT ACGAACGATAAAAGGGCAGTGAAATTAGATAATT T$_{09}$*
47	83	285	ATATAA GTACAAACAAACAGTGAGAAGATAACGAGACTGAGTA TATATTTT*
50	88	255	ATAT AAAATGTACAAATAAGCAGTAGAAGAGCAGTGAAATTGA TGATAT$_{09}$
47	89	237	ATAT AGAAATGTACAAACGAGCAATAAGGGAACAGTGAAATTAGATG TTTAAT$_{09}$
51	88	152	ATAT AAAATGTACAGACAAGCAGTGAAGAGACAGTGAGATTA TACAGTTTT
39	77	112	ATATATAAAATGTACAT ACGAACGATAAAAGGGCAGTGAAATTAGATAATTA T$_{12}$
51	86	70	AAAATC AATGTACAAATAAGCGATAAAGAGACAGTGAGATCA TGAT$_{15}$
48	89	53	ATAT AGAAATGTACAAACGAGCAATAAGGGAACAGTGAAATTAGAT T$_{12}$
47	83	31	ATATAA GTACAAACAGACAGTGAGAAGATAACGAGACTGAGTA TATTTAT$_{17}$
78	118	573	ATATAT AATCACAAACAAATAGAAAATGAGAGAGGTGTATGA TACTAT$_{15}$
93	123	1,321	ATATAT AAACAAATCACGAACGAGTAGAAAAT TGGAAGAATGTAT$_{12}$
78	118	685	ATATAT AATCACAAACAAATAGAAAATGAAAGAGGTGTATGA TACTATTTT
89	123	251	ATATT AAACAAATCACAAATGAGTAGAGAACGAGA TTGATGTATAT$_{16}$
86	121	243	ATAC ATAAATCACAAACGAATAAGGGGCAGAAGAG TTGTATGATAT$_{05}$
76	124	120	ATATAT AAAACAAATCATAAACGAATAAAGAGTGAGGAAGGTGTATAAAT TTT
105	140	15,770	A TAGATAACAAACATAAGAGCAAGTCACGAG GTAGATATTGATAT$_{14}$
105	140	692	A TAGATAACAAACATAAGAGCAAGTCACGAG ATAGATATTGAT$_{12}$
98	142	442	AT ATTAAATAACAAACATAGAAATAGATCACAGATAGATGA TTGATGAAT$_{06}$
122	166	27,885	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATAAGA TAT$_{13}$
124	166	212	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATAA T$_{12}$
125	166	139	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATA T$_{17}$
127	166	114	ATAT ACAAAAATCCAATGAAAAATAAAGACTGAGTGATGGATG CAAT$_{15}$
123	162	309	ATATAT AAATCCAATAAGAAATGAAAGCTAGATAGTGAGTATAAG T$_{16}$
124	166	421	ATATAT ACAAAAATCCAATGAAAAATAAAGACTGAGTGATGGATGTAA TTATTTAT$_{14}$
129	167	370	ATAT AACAAAAATCTAATAGAGAACAGAAACTAAGTAATGAG ATAT$_{09}$
123	161	348	ATATTAT AATCCAATAGAAAGCAGAGACTAAATGATAGATGTAAA T$_{09}$*
132	168	110	ATATAT AAACAAAAATTCAATAGAAAACGAGAACTGAGTAAT TGATATGT$_{05}$
125	167	53	ATATAT AACAAAAATCCGATAAAAGATGGAAACTAAGTGATAGATATA T$_{13}$
154	196	1,492	ATAT ATACAACAATAAACTCGTATTAAGTGAGAGATGAAGATTTAAT T$_{13}$
156	199	136	ATAT TAAACACAACGATAGATCTATATTAAGTAGAAGATAGAAATTTA T$_{16}$
162	200	2,176	ATAT ATAAACACAACAGTAGACTCGTATTAGATAGAAGATAGA TATCAT$_{15}$
156	200	114	ATAT ATAAACACGACAATAGATTCATATTAAATAGAAGATAGAGATTTA T$_{12}$
157	200	98	ATAT ATAAACACGACAATAGATTCATATTAAATAGAAGATAGAGATTT T$_{07}$*

5'	3'	Reads	CR3 gRNA Sequences cont.
190	230	2	ATATACA ACATATCAAGTGGTAAGATAAGAGAAGAAAGTAGATGTAAT T_{06}
190	230	1,243	ATATACA ACATATCAAGTGATAAGATAAGAGAAGAAAGTAGATGTAAT T_{12}
192	232	279	ATAGATA ACACATATCAGATGATAAGGTAAAGAGAGAGAATAGATATA T_{13}*
226	277	19	ATAT AATAACAATATAAACGAACTGGATGATGTATAGTACTTTGATATATA AT_{12}
231	265	2	ATATAACAATAC AAATGAGCTAGATAATGGATGATAGTTTGATAT T_{12}
237	265	1	TATAAAATAACAATAC AAGTGAATTAGATAATGGATGATG TTTCAAT$_{15}$
241	279	936	AT AGAATAACAATATAAACGAACTAGATGATGGATAGTA TCT_{17}
234	265	528	ATATAACAATAC AAATGAACTGAGTAATGGATAGTAGTTTGA $CAAT_{09}$
293	308	2	ATATAT AAATTATTTGCATACT GAGATTAGTGATAGTAAAGTGATTAAT$_{13}$
268	312	125	AT ATAAAAATTATTTGCATGCTTAAGTGAGTTATAGAGGTAATGATA AAT_{07}

D) C-Rich Region 4

5'	3'	Reads	CR4 gRNA Sequences
25	64	121	ATAAT AAAAATGCACAACTAGAATTGAAGTAAAGTGATGATA $TATAT_{14}$
25	64	596	ATAAT AAAAATGCACAACTAGAATTGAAGTAAAATGATGATA $TATAT_{14}$*
25	64	308	ATATT AAAAATGCACAACTAGAATTGAAATAAAGTGATGGTA $TATAT_{13}$
25	62	83	ATATAATT AAATGCACAGCCAAAGTTAAGGTAGAATAGTGATA TAT_{14}
48	103	296	ATA TATATAAAACACAGACATACTAAGTAAGAGAAAGAGAGGTGTATGATT T_{12}
48	103	175	*ATA TACATAAAACACAGACATACTAAGTAAGAAAAAGAGAGATGTATGATT T_{12}*
65	98	108	ATATAT TAAAACACAAATACATCAGATAGAAAGAGA TTGAGTGTATAAT$_{17}$
51	103	18	ATA TACATAAAACACAGACATACTAAGTAAGAAAAAGAGAGATGTATG T_{21}
59	103	11	ATA TACATAAAACACAGACATACTAAGTAAGAAAAAGAGAGAT T_{15}
52	103	10	ATTATA TACATAAAACACAGACATACTAAGTAAGAAAAAGAGAGATGTAT T_{15}
87	134	7,793	ATATAT AAACAACAATAGAGTATATCATAGACTGTATATGAAGCATAAAT T_{10}
89	134	426	ATATATAT AAACAACAATAGAGTATATCATAGACTGTATATGAAGCATAA CT_{08}
88	134	208	T ATATAT AAACAACAATAGAGTATATCATAGACTGTATATGAAGCATAAA AT_{13}
93	139	200	AAACAAAACAACAGTGAAATATACCGTAGATTGTATGTGAAAT TATATAT$_{06}$
90	134	143	ATATATAT AAACAACAATAGAGTATATCATAGACTGTATATGAAGCATA T_{13}
93	142	2,052	A AAAAAACAAAACAACAGTGAAATATATCGTAGATTGTATGTGAAAT TAT_{05}
90	138	584	ATAT AACAAAACAACGATAAGATGTATCATGAGCTGTATATGAGATATA T_{25}
91	138	139	ATAT AACAAAACAACGATAAGATGTATCATGAGCTGTATATGAGATAT T_{13}
127	171	6,971	ATAT ATATACTCACACAAATAGATGACAGAGATAGAAAGTAAGATGATA TAT_{15}
128	171	204	ATAT ATATACTCACACAAATAGATGACAGAGATAGAAAGTAAGATGAT T_{13}
154	192	3,603	ATATTA AACTATAACAAGGCAGATAGAACGTACCTATATAGATAA T_{14}
174	200	6	AT AAATAAACAACTATAATAAGACAAGTG CGATGTACT$_{10}$
166	196	5	ATATA AAACAACTATAACAAAATAGATAGAATGTGC $GTAT_{06}$
171	200	3	AAATAAACAACTATAATGAAGTGAGTGAAG $GTACGT_{12}$
169	196	1	ATATA AAACAACTATAACAAAATAGATAGAATG $GGCGTAT_{06}$
171	196	1	A AAACAACTATAATGAAGCAGATAGAA $GGTACGTATAT_{17}$
186	232	31	ATAAAAAATCACAGCCTAAAATGACGAGAGAAAGTAAATGGTTATA $TAGAT_{05}$
186	225	22	ATATAT ATCACAACCTAAGATAACGAAAAAGAGCAGATAATTGTA TATTTT
186	228	7	ATATAT AAAATCACAACTTAGAATGACGAAAGAGAATAGATAGTTGTA $TATTGT_{22}$
186	228	4	ATATAT AAAATCACAACTTAGAATGACGAAAGAGAATAGATAGTT $TTAAAAAAT_{16}$
186	232	166	ATAAAAAATCACAGCCTAAAATGACGAGAGAAAGTAAATGGTTATA T_{16}
186	227	152	ATAT AAATCACAACCTGAAATAGCAGAGAAGAGTAAATGATTATA T_{13}
189	232	25	ATAAAAAATCACAGCCTAAAATGACGAGAGAAAGTAAATGGTT TTT*
189	227	12	ATATAT AAATCACAACCTGAAATAGCAGAGAAGAGTAAATGATT T_{09}*
213	261	14,358	ATAT ATAAACTATACAATTGAAGCACTGATAGAAGGTTGTGATTTAA T_{12}
213	261	264	ATAT ATAAACTATACAGTCAAGACACTGATGAGAGATCGTGATTTAA T_{13}
216	261	114	ATAT ATAAACTATACAATTGAAGCACTGATAGAAGGTTGTGATTT TTT
222	261	112	ATAT ATAAACTATACAATTGAAGCACTGATAGAAGGTTG ATTTAATAT$_{08}$
213	258	5,062	ATATAT AACTATACAGTCGAGACATCAATGAGAGATTATGACTTAA T_{13}
213	259	1,370	CCCTCGGCCGCAGCGATATA AAACTATACAATTAGAGCATCAGTAGAAGATTGTGACTTAA T_{14}
213	259	1,074	ATATA AAACTATACAATTAGAGCGTCAGTAGAAGATTGTGACTTAA T_{15}*
213	258	490	ATATAT AACTATACAATTGAGATATCAGTGAAGAATTGTGATTTAA T_{11}
215	258	340	ATATAT AACTATACAGTCGAGACATCAATGAGAGATTATGACTT T_{14}
223	258	183	ATATAT AACTATACAGTCGAGACATCAATGAGAGATT T_{05}
251	300	23	AAAAAAAAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATAT T_{11}
242	284	9	AAAAAAA AAAATTATAACGTCATAGAAGAGATAGACTATATGATTAA T_{09}
251	295	77	ATAT AGAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATAT TTTT
250	295	64	AT AGAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATATA T_{06}
253	295	25	ATAT AGAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTAT T_{16}
253	298	22	AAAAAAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTAT TTT
244	295	20	AT AGAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATATAGTT T_{12}
250	297	17	AAAAAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATATA T_{12}
251	301	16	AAAAAAAAAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATTATAT T_{05}
255	295	11	ATAT AGAAAATAAACAGAATTGTGACGTCATAAAGAAGATAGATT TTTT
280	320	643	AAAA AAAAAACACAAGGCGAGATAGAGAAAAGAGATAAATAAGAT GT_{10}
279	320	93	AAAAA AAAAAACACAAGGCGAGATAGAGAAAAGAGATAAATAAGATT T_{05}

5'	3'	Reads	CR4 gRNA Sequences cont.
307	351	725	ATAAAAT ACCAAACAGATCAGATAAAGGCAGTGATATAGAGAATATAAAAT T_{11}

311	353	675	ATAT AAACCAAACAAGCTGAATAAGAACAGTGATATAGAAGATATA TAT$_{14}$
306	351	259	AAAAT ACCAAACAGATCAGATAAAGGCAGTGATATAGAGAATATAAAATA T$_{12}$
309	351	214	ATAAAAT ACCAAACAGATCAGATAAAGGCAGTGATATAGAGAATATAAA T$_{12}$
324	354	140	AT AAAACCAGACAAACTGAATGAAGACAGTAACGTAGAAGATAT T$_{10}$
307	352	485	TTGTTTGGTTGATTAAAT AACCAAACAAGTCGAGTAGAGACAGTGATATAAAAGGTATAAAAT T$_{09}$*
310	353	374	ATAT AAACCAAACAGACCAAGTGAAGATGGCAGTATAAGAGATATGA TATAT$_{11}$
309	352	134	AAAT AACCAAACAAGTCGAGTAGAGACAGTGATATAAAAGGTATAAA T$_{16}$*
310	352	42	AAAT AACCAAACAAGTCGAGTAGAGACAGTGATATAAAAGGTATAA T$_{12}$
312	352	38	AAAT AACCAAACAAGTCGAGTAGAGACAGTGATATAAAAGGTAT T$_{13}$
343	388	863	ATAT ATAACACAAAACATAACAGAGAGTATAGAGAGAAATTGAATGA T$_{15}$
340	390	2,308	ATAT AAATAACACAAAATACGACGAGAAATATAAGAGAGAATTGAGTAAATT T$_{05}$*
343	388	255	ATAT ATAACACAAAACATAACAGAGAGTATAGAGAGAAATTGAATGA T$_{15}$*
374	417	411	AAAA AAAAAACAACATAGAAAGTGAATCAGAGAATGACATAAGATATA T$_{11}$
377	417	447	A AAAAAACAACATAGAAAATAAGTCAGAGAGTAATATGAGAT TGTTATAAT$_{05}$
374	415	423	ATATAT AAAACAACATAAGAGATGAATCAAGAGGTAATATAAGATATA T$_{17}$
404	457	4,211	AAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGGTATAA T$_{14}$
405	456	620	AAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGGTATA T$_{13}$
406	457	392	AAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGGTAT T$_{12}$
405	457	406	AAAAAAAAAAACAAAAACAGAAAAACTGTCTAAGATAGAGAATGATATA T$_{15}$
409	457	185	AAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGG GATAAAT$_{08}$
404	458	227	AA AAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGATATAA T$_{08}$*
405	456	186	AAAAAAAAAACAAAAACAAGAAGACTATCTGAGGTAGAAAATGATATA T$_{14}$
411	443	97	ATATAAACAT AAACAAAGAGACCATCCGAAATAGAGAAT T$_{15}$
405	457	34	AAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGATATA T$_{09}$
406	458	25	A AAAAAAAAAAAACAAAGACAAAGAAATCACTCAGAATAGAAGATGATAT T$_{09}$*
442	489	3,197	ATAACAACCACAGATAGAAATGAACGTAAATGAGAGAGAAAGATGAAA T$_{13}$
440	489	238	ATAACAACCACAGATAGAAATGAACGTAAATGAGAGAGAAAGATGAAAGT T$_{11}$
442	487	730	ATAA AACAACCACAAGTAGAAATAGACATAAATGAGAGAGAAAGATAAAA T$_{12}$
442	486	246	ATAT ATAACCACAAATAGAGACAAATGTAAGTAAAGAGAGAAAGTAAAA T$_{19}$
443	487	208	ATAA AACAACCACAAGTAGAAATAGACATAAATGAGAGAGAAAGATAAA T$_{06}$*
446	487	149	ATAA AACAACCACAAGTAGAAATAGACATAAATGAGAGAGAAAGAT T$_{11}$
453	497	1,153	AT ACAAAATAACAACAATCATAAGTAAGAATAGATGTAGATGAGAAA T$_{11}$
453	494	282	ATATAT AAATAACAACAATCACAGATGAAGGTAGATATAAGTGAGAAA T$_{14}$
453	497	2,511	AT ACAAAATAACAATGATCACAAATAAGAGTGAATGTAGATAGAGAA T$_{15}$*
453	494	306	ATATAT AAATAACAACAATCACAGATGAAGGTAGATATAAGTGAGAAA T$_{13}$*
475	519	766	ATATAT AATAACAACAATAATCAGATTAGCAGAGTAATGATAGTTATAAAT T$_{13}$
480	524	390	ATACAAATAACAACAATAGCCAAGTTAATAGAGTGATGATGATTA AT$_{14}$
478	524	1,913	ATACAAATAACAACAATGATCAGACTAATAGAGTAATAGTGATTATA T$_{14}$
481	524	105	ATACAAATAACAACAATGATCAGACTAATAGAGTAATAGTGATT T$_{12}$
479	524	101	ATACAAATAACAACAATGATCAGACTAATAGAGTAATAGTGATTAT T$_{15}$
504	554	146	AA AAAAAACGCATATAAGTGAGCCTATACATAGATAATGATGATAATT T$_{03}$
503	547	53	AAAAAAAAAAA GCATATAAATAGATCTATATATGAGTGATAGTGACAATTA T$_{15}$
504	547	50	AAAAAAAAAA GCATATAAATAGATCTATATATGAGTGATAGTGACAATT T$_{11}$
507	554	98	AAAAA AAAAAACGCATATGAGTAGATCTGTACATAGATAGTGATGATA T$_{12}$
507	556	84	ATAGAAAACGCATATGAGTAGATCTGTACATAGATAGTGATGATA CT$_{06}$
508	554	37	CAACTCGACTGCGTGAA AAAAAACGCATATGAGTAGATCTGTACATAGATAGTGATGAT TTT
542	575	511	AAATTTATAAAACCAAT CCGTAATTATCTGAGATGAGAAGTGTATA AT$_{12}$
542	584	997	ATATTTATAAAACC ATACTGTGATTATCTAAGGTAGAAAGTGTGTA AT$_{21}$
542	596	352	ATA TTTATAAAACCAATATCATAGTTATCTGAGGTAGAGAGTGTATA ATTTAATGTAT$_{18}$
537	568	62	CATACC TAATTATCTGAAGTAGAAGATGTATGTAAAT T$_{14}$

E) Cytochrome b

5'	3'	Reads	CYb gRNA Sequences
32	59	327	AAATAATAGGGATTTATGATGAGATATG CTGTGGATAT$_{14}$
32	60	255	AAAATAATAGGGATTTATGATGAGATATG CTGTGGATAT$_{12}$
32	61	190	AAAAATAATAGGGATTTATGATGAGATATG CTGTGGATAT$_{13}$
32	64	176	AA AAAAAAAATAATAGGGATTTATGATGAGATATG CTGTGGATAGT$_{09}$
32	62	65	GT AGAAAATAATAGGGATTTATGATGAGATATG CTGTGGATAGT$_{12}$
32	64	35	AA AAAAAAAATAATAGGAGTTTATGATGGAATATG CTGTGGATATTTT*
32	64	17	AA AAAAAAAATAATAGGAGTTTATGGTGAGATATG CTGAGAGTAT$_{05}$
53	91	18,713	AAAAAA AAAAGACAGTGTGAATTTCTGAGTAATAAAGGGAATAAT T$_{11}$
51	91	10,649	AAAAAA AAAAGACAATATAGATTTCTGGGTGATAAAAGGGATAATAA CT$_{11}$
52	91	908	ATAA GAAAGACAATATAGGTTTCTGGGTAATGGAGAGAATAATA T$_{16}$
54	91	339	AAAAAA AAAAGACAATGTAGATTTCTGAGTAATGGGGAGGATAA CTATTTATTTT*
54	91	8,406	AAAA AAAAGACAATGTAGATTTCTGAGTAATGGGGAGGATAA CTAT$_{05}$*
54	91	2,265	AAAAAA AAAAGACAATGTAGATTTCTGAGTAATAGGGAGGATAA CTAT$_{16}$
53	91	204	AAAAAAA AAAAGACAATGTAGATTTCTGAGTAATGGGGAGGATAAT T$_{07}$
56	91	95	AAAAAAA AAAAGACAATGTAGATTTCTGAGTAATGGGGAGGAT T$_{05}$*

F) Maxicircle Unidentified Reading Frame II

5'	3'	Reads	Murf II gRNA Sequences
30	79	2,605	ATAG AAAGCACAAAAATAAAATTAAATTAGAGTAATTGGATGTTAAAATT T$_{11}$
30	79	125	ATAG AAAGCACAAAAATAAAATTAAATTAGAGTAATTGAATGTTAAAATT T$_{08}$
34	79	17	ATAG AAAGCACAAAAATAAAATTAAATTAGAGTAATTGAATGTTAA CAT$_{12}$
33	79	15	ATAG AAAGCACAAAAATAAAATTAAATTAGAGTAATTGAATGTTAAA T$_{09}$

G) NADH Dehydrogenase 3

5'	3'	Reads	ND3 gRNA Sequences
33	76	313	ATATATT ATAAACCATGATATCGAAAATGGGTGTAGAAATGATGATA T_{12}
31	79	70	ATAT ATAATAAACCACAGTATCAGAGACAGATATAGAAGTGATGATAGT T_{13}
30	73	625	ATAT TAAACCACAATATCAGAAATAAGTGTAGAAATAGTGATAATA T_{12}
31	79	511	ATAT ATAATAAACCACAGTATCAGAGACAGATATAGAAGTGATGATAGT T_{09}
42	79	397	ATAT ATAATAAATCACAGTATCAGAGACAGATATAGAA TGATGATAGT$_{12}$
33	79	26	ATAT ATAATAAACCACAGTATCAGAGACAGATATAGAAGTGATGATA T_{16}
63	113	1,006	ACATAAGAAACATAAAGAAAAATCTGTGAGTAGAGTGATAAGTTATAAT T_{11}
65	113	411	ACATAAGAAACATAAAGAAAAATCTGTGAGTAGAGTGATAAGTTATA T_{15}
64	113	84	AT ACATAAGAAACATAAAAAGAAATCTGTAAGTAGAGTAGTAAGTTATAA GT$_{14}$
63	108	59	ATACAT AAAAACATAAAAAGAAACTTATAAGTAGAGTGATAGATTATAAT T_{09}
65	108	33	ATACAT AAAAACATAAAAAGAAACTTATAAGTAGAGTGATAGATTATA TTTT
66	108	27	ATACAT AAAAACATAAAAAGAAACTTATAAGTAGAGTGATAGATTAT T_{12}
68	108	14	ATACAT AAAAACATAAAAAGAAACTTATAAGTAGAGTGATAGATT T_{10}
99	143	826	AT ATGAAAACAATCAAAGAAGTGTGATAGAAAGTATAAAAGGTATAA T_{11}
98	141	542	ATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATATAAA T_{12}
101	141	501	ATATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATAT T_{09}
98	141	240	ATATAA GAAAACAATCAGAGAAATGCAGTAAAAGATATAAGAGATATAAA T_{13}
100	141	199	ATATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATATA T_{11}
99	141	188	ATATAA GAAAACAATCAGAGAAATGCAGTAAAAGATATAAGAGATATAA TCT$_{13}$
100	141	168	ATATAA GAAAACAATCAGAGAAATGCAGTAAAAGATATAAGAGATATA T_{18}
99	141	52	ATATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATATAA T_{10}
98	141	24	ATATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATATAAG T_{09}
100	141	21	ATATAA GAAAACAATCAGAGAAATGCGGTAAAAGATATAAGAGATATG T_{20}
130	170	507	ATATAT CAAATCACACGAAGATCATAGATGACGATGAAGGTAGTTAA TAT$_{18}$
130	170	413	ATACAT CAAATCACACGGAAATCATAGATGGCAATGAAGATAGTTAA T_{11}
130	174	2,397	ATA TATACAAACTACACGGAGATCACAAATAACAGTGAGAATGATTAA T_{12}
132	174	133	ATA TATACAAACTACACGGAGATCACAAATAACAGTGAGAATGATT T_{15}
158	205	412	T ATGTATAAAACATTAAATGTGAGTTTGTGTCGTGTAGATTATGTGAA T_{15}
155	189	236	AAATAAAACACC AACGTGAATTTATATTGTATAGATCGTATGAGAAT AT$_{15}$
158	205	1,756	ATAT ATGTATAAAACATTAAATGTGAGTTTGTGTCGTATAGATTATGTGAA T_{14}
190	229	4,864	ATATAT AACAACTAACAGAATATAGATCTGATGTATAAGATATCA T_{14}
190	229	123	ATATAT AACAACTAACAGAATATAGATCTGATGTATAAGATATCG T_{13}
190	233	227	ATAT AACAAACAACTGATGAGACATAGATTCGATGTATAAGATATCA T_{13}
198	234	139	ATAT AAACAAACAACTAATAAAATGTAAATCTGATGTGTGA TATACT$_{08}$
191	233	40	ATAT AACAAACAACTGATGAGACATAGATTCGATGTATAAGATATC T_{11}
223	263	3,225	TTAT ATACAAATAATGGGATTTAACGATATAGAGGATGAATGATT T_{11}
222	263	3,165	ATAT ATACAAATAATGGGATTTAACGATATAGAGGATGAATGATTA T_{15}
222	264	384	AAAAAT AACACAGATAATGGAATTTAATGATATGAGAAATGGATGATTA T_{05}
223	264	164	AAAAT AACACAGATAATGGAATTTAATGATATGAGAAATGGATGATT TCT$_{17}$
223	263	106	ATAT ATACAAATAATGGGATTTAACGATATAGAGAGTGAATAATT TTTT
222	263	232	AAAT ATACAGATAATGGGATTTAACGATGTGAGAGATAGATAATTA T_{16}
223	263	177	AAAAAT ATACAGATAATGGGATTTAACGATGTGAGAGATAGATAATT T_{05}
220	263	32	AAAT ATACAGATAATGGGATTTAACGATGTGAGAGATAGATAATTAAT T_{13}
253	299	260	ATAT GATAAAACAACACTATTATGAGAACGAGTGATAGAATATAGATAAT TTCT$_{14}$*
253	299	106	ATAT AATAAAACAACACTATTACAAAGATAGACAGTGAGATATAGATAAT T_{13}
253	298	83	ATACAT ATAAAACAACACTATCATAGAAGCAGACAGTGAGATATGAGTAAT T_{15}
253	299	133	ATAT AATAAAACAACACTATTACAAAGATAGACAGTGAGATATAGATAAT T_{07}
252	299	89	AT AATAAAACAACACTATTACAAAGATAGACAGTGAGATATAGATAATG AT$_{06}$
253	299	57	ATAT AATAAAACAACACTATTACGAAGATAGACAGTGAGATATAGATAAT T_{12}
252	299	22	AT AATAAAACAACACTATTACGAAGATAGACAGTGAGATATAGATAATG AT$_{16}$
284	329	786	ATAT AAAACCACAAAAATAGAAAGCTATAATAGAGATAGAATAATGTTA A$_{07}$
288	320	435	ATTTTAAGTT AGAGTGAGAAATTGTAGTGGAAATAAGATGATA AAAT$_{11}$
285	329	282	ATAT AAAACCACAAAGGTAGAAGATCGTAATAGAGATAGAATAATATT T_{09}
300	333	268	AT AACAAAAACCACAGAGATGAGAGATTGTAATAAG TATAGTGATAAT$_{13}$
285	329	168	ATAT AAAACCACAAAAATAGAAAGCTATAATAGAGATAGAATAATGTT T_{12}
282	329	64	ATAT AAAACCACAAAAATAGAAAGCTATAATAGAGATAGAATAATGTTATT CT$_{09}$
285	328	3	AT AAACCACAAAGATAGAAGGCCATAATAGAGATAAAATAATATT T_{10}

5'	3'	Reads	ND3 gRNA Sequences cont.
322	369	48,018	ATAT AATACCACATGAATCTTATATGTACGATGGAAGATGAGAATTAT T_{13}
324	369	5,034	ATAT AATACCACATGAATCTTATATGTACGATGGAAGATGAGAATT T_{11}
321	369	3,062	ATAT AATACCACATGAATCTTATATGTACGATGGAAGATGAGAATTATG TTTCT$_{06}$
322	369	142	ATAT AATACCACATGAATCTTATATGTACGATGGAAGATGAAAATTAT TCT$_{10}$
322	370	721	AT AAATACCACATGAATCTTATATGTACGATGGAAGATGAGAATTAT T_{12}
321	370	291	AAATACCACATGAATCTTATATGTACGATGGAAGATGAGAATTATG CAGT$_{08}$
320	368	149	ATAT ATATCACACAAATTCTATACATATAATAGAGAATGAGAGTTACAA T_{06}
324	370	43	AT AAATACCACATGAATCTTATATGTACGATGGAAGATGAGAATT T_{05}*
347	388	95	ATATGTAAAC AATATACGTGATTTCAGAGATACTATGTGAATTCTAT T_{22}
345	388	82	ATATGTAAAC AATATACGTGATCTCAGAGATACTATGTGAATTCTATGT T_{07}
355	388	1,307	ATATAAACAAAC AATATATGTGGTTTCGAAAGTGTCATGTGAATT AT$_{12}$
350	388	633	ATAGAT AATATACGTGATCTTAGAGGTACCATGTGAGTCT GAGTTAT$_{12}$
349	388	376	AGTATGCGTGATTTTAGAGATATTGTATGAATTTT T$_{08}$
355	388	104	AGTATGCGTGATTTTAGAGATATTGTATGAATT ATAT$_{15}$*
402	438	417	ATATA TATAATACAACAAGGAGCGTCATAAGTAAAGTGAA TTCGTTATAT$_{13}$
402	438	128	ATAT TATAATACAACAGAAAATGTCATAAGTGAGATGAA TTCGTTATAT$_{09}$
402	435	753	ATATATAA AATACAACAAGAGACGTCGTAAATAGAGTAAA TTCGT$_{08}$
403	438	126	ATAT TATAATACAACAGAAAATGTCATAAGTGAGATGA TTCGTTAT$_{06}$

H) NADH Dehydrogenase Subunit 7

5'	3'	Reads	ND7 gRNA Sequences
36	69	100,761	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAG TGATATTAT$_{14}$
35	69	765	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAGG GATATTAT$_{11}$
36	71	240	AAAT ATACAAATGTAAAGAAGCTATCAGAGGTAATATAAG TGATATAAT$_{13}$
36	69	121	ATATA ATAAATGTAAAGAGACTATTGAGAGTGACATAAG TGATATTAT$_{13}$
38	71	113	ATAT ATACAAATGTAAAGAGACTATCGAGAGTGACATA TGTGATATTAT$_{11}$
28	71	35,079	AAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATATT AT$_{12}$
27	71	20,487	AAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATATTT T$_{08}$
31	71	546	ATAAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGAT TTTT
34	71	402	AAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAAT T$_{15}$
36	69	354	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAG TGATATTAT$_{12}$
24	71	327	ATAAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATATTTGTT TTTT
24	71	298	AAAT ATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATATTTATT T$_{11}$
28	71	194	AAAT ATACAAATGTAAGAGAACTGCCAAAAGTAACGTAGAATGATATT ATCT$_{09}$
59	91	4,376	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAA TGT$_{13}$
58	91	1,242	ATATATAAATACA GTGGACTAAATGTGAAGCGATATAAATGTGAAA T$_{12}$
108	137	13	ATATAATAAACAACATAAAGTGCCATGTACT#CGAGAT TTTTAG
95	132	1	AT ATAAACAACATAAAATACTATGTGATGTAGGAT CTGTGAATTAAT$_{09}$
124	170	3,944	ATAC ATCAATATAAACGATAGATTTACCGTAGAAGTATAGTGAATAAT T$_{14}$
124	168	3,640	ATATA CAATATAAACAGTAGATTCACTGCAGAAGTATGATAGATAAT T$_{11}$
139	170	1,507	AT ATCAATATAAACAATAAGTTCGTCATAGA TTTACAGTAGATAAT$_{12}$
121	166	396	ATACAT ATATAAACAATGAATTCACTGTGAAGATACGATAGATGATATA T$_{14}$
139	170	384	AT ATCAATATAAACAATGAGTTCGTCATAGA TTTACAGTAGATAAT$_{12}$
121	166	178	ATAT ATATAAACAATAAATTCATCGTGAAGGTACAGTAAATGATATA T$_{16}$
122	166	26	ATAT ATATAAACAATAAATTCATCGTGAAGGTACAGTAAATGATAT T$_{15}$
152	190	988	AAATTACGATGCAT AATAATCTATGGTACAGTTGATATGAGTGATAA T$_{12}$
147	199	563	ATATATA CACGATGCAGATAATCTATAGTATGATTGATATAAGTGATAAATTT T$_{09}$
150	199	200	ATATATA CACGATGCAGATAATCTATAGTATGATTGATATAAGTGATAAAT AT$_{12}$
151	199	2,089	AAA TACGATGTAAATAACCTGTAGTATAGTTAGTGTAGATAGTGAA AT$_{12}$*
147	179	295	ATATATAAATGCAAATAACG TGTAATACAGTCAATATAGATGATAAATTT T$_{09}$
152	190	143	AAATTACGATGCAT AATAATCTATGGTACAGTTGATATGAGTGATAA T$_{13}$
246	269	2,470	ATATCAAC ACATAATCTGACTTGTCGGAGTAT CTAAAGGAATAAAT$_{14}$
246	269	837	ATATCAAC ACATAATCTGACTTGTCGGGGTAT CTAAAGGAATAAAT$_{12}$†
261	311	3,259	ATAT AACATAAAGACAATAAGTGCTTATTACAGTGAACATTGATATAATTT T$_{08}$
261	293	2,874	ATATATAAGACAAC GATGCTCATTATGATAGATACTGATGTAATTT T$_{10}$
261	310	146	ATAT ACGTAAAGACAATAGGTGTTCATTGCAGTAGATATTGATGTAATTT T$_{10}$
260	293	133	ATATATAAGACAAC GATGCTCATTATGGTAGATACTGATGTAATTTA T$_{07}$
261	293	1,105	ATATATATAAGACAAC AGTGCTCGTTACAGTGAATATTGATGTAATTT T$_{11}$*
297	338	888	AT ATAAATAACATCGCAACGTATATTCGAAATGTAGAGATA CAT$_{13}$
295	338	765	ATATA ACAAACAACATCGTAATATGTGCTCGGAGTATAGAGATAAT TAAATAT$_{13}$
292	334	91	ATACT ACAACATCGCGATATATACTTGGAATGTAAAGGTGATAAA GT$_{11}$
292	324	394	ATATATAACAACATCG AGCATATACTCAGAATATAAAGGTGATAAA GT$_{12}$
292	324	178	ATATATAACAACATCG AGTATATACTCAGAATATAAAGATGATAAA GT$_{09}$
293	324	90	ATCGG AGCATATACTTGAGATATAAAGATGATAA T$_{11}$
327	373	2,128	A TATAATTAATAAACGTATAAATGTGCAGTGTGACGATGAATGATATT AT$_{13}$
327	365	1,225	ATACAT ATAAACGTATAGGTGCATAATGTAACGATGAATGATGTT AAT$_{11}$
327	365	20,110	ATACAT ATAAACGTATAGATGCGCAATGTAACGATGAATGATGTT AAT$_{12}$
327	373	997	A TATAATTAATAAACGTATAAATGTGCAGTGTGACGATGAATGATATT T$_{13}$
330	378	517	AAA TTACAATTAATAGACGTATAAGTGCATAGTGTAGTGATGGATAAT T$_{17}$
329	378	111	AAA TTACAATTAATAGACGTATAAGTGCATAGTGTAGTGATGGATAATG AT$_{13}$
352	398	283	ATATA GAAAACTACAGGTAAATTCTGCAATTAGTAGACGTGTAAAT AT$_{07}$
353	402	3,043	ATATA TATTAAAACTATGGGTAGATTCTGTAATTAATAGACGTGTAAA AT$_{13}$*
352	385	602	ATATATAAAACTACGA GTAGATTCTATGATTGATGAACGTGTAAAT T$_{11}$*
354	402	556	ATATA TATTAAAACTATGGGTAGATTCTGTAATTAATAGACGTGTAA T$_{12}$
352	402	173	ATATA TATTAAAACTATGGGTAGATTCTGTAATTAATAGACGTGTAAAT TTCTTTT
390	424	2,283	ATATTAA ATACATGATATACGCGATAGATTATTAGAGTTATG AGTTAAT$_{16}$
391	427	200	AAATT ACCATACATGATATATACAGTGAACTATTAGAATTAT AGGTAATGT$_{06}$
390	424	162	ATATTAA ATACATGATATGCGCAGTAGACTATTAAAGTTATG AGTTAAT$_{12}$
391	424	298	AAATTACA ATACATGATATATACAGTGAACTATTAGAATTAT AGGTAAT$_{14}$

5'	3'	Reads	ND7 gRNA Sequences cont.

5'	3'	Reads	ND7 gRNA Sequences cont.
412	452	15,183	ATATATAA GGAGACAAATGATCTAGATTCGAGACTGTATATGATATAT T_{13}
414	451	223	ATATATAACAAC GAGACAGATAATCTAGATTTGAAGTTATATGTGATAT T_{12}
416	452	199	ATATATAA GGAGACAAATGATCTAGATTCGAGACTGTATATGAT T_{13}
416	464	5	AT ATTATAATAACGGAGATGAGCAATTTAGATTCAGAGTTATATGTGAT T_{13}
407	450	3,126	ATATA AGACAAACAATCTAAATCTGAGACTGTATATGATATGTATAAT T_{10}
414	458	2,332	ATAT ATAACGGAGACAGATAGCTTGAATCTAAGGTTGTATGTGATAT T_{12}
410	450	132	ATATA AGACAAACAATCTAAATCTGAGACTGTATATGATATGTAT T_{13}
453	485	3	ATAAAATGTCATCAATTA GTTACGTTCTTTGAGTGATTGTGATAAT AT_{13}
453	485	207	ATAAATAACATCAATTA GTTACGTTCTTTGAGTGATTGTGATAAT AT_{13}
486	530	2,523	ATATAA GCATACGACAATTACGATATGAGTCAGAGAATGTTGTTAATTT T_{11}
499	526	2,002	ATATATAA ATACGACAATCATAACGTGAATCAGAGA CTGTGATTAAT$_{11}$
486	530	1,136	ATATAA GCATACGACAATTACGATATGAGTCAGAGAATGTTGTTAATTT T_{11}
486	528	344	ATATA ATACGACAATCACGATATAGATTAAAAGATGTTGTTAATTT T_{10}*
508	548	1,129	ATAT AAATCATGAAAGCTGAGTGTGTATGGCAGTTACGATATA TGTTAAT$_{14}$
508	544	116	ATATAAAA CATGGAAGCTAAGTGTGTATGATGATTATGATATA TGATTAAT$_{06}$
508	553	1	ATA TAATAAAACCATGAGAGTTGGATGTATATGATGATCGTGATATA T_{17}
526	569	3,050	ATATAT ATCATCAAGAATATCTAATGAGACTATGAGAGTTAAATGTATA AT_{12}
526	569	4,850	ATATAT ATCATCAAGAATATCTAATGAGACTATGAGAGTTAAATGTATA AT_{13}
527	569	351	ATATAT ATCATCAAGAATATCTAATGAGACTATGAGAGTTAAATGTAT T_{06}
540	576	544	ATAATAAT AAACAAAATCATTGAGAGTATCTGATAGAGTTATGA TAGTCAT$_{05}$
531	574	291	AAAT ACAAAATCATCAGGGATACTTGGTAAGATTGTGAAAGTTAAGT T_{16}
540	571	8,794	ATATAAT AAATCATCAAGAATATCTGATAGAACTGTGA TTGCTAAT$_{14}$*
540	571	1,694	ATACAAT AAATCATCAAGAGTATCTAATAGAACTGTGA TTGCTAAT$_{14}$
564	615	1,896	AT ATATTATCAACAACAATAGAAGATTGGCGAAATTAGAGATAGAATTATT T_{12}
567	615	360	AT ATATTATCAACAACAATAGAAGATTGGCGAAATTAGAGATAGAATT T_{09}
562	603	267	ATATT ACAACAACAAGAAATCAATGAAGTCAGAGATAAAGTTATTAA T_{15}
564	596	173	TAGATATCAACAACAT CAGAGAATCAATGAAACTAGAGATAGAGTTATT T_{10}
567	611	1,316	ATACA TATCAACAACGACAAGAGATCAGTGAAATTAGAAGTAAAGTT T_{13}*
564	615	1,043	AT ATATTATCAACAACAATAGAAGATTGACGAAATTAGAGATAGAATTATT T_{10}
564	611	754	ATACA TATCAACAACGACAAGAGATCAGTGAAATTAGAAGTAAAGTTATT T_{09}*
567	615	703	AT ATATTATCAACAACAATAGAAGATTGACGAAATTAGAGATAGAATT T_{15}
563	611	188	ATACA TATCAACAACGACAAGAGATCAGTGAAATTAGAAGTAAAGTTATCA T_{06}
564	611	146	ATACA TATCAACAACGACAAGAGATCAGTGAAATTAGAAGTAAAGTTATC T_{12}
568	603	131	ATATT ACAACAACAAGAAATCAATGAAGTCAGAGATAAAGT AT_{08}
584	630	831	ATAAT TAACAAACAAATATGATATTATCAGTGACAGTGAGAAATTGATA T_{15}
596	642	20	ATA TATAACAATCCATAGCAGATAGACGTGATATTATTGATGATAGT TTAAAT$_{18}$
596	640	512	AAAT TAACAATCCATAACAAGTGAGCGTGATATTGTCAGTGATAAT T_{11}
596	648	159	ATAAATCATAACAATCTATAATAGACGAGCGTGATATTGTCAATGATAAT T_{10}
596	642	32	ATA TATAACAATCCATAGCAGATAGACGTGATATTATTGATGATAAT T_{12}
629	670	1,258	ATAT GATAAACGATTACCTACAGATAATGAGTCATAGTGATTTATA T_{13}
632	674	463	ATAT ATAAAATAAACGATTACCTGTGAATGATAGATTATGATGATTT T_{11}
630	674	419	ATAT ATAAAATAAACGATTACCTGTGAATGATAGATTATGATGATTTAT T_{26}*
630	669	299	ATACAT ATAAACGATTACTCATAGATAGCAAGTCATAGTGATTTAT T_{08}*
630	671	101	ATTT AAATAAACGATTACTTACGAGTGACAGATTGTGATGATTTAT T_{14}
629	671	65	ATATTT AAATAAACGATTACTTACGAGTGACAGATTGTGATGATTTATA T_{15}
656	699	59	ATATAT ATGACAAACTACGTAAGTGCAGATAAAGTAAGTGATTATTT T_{12}
666	701	2,991	ATAT AGATGACAAACCATGTAGACGTGAATAAGATAG TGATTGATACAT$_{12}$*
679	727	2	AAAAAA AAAAACTAAATCATATAAATTAAAGGGTGATGAACTGTGTAAAT T_{13}
679	711	1	ATAAGTCAGAAAGTGATAGATCGTGTAAAT T_{15}
679	725	1,728	AAATTAAATCATATAAGTTGGAAGATGGCAGATTGTGTGAAT TGTAGT$_{16}$
679	727	373	AAAAA AAAAACTAAGTCATATAAGTCAGAAAGTGATAGATCGTGTAAAT TTT
711	758	3,229	ATATAT ACGAGACAAAATATCACTTAGATTATTAGAGATTGAGTTATA AT_{13}
710	741	22	AA TTAGACTATTAGAGATTGAGTTATAT TTT
709	741	2	AAAACGAGACAAGATACCAA TTAGACTATTAGAGATTGAGTTATATA T_{14}
725	764	55,757	ATATAA TAACGAACGAGGCAGAGTATCATTTAGACTATTAGA TTCAAT$_{14}$
731	759	2,867	ATATATAAT GACGAGATAAGACATCACTTAGACTGT AGAGAT$_{14}$
722	765	8,175	ATATA TTAACGAACGAGATAGAACATTGCTTGAGTTATTGAGAAT AT_{14}*
719	765	552	ATATA TTAACGAACGAGATAGAACATTGCTTGAGTTATTGAGAATT T_{12}
735	765	115	ATA TTAACGAACGAGATAGAACATTGCTTGAGTT TT

756	794	48,263	AT ACTAAATAAACGACGATCTTATACTGTATCTGATGAATG TGATATTAAT$_{14}$
757	794	318	AT ACTAAATAAACGACGATCTTATACTGTATCTGATGAAT ATGATATTAATTTT
751	794	174	AT ACTAAATAAACGACGATCTTATACTGTATCTGATGAATGGGATA TTAAT$_{14}$
755	794	167	AT ACTAAATAAACGACGATCTTATACTGTATCTGATGAATGA TATTAATTGT$_{07}$
756	794	2,829	AT ACTAAATAAACGACGATCTTATACTGTATCTGATGAATG TGATAT$_{14}$*
772	806	367	ATATA TCATAACAACTAGATAAACGATGATCTCACA GTATAGTTAAT$_{12}$
782	816	77	ATAACTTATAACAACTAGATAGACGG GAATCTTATATTGTAGTTAAT$_{16}$
778	830	25,793	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTGAATGATGAT AT$_{12}$
777	822	5,729	ATATAT TAAAATACAACTTATGATGACTAAGTGAATGATGATT CAAT$_{10}$
775	830	2,644	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTGAATGATGATTTT T$_{08}$
781	833	2,373	AAATA ATAACAAAACATGAGATATAACTTGTAGTGATTAGATGAATGAT T$_{11}$
780	833	2,079	AAATA ATAACAAAACATGAGATATAACTTGTAGTGATTAGATGAATGATA T$_{13}$
778	833	903	AAATA ATAACAAAACATGAGATATAACTTGTAGTGATTAGATGAATGATAGT AT$_{13}$
778	830	450	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAGGTGAATGATGAT ATTAAAT$_{13}$
778	830	331	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTGAACGATGAT ATTAAAT$_{14}$
778	830	134	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTAAATGATGAT AT$_{07}$
792	845	98	ATAT AAAACAATAATCATGATGAGATATAAAGTGCAGTTTGTGATAATT T$_{10}$
790	839	52	ATAT ATAATCATAACAAGATGTAGAGTACGATTTATAGTGATTAA T$_{12}$
790	839	343	ATAT ATAATCATAACAAGATATAGAGTACGATTTATAGTGATTAA T$_{11}$
792	844	190	ATAT AAACAATAATCATGATGGGATATAAAGTGCAGTTTGTGATAATT T$_{10}$
790	839	99	ATAT ATAATCATAACAGAGCATAGAATACAGTTTATAGTGATTAA TTGT$_{05}$*
790	844	89	ATAT AAACAATAATCATGATGGGATATAAAGTGCAGTTTGTGATAATTAA TAT$_{05}$
843	877	26,437	ATAT ATAAACGGTCAAATGTATTACTTATAAGATAGAG TTGATAAT$_{14}$
834	877	1,279	AT ATAAACGGTCAAATGTATTGCTTGTAAGGTAGAGGTGATAATT T$_{13}$
829	865	188	ATATATAACGATC AATGCATTATCTATGAAGCAAGAACAGTGATTATAAT T$_{15}$
831	857	121	AATGTGTG ACCTATAAAGTGAGAATAATGATTATA T$_{12}$
832	865	23	ATATATAACGATC AATGCATTATCTATGAAGCAAGAACAGTGATTAT TTTT
832	857	20	AATGTGTG ACCTATAAAGTGAGAATAATGATTAT T$_{11}$
834	879	1	ATAT AAATAAACGATTAGATGTATCACTTATAGAATAAAAATGATGATT T$_{15}$
863	902	2,491	ATATA ATATGCATATCAGATAGACGTAGAAATAAGTGATTAAAT T$_{10}$
866	902	126	ATATAATAA ATACGCATATTAGATGAGTGTAAAAGTAAGTGATTA T$_{10}$
872	916	3	ATATAA AATCAACAAATTCGTATGTATATCGAGTAAATGTAGAAATAAAT T$_{09}$
872	919	3,227	A TACAAATCAACAAATTCGTGCGTATATTAAGTGGGTGTAGAAGTGAAT T$_{17}$*
867	919	528	A TACAAATCAACAAATTCGTGCGTATATTAAGTGGGTGTAGAAGTGAATGATT T$_{15}$
872	919	295	A TACAAATCAACAAATTCGTGCGTATATTAAGTGAGTGTAGAAGTGAAT TAAAT$_{19}$
871	919	142	A TACAAATCAACAAATTCGTGCGTATATTAAGTGGGTGTAGAAGTGAATG T$_{15}$
901	939	1,365	ATATAT ACTAATAAAAAGGCATTGCTTACAGATTGATAGATTTAT T$_{12}$
899	931	292	ATATATTACCAACAT AGAGACATTGCTTATGAGTTAACAGATTTATAT T$_{12}$
901	939	249	ATATAT ACTAACAAAAAGATATTGCTTATAGATCAGTGAATTTAT T$_{11}$
907	947	24	ATAAAGAAACCAACAGAAGAATATTGCTTGTAAGTTAATGA TCTTGTAT$_{10}$
907	951	64	CCAAGGCA AAAGATAAAGAAACCAACAGAAGAATATTGCTTGTGAGTTAATGA TCT$_{17}$
932	978	44,852	ATATAT TCAAACAAACAGATAGAACCGGAGACGAGAAGATTGATAA T$_{13}$
934	988	11,706	ATATAAATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTAAT T$_{12}$
931	978	3,513	ATATAT TCAAACAAACAGATAGAACCAGAGACGAGAAGATTGATGAA T$_{12}$
933	986	2,467	ATAAATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTAATA T$_{12}$
940	978	182	ATATAT TCAAACAAACAGATAGAACCGGAGACGAGAAG CTTGATAAT$_{05}$
936	988	175	ATATAAATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTA TTAT$_{15}$
941	978	137	ATAT TCAAACAAACAGATAGAACCGGAGACGAGAA TATTGATAATTTAAT$_{14}$
937	988	97	ATATAAATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATT TAT$_{13}$
934	983	18,961	ATATAT AATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTAAT T$_{12}$*
933	983	1,729	ATATAT AATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTAATA T$_{05}$
936	983	472	ATATAT AATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATTA T$_{14}$
937	983	397	ATATAT AATAATCAAACAGACGAATGAAACTAGAGATAGAGAAATT T$_{11}$

5'	3'	Reads	ND7 gRNA Sequences cont.
959	1000	173,548	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGATGAAA AT$_{14}$

952	1000	748	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGATGAAATT T_{09}
960	1000	472	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGATGAA T_{13}
966	1000	412	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATA TATGAAAAT$_{13}$
964	1000	372	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGA AAAAAT$_{14}$
963	1000	371	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGAT T_{15}
961	1000	265	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGATGA T_{12}
962	1000	234	ATAAA GGTAATATCACAGTGTAGATAGTCGGATAGATAGATG TAAATTTAT$_{09}$
959	1000	3,928	ATATAAA GGTAATATCACAGTGTAGATAGTCGGATAAATAGATGAAA AT$_{11}$
983	1017	101	ATATATAAATAACATAT TAATGGTCTTGATGGTGATATTGTAGTATAA T_{12}
1001	1032	3	ATATA TATAAAATAACGTGATGATGGTCTCGAT AGTAATTGATAAT$_{08}$
1000	1043	1	ATAT ACACCACAAACTGTAGAATGACGTGATAGTGGTTTCAGTG ATAAT$_{15}$
1015	1043	12	AT ACATCACAAACTATAAAGTAATATGATAA GGGTTTCGAT$_{15}$
1013	1043	208	ATAT ACACCACAAACTGTAGAATGACGTGATAGTG AT$_{18}$
1032	1067	6,797	ATATATACAAGC AATGATGTACTCGGTAAATAGTGACACTGTGAATT T_{12}
1030	1067	2,406	ATATATACAAGC AATGATGTACTCGGTAAATAGTGACACTGTGAATTAT T_{12}
1032	1078	751	A AATATAAGCAAATGATGTATTCGGTGAGCAGTGATATCGTAAATT T_{10}
1057	1085	44	ATTAGAAA GGTGTTCGGTATAGGTAGATGATATAT AT$_{10}$
1055	1085	40	CTCATTAGAAA GGTGTTCGGTATAGGTAGATGATATATTT TTT
1089	1121	1	AT ATATGATAACAAACAATACTTACTTT GAGGTGTTTTGTGAGTAAAT$_{14}$
1087	1113	412	ATA AATAACAAACAATATTCGTCTTTG AGATATTCGATATAAGTAAAT$_{12}$†
1099	1143	81,287	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAGTGATAT AT$_{14}$
1094	1143	6,615	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAGTGATATTT T_{10}
1099	1142	5,793	ATATAAA AGAACATAAACTGACACGAGGGTATAGTGATGAATGATAT AT$_{13}$
1099	1131	705	ATATAAGAGAACATAAA CAGCATAGAGGCATAGTAACAAGTGATAT AT$_{14}$
1101	1143	550	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAGTGAT T_{13}
1105	1143	520	TATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAG GGATATAT$_{13}$
1106	1143	318	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAA T_{11}
1108	1143	303	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAAC T_{11}
1094	1142	289	ATATAAA AGAACATAAACTGACACGAGGGTATAGTGATGAATGATATTT T_{10}
1099	1143	259	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAGCGATAT ACT$_{12}$
1107	1143	177	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACA T_{12}
1099	1143	163	ATATAA GAGAACATAAACTAGCATAGAAGCATAGTAACAAGTGATAT AT$_{05}$
1099	1143	162	ATATAA GAGAACATAAACTAGCATAGAGGCATAGTAACAAATGATAT AT$_{05}$
1099	1143	128	ATATAA GAGAACATAAACTAACATAGAGGCATAGTAACAAGTGATAT AT$_{14}$
1099	1143	98	ATATAA GAGAACATAAACTAGCATAGAGACATAGTAACAAGTGATAT AT$_{14}$
1099	1143	356	ATATAA GAGAACATAAATTGACATGGAAGCATAGTAATAAGTGATAT AT$_{12}$*
1095	1143	212	ATATAA GAGAACATAAATCAGTGCAAAGGTATAGTAGTGAGTGATATT AAT$_{09}$*
1108	1145	569	ATATAAACGTAT ACGAGAGCATAGATCAGTGTGAGAATGTAGTAAT T_{14}
1094	1148	295	ATATAAAA TAAACGAGAATATAAACTGATGTAGAGATATAGTGATAAGTAATATTT T_{08}
1107	1157	82	AT ATGTAAATGTAAACGAGAATATAGATTGATGTAGAGATATAGTAATA T_{13}
1107	1146	90,559	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATATGATAATA T_{14}
1108	1146	10,003	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATATGATAAT T_{14}
1111	1146	2,224	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATATGAT T_{10}
1114	1146	2,220	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATAT T_{14}
1110	1146	1,887	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATATGATA T_{16}
1108	1154	978	ATATA TAAACGTAAATGAGAATATGAATCAGTGTGAAAATGTAATAAT T_{07}*
1107	1146	400	ATTAAACGT AACGAGAATGTGAACTGACATAGAGATATGATAATG T_{13}
1120	1146	266	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGAT T_{13}
1113	1146	243	ATATTAAACGT AACGAGAATGTGAACTGACATAGAGATATG TTTT
1108	1152	232	ATAT AACGTAAACGAGAGCATAAATTGATGTGAAGATGTGATAAT TTCTTTT
1128	1167	3,079	ATATAT AATCCGTACAATGCGAACGTAGACGAGAATATGAGTTAAC T_{14}
1136	1182	1,834	ATAT ATAAATATGCAAGAAATCTGTATGATGTAGATGTGAATGAGAATAT T_{10}
1138	1182	866	ATAT ATAAATATGCAAGAAATCTGTATGATGTAGATGTGAATGAGAAT T_{11}
1128	1167	337	ATATAT AATCCGTACAATGCGAACGTAGACGAGAATATGAGTTAAT TTT
1131	1167	203	ATATAT AATCCGTACAATGCGAACGTAGACGAGAATATGAGTT T_{19}
1150	1197	213	AT ATAAACATCCAATAGACGAGTATGTGAGAGATTTGTATGATGTAAAT T_{10}
1150	1195	1,077	AAATATCCAATAAACAGATATGTAGAAGGTCCGTATAATGTGAAT ATAT$_{13}$

5'	3'	Reads	ND7 gRNA Sequences cont.
1181	1218	4,103	ATATATAA ATGCAATAGAAGATCACGCAAATAGATATCTGATAAAT T_{13}
1183	1224	1,399	ATA TAAATCATGCAGTAAAAGACTATGTAGATGGATATTCAGTGA TAAT$_{14}$

1183	1223	1,017	AAATA AAATCATGCAGTAGAGAACCGTGTAAGTGAGTATCTGATAA TTAT$_{11}$
1183	1224	138	ATA TAAATCATGCAGTGAAGAGCTACGTAAATGGATATTCAGTGA TAAT$_{11}$
1183	1224	2,561	ATA TAAATCATGCAGTAAAAGACTATGTAGATGAATATTCAGTGA TAAT$_{14}$*
1210	1257	123	ATAT ACAACATCAATATTACTTAGAACGGTAACTAGATTGTGTAATAA T$_{14}$
1233	1257	167	ATAT ACAACATCAATATTACTTAGAACG ATAACTAGATTGTGTAAT$_{10}$†
1233	1257	83	ATAT ATAACATCAATATTACCTAGAGTG TCAGTTAGATTATGTGATAAAT$_{14}$†
1240	1268	95	AAACTAACGATATT CGGATCTGAGAGTAACATTGATATTATTT T$_{07}$
1251	1282	63	ATATATAT AACTAACGATCTATGGGTTTAAAGACAGTGT GAAT$_{12}$
1242	1283	6	AC AAACTAACGATTTACGGATTTAGAGACAGTGTTAATGTTAT AT$_{13}$
1242	1270	296	A TACGGATTCAGAAGTGATATTGATGTTAT AT$_{15}$
1240	1268	216	AAACTAACGATATT CGGATCTGAGAGTAACATTGATATTATTT T
1240	1270	54	ATATAAACTAACGA TACGGATTCAGAAGTGATATTGATGTTATTT T$_{16}$
1239	1268	25	ATT CGGATCTGAGAGTAACATTGATATTATTTA T$_{17}$
1240	1268	23	ATT CGGATCTGAGAGTAACATTGATATTGTTT TT
1241	1268	12	GATATT CGGATCTGAGAGTAACATTGATATTATT AT$_{15}$
1269	1320	6	ATA TAAACAATCCTACAATGATCTCGTGTATAAGACTGATGATTTA AT$_{12}$
1269	1320	1,074	ATA TAAACAATCCTACAATGATTTCGTGTATAAGACTGATGATTTA AT$_{17}$

I) NADH Dehydrogenase 8

5'	3'	Reads	ND8 gRNA Sequences
34	58	2	GTGGG ATATGAAAGTAAGAGAATAAAAAAA ATTAAT$_{13}$
29	68	1	AA AAAAAAAAACATACAGAAATAGAAAGATAAGAAAGTGATA TATTAT$_{08}$
28	56	9	AAA ATAGGAGTAGGAGGATGAGAAAATGATAG AGGGATTTT*
55	98	2,577	ATAT AAAACAAACAAAAAGAAGAAACAAGAAATTGAAGAGAGATATAT T$_{13}$
55	98	274	ATAT AAAACAAACAAAAAGAAGAAGCGAGAAATTGAAGAGAGATATAT T$_{12}$
54	97	236	ATAT AAACAAACAGAAAAAGAAAACAAAGAGTCAGAGAAAGATATATA T$_{17}$
55	97	100	ATAT AAACAAACAGAAAAAGAAAACAAAGAGTCAGAGAAAGATATAT T$_{10}$
57	98	98	ATAT AAAACAAACAAAAAGAAGAAGCGAGAAATTGAAGAGAGATAT T$_{05}$
57	97	65	ATAT AAACAAACAGAAAAAGAAAACAAAGAGTCAGAGAAAGATAT T$_{05}$
54	97	57	ATAT AAACAGACAGAAAAAGAAAACAAAGAGTCAGAGAAAGATATATA TGTAATTATTTT*
59	97	29	ATAT AAACAAACAGAAAAAGAAAACAAAGAGTCAGAGAAAGAT T$_{09}$
87	136	6,039	ATAAATAGTAACACAATGAGCAGAGTACGTATAAGAATGAGTAAA T$_{14}$
84	133	633	A AAATAGTAATACAACAGACAGAGCATATATAGAAATAAGTGAGAAA T$_{16}$
87	133	593	AAATAGTAACACAATGAGCAGAGTACGTATAAGAATGAGTAAA T$_{15}$
87	135	394	TAAATAGTAACACAATGAGCAGAGTACGTATAAGAATGAGTAAA T$_{12}$
84	131	380	AT ATAGTAATACAACAAACGAGATACGTATAGAAATAGATGAGAAA T$_{13}$
87	136	292	GTAAATAGTAACACAATGAGCAGAGTACGTATAAGAATGAGTAAA T$_{12}$
96	136	191	ATAAATAGTAACACAATAGACGAGATACGTGTAGAA TAAGTGATTTAAT$_{11}$
86	139	11,689	ATA TAAACAAATAGTAACATGACGGATAGAACGTATATGAGAATGAGTAAAA T$_{12}$*
85	139	1,938	ATA TAAACAAATAGTAACATGACGGATAGAACGTATATGAGAATGAGTAAAAG T$_{13}$
86	139	1,171	A TAAACAAATAGTAATATGACGAATGAAGCGTATATGAGAATAAGTAAAA TTTT*
87	139	816	ATA TAAACAAATAGTAACATGACGGATAGAACGTATATGAGAATGAGTAAA TTTAT$_{14}$
88	132	612	ATAT AATAGTAACACAACGAATAGAACATGTATAGAGATGAATGA TAT$_{17}$
84	131	598	ATAT ATAGTAACACAATAAATGAGACATATATGAAGATGAATGAGAAA TTTT*
92	133	267	ATATA GAATAGTAACACAGCAGATAAGATACATATAGAGATAA TGACAGT$_{07}$
92	132	267	ATATAT AATAGTAACACAGCAGATAAGATACATATAGAGATAA TGACAGT$_{14}$
86	138	221	AA AAACAAATAGTAATATGACGAATGAAGCGTATATGAGAATAAGTAAAA T$_{06}$
85	138	196	AA AAACAAATAGTAATATGACGAATGAAGCGTATATGAGAATAAGTAAAAA T$_{10}$
111	153	70,659	ATATATAAATGGTA AACTCAATGGGTGGATAAGTAGTAATGTGATGAAT T$_{13}$
111	153	479	ATATATAAATGGTA AACTCAATGGGTGGATAAATAGTAATGTGATGAAT TTT
111	153	245	ATATATAAATGGTA AACTCAATGGGTGGATAAGTAGTAATGTGATAAAT T$_{12}$
111	153	147	ATATATAAATGGTA AACTCAATGGGTGGATAAGTAGTAATATGATGAAT TGTAATAT$_{15}$
115	153	127	ATATATAAATGGTA AACTCAATGGGTGGATAAGTAGTAATGTGAT T$_{14}$
113	153	121	ATATATAAATGGTA AACTCAATGGGTGGATAAGTAGTAATGTGATGA T$_{15}$
111	152	804	A TATACAATGGTT ATTCAGTGGGTAGACAGATAGTGATATGATAGAC TTAT$_{12}$
117	170	4	ACAT ATAAACTAACAATGGTTGATTTAGTGAGTGAATGAGTAGTAATATT T$_{11}$
158	186	169,806	ATAT GAACGCAAAGATGGATTACCACGAGTTAGTAAATTGATGAT AT$_{14}$
161	187	124,070	ATATAAT AAACGCAAAATATGGTTACTATGAACTGATAGATTAAT T$_{12}$
161	187	799	AAACGCAAAATATGGTTACTATGAACTGATAGATTAAT T$_{11}$
161	198	66	ATATAAT AAACGCAAAAAATGGTTACTATGAACTGATAGATTAAT T$_{14}$
161	207	2,257	ATATA TAATAAAGACGCAAAAGATGGTTACTGTGAATTGATGAGTTAAT T$_{12}$*
161	199	966	ATATATAGT AAAACGCAGAAAATGGTTACTATGGATTGATGAATTAAT T$_{16}$
160	199	804	ATATAGT AAAACGCAGAAAATGGTTACTATGGATTGATGAATTAATG T$_{15}$*
186	230	24,763	ATA CAATACAACGCTCTGAATCATATCGATAAAAGTGTGAGAAAT T$_{13}$
186	229	182	AATACAACGCTCTGAATCATATCGATAAAAGTGTGAGAAAT TTAAT$_{12}$
186	230	109	ATA CAATACAACGCTCTGAATCATATCGATAAAAGCGTGAGAAAT T$_{11}$
186	220	204	ATA CAATACAACACTCTGAATCATATCGATAAAAGTGTGAGAAAT TTAAT$_{10}$
194	230	160	ATA CAATACAACGCTCTGAATCATATCGATAAAAGTG GGAGAAAT$_{11}$
195	230	149	ATA CAATACAACGCTCTGAATCATATCGATAAAAGT TGAGAAATTTAAT$_{11}$
196	230	199	ATA CAATACAACGCTCTGAATCATATCGATAAAAG AGTGAGAAATTTAAT$_{05}$
186	228	161	T ATACAACGCTCTAAATTATACCAGTGAAAATGCGAGAAAT T$_{14}$
187	228	52	ATATTAT ATACAACGCTCTAAATTATACCAGTGAAAATGCGAGAAA AT$_{11}$*
189	229	31	ATATA AATACAACGCTTTAGATCATATCAGTGAGAGTGTGAAA T$_{13}$
213	244	173	A ACATAAACGACAGGTAATATGATGTTCTGAAT GATATCGATAAT$_{06}$
209	239	19	ACAT AACGACAAGTGATATAACGTTTTAAGTTACA GTGATAAT$_{19}$
219	245	10	AAATA TACATAAACGATGAGTAATATGACGTT GTAGAT$_{13}$
219	254	1	AAATTAAATCACATAAATGACGAGCGATACAGTGCT GTAGATGATACT$_{06}$

5'	3'	Reads	ND8 gRNA Sequences cont.
240	288	3	TATGAACATCCGATACTGAACTAGGGTAGATTGAGTTATATA T_{10}
237	267	2	CATCCAATGTAT AATTAGGGTAGATTGAGTTATGTAAAT T_{15}
246	271	340	ATACGAACATCCATA GTTAGATTAGGGTAAATTGAGT GATGTAAT$_{12}$†
259	285	325	ATATAA GAACATCCGATGCTAGATTA TGGTAGATTGAGTTATATGATTGTAATAT$_{05}$†
270	291	268	ATATA TAACACGAACATCTGATGT ATAAACTGAAGTAAATTGAT$_{16}$†
276	318	51,820	ATATAC AACGATGATCACTGAGATTTTACCTAATATGGATGTT AAT$_{13}$*
276	319	28,492	ATATAT AAACGATGACTACTAGAATTCTACTCAATGTGAATGTT AAT$_{14}$
277	317	1,829	ATATAT ACGATGACTACCAAAATTCTATCTGATATGAATGT GATAATAT$_{11}$
279	318	594	ATATAC AACGATGATCACTGAGATTTTACCTAATATGGAT T_{17}
275	318	384	ATATAC AACGATGATCACTGAGATTTTACCTAATATGGATGTTT T_{09}
279	319	334	ATATAT AAACGATGACTACTAGAATTCTACTCAATGTGAAT T_{15}
276	314	215	GATGACTACTAGAATTCTACTCAATGTGAATGTT AATTTATGATAT$_{14}$
276	311	186	ATATACAACT ATGATCACTGAGATTTTACCTAATATGGATGTT AAT$_{11}$
284	321	140	ATATATA CAAAACGATGATTACCGAGATTTCATTTAATATGA TTGTCTAATTTT
287	319	139	ATATAT AAACGATGACTACTAGAATTCTACTCAATG AATGTTAATTTT
275	319	109	ATATAT AAACGATGACTACTAGAATTCTACTCAATGTGAATGTTT T_{12}
289	318	4,761	ATAATAT AACGATGACTATCAGAACTTTACTCGA GATGAATGT$_{13}$
276	318	1,264	ATATAT AACGATGACTACTGAGACTCTATCTGACGTGAATGTT AAT$_{12}$
275	318	175	ATATAT AACGATGACTACTGAGACTCTATCTGACGTGAATGTTT T_{07}*
310	353	13,915	ATATTA GATGATAACTCAGTGTAGATTGATCTGTAGAATGAT AATAT$_{13}$
301	344	636	ATAT ATAACTCAATGTAGATCGATTTGTGAGATGATGATTGTCAA T_{11}
310	350	312	ATGATAACTCAGTGTAGATTGATCTGTAGAATGAT AATATATAATGT$_{05}$
316	344	266	ATATACA ATAACTCAATGTGAATTAATCTGTGAGAC TAT$_{14}$
301	344	4	AT ATAACTCAATGTAGATCGATCCGTAGAATGATGATTGCTAA T_{11}
331	364	1,709	ATAT ATAAATACAACGGTGATAATTCGATGTGA TGATATCTGTAAT$_{16}$
325	358	137	ATAA ATAACGACGATGATTCAGTGTAAATTGGT ATGTGAAATGATGT$_{12}$
325	355	21	ACGACGATAGCTTGATGTAGGTTAGT GTGTGAGATAATGATAT$_{17}$
338	382	1,964	ATATAA ATGCATACAAGAATCATAGTAAGTACAGTGATGATAATTT T_{09}
350	386	1,269	ATATA AAACATGTATACAAAAATTACAGTAAATGCGACGA AAATAAT$_{13}$
339	382	1,073	ATATAA ATGCATACAAGAATCATAGTAAGTACAGTGATGATAATT AT$_{14}$
341	373	13,719	ATATATAAATGCATAC AAGGCCATGATAGATACAATGATGATAA AT$_{11}$
338	385	2,361	ATGTAC AACATGCATACAGAGACTATAATAGATACAGTGATGATAATTT TTATTTT
342	385	159	ATGTAC AACATGCATACAGAGACTATAATAGATACAGTGATGATA T_{19}
343	385	133	ATGTAC AACATGCATACAGAGACTATAATAGATACAGTGATGAT T_{07}
372	417	2	ATAT ATAAATGCGTAATGGTATCTTTCGGGTAGATATGTGTATAAA T_{14}
390	414	894	AAATTGTATATAT AATGCGTAATGGTGTTTGTTG AGCGAGTGTGTGTTAT$_{15}$†
391	431	23	AT ATATATAACAAACAATGGGTGTATAATGGTATCTGAT TGTGCATTAATTTAAAAT$_{14}$
405	434	9	AT AAAACACATAATAAATGATGAGTGCGTGAT ATAGGTATAAT$_{13}$
411	442	1,950	AAAAA AAACAACAAGAACACATAGCAGATAGTGGGTG ACGTATAT$_{12}$
411	443	923	A TAAACAACAAGAACACATAGCAGATAGTGGGTG ACGTATATGAT$_{13}$
407	441	215	AAAT AACAACAAAAACATATAATAAATGATGGATGTGTA TAAGGTATAT$_{15}$*
426	466	24,242	ATAAAACTTGGA CGCTAATAGATATATGGTTAGATAATGAGAATATAT T_{13}
425	466	5,218	ATAAAACTTGGA CGCTAATAGATATATGGTTAGATAATGAGAATATATA T_{14}
423	466	2,651	ATAAAACTTGGA CGCTAATAGATATATGGTTAGATAATGAGAATATATAGT TAAT$_{11}$
428	466	1,071	ATAAAACTTGGA CGCTAATAGATATATGGTTAGATAATGAGAATAT T_{10}
423	466	203	TAAAACTTGGA CGCTAATAGATATATGGTTAGATAATGAGAATATATGGT TAAT$_{14}$
426	482	13	AT AAAACTTGGGCGCTAATAGATATATGGTTAGATAATGAGAATATAT T_{12}
423	462	29,495	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGATATATAAT T_{13}
426	462	11,472	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGATATAT T_{18}
425	462	7,969	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGATATATA T_{12}
428	462	4,922	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGATAT TCTTTT
430	462	582	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGAT T_{14}
426	464	526	ATATATAACTTGGGCA TCAATGAGTATATGGTTAAGTGATGAAGATATAT T_{10}
426	480	127	ATATAT AACTTGGGCGTCAATGAGTATATGGTTAAGTGATGAAGATATAT TTT
427	462	111	ATACTTGGGCACA CAATAGATATATGGCTGAATAATAGAGATATA ATCTTTT
465	503	1	ATAT TAAAACAACAATCAAATGATAGAAGCTTGGGTG ACTGTGAATATTTT
477	512	176	ATATA TAAATAACATAAGACGATAACTGAATAATAGAAATT AAGTGTCAAT$_{14}$
482	510	832	ATATATT AATAACATAGAGCAATGACCGAGTAATGA TAT$_{12}$
477	512	237	ATATA TAAATAACATAAAACGATAACTGAATGATAGAGATT AAGTGTCAAT$_{09}$

5'	3'	Reads	ND8 gRNA Sequences cont.

489	531	21,576	ATAG ATAAAACACAAATAAAGGTCAAGTGATATAGAGTGATGATTAA T_{14}
489	528	5,835	ATAAATAT AAACACAAATAAAGGTCAAGTAATGTAGAGTGATAATTAA T_{12}
500	540	2,861	ATATAT AATAACTACACGAGACATGAATAGAAATTAAGTGATGTAAA T_{12}
494	536	1,242	ATATAT ACTACATAAAACACAGATAAGAATCAGATAGTGTGAGATAATA T_{15}
495	539	884	ATAT ATAACTACACAAAATATAGGTAAAAGTTAGATGATGTGAAATGAT T_{11}
495	536	151	ATATAT ACTACATAAAACACAGATAAGAATCAGATAGTGTGAGATAAT T_{15}
494	539	1,101	ATAT ATAACTACACAGAACATAGATAAGAGTCAGATAGTATAAAGTGATA T_{19}
495	539	405	ATAT ATAACTACACAGAACATAGATAAGAGTCAGATAGTATAAAGTGAT T_{15}*
500	541	198	ATA AAATAACTACACAGAATACGAGTAAAGATTGAATGATGTAAA TAAT$_{18}$
495	539	127	ATAT ATAACTACACAGAGCACAAATGAAAGTTAAGTAATGTGAAATAGT T_{23}
523	567	413	ATAAA TATAAACACAGTAAATCACTCGAGATAGATAGTTATATGAGATAT T_{11}
554	598	20	ATATA TAATTTCACCGTGAATTTCTTTAGATTGTAGATATGATAA T_{13}
554	598	7	ATATA TAATTTCACCGTGGATTTCTTTAGATTATAGATATGATAA T_{06}
554	598	5,311	ATATA TAATTTCACCGTGGATTTCTTTAGATTGTAGATATGATAA T_{05}
554	598	386	ATATA TAATTTCACCGTGGATTTCTTTAGATTGTAGATATGGTAA T_{15}*

J) NADH Dehydrogenase 9

5'	3'	Reads	ND9 gRNA Sequences
33	71	18	ATAT AAACATAAACGAAATGAGCATAGAAGTATATGTATGATG ATATTAT$_{14}$
29	72	93	ATAT AAAACATAAACGGAATAAATATAAGAGTGTATATGTGATATAAT T$_{15}$
25	72	32	ATAT AAAACATAAACGGAATAAATATAAGAGTGTATATGTGATATAATGTTT T$_{11}$
60	105	540	A TATAACACAAACAATAGAATAAAGTTAAGTGAGAATATGAGTGA TTTAT$_{11}$
60	101	472	ATAT ATACAAACAATAAAATAGAATTAGATAGAGGTATAAGTGA TGTAAAT$_{11}$
87	124	1	ATATCAAC AAACAAATAGAGCACTGTCTATGATATAAGTGATAA TTCT$_{09}$
87	124	536	AAATATCAAC AAACAGACGAGACATCATCTACAGTATAAGTGATAA TAT$_{10}$
87	124	453	AAATATCAAC AAACAAATAGAGCACTGTCTACGATATAAGTGATAA T$_{05}$
117	160	1,145	ATAT ATAAAACAATAAAGAAATAGAAGGCTACAGTTAATGAGATAAAT T$_{13}$
130	167	137	AAAATTAACAAAACAATAAAAGAACGAGAAATTACAGT GAATGAAGTAAAT$_{07}$
117	162	906	ATA TAACAAAACAATAAAGAAATGAGAAACTGTGATTGATAAGATAAAT T$_{12}$*
119	162	65	ATATA TAACAAAACAATAAGAAGACAGAGAGTTACAGTTAATAAGATAA TTAT$_{19}$*
116	162	55	ATA TAACAAAACAATAAAGAAATGAGAAACTGTGATTGATAAGATAAATA T$_{05}$
121	162	30	ATA TAACAAAACAATAAAGAAATGAGAAACTGTGATTGATAAGAT T$_{11}$
149	193	1,196	ATAT AATAAAAACATACAATAAAATAGAAGGAAACTAATAAGATGATAG T$_{15}$
147	187	5,984	ATATAT AACATACAATAAGACGAAGAGAAACTAATAGAGTGATAAGA T$_{15}$*
150	193	463	ATAT AATAAAAACATACAATAGAATGAAAGGAAACTAATAAGATGATA TAT$_{11}$
176	216	3,869	ATACAATAT AAAACAAAAGTCACAGATTAAGGGATAGAGATGTGTGATAA T$_{14}$
177	216	357	ATACAATAT AAAACAAAAGTCACAGATTAAGGGATAGAGATGTGTGATA T$_{11}$
176	216	620	ATACAATAT AAAACAAAAATCGTAGATTAAAAGATAGAGATATATGATAA T$_{14}$
176	221	63	ATAT ATATAAAAACAAAAGTCACGAATTAGAAAGTAAAGATATGTGATAA TTTT
201	242	87	ATATAA AATCAATAACAGATCATGATGATATAGAGATAGAAGTTATAA T$_{15}$
204	249	652	AAA AAAAATCAATCAATGATAGATCACAGTGGTATAGGGATGAGAATTA AT$_{11}$
205	249	485	AAAA AAAAATCAATCAATGATAGATCACAGTGGTATAGGGATGAGAATT T$_{11}$
202	248	101	ATAT AAAATCAATCAATAGCAAGTTACGATAATATAGAAGTGAGAATTATA T$_{13}$
203	248	94	ATAT AAAATCAATCAATAGCAAGTTACGATAATATAGAAGTGAGAATTAT T$_{13}$
205	248	77	ATAT AAAATCAATCAATAGCAAGTTACGATAATATAGAAGTGAGAATT T$_{10}$*
203	249	58	AAAAA AAAAATCAATCAATGATAGATCACAGTGGTATAGGGATGAGAATTAT T$_{05}$
239	279	569	AAATTAT ACAACATAAGACGACGAAGATAAAGGTCATAGAGATTAATT T$_{12}$
239	268	561	AACATAAAT CGGTGAAGATAGAGACTATGAGAATTAATT T$_{07}$
240	268	346	ATAAAT CGGTGAAGATAGAGACTATGAGAATTAAT ATGT$_{15}$
238	279	86	AAAATTAT ACAACATAAGACGACGAAGATAAAGGTCATAGAGATTAATTA T$_{11}$
239	288	23	AT AAATATACAACAATATAGAACGATGAAAGTGAAGATTATAAGAGTTAATT T$_{12}$
239	286	26,097	ATATATAACAACATAGAACGATGAGAATAGAGATCATGAAAGTTAATT T$_{11}$*
239	286	7,153	ATAC ATATACAACAATATAGAACGATGAAAGTGAAGATTATAAGAGTTAATT T$_{11}$
238	286	1,828	ATATATAACAACATAGAACGATGAGAATAGAGATCATGAAAGTTAATTA T$_{11}$
243	286	1,812	ATATATAACAACATAGAACGATGAGAATAGAGATCATGAAAGTT T$_{10}$
239	279	553	AAATTAT ACAATATAAGACGACGAAGATAAAGGTCATAGAGATTAATT T$_{10}$*
242	286	499	ATAC ATATACAACAATATAGAACGATGAAAGTGAAGATTATAAGAGTTA T$_{11}$
247	289	451	AAAATATACAACAATATGAAGCGACGAAAATAGAGACTATAGA TATTAT$_{06}$
239	268	300	ATAAAT CGATGAAGATAGAGACTATGAGAATTAATT T$_{13}$*
246	286	262	ATATATAACAACATAGAACGATGAGAATAGAGATCATGAAA T$_{11}$
243	286	236	ATAC ATATACAACAATATAGAACGATGAAAGTGAAGATTATAAGAGTT T$_{17}$
236	282	197	ACAACAATATAGAACGATGAAAGTGAAGATTATAAGAGTTAATTAAT T$_{13}$
272	312	43	ATAA GAACACACAGAGACAAGCAGAGTAAAGTGTATAGTGACATA T$_{06}$
273	314	5,869	ATATAT ACGAACACACAGAAATAAGCAAGGTAGAATATATGATGATAT T$_{13}$*
271	314	250	ATAT ATGAACACACAGAAACAGATGAGATAGAGTATACAGTGATATAA T$_{17}$*
275	314	157	ATATAT ACGAACACACAGAAATAAGCAAGGTAGAATATATGATGAT T$_{12}$
272	314	93	ATAT ATGAACACACAGAAACAGATGAGATAGAGTATACAGTGATATA T$_{14}$
303	339	44,686	ATAT ACAAACAACACAAGATAGAGCACAAATGAATGCACAG TGATAAT$_{13}$
298	343	2,681	ATAAACAAACAACACAGAGCAGAATACAAGTGAGTATATGAGAATA T$_{11}$
298	343	1,741	GTAAACAAACAACACAGAGCAGAATACAAGTGAGTATATGAGAATA T$_{13}$
297	339	270	ATAT ACAAACAACACAAGATAGAGCACAAATGAATGCACAGGGATAA TAT$_{11}$
304	339	139	ATAT ACAAACAACACAAGATAGAGCACAAATGAATGCACA CTGATAAT$_{13}$
305	339	6,792	ATAT ACAAACAACACAAGATAGAGCACAAATGAATGCAC TGAGATAAT$_{15}$
301	338	260	ATATAT TAAACAACACGAAACGAAGCATAGATGGATATATGAGA TTCTTTT

5'	3'	Reads	ND9 gRNA Sequences cont.
319	366	36	AT ATTAAAACACAATCCGAAGAGTACAAATAGACAGTATAAGATAAAAT T_{11}
319	366	17	AT ATTAAAACACAATCCGAAGAGTATAAATAGACAGTATAAGATAAAAT $TAAT_{19}$
343	368	157	AT AAATTAAAACACAATCTAAGAAATA GAGATAGACAGTATAAGATAAAAT$_{08}$†
343	369	45	AAAATTAAAACACAATCTAAGAAATA GAGATAGACAGTATAAGATAAAAT$_{08}$†
352	392	10,040	ATATT AAACGCATAACAGAAATGATTAGAACTGAGATATGATT AAT_{14}
349	392	270	ATATAT AGACGCATAATAAAAGCAACTGAGGCTAGAATATGATTTAA T_{19}
349	392	232	TAT AAACGTATAACAAAAGCAGTTAAAGCTAGAATGTGATTTAA T_{14}
352	392	5,765	ATATT AAACGCATAACAGAGACAATTAGAACTGAGATATGATT AAT_{10}*
349	392	2,338	ATATAT AAACGCATAATAAGGGCAACTGAAACTAGAGTATGATTTAA $TTTCT_{12}$
351	392	691	ATATT AAACGCATAACAGAGACAATTAGAACTGAGATATGATTT T_{19}
356	392	575	ATATT AAACGCATAACAGAGACAATTAGAACTGAGATAT TTT
351	392	189	ATAT AAACGCATAATAAGGGCAACTGAAACTAGAGTATGATTT T_{16}
358	392	167	ATATT AAACGCATAACAGAGACAATTAGAACTGAGAT T_{11}
382	421	1	AATCAAAATATTCGTGTTCTGATGATAGAGATGTATAAT T_{10}
380	418	317	ATATA TAAAACATTCGCGTTCTAATGATAGAGATGTATGATAA T_{12}
409	447	761	ATATAAA ATTACCAACAGAATAGAAGCTAGACAAGTTAAGATATTT T_{08}
409	447	667	ATATAAA ATTACCAACAGAATAGAAGCTAGACAAGTTAAGATATTC T_{12}
410	447	82	ATATAAA ATTACCAACGAAATAAAGACTAAGTGAATTAAGATGTT $ATAT_{05}$
439	484	72	ATAT AAACCAATCAACGAGTAAATGATGTAGAGTATTATTGTCAAT T_{13}
438	483	185	ATATAT AACCAATCAATAAGTGAGCGATGTAAGATGTTATTGTTAATA T_{10}
437	472	130	ATATA TAACAAATAAACGATGTGAGATATCATTACCAGTGA $TTTAAT_{08}$
438	483	46	ATATAT AACCAATCAATAAGTGAGCGGTGTAAGATGTTATTGTTAATA T_{14}
439	483	34	ATATAT AACCAATCAATAAGTGAGCGATGTAAGATGTTATTGTTAAT T_{13}
447	472	16	ATATA TAACAAATAAACGGTGTAGAGTGTCA $GTAT_{16}$
442	483	25	ATATAT AACCAATCAATAAGTGAGCGATGTAAGATGTTATTGTT T_{12}
468	514	234	ATAT ATAACACTTCAACGGAGAGAGACCAATGAGAGATCAGTTAATA T_{13}
469	514	2,106	ATAT ATAACACTTCAACGAGAAGAAGCCAGTGAGAAATCAGTTAAT T_{11}
469	514	1,535	ATAT ATAACACTTCAATAGAAAGAAACTGACAGAGAATTGATTAAT T_{13}
472	514	59	ATAT ATAACACTTCAACGAGAAGAAGCCAGTGAGAAATCAGTT $TTTAATTGT_{14}$
502	549	746	ATATAAAATAACAATACAAGTGAATCAGATAATGGATGATGTTTCAAT T_{13}
509	542	1,407	ATAT ATAACAATACAAATGAACTGAGTAATGGATAGTA GTTTGACAAT$_{14}$*
497	539	935	ATAT ATAATACAAACAAACTGAGTGATGGATAGTGCTTTAATGAGAA TAT_{14}*
501	547	576	ATAGAATAACAATATAAACGAACTAGATGATGGATAGTATTTTGATA TAT_{16}*
501	547	339	ATAGAATAACAATATAAACGAACTAGATGATGGATAGTATTTTGATA TATGCCACG
501	547	55	GTAGAATAACAATATAAACGAACTAGATGATGGATAGTATTTTGATA TATTTT
508	547	36	ATAGAATAACAATATAAACGAACTAGATGATGGATAGTAT CT_{17}
534	566	15	AAACGTACATATG ATCTCTTCTGCTAGTATATAAGATGATAATA AT_{14}
533	566	11	AACGTACATATG ATCTCTTCTGCTAGTATATAAGATGATAATAT T_{11}
535	566	89	ATATG ATCTCTTCTGCTAGTATATAAGATGATAAT $TTCT_{16}$
533	566	72	AAATAAACGTACATATG ATCTCTTCTGCTAGTATATAAGATGATAATAT T_{08}
534	566	24	ATG ATCTCTTCTGCTAGTATATAAGATGATAATA AT_{15}
535	578	13	ATATT AACGTACATACTGTCTTCTCTATTGACATATAAGATGATAAT T_{13}
543	580	39	ATAT TAAACGTACATATTATTTTCTCTACTGATATACAAG $TGATAAT_{09}$
568	604	76	ATATA TATGCAACAACAGAGATAATATTGTAGATGTATATGT $GATATCGT_{11}$
569	600	78	ATATA TAACAACAAAAGTGACATTGTAAACGTATATG ATGAGTTTT
582	612	66	TATT AATTGGTATGTGACAATAGAAGTGATATT T_{09}
582	611	66	AATGCAGATAATT ATTGGTGTGCAATGATAGAGGTAATATT T_{12}*
580	611	17	AATGCAGATAATT ATTGGTGTGCAATGATAGAGGTAATATTGT T_{12}*
597	625	7	A AATGCAAATAGAAGTTGGTGTGCAAT TATAGAGATGATAT$_{09}$
609	644	274	ATATATAAGAATTACAATGGT GTATTAAGTGAAGTAATGTAAATGAGAATT T_{12}
609	640	1	GATA GTTAGATAGAATAATGTAAGTGAAAATT T_{12}
615	659	1	ATATATAA AGGATTACAGTGGTGATATTGAATAGAGTGATGTAAATG T_{12}
618	658	1,673	ATATAT GAATTACAACGGTGATATTGAATAGAATAATGTGA $TTGAAAT_{14}$*

K) Ribosomal Protein Subunit 12

5'	3'	Reads	RPS12 gRNA Sequences
43	78	12,531	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGA TTGATATAT$_{14}$
35	76	2,218	ATA AACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGAT TTAT$_{12}$
43	78	1,663	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGA TTGATATAT$_{12}$
38	78	423	ATAT ATAACAACCATACAGAAATCGTATATGTGAGAGAAGTGA TTTCT$_{15}$
63	109	5,122	AT ATAATATAAAACAAATAAGACAGAGTGTAGATAGTAATTGTATGA TAT$_{12}$
66	99	32	AT ATAAATAAAACAAAACGTAAATGATAACTGTG AGATGAT$_{07}$
74	106	1,212	ATAT ATATAAAACAAATAGAATAGAACGTAGATGAT TACTGTATAAT$_{12}$
73	115	120	A TATATAATAACATAAAACAAATAGAACGAGATGTAAATGATA TCTAT$_{13}$
73	115	1,879	ATA TATATAATAACATAAGACAGATAGAACGAGATGTAAATGATA T$_{21}$
74	115	1,091	ATA TATATAATAACATAAGACAGATAGAACGAGATGTAAATGAT T$_{13}$
77	115	466	ATATA TATATAATAACATAAGACAGATAGAACGAGATGTAAAT T$_{05}$*
79	115	311	ATA TATATAATAACATAAGACAGATAGAACGAGATGTAA T$_{14}$
96	121	233	AATCA CGGATTTATATAGTAACGTAAAATGA TATTAT$_{12}$
92	121	90	AACTGGGC-ATCT CGGATTTGTATAGTGATATAAAGTGAATAA TTTT
96	121	1,542	ATATAAACTGTGCAATCGA TGGACTTATATAGTGATGTAAGATGA TAAT$_{12}$†
96	121	1,066	ATATAGAACTAGGCAGTCA CGGATTTGTATAGTAATGTAGAATGA TAT$_{14}$†
93	121	793	ATATAGAACTGGCAATTT CGGATTTATATAGTGACATGAGATAGATA T$_{15}$*†
94	121	219	ATATAGAACTGGCAATTT CGGATTTATATAGTGACATGAGATAGAT T$_{10}$†
96	131	2	ATATAGAATTA GGCAATCGCGGATTTATATAGTAACGTAAAATGA TAT$_{10}$
119	158	3	ACT TACAATACACGTTGGTTATCGGAGTTAGGTGATTGTG ACTTAT$_{10}$
139	170	44	ATATA ACGGCATATAGTATACGTCGGTTACTAGGATTGTGTAATTTT
132	164	18	CATATAAA GGCATATAGTATACGTCGGTTACTGGGATTGTGTAAT$_{07}$
133	158	56	ATACT TACAATACACGTTGGTTATCGGAGTT AGATGAT$_{13}$
169	208	4,724	ATAAAT AACAACGCAATATCCGAGTAAGATTGTATAAGTGAGATAT AT$_{12}$
164	195	146	ATAACGCAACA TCAGATGAGATTATATAAGTGAGATATG ATATAT$_{11}$
158	207	67	ATAT ATAACGCAACATTCGAATGAGATTATGTAGATGAAATATGGTAT TAT$_{05}$
164	201	896	ATA TAACATCCAAACAAGATTATATAGGTAGAGTATG ATGTATAATTTTAT$_{22}$
156	207	192	ATAT ACAACGTAACATTCAGATAAGATTGTGTAGATAGAATATGGTATAT T$_{12}$
158	207	104	ATAT ACAACGTAACATTCAGATAAGATTGTGTAGATAGAATATGGTAT T$_{06}$
198	235	4,950	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGATGATGTAAT T$_{11}$
194	235	3,025	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGATGATGTAATATTT T$_{11}$
196	235	222	TAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGATGATGTAATAT AT$_{14}$
198	235	444	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGACGATGTAAT T$_{12}$*
194	235	338	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGACGATGTAATATTT T$_{10}$
198	229	36	ATAACTT GACTAATAGAGTAGTGAGAGAGACAGTGTAAT T$_{08}$
200	235	25	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGACGATGTA T$_{15}$
196	235	20	ATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGACGATGTAATAT AT$_{12}$

5'	3'	Reads	RPS12 gRNA Sequences cont.
203	246	341,382	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATAAT AT_{14}
203	246	1,324	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAAATAAT AT_{15}
205	246	1,091	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATA T_{13}
209	246	964	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGA TAATAT$_{12}$
206	246	922	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGAT T_{14}
200	246	609	ATAT ATAATGACATAATTAGACTGATAAGATAACGAGAAAAGTGATGTA T_{12}
203	246	544	ATATAT ATAATGACGTAACTGAGCTAATGAAGCAATGAGAGAGATAAT AT_{13}
203	246	530	ATATAT ATAATGACGTAACTGAGCTAATGAGACAATGAGAGAGATAAT AT_{12}
208	246	484	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAG TTAATAT$_{11}$
204	246	430	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATAA AAT$_{13}$
203	246	370	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGGGATAAT AT_{12}
207	246	370	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGA AATAT$_{13}$
203	246	312	ATATAT ATAATGACGTAACTGAACTAATGAGGCAATGAGAGAGATAAT AT_{12}
210	246	282	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAG CGATAATAT$_{14}$
211	246	268	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGA T_{12}
203	250	267	AT ATAAATAATGACATAACTAGGTTAGTAAAGTGACGAAGAAGATAAT ATTATTTT
203	246	196	ATATAT ATAATGACGTAACTGAGTTAATGAGGCAATGAGAGAGATAAT AT_{13}
203	246	193	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAACGAGAGAGATAAT AT_{21}
205	246	177	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAAATA T_{15}
212	246	176	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAG CGAGATAATAT$_{14}$
203	246	173	ATATAT ATAATGACGTAACTGAGCTAATAAGGCAATGAGAGAGATAAT AT_{14}
203	246	157	ATATAT ATAATGACGTAACTGAGCTAATGGGGCAATGAGAGAGATAAT AT_{12}
203	246	141	ATATAT ATAATGACGTAACTGAGCTAATGAGGCGATGAGAGAGATAAT AT_{19}
203	246	130	ATATAT ATAATGACGTAACTAAGCTAATGAGGCAATGAGAGAGATAAT AT_{13}
203	246	119	ATATAT ATAATGACGTAACTGAGCTAATGAGGTAATGAGAGAGATAAT AT_{12}
203	246	117	ATATAT ATAATGACGTAACTGAGCCAATGAGGCAATGAGAGAGATAAT AT_{16}
203	246	116	ATATAT ATAATGACGTAACTGGGCTAATGAGGCAATGAGAGAGATAAT AT_{13}
203	245	56	ATAAAAT TAATGACATAACTAAATTGATAGGGTAATGAGAGAGATAAT AT_{09}
234	264	35	TAG TGCCTTCTATAGTAGATGATGATATA $TGAT_{14}$
234	280	24	ATATA AGATCAACAAAACTGCCATTTTCTGTAGTAAGTGATGATATA T_{14}
248	281	14	ATA TAAATCAACAGAACTGCCATCTTTTGTAGTA TAGTGATATAAT$_{13}$
234	280	434	ATATA AGATCAACAAAACTGCCATTTTCTATAGTGAGTGATGATATA T_{16}
248	282	270	ATAT GTAAATCAACAGAACCGTCATCTTTTGTAGTA TAGTGATATAAT$_{12}$
288	322	16,349	ATA TACAATACGTGTATGATATTTTATACT AGGTAGATCAGTGAAATT T_{12}
267	322	10	ATA TACAATACGTGTATGATATTTTATACTGGGTAGATCAGTGAAATT T_{07}
269	308	3,731	ATA TATAATACTTTACATCGGGTAAATTGACGAGA ACATGAT$_{11}$
288	322	909	ATA TACAATACGTGTGTAATATTTTATACT AGGTAGATCAATGAAAT$_{15}$
309	349	128	AAATAT AACATATCTTATATCTGAATCTAACTTGTAATATGTG AAT_{16}
309	336	195	AAATAAAACATATCTGAT TCTAAATCTAACTTGTAATGTGTG AAT_{25}†

*Indicates that the tail sequence was shortened where random nucleotides after the poly U tail had been indicated.

†Indicates that the gRNA was identified under conditions of reduced stringency.

APPENDIX D. Identified CR3 mRNA and gRNA transcripts.

A-C: Major CR3 mRNA and gRNA sequence classes. The CR3 mRNA transcriptome was generated using the TREU667 cell line. Identified sequences were then used to search gRNA transcriptomes from four different cell lines: EATRO 164 Bloodstream (BS), EATRO 164 procyclic (PC), TREU 927 procyclic and TREU 667 procyclic. ORF = previously identified Open Reading Frame (purple protein sequence). ARF = Newly identified Alternative Reading Frame (green protein sequence). Alternatively edited nucleotides are shown in Red. Inserted U-residues are lowercase while deleted U-residues are shown as asterisks. Canonical Watson-Crick base pairs (|); G:U base pairs (:). Previously identified start codons are doubled underlined. Potential upstream AUG start codons are indicated by wave underlines. gRNAs were sorted based on guiding sequence class. Sequence variations observed in the 3'-U-tail were ignored in assigning class. Transcript copy number (Reads), were determined by adding all gRNAs of the same sequence class. Only major sequence classes are shown (defined as containing greater than 100 transcript copies). In the case of rare transcripts, the identified gRNA are shown regardless of copy number. gRNA transcript numbers varied greatly between the different cell lines. Interestingly, the most abundant mRNA (CR3 Form C, 7147 reads) had the fewest identified gRNA reads.

A. CR3 Form A	mRNA Sequence	Reads
	AUGUGUAUGAUAUAUAAuuAAuuAuuuuuCAuuuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuGuA	2114
	AUGUGUAUGAUAUAUAA--AAuuAuuuuuCAuuuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuGuA	1217
	AUGUGUAUGAUAUAUAA--AA--AuuuuuCAuuuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuGuA	480
Fully Edited Form		**Reading Frame**
M C M I Y N ST M F D C L V L L F F Y C L F V AUGUGUAUGAUAUAUAAuuAAuuAuuuuuCAuuuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuGuA |||||||:|:|||||:|:|||:|:||| |||:|:|||||||||| $_N$UUAAUUAGUGAAAGUGAGAUAUAGACU----AACGAGCCAAAACAACAUGUAUA		ORF
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	ATATGTACAACAAAACCGAGCAATCAGATATAGAGTGAAAGTGATTAAT$_N$	21297
TREU 927 PC	ATATAACAACAAAACTGAACAATCAAATGTAGAGTGAAAGTGATTAAT$_N$	3398
TREU 667 PC	ATATAACAACAAAACTGAACAATCAAATGTAGAGTGAAAGTGATTAAT$_N$	452

B. CR3 Form B	mRNA Sequence	Reads
	AUGUGUAUGAUAUAUAuuAuuAuuuuAuuuuAuuuuCAuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuG	413
	AUGUGUAUGAUAUAUA--AuuAuuuuAuuuuAuuuuCAuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuG	460
Fully Edited Form		**Reading Frame**
M C M I Y I I I Y L F S L C L I V W F C C F F I V C L AUGUGUAUGAUAUAUAuuAuuAuuuuAuuuuAuuuuCAuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuG |||:|||:|||:|:|:|::|:||||||:|::|| |||||||||| $_N$UAUGAUAGUAAGUGAGUGGAGGUAAUAUAGGCU----AACAAACCAAUAUAUA		ARF +1
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	ATATATAACCAAACAATCGGATATAATGGAGGTGAGTGAATGATAGTAT$_N$	5
TREU 927 PC	ATATATAACCAAACAATCGGATATAATGGAGGTGAGTGAATGATAGTAT$_N$	14164
TREU 927 PC	ATATATAACCAAACAATCGGATATAATGGAGGTGAGTGAATGATAGTATAT$_N$	642
TREU 927 PC	ATATATAACCAAACAATCGGATATAATGGAGGTGAGTAAATGATAGTAT$_N$	114
TREU 667 PC	ATATATAACCAAACAATCGGATATAATGGAGGTGAGTGAATGATAGTAT$_N$	1936
TREU 667 PC	TATATAACCAAACAATCGGATATAATGGAGGTGAGTGAATGATAGTATAT$_N$	236

C. CR3 Form C	mRNA Sequence	Reads
	AUGUGUAUGAUAUAUAAAAACA--AuGuGuA*****UGuuGuuGuuuuGuuuuG***AuuuuGGuuGuACAuuuuuuuuuG	7147
	AUGUGUAUGAUAUAUAAAAACA--A-GuGuA*****UGuuGuuGuuuuGuuuuG***AuuuuGGuuGuACAuuuuuuuuuG	889
	AUGUGUAUGAUAUAUAAAAACAuuA-GuGuA*****UGuuGuuGuuuuGuuuuG***AuuuuGGuuGuACAuuuuuuuuuG	565*
	AUGUGUAUGAUAUAUAAAAACA-uAuGuGuA*****UGuuGuuGuuuuGuuuuG***AuuuuGGuuGuACAuuuuuuuuuG	505
Fully Edited Form		**Reading Frame**
	M C M I Y K N N V Y V V V L F W F W L Y I F F V AUGUGUAUGAUAUAUAAAAACAAuGuGuA*****UGuuGuuGuuuuGuuuuG***AuuuuGGuuGuACAuuuuuuuuuG |||:|:|| |:|:::|:||:|:::||: |:||:||:|||#||||| ₙUUAUAUAU-----AUAGUGAUAAGAUGGAAU---UGAAGCCGACACGUAAAUUAAUAUA	ARF +1
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	ATAATAAAAATGCACAACTAGAATTGAAGTAAAGTGATGATATATATₙ	122
EATRO 164 PC	AAAATGCACAACTAGAATTGAAGTAAAGTGATGATATATATATₙ	3
EATRO 164 PC	ATATATAAAATGTACAACCAGAATTAAGATAAAGTGATGATGTATATATₙ	1
EATRO 164 BS	ATAATAAAAATGCACAACTAGAATTGAAGTAAAATGATGATATATATₙ	453
EATRO 164 BS	ATATTAAAAATGCACAACTAGAATTGAAATAAAGTGATGGTATATATₙ	224
EATRO 164 BS	ATAATAAAAATGCACAACTAGAATTGAAGTAAAATGATGATATATATATₙ	2
TREU 927 PC	ATAATAAAAATGCACAACTAGAATTGAAATAAAGTGATGGTATATATₙ	166
TREU 927 PC	ATATAATTAAATGCACAGCCGAAGTTAAGGTAGAATAGTGATATATATATₙ	117
TREU 927 PC	ATATATAAAATGTACAACCAGAATTAAGATAAAGTGATGATGTATATATₙ	86
TREU 667 PC	ATAATAAAAATGCACAACTAGAATTGAAATAAAGTGATGGTATATATₙ	53
TREU 667 PC	ATATAATTAAATGCACAGCCGAAGTTAAGGTAGAATAGTGATATATATₙ	15
TREU 667 PC	ATATATAAAATGTACAACCAGAATTAAGATAAAGTGATGATGTATATATₙ	14
TREU 667 PC	TATAATTAAATGCACAGCCGAAGTTAAGGTAGAATAGTGATATATATATₙ	5

* no gRNAs identified.

APPENDIX E. ND7 5'-most gRNA populations and the predicted mRNA sequences generated.

A-J: ND7 gRNA major classes and predicted editing patterns. ND7 terminal (5' most) gRNA populations and the predicted mRNA sequence generated. Predicted sequences presented are based on the most abundant gRNAs that generate each reading frame found in the four gRNA transcriptome databases. Initial characterization of the ND7 transcript was done using the EATRO 164 cell line and is unusual in that it is edited in two distinct domains [20]. While the 5' domain was edited in both life cycle stages, complete editing of the 3' domain was only detected in bloodstream stage parasites. Interestingly, the most abundant EATRO 164 PC (procyclic or insect form) gRNA would generate a sequence that brings the 5' most AUG into a +2 frame. The ARF is 65 AA long and involves the entire 5' editing domain. In contrast, the most abundant gRNAs in the EATRO 164 Bloodstream stage library (EATRO 164 BS), would generate sequences that use the originally described ND7 ORF). While gRNA transcript numbers again varied greatly between the different cell lines, all three cells lines had gRNA sequence variants that allowed access to both reading frames.

A. ND7 Form A	**Predicted mRNA Sequence**	**Reading Frame**
	M I S I I L C Y F W ST M T T W ST M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAAGUAUCAuuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG :|||||||:|:T|||:|||:||:||||:|:|||||||||| $_N$UAUAGUAAGAUGCAAUGAAAGCCGUCAAGAGAAUGUAAACAUAUAAA	ORF
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 BS	AAATATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATAT$_N$	20487
TREU 927 PC	AAATATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATAT$_N$	10537
TREU 667 PC	AAATATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATAT$_N$	787

B. ND7 Form B	**Predicted mRNA Sequence**	**Reading Frame**
	M T T W Y S I I L C Y F W ST M I ST M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAuAGUAUCAuuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||:|||||||:||:||||:|||:||:||||:|:|||||||||| $_N$UAUUAUAGUAAGAUGCAAUGAAAGCCGUCAAGAGAAUGUAAACAUAUAAA	ORF
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 BS	AAATATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAATGATATTAT$_N$	35079
TREU 927 PC	AAATATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATATTAT$_N$	38432
TREU 667 PC	AAATATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATATTAT$_N$	4365

C. ND7 Form C	**Predicted mRNA Sequence**	**Reading Frame**
	M T T W ST M I S T F M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAAGUACAuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||||:|:|:||:|:|:||:|:|||||||||||| $_N$UAAAUGUAGUGAAGAUUGUCGGAGAAAUGUAAACAUAGCAUAUACA	ORF
Cell Line	**gRNA Sequence**	**Reads**
TREU 927 PC	ACATATACGATACAAATGTAAAGAGGCTGTTAGAAGTGATGTAAAT$_N$	75654

D. ND7 Form D	**Predicted mRNA Sequence**	**Reading Frame**
	M T T W ST M I I V S F M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAAuuGUAuCAuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||:|||||:|||:|||::|:||:|||:|||||||||| $_N$UAAUAUAGUGAAUAUAAUGGAGACUAUCGAAGAAAUGUAAACAUAUAAA	ORF
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	AAATATACAAATGTAAAGAAGCTATCAGAGGTAATATAAGTGATATAAT$_N$	240
TREU 927 PC	ATATACACAAATGTAAAGAGACTATCGAGAGTGACATAAGTGATATAAT$_N$	477
TREU 667 PC	AAATATACAAATGTAAAGAAGCTATCAGAGGTAATATAAGTGATATAAT$_N$	1152

E. ND7 Form E	Predicted mRNA Sequence	Reading Frame
M T T W ST M I M T F F M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAAuG*ACAuuuuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG ||: |#||:::|||||::|:|:|:::||||:|:||||||||:| $_N$UAU−UAUAGGGAAUACGGUGAGAGUUAUCAGAGAAAUGUAAAUAAUAUA		ORF
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	ATATAATAAATGTAAAGAGACTATTGAGAGTGGCATAAGGGATATTAT$_N$	765

F. ND7 Form F	Predicted mRNA Sequence	Reading Frame
M T T W ST M I S T F M L F L V V F L H L Y R F T F G P Q AUGACUACAUGAUAAGUACAuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||||::|:|:|||:|:::||::|:|||||||||||||||| $_N$UAAAUGUAGUGAAGAUUGUCGGAGAAAUGUAAACAUAGCAUAUACA		ORF
Cell Line	**gRNA Sequence**	**Reads**
TREU 667 PC	ACATATACGATACAAATGTAAAGAGGCTGTTAGAAGTGATGTAAAT$_N$	6623

G. ND7 Form G	Predicted mRNA Sequence	Reading Frame
M T T W Y S I I Y V I F G S F F T F V S F Y I W S T A AUGACUACAUGAUAuAGUAuCAuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||:|||||:|||||::|:|:|:::||||:|:||||||||:| $_N$UAUUAUAGUGAAUACGGUGAGAGUUAUCAGAGAAAUGUAAAUAAUAUA		ARF +2
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 PC	ATATAATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATTAT$_N$	100761
EATRO 164 BS	ATATAATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATTAT$_N$	354

H. ND7 Form H	Predicted mRNA Sequence	Reading Frame
M T T W ST M I S T F Y V I F G S F F T F V S F Y I W S T A AUGACUACAUGAUAAGUACAuuuuAuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||:||:||||:|||:||:||||:|:|||||||||| $_N$UAAGAUGCAAUGAAAGCCGUCAAGAGAAUGUAAACAUAUAAA		ARF +2
Cell Line	**gRNA Sequence**	**Reads**
EATRO 164 BS	AAATATACAAATGTAAGAGAACTGCCGAAAGTAACGTAGAAT$_N$	402

I. ND7 Form I	Predicted mRNA Sequence	Reading Frame
M T T W ST M I S T M L F L V V F T F V S F Y I W S T A AUGACUACAUGAUAAGUACAAuGuuAuuuuuGGuAGuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG |||:|||:|:|||::||:||||:|::||||:|||||||||| $_N$UAUAGUAAGAGUCAUUGAAGAUGUGAGUAUAGUAAAAUGUAAAAUAUA		ARF +2
Cell Line	**gRNA Sequence**	**Reads**
TREU 667 PC	ATATAAAATGTAAAATGATATGAGTGTAGAAGTTACTGAGAATGATAT$_N$	12929

J. ND7 Form J	Predicted mRNA Sequence	Reading Frame
M T T W ST M I S T F I V I F G S F F T F V S F Y I W S T A AUGACUACAUGAUAAGUACAuuuAuuGuuAuuuuuGGuAGuuuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAG ||||||::|:|:||:|:::||::|:|||||||||||||| $_N$UAAAUAGUAGUGAAGAUUGUCGGAGAAAUGUAAACAUAGCAUAUACA		ARF +2
Cell Line	**gRNA Sequence**	**Reads**
TREU 927 PC	ACATATACGATACAAATGTAAAGAGGCTGTTAGAAGTGATGATAAAT$_N$	251

APPENDIX F. RPS12 5'-most gRNA populations and the predicted mRNA sequences generated.

A-E: RPS12 gRNA major classes and predicted editing patterns. RPS12 terminal (5' most) gRNA populations and the predicted mRNA sequence generated. RPS12 differs from both CR3 and ND7 in that the alternative edit that shifts the reading frame occurs just downstream of the previously identified start codon (double-underlined). We do note that the identified alternative gRNAs are rare in all of the gRNA libraries except TREU 667.

A. RPS12 Form A	Predicted mRNA Sequence	Reading Frame
	M W F L Y G C C L R F V L F V CAAACUAAAGUAAuAuAuuAGuuuuuuGCGuAuGuGA*UUUUUGUAUG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGu ||||||:|:|:||:|:||||:||||:|||| |:||::|||| |||||||| $_N$UAUAUAGUUAGAAGAUGCAUGUACU-AGAAGUAUAC-CAACAACAUAUA	ORF

Cell Line	gRNA Sequence	Reads
EATRO 164 PC	ATATACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATAT$_N$	12531
EATRO 164 BS	ATATACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATAT$_N$	1663
TREU 927 PC	ATATACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATAT$_N$	505
TREU 667 PC	ATATACAACAACCATATGAAGATCGTGTACGTAGAAGATTGATATAT$_N$	936

B. RPS12 Form B	Predicted mRNA Sequence	Reading Frame
	M W F L Y G C C L R F V L F V CAAACUAAAGUAAuAAAuuuuGuuuuuuuuGCGuAuGuGA*UUUUUGUAUG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGu ||||||:|:::::||:|:|||:||||:| :||:||||:| ||||| $_N$UAUUUAGAGUGGAAGAGACGUAUACAUU-GAAGACAUGC-CAACAAAUA	ORF

Cell Line	gRNA Sequence	Reads
EATRO 164 PC	ATAAACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGATTTAT$_N$	2218
TREU 927 PC	ATAAACAACCATACAGAAGTTACATATGCAGAGAAGGTGAGATTTAT$_N$	300
TREU 667 PC	ATAAACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGATTTAT$_N$	3834

C. RPS12 Form C	Predicted mRNA Sequence	Reading Frame
	M W F C M V V V Y V L F Y L F CAAACUAAAGUAAAAAGuuuuuuuuuuuuuGCGuAuGuGA**UUUUGUAUG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGu |||:|:|:|:|:|||::|||||:||| ||:|||||| |||:| $_N$UUUUAGAGAGAGAAAGUGCAUAUACU--AAGACAUAC-CAAUAUAUA	ARF +1

Cell Line	gRNA Sequence	Reads
EATRO 164 BS	ATATATAACCATACAGAATCATATACGTGAAAGAGAGAGAT$_N$	144

D. RPS12 Form D	Predicted mRNA Sequence	Reading Frame
	M L F F F R M W F C M V V V Y V L F Y L F CAAACUAAAGUAuuAuAuAAuGuuGuuuuuuuuuuCGuAuGuGA**UUUUGUAUG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGu |:||||||||:|::|||:|:|:||:|||::|| |||||||| |||||| $_N$UGAUAUAUUAUAGUAAAGAGAGAGUAUAUGCU--AAAACAUAC-CAACAAGAUAUA	ARF +1

Cell Line	gRNA Sequence	Reads
TREU 927 PC	ATATAGAACAACCATACAGAAATCGTATATGCGAGAGAAATGATATTATAT$_N$	22

E. RPS12 Form E	Predicted mRNA Sequence	Reading Frame
	M W F C M V V V Y V L F Y L F CAAACUAAAGUAAuuuAAAuuuuGuuuuuuuuuGCGuAuGuGA**UUUUGUAUG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGu ||||||||:|::|:|:|||::||||||| ||:||||| |||:| $_N$UAAAUUUAGAGUAGAGAAAGUGCAUAUACU--AAGACAUAC-CAAUAUAUA	ARF +1

Cell Line	gRNA Sequence	Reads
TREU 667 PC	ATATATAACCATACAGAATCATATACGTGAAAGAGATGAGATTTAAAT$_N$	2664

APPENDIX G. Alignments of *T. brucei* and *T. vivax* edited mRNAs ATPase 6 (A), COIII (B), CR3 (C), CR4 (D), ND3 (E), ND7 (F), ND8 (G), ND9 (H), and RPS12 (I).

Uppercase letters indicate nucleotides originally encoded in the DNA, lower case u's indicate uridines inserted during editing and asterisks indicate uridines removed during editing.

A. ATPase 6 - Pan-edited non-dual coding

```
AAAAAUAAGUAUUUUGAUAUUAUUAAAGUAAAuAuGuuuuuAuuuuuuuuuuuGuGAuuuA T. brucei
---------------------------------AuGuuuuuGuuuuuuuuuuuGuGAuuuG T. vivax

UUUUG-GuuGCGuuuGuuA----uuAuGuAuGuAuuAuuGuGuAuGAuCuAGGuuAuGuu T. brucei
UUUUG*GuuGCGuuuGuuA****UUAuGUGuGuAuuAuuGuGuGuGAuCuAGGuuAuGuu T. vivax

uuAuuGuGuAuuuuAA---uUGuuuAAuGuuGAuuuuuG-AuuuuuuAuuAuuuuGuuuG T. brucei
uuGuuGuGuAuuuuAA***UUGuuuGAuGuuAAuuuuuG*AuuuuuuGuuGuuuuGuuuG T. vivax

*UUUGAuuuGuAuuuGuuuGuuGGuuuGuG***UUUGuuuuuAuuGuuGuGGuuuAuGuu T. brucei
-uuuGAuuuGuAuuuGuuuAuuGGuuuAuG---uuuAuuuuuGuuAuuGuGGuuuAuGuu T. vivax

GuuuA----AuuuAuAuAGuuuAAUUUUGuAuuA*UUGuAuuAC----uUAUUUG***AA T. brucei
GuuuA****AuuuGuAuAGuuuGAUUUuGuAuuA*UUGuAuuACC****UAuuuG--*AA T. vivax

uuuG*UAuuUGuuGuuuuGuAuuGuuuuuuuAuuGuA-----uAuuG-CAuuuuuAuuuu T. brucei
uuuG-uAUUUGuuGUUUuGuAuuGuuuuuuuAuuGuA*****UA-uGuCAuuuuuGuuuu T. vivax

uGuuuuGuuuuuuA-uGuGAuuuuuuuuuGuuuAAuAAuuuGuUA-GuuGGuGAuA**** T. brucei
uGuuuuGuuuuuuGuuG-GAuuuuuuuuuGUUUAAuAGuuuGuuGUG-uGGuGAuA---- T. vivax

GuuuuAuGGAuG--uuuuuuuuAUUC**GuuuuuuGuuGuGuuuuuuAGAGuGuuuuuCu T. brucei
GuuuuAuGGAuGuuuuuuuuuuG-*C--GuuuuuuGuuGuGuuuuuuAGAAuGuuuuuCu T. vivax

uuGuuGuGuC----GuuGuuuGuCGACGuuuuuGCGuuuG-UUUUGuAAuuuAuuAuCAu T. brucei
uuGuuAuGUC****GuuGuuuGuCAACAuuuuuACGuuuG*UUUUGuAAuuuAuuGuCAu T. vivax

CCCAuUUUUUAuuGuuGAuGuuuuuuG-A-uuuuuuuUAuuuuA-uuuuuGuuuuuuuuu T. brucei
CCCAUUUUuuAuuGuuAAuGuuuuuuGuAuuuuuuuuUAuuuuAuuuuuuG-uuuuuuuu T. vivax

uuuA---------uG----GuGuuuuuuG--uuA-uuGAuuuAuuuuAuuuAuuuuuGuG- T. brucei
uuuGuuuuuuuuuGuuuuGuG-----uAuuuuAuuuGGuuuAuuuuGuuuG--uuuGUGu T. vivax

-uuuuGuuuuuGuuuAuuAuuuuAU--G-uGuuuuuAuAuUUGuuGGAuuuAUUuGCC-- T. brucei
uuuuuGuuuuuGuuuGuuGuuuuAUuuGuuG---uuGuAuuuAuuGGAuuuAuuuGCC** T. vivax

***GC-CA-uAuuAC****AGuuAuuuAuuuuuuGuAAuAuGAuuuuGCAGuuGAuAAuG T. brucei
***GC**GuuAuuAC----AGuuGuuuAuuuuuuGuAAuAuGAuuuuGCAAuuGGuAAuG T. vivax

--G**AuuuuuuGuuGuuuuuGuuG-uuuGuuuAGuuuuGuAuuuGAuuuuuGAuAGuuA T. brucei
**G**AuuuuuuAuuGuuuuuGuuGuuuuG-uuAG------------------------- T. vivax

uuAuAuuGuuGuuGAAAuuuG**GuuUGuuA**UUGGAGUUAUAGAAUAAGAUCAAAUAA T. brucei
------------------------------------------------------------ T. vivax

GUUAAUAAUA T. brucei
---------- T. vivax
```

B. COIII - Pan-edited non-dual coding

```
GGUUAUUGAGGAUUGUUUAAAAUUGAAUAAuuAuuAuuuuuuuuAuGuuuuuGuuuC**** T. brucei
---------------------------------------------------------- T. vivax

*GuuGuAuAuuuGuuGGuGuuA-****GuGGuGuuuuuGuuuuuuuuAuCuuuACCuGCCA T. brucei
---------------------------------------------------------- T. vivax

uuGuuAuuGuGuAuuGGuuAuuuuGuuuGuuG****GGAuuuAuuuGuuuAuuGUUUG** T. brucei
---------------------------------------------------------- T. vivax

**GuAGuuuuuuAuuuGuuGAuuGuG****GuuuuAuuuuuuuuuuuuGuuGGuuuuuGuA T. brucei
---------------------------------------------------------- T. vivax

uuuGuuuGuuGuuGuuAuuGuuAGAuuuGuuuuGuGAuuuuuuuACGuGGuuuAuuuGAuu T. brucei
---------------------------------------------------------- T. vivax

uuuGuGuuuuAuuACGuuGuAuCCAGuAuuGuuuuuuAuGGuuuuuAuGuAG*UGAGuuu T. brucei
---------------------------------------------------------- T. vivax

GuuuuAuuuAuGGCGuuuuuuG**UUGuAuuAuuuGGuuuAuGuuuAuuuuuGuGuuGuG T. brucei
---------------------------------------------------------- T. vivax

AGuuuGCUUUCGuuuuuuGuuuACCuuAuAuGuuuuGuuGuuuAuuAuGuGAuuAuGGuu T. brucei
---------------------------------------------------------- T. vivax

uuGuuuuuuAuuGG*UAuuuuuuAGAuuuAuuuAAuuuGuuGAuAAAuACAuuuuAUUUG T. brucei
---------------------------------------------------------- T. vivax

uuUGuuAGuGGuuuAuuuGuuAAuuuuuuuuGuuuuGuGUUUUUGGuuuAGGuuuuuuuGu T. brucei
---------------------------------------------------------- T. vivax

uG**UUGuuGuuuuGuAuuAuGAuuGAGuuuGuuGuuuG****G--uuuuuuGuuuuuG- T. brucei
---------------------------------uuG-uuG--uuGuuuuuuuuGuuuuuGu T. vivax

uGAAACCA--GuuA---UGAGA**GUUUGCAuuGuuAuuuAuuACAuuAAGuuGuGG*** T. brucei
uGAA-uCA**GuuG***UGGGA**AuuuACGuuGuuAuuuAuuACGuuGAGuuGuGG--- T. vivax

*UG-uuuuuGGuuCuAuuuuAuuuuuAuuG---GAuuuAuUACAuuuuA**UGCAuGuuu T. brucei
-uG*UUUUUGGuuCuAuuuuAuuuuuAuuG***GAuuuGuuGCAuuuuA--uGCAuGuuu T. vivax

uuuuAGGuGuuuuGuuGuuG-uuuAuuuG-uuuuAuG--CGuuuGuuuAAuuuuuuGuGu T. brucei
uuuuAGGuGuuuuAuuGuuGuuuuA--uGuuuUUAuG**CGuuuGuuuAGuuuuuuAuGu T. vivax

AuGGAuACACGuuuuGuuuuuuuGuAuuGuGuuuGuuuAuAuuGACAuuuuGuuGA-UUU T. brucei
AuGGAuACACGuuuuGuuuuuuuGuAuGuuGuuuGuuuGuAuuGACAuuuuGuuGAUUU* T. vivax

AGuuuGAuuuuuuuuuAuuGCGAuuuGuuuAuuuuGAuGuuuuAuG----uGuuAuGuAuu T. brucei
GGuuuGGuuuuuuuuuGuuGCGAuuuGuuuAuuuuGAuGuuuuAuG****UGuuAuGuAuu T. vivax

uGuGuGuGuAAuuuuAuuGGuGuuuuUUUAGUUGuuGAuuA*GuuAAuuuGuAuuGGUAG T. brucei
uGuGuGuGuAG---------------------------------------------- T. vivax

UUUGUAGGAAG-- T. brucei
------------ T. vivax
```

C. CR3 - Pan-edited dual coding

```
AGAAAUAUAAAUAUGUGUAUGAUAUAUAAAAACAAuGuuuGA****UUGuuuGGuuuuG- T. brucei
---------------------------------AuGuuuGA---*UUGuuuAGuuuuGU T. vivax

--uuG-uuuuUUUAuuGuuuGuuuGuACAuuuuuuuuuGuuuuuuAuuuGuuuGuG----- T. brucei
U***Guuuuuuuu AuuG-uuGuuuGuACAuuuuuuuuuGuUUUUUGuuuAuuuGuG***** T. vivax

***A**UUUGuuuuuAuGuuuGuuA-*UUUA------GuuuuuGuuuuuuAuuGGAuuuu T. brucei
***A--uuuGuuuuuAuGuuuGuuG--uuuGuuuuuuG--uuUA---uuuG-uGGAuuuu T. vivax

uGuuuuuuAuuuAA----uA--uGGGuuuA---uUGuuGuGuuuAuuuuuuuuuuuuuAuu T. brucei
uGuuuuuuGuuuAA****UA**UGGGuuuG***UUGuuGUGuuuAuuuuuuuuuuuuuGuu T. vivax

uuAuCAuuuGAuAuGuuGuuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCGUUAUUAGAUU T. brucei
uuAuCAuuuGAuAuGuuAuuAuCGuUUUUGuuAuuAuAuAuAAGuUUUCGUUAUUAA--- T. vivax

AAAAAAGUAUGCAAAUAAUUUUUGU T. brucei
------------------------- T. vivax
```

D. CR4 - Pan-edited dual coding

```
UAAUUUAUUGUUAUCUUUGUGUAUUUAUUAuuAuuuuAuuuuAA---uuuuG---GuuGu T. brucei
-----------------------------------------AUUuuuuuGuuuGuuGu T. vivax

GC--***Auuuuuuuuuuuuuuu AuuuG***GuG*UGuuuGuGuuuuA*UGuA*C*A---- T. brucei
GC*****Auuuuuuuuuuuuuuu AuuuG---GuG-uGuuuGuGuuuuA-uGuA--uA**** T. vivax

*GuuuAuGGuAuAuuuuAuuGuuGuuuuGuuuuuuGuuuuuGuuGUUUG-uuUGuG--uG T. brucei
*GuuuGuGGuAuAuuuuGUuGuuGuuuuuuuuuuuuGuuuuuuuuGuuuG***UGuGuuuG T. vivax

GGuA--uGuuuuAuuuGuuuuGuuAuA----GuuGuuuGuuuuuuuuuGuuGuUUUG*GG T. brucei
GGuA**UGuuuuAuuuGuuuuGuuAuA****GuuGuuuGuuuuuuuuuuuuuGuuuuG-GG T. vivax

uuGuG----AuuuuuuAuuG**G--uGuuuuG***AuuGuAuA---GuuuAuuuuuuuuG T. brucei
uuGuG----AUUUUuuGuuA--GuuuA--UuG---AuuGuAuA***GUuuGuuuuuuuuG T. vivax

uGAC-GuuAuAAuuUUGuuuAuuuuuuuuuuuuuAuuuuGuuuuGuGuuuuuuG---uAuu T. brucei
uGAC*GuuAuAAUuuuGuuuAUUUUuuuuuuuuuGUuuuGUUuuGuGuuuuuuGuuuuGuu T. vivax

G*UUGuuuuuA-uUUGGuuuGuuuGGuuuuuuuuuG***UAuuuuuuGUUGuGuuuuGuG T. brucei
Guuuuuuuuu G----GUUU*GuuuGGuuuuuuuuuG---uAuuuuuuGuuAuGUuuuGuG T. vivax

uuAuuuuuuGAuuuAuuuuuuAuGuUGuuuuuUGuuuuG--GG***UG*G-uuuuuuuGu T. brucei
uuAuuuuuuGAuuuGuuuuuuAuGuuGuuuuuuG---uAuuGG----GUGGuuuuuuuGu T. vivax

uuuuGuuuuuuuuuuuuGuuuAuGuuuGuuuuuA---uuuGuGGuuGuuG--uuAuuuuG T. brucei
uuuuGuuuuuuuuuuuuGuuuAUGuuuA---uuGuuuuuuGuAGuuAUUG**UUGuuuuA T. vivax

uuAGuuuGGuuGuuGUUG-uuAuuUGuG---uA--uA****GGUUUAuuUAuA*UGCGuu T. brucei
uuAGuuuGGuuGuuGuuGUU*AuuuGuGU***AUuuA--UUGGuuuAuuuAuA-uGCG-- T. vivax

uuuuAuuuuA------GAuAAuUAuG****G****UA**UUGGUUUUAUAAAAUGUUUUU T. brucei
uuuuG---uGUUUU**AA---------------------------------------- T. vivax

UCU T. brucei
--- T. vivax
```

E. ND3 - Pan-edited dual coding

```
UCAAAAAAUCCUCGCCUUUUUACUUUA-GUUUGUUAUCAuuA--uuuuuAuAuuuGuuuu T. brucei
-------------------------AUG----UUAuCA--AUuuuuuuGuAuuuG---- T. vivax

UG---*A*UA-uuGuGGuuuA**UUAuuuuAuuuA-uAGG--uuuuuuuuuAuGuuuuuu T. brucei
-GUUUuG-uGuuuGuGGuuuA--UuAuuuuA-uuGuuAGGuuuuuuuCU**AuGuuuuuu T. vivax

AuGuuuuuuAuuGCAuuuuuuuG----AuuGuuuuCGuuGuuGuuuGuGGuuuuCGuGuG T. brucei
AuGuuuuuuGuuACAuuuuuuuG****AuuGuuuuCGuuGuuGuuuAuGAuuuuCAuGuG T. vivax

---GuUUGuA---uGAuAuG--A-----AuUCACGuuuG*GUGuuuuA--uACAuuGGAu T. brucei
***GuuuGuA***UGAuAuG**A*****AuuCACGuuuG-GuGuuuuA**UACAuuAGAu T. vivax

uuAUGuuuuGuuA-------------GuUGuUUGuuuuuuGuAuuGuu---AAAuuCCAu T. brucei
uuAuGuuuuGuuA*************GuuGuuuGUuuuuuGuGuuAUU***GAAUuCuGu T. vivax

UAuuuGu----GuUUUGuuGuuuGuuuuuG--UG----------A-uA*GuGuuGuuuuA T. brucei
uAuuuGU****GuuuuGuuAUUUG----uAuuuGU*********GuuG-GuAuuGuuuuA T. vivax

uuuuuGuuAU----GGuuuuuuG-uUUUUGuGGuuuuuGuuuuuuGuuG-uA-uGuA--- T. brucei
uuuuuGUUAU****GGuuuuuuA*UUUUUGuGGuuuuuGuuuuuuG-uGuuGuuA--**U T. vivax

uAG****GAuuUGuGuGGuAuuuuuGGGA-UCAC*GuAuAuuuGuGuGGuGUAAuuuuAu T. brucei
UAG----GGuuuGuGUGGuAuuuuuGAGA*UCA-uGuAUAUU------------------ T. vivax

uuuGuuuAuGA**UGuuuUUUGUUGUAUUAUACAUAUUAUAUUAAUAAAUAUAUAAAA T. brucei
---------------------------------------------------------- T. vivax
```

F. ND7 - Pan-edited dual coding

```
UGAUACAAAAAAACAUGACUACAUGAUAAGUAuCAuuuuAuG-uuAuuuuuG--GuAGuu T. brucei
---------------------------------------AuGuuuAuuuuuGuuGuAGuu T. vivax

uuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAGCAuCCCG***CAGCACAuG**G- T. brucei
uuuuuGCAuuuGuAUCGUuuuACAUuuG-GCCCACAGCAuCCCG---CAGCACAuG-*G* T. vivax

uGuuuuAuGuuGuuuAuuGuAuuuuuGuGGuGA*AuuuAuuGuuuA**UA---UUGAuUG T. brucei
UGuuuuAuGUuGuuuAUUGuAuuuuuGUGGuGA-AuuuAuuGuuuA--uA***UUGAuuG T. vivax

uAuuAuA***G*GuuA--UUUGCAUCGUGGUACAGAAAAGUUAUGUGAAUAUAAAAGUGU T. brucei
uAuuAuA---G-GuuA**UUUGCAUCGAGGUACAGAAAAGUUAUGUGAGUAUAAGAGCGU T. vivax

AGAACAAUGUCUUCCGuAUUUCGA---CAGGUUAGAuuAuGuuA---*GuGuuuGuuGuA T. brucei
AGAGCAGUGUCUUCCGuAUUUuGAU***AGAuuAGAuuAuGuuA****GuGuuuGuuGuA T. vivax

AuGAGCAuuuGuuGuCuuuA***UGuuuuGAGuA--uAuGuuGCGAuGuuGuuuGuCGuu T. brucei
AuGAACAuuuAuuGuCuuuA---uGuuuuGAGuA**UAuGuuACGGuGuuGuuuAuCAuu T. vivax

ACGuuGuGCAuuuAuGCGuuuAuuAAuuGuA****GAAuuuAC***CCGuAGuuuuAAuG T. brucei
GCGuGuuGCAuuuAuGCGuuuAuuGAuuGuA----GAGuuuACU**C*GUAGuuuuAAuG T. vivax

GuuuGuuGuGuAuAuCAuGuAuGGuuuuGG*AuuuAGGuuGuuuGuCUCCGuuG*UUAuG T. brucei
GuuuAuuGuGuGuGuCGuGuAuGAuuuuAG-AuuuAGGuuGuuuAuCCCCGuuA-UuAuG T. vivax

AuCAuuuGAGGAA-***CG*UGA-CAAAuuGAuGACAuuuuuuGAuuuAuG**UUGuGGu T. brucei
GuCAuuuGAGGAG****CG-uGAU*AAGuuAAuGACGuuuuuuGAuuuGuG--uuGuGGu T. vivax

uGuCGuAuGCAuuuGGCUUUCAuGGuuuuAuuA-*GGuAUUCUUGAUGAuuuuGuuuuuG T. brucei
uGuCGuAuGCAuuuGGCUUUCAuGGuuuuAuuG**GGuAuuCUUGAuGAuuuuGuuuuuG T. vivax

GuuuuGuuGAuuuuuuGuuGuuGuuGA***UAAuAuCAuGuuuGuuuGuuAuGGAuuGuu T. brucei
```

```
GuuuuGuuGAUUUuuuGuuGuuAuuGA---uAAuAuCGuGuuuGuuuGuuAuGGAuuGuu T. vivax

AuGAuuuGuuAuuuG--uGGGuAA---UCGuuuAuuuUAuuuGCGuuuGC***GuGGuuu T. brucei
AuGAuuuAuuGuuuG**UGGGuAA***UCGuuuGuuuuAuuuGCGuuuGC---GuGGuuu T. vivax

GuCAuuuuuuGAuuuA---uAuGAuuuA**GuuuuuA**A**UAGuuuAAGuGGuGuuuu T. brucei
GuCAuuuuuuGAuuuG***UAuGAuuuG--GuuuuuA--A--uAGuuuGAGuGGuGuuuu T. vivax

GuCuCGuuCGuuAGGuAuGGuGuGAGAuuGUCGuuuAuuuAGuuGuuA****UGA***** T. brucei
GuCACGuuCAuuGGGuAuGGuAuGAGAuuGCCGuuuAuAuuGuuGuuA-***UGA----- T. vivax

GuUG-uAuuuuA---uGuuuuGuuAuGAuuAuuGuuuuuGuuuuAuA-GGuGAuGCAuuu T. brucei
GuuG*UAUUUUA***UGuuuuGuuAuGAuuAuuGuuuuuGuuuuAuA*GGuGAuGCAuuu T. vivax

GA*UCGuuuAuuuuuACGuuuGuuuGAUAuGCGuAuGAGuuuGuuGAuuuGuAAGCAA-u T. brucei
GAC*CGuuuGuuuuuGCGuuuGuuuGAuAuGCGuAuGAGuuuGuuGAuuuGuAAGCAA*U T. vivax

GuuuuuuuGuuGGuuuuuuuuGuuuuuG*****GuuuuGuuuGuuuGuuuG**AuuAuuuA T. brucei
GuuuuuuuGuuGGuuuuuuuuGuuuuuG-----GAuuuGuuuGuuuGuuuG--AuuAuuuG T. vivax

uAuuGuGAuAuuACCAuuG****AGACCAuuAuuAuGuuAuuuuAuAGuuuG--uGGuGu T. brucei
uAuuGuGAuGuuACCAuuG----AGACuAuuAuuAuGuuGuuuuAuAGuuuA**UGGuGu T. vivax

uGuuGuuuGCCGGGuAuAU*-----CAuuuGC*UUGU-GuuGAACACCCCAAAG-----G T. brucei
uGuuGuuuACCAGGuAuAU******CAUUUGC-UUGU*GuuGAGCAuCCCAAGG*****G T. vivax

uGA***GuAuuGuuuGuuAuuAU****GuuuuuGuGuuGGuuuAuGuuCUCGuuuACGuu T. brucei
uGA---GuAuuGuuuGuuAuuAU****GuuuuuGuGuuGGuuuGuGUUCCCGuuuGCGuu T. vivax

uGCGuuGuGCGGAuuuuuuGCA--*UA--UUUGuuuAuuGGAuGuuuGuuuGCGuGGuuu T. brucei
uGCGuuGuGCGGAuuuuuuACA***UA**UUUGuuuGuuGGAuGuuuGuuuACGuGGuuu T. vivax

uuuAuuGCAuGAuuuAGuuGC--***C*GuuuuA--GGuAAuAuuGAuGuuGuuuuuGGA T. brucei
uuuAuuGCAuGAuuuAGuuGC*****C*G--uuAuuGGuAAuAuuGAuGuuGuuuuuGGA T. vivax

uCC--GUAGAUCGuuA*GuuuuAuAuGuG**A******GGUUAUUGuAGGAUUGUUUAAA T. brucei
uCU**GuGGAUCGuuA*G---------------------------------------- T. vivax

AUUGAAUAAAAA T. brucei
------------ T. vivax
```

G. ND8 - Pan-edited non-dual coding

```
-------CAAUUUAAUAAUUUUAAGUUUUGGUUGAUUAuuAuuuuuuuAuuuuuuuAuuu T. brucei
------------------------------------------------------------ T. vivax

uuGuAuGuuuuuuuuuGAuuuuuuGuuuuuuuUUUUUGuuuGuuuuuAuAuGuGUuuuGu T. brucei
----AuGuuuuuuuuuGAuuuuuuGuuuuuuuuuuuuGuuuGuuuuuAuAuGuGuuuuGu T. vivax

uuGuuGuGuuA****CuA*UUUGuuuA-***CCCAuuGAGuuAACCA--uuGuuAGuuuA T. brucei
uuGuuGuGuuA---CC-A-UUuGuuuA****CCCAuuGAAuuAAC-AuuuuG-uAGuuuG T. vivax

uuGGuuC--GuGGUAA---C-C---AuuuuuuGCGUUUUuA***UUGGuGuGGuuuAGAG T. brucei
uuGA-*CCCGuGGuAA***C*C***AuuuuuuGCGuuuuuA--*UUGAuGuGGuuuAGAA T. vivax

CGuuGuAuuGCuuGuCGuuuAuGuGAuuuAAuuuG-C-----CCuA****GuuuAGCAuu T. brucei
CGUuGuAuuGCuuGuCGuuuAuGuGAuuuGAuuuGuC*****CC-A----GuuuAGCAuu T. vivax

GGAuG***UUCGuGuuGGGuGG---AGuuuuGGuGGuCA**UC*GuuuuGCG--GAuuG- T. brucei
AGAuG---uuCGuGuuGGGuGG***AGuuuuGGuGGuCA--uC-GuuuuGCA**GAuuG* T. vivax
```

```
-AuuuACAuuGAGuuA-*UC**GU-**C----GuuGuAuuuAuuGuGGuuuuuGuAuGCA T. brucei
*AuuuACAuuGAGuuA***C-CG***AC****GuuGuAuuuAuuGuGGuuuuuGuAuGCA T. vivax

uGuuuGCCCGACAGAU****G---CC----AuuA-----CGCA---UUCAuuGuuuGuuA T. brucei
uGuuuGuCCAACAGAu----G***CC****AuuA*****CACA***UUCAuuGuuuGuuA T. vivax

uGuGuuuuuGuuGuuuA-------GCC**A**UGuAuuuAuuG*GCGC***C***C---- T. brucei
uGuGuuuuuGuuGuuuA*******GCC**A--UGuAuuuAuuG-GCGC---C---C**** T. vivax

AAGuuuuuAuuGuuuGG---uuGuuGuuuuAuGuuAuuuGAuuuuuAuuuGuGuuuuGuG T. brucei
AAGuuuuuGuuAuuuGG***UUGuuGuuuuAuGuuGuuuGAuuuuuAuuuGuGuuuuGuG T. vivax

uAGuuAuuuAuuuuGGGuGAuuuAuuGUGuuuAuGAuuuAA***AGAA**AuuCACGGUG T. brucei
uAG---------------------------------------------------------- T. vivax

AAAUUAAAUUUUGACUAAAU T. brucei
-------------------- T. vivax
```

H. ND9 - Pan-edited dual coding

```
UUAAUAUCAACUUAAUUUUUUUUAUAAACAuuAuAuuAUGuGuA--uAuUUUUAuGuuuA T. brucei
-----------------------------------AuG-G-GuuuGuuGuuGuGuuuA T. vivax

uuuCGuuuAuGuuuuuGuuuAAuuUUAuuuuA**UUGuuuGuGuuGuA----------GA T. brucei
uuuCGUUuGuGUUUUUGuuuGAUuuuGuuuuA--UuGuuuGUGuuGuA**********GG T. vivax

uGGuGuuUUGuuuGuuuuGuuGA-uuGuAGuuuuuuGuuuuuuuAuuGuuuuGuuA-Guu T. brucei
UGGuGuuuuGuuuGuuuuGuuGA*UUGuAGuuuuuuGuuuuuuuAUUGUUUuGuuA*Guu T. vivax

uuuuuuuGuuuuAUUGuAuGuuuuuAuuuuuuAAuuuGuG-AuuuuuGuuuuuAuAuuGU T. brucei
uuuuuuuGuuuuAuuGuAuGuuuuuAUUuuuuAAuuuAuG*GuuuuuGuuuuuGuAuuGu T. vivax

UGuG-AuUUGuuAuuGAuuGAuuuuuGuGGuuuuuGuuuuuGuCGuuuuAuGuuGuuGUA T. brucei
uGuG*AuuuAuuGuuGAuuGAuuuuuGuGGUuuuuGuuuuuGuCGuuuuAuGuuAuuAuA T. vivax

uAuuuuAuuuuGuuuGuuuuuGuGuGuuCGuuuGuGuuuuGuuuuGuGuuGUUUGuuuGU T. brucei
uAuuuuGuuuuGuuuGUuuuuGuGuuuuCGuuuAuGuuuuGuuuuGuGUUGUUuGuuuuu T. vivax

AuuuuuuGGAuuG----uGuuuuA-*GuuuuA**GuuGuuuuuGuUAuGC---GuuuuuG T. brucei
GUUUUuuGGAuuG****UGuuuuA**GuuuuA--GuuGuuuuuGuuAuGC***GuuuuuA T. vivax

uuGuuGGA----ACGC*GAAuGuuuUGAUUUGuuuGGuuuuUAuuuuG--uuGGuAAuGA T. brucei
UUGuuAGA****ACG-uGAGuGuuuuGAuuuGuuuGGuuuuuAuuuuG**UUGGuAAuGA T. vivax

uAuuuuACAUCGuuuAuuuGuuG--AuuG****GuuuuuuGuuG-Guuuuuuuuu GuuGA T. brucei
uGuuuuACACCGuuuAUuuGuuG**AuuG----AUUuuuuGuuG*Guuuuuuuuu GuuGA T. vivax

-AGuGuuAUCCA--uuAuuuGGuuuGuuuGuAuuGuuAuuuuGuG---uGuuG**GuG-G T. brucei
*AGuGuuAuCCA**UUAuuuGGUuuGuuuGuAuuAuuGuuuuGuGuuuuA--GuuA-AuG T. vivax

A--GGA-GAUA-GuAuGuACGuuuACAAuGuuA--uuuuuGuuGuuGCAuACC**AAuuU T. brucei
A-**GAUGAuAuGuA----CGuuuACAAuG--GuuuuuuuGuuGuuGCAuACC**AAuuu T. vivax

UUAuuuG*CA-----uuAuuuuAuuuA***AuA**UCACCGuUGUAAUUCUAAAUUUCUC T. brucei
uuAuuuG-CAUUuuuuuG-uuuA--*G------------------------------- T. vivax

ACUUCC T. brucei
------ T. vivax
```

I. RPS12 - Pan-edited dual coding

```
CUAAUACACUUUUGAUAACAAACUAAAGUAAAuAuAuuuuGuuuuuuuuGCGuAuGuGA* T. brucei
----------------------------------------------------AuGuGA* T. vivax

UUUUUGUAUG*GuuGuuGuuuAC----*GuuuuGuuuuAuuuGuuuuAuGuuAuuAuAuG T. brucei
UUUUUGuAuG-GuuGuuGUUUGC*****GuuuuGuuuuGuuuGuuuuAuGuuAuuAuAuG T. vivax

AGuCC----G**CGAuuGCCCAGuuCCGGuAACCGACGuGuAuuGuAuGC**C****GuA T. brucei
AGUCC****C--CGAuuGCCCAGuuCCGGuAAuCGACGuGuGuuGuAuGC--C--**GuG T. vivax

uuuuAuuUAuAuAAuuuuGuuuG-GA-uGuuGCGuuGuuuuuuuuGuuGuuuuAuuG--- T. brucei
uuuuAuuuGuAuAAuuuuG--uGuGGuuGuuGCGuuGuuuuuuuuGuuG---uG-uGUUU T. vivax

---GuuuA---GuuA--uG**UCAuuAuuuAuuAuAGA--***G----GGUGGuGGuuuu T. brucei
UuuG---GuuuG-CAUUUG--UCGuuAuuuAuuAuAGA*****G****GGuGGuGGuuuu T. vivax

GuuGAuuuACCC--***G****GuG*UAAAGuAuuAuACA*CG**UAuuG--uA--AGuu T. brucei
GuuGAuuuACCC*****G--**GuA-UAAAGuAuuAuACA-CG--uA-uGuuuAuuAAuu T. vivax

AGA*UUUAGAuAUAAGAUAUGUUUUU T. brucei
AA---------------------- T. vivax
```

APPENDIX H. Alignments of protein sequences of pan-edited dual-coding genes in *L. tarentolae*, *L. amazonensis*, *P. serpens*, and *Perkinsela CCAP1560/4* with *T. brucei* and *T. vivax* sequences.

A: *CR3*, B: *CR4*, C: *ND3*, D: *ND7* 5' Editing Domain E: *ND9*, F: *RPS12*, G: *ND8* (Nondual-coding). Absent sequences were unavailable. All ORF alignments show published protein sequences. All ARF alignments show +1 or +2 (*ND7* only) reading frame translations of the full length mRNA sequences. In the ARF alignments of *CR3*, *ND7* and *RPS12*, translations were made using the alternative *T. brucei* mRNA sequences shown in Figure 8. Two alignments of CR3 ARFs are presented to display the *P. serpens* +2 reading frame which has no stop codons. *L. tarentolae* CR4 published protein sequence shows limited homology in the C terminus to all other CR4 protein sequences. The edited mRNA has two editing sites where 13 U residues are inserted. If the second of these insertion sites is shortened to 12 U residues, the translation of this mRNA has much better homology to other CR4 proteins. Alignments with translations of the two different sequences (13U and 12U) are both shown, with the location of the altered site highlighted in red. While *ND8* does not appear to be dual-coding, this alignment was included as well, for comparison of the conservation of a nondual-coding gene with that of the dual-coding genes. It should be noted that *ND8* is the only nondual-coding gene that is pan-edited in *L. tarentolae*, *L. amazonensis*, and *P. serpens*, and *ND7*, *A6* and *COIII* are only partially edited in these species. !=Termination codon

A. CR3

CR3 ORF Alignment

```
CR3ORF T. brucei        --MFDCLVLL-FFYCLFVHFFCFLFVCDLFLCLLFSFCFLLDFCFLFNMGLLLCLFFFFI
CR3ORF T. vivax         --MFDCLVLL-FFLLLFVHFFCFLFICDLFLCLLFVFCLFVDFCFLFNMGLLLCLFFFFV
CR3ORF L. amazonensis   --MFDFVIIMFL-FMSFVHFFCFLFIVDLLFCLMFFVFFLYDFCFVCNLGFCCCLFFFFL
CR3ORF P. serpens       IFLFDFVLFLVLFLLFFVHFFCFLFIIDLFCCFLLLFFLVFDFCFCCCFGFVSCLFLFFV
                          :** :::: :    *********: **: *::: . :. ****   :*:  ***:**:

CR3ORF T. brucei        LSFDMLLSFLLLYISFRY!
CR3ORF T. vivax         LSFDMLLSFLLLYISFRY!
CR3ORF L. amazonensis   LSIDMILSFILLYVSFRY!
CR3ORF P. serpens       FHFDMVLSFILLFVSFRY!
                        : :**:***:**::****
```

CR3 ARF Alignment with P. serpens +1 RF

```
CR3ARF T. brucei         ---------RNINMCMIYKNNVYVVVLFWFWLYIFFVFYLFVICFYVCYLVFVFYWIFVF
CR3ARF T. vivax          --------------CLI----V!FCCFFYCCLYIFFVFCLFVICFYVCCLFFVYLWIFVF
CR3ARF P. serpens +1     !!VYNI!KHNILYFCLI---LFCF!CYFYCFLCIFFVFYLLLICFVVFYYCFF!CLIFVF
CR3ARF L. amazonensis    --K!NNMY!V!IYICLI----SLL!CFCLWVLYIFFVFYLLLICYFVWCFLFFFYMIFVL
                                    *:*      .       * ***** *::**: *    *.   ***:

CR3ARF T. brucei         YLIWVYCCVYFFFLFYHLICCYHFYYCI!VFVIRLKKYANNFC-
CR3ARF T. vivax          CLIWVCCCVYFFFLFYHLICYYRFCYYI!VFVI-----------
CR3ARF P. serpens +1     VVVLVLWVVYFYFLFFILIWFYLSYYYL!VSVIKSI!KHKLIS-
CR3ARF L. amazonensis    CVI!VFVVVCFFFFCYPLIWFCRLFYYMLVSDIKII!LLFL!!K
                          :: *   * *:*: : **      * : *  *
```

CR3 ARF Alignment with P. serpens +2 RF

```
CR3ARF T. brucei         -------RNINMCMIYKNNVYVVVLFWFWLYIFFVFYLFVICFYVCYLVFVFYWIFVFYL
CR3ARF T. vivax          ------------CLI----V!FCCFFYCCLYIFFVFCLFVICFYVCCLFFVYLWIFVFCL
CR3ARF P. serpens +2     ----------SKCIIYKNIIFYI-FVWFC----FVFSVIFIVFCAFFLFFIYYWFVLLFF
CR3ARF L. amazonensis    K!NNMY!V!IYICLI----SLL!CFCLWVLYIFFVFYLLLICYFVWCFLFFFYMIFVLCV
                                   *:*        :        *** ::.* : .  :.*.:  :.:: .

CR3ARF T. brucei         IWVYCCVYFFFLFYHLICCYHFYYCI!--------VFVIRLKKYA--NNFC------
CR3ARF T. vivax          IWVCCCVYFFFLFYHLICYYRFCYYI!--------VFVI------------------
CR3ARF P. serpens +2     IIVFFSVWFL--FLLLFWFCELFIFIFCFSFWYGFIFHIIICKFPLLKAFKNISL!V
CR3ARF L. amazonensis    I!VFVVVCFFFFCYPLIWFCRLFYYML--------VSDIKII!LL--FL!!K-----
                         * *   * *:     *:   .:   :         :  *
```

B. CR4

CR4 ORF Alignment with L. tarentolae 13U translation (Published Sequence)

```
CR4ORF T. brucei          ILILVVHFFFFYLVCLC-----FMYSLWYILLLFCFLFLLFVCVGMFYLFCYSCLFFFVV
CR4ORF T. vivax           IFLFVVHFFFFYLVCLC-----FMYSLWYILLLFFFLFFLFVCLGMFYLFCYSCLFFFFV
CR4ORF L. tarentolae 13U  -------------KCCCFWFFYVLFCVLYILFLFFFLFIWFVCYGLFYLYCICLFICFSL
CR4ORF L. amazonensis     ---ISNILLFLYIFIYICWLIF-MYSCWYILILFFFLFLLFVVYGLFYLYCIVCLFILCL
                                          ::.  ***:** ***: **  *:***:*   :: : :

CR4ORF T. brucei          LGCDFLLVFWLYSLFFLWRYNFVYFFFLFCFVFFVLLF-LFGLFGFFLYFLLCFVLFFDL
CR4ORF T. vivax           LGCDFLLVYWLYSLFFLWRYNFVYFFFLFCFVFFVLLF-FFGLFGFFLYFLLCFVLFFDL
CR4ORF L. tarentolae 13U  LCCDFVVVFWLYSVFFVYRYNYFFFFVYFLGVYFFVIILICIWFFIFFFLCLCFDFL--F
CR4ORF L. amazonensis     LCCDFVVVFWLYSVFFIYRYNFVFFFFFLWFVFIFLIIFIFGFGFLFFFLVLCLVFYFEF
                          * ***::*:****:**::***:.:**. :  *::.::: :     :*::: **: :   :

CR4ORF T. brucei          FFMLFFVLGGFFVFVFF---------FCLCLFLFVVVVILLVWLL-----LLFVYRFIYM
CR4ORF T. vivax           FFMLFFVLGGFFVFVFF---------FCLCLLFFVVIVVLLVWLL-----LLFVYRFIYM
CR4ORF L. tarentolae 13U  WIFVYVVFCFLWIFVVCDVYFIFYIIFCFNCVGVLLVVIYICVSIFLYDVLYFNFNWIIL
CR4ORF L. amazonensis     LFMLFFVFCGFLLFVMFILFFVSFF---------VLIVLLFCWMLF-----IFVFRFICM
                           ::::.*:  : :**.                  :::*: :   :       * :.:* :

CR4ORF T. brucei          RFLF!
CR4ORF T. vivax           RFVF!
CR4ORF L. tarentolae 13U  KF---
CR4ORF L. amazonensis     RFVF!
                          :*
```

CR4 ORF Alignment with L. tarentolae 12U translation (Hypothetical Sequence)

```
CR4ORF T. brucei          ILILVVHFFFFYLVCLC-----FMYSLWYILLLFCFLFLLFVCVGMFYLFCYSCLFFFVV
CR4ORF T. vivax           IFLFVVHFFFFYLVCLC-----FMYSLWYILLLFFFLFFLFVCLGMFYLFCYSCLFFFFV
CR4ORF L. tarentolae 12U  -------------KCCCFWFFYVLFCVLYILFLFFFLFIWFVCYGLFYLYCICLFICFSL
CR4ORF L. amazonensis     ---ISNILLFLYIFIYICWLIF-MYSCWYILILFFFLFLLFVVYGLFYLYCIVCLFILCL
                                          ::.  ***:** ***: **  *:***:*   :: : :

CR4ORF T. brucei          LGCDFLLVFWLYSLFFLWRYNFVYFFFLFCFVFFVLLFL-FGLFGFFLYFLLCFVLFFDL
CR4ORF T. vivax           LGCDFLLVYWLYSLFFLWRYNFVYFFFLFCFVFFVLLFF-FGLFGFFLYFLLCFVLFFDL
CR4ORF L. tarentolae 12U  LCCDFVVVFWLYSVFFVYRYNYFFFFVYFLGVYFFVIILICIWFFIFFFYVCVLIFYFEF
CR4ORF L. amazonensis     LCCDFVVVFWLYSVFFIYRYNFVFFFFFLWFVFIFLIIFIFGFGFLFFFLVLCLVFYFEF
                          * ***::*:****:**::***:.:**. :  *::.::::      :*:: :  ::::*::

CR4ORF T. brucei          FFMLFFVLGGFFVFVFFFCLCLFLFVVVVILLVWLLLLFVYRFIYMRFLF!---------
CR4ORF T. vivax           FFMLFFVLGGFFVFVFFFCLCLLFFVVIVVLLVWLLLLFVYRFIYMRFVF!---------
CR4ORF L. tarentolae 12U  LFMLFFVFCGFLLFVMFILFFISFFVLIVLVFCWLLFIFVFRFFCMTFCILILIGLF!NL
CR4ORF L. amazonensis     LFMLFFVFCGFLLFVMFILFFVSFFVLIVLLFCWMLFIFVFRFICMRFVF!---------
                          :******: **::**:*: : : :**::*::: *:*::**:**: * * :
```

CR4 ARF Alignment with L. tarentolae 13U translation (Published Sequence)

```
CR4ARF T. brucei          -----IYCYLCVFIIILF!FWLCIFFFFIWCVCVLCTVYGIFYCCFVFCFCCLFVWVCFI
CR4ARF T. vivax           -----------------FFCLLCIFFFFIWCVCVLCIVCGIFCCCFFFCFFCLCVWVCFI
CR4ARF L. tarentolae 13U  QIH!NTYMYNCKSVV-VFGFFMY---------YFVC---CIFYFCFFFCLFDLCVMVYFI
CR4ARF L. amazonensis     ------DIKNIK!VI-FYYFYIFL---FTFVGWFLCIVVGIFWFYFFFYFCYL!FTVYFI
                                       :   :           .:*    **   *.* :  * . * **

CR4ARF T. brucei          CFVIVVCFFLLFWVVIFYWCFDCIVYFFCDVIILFIFFFYFVLCFLYCCFYLVCLVFFC-
CR4ARF T. vivax           CFVIVVCFFFLFWVVIFC!FIDCIVCFFCDVIILFIFFFCFVLCFLFCCFFLVCLVFFC-
CR4ARF L. tarentolae 13U  YIAFVCLFVLVCYVVILLLCFDCIVFFLFTVIIIFFFLFIFWVFIFLLLFWFVFGFLFFF
CR4ARF L. amazonensis     CIALFVCLFYVCYVVILLLCFDCIVFFLFIDIILFFFFFFYGLCLFF!LFLYLDLVFYFF
                           :.:.  :. : :***:   :**** *:   **:*:*:* : : ::   *  :  .::

CR4ARF T. brucei          IFCCVLCYFLIYFLCCFLFWVVFLFLFFFFVYVCFYLWLLLFC!FGCCCYLCIG------
CR4ARF T. vivax           IFCYVLCYFLICFLCCFLYWVVFLFLFFFFVYVYCFL!LLLFY!FGCCCYLCIG------
CR4ARF L. tarentolae 13U  FYVCVLIFYF-----------------------EFLFMLFFVFCGFLLFVMFILFFISF
CR4ARF L. amazonensis     F!FCVWFFILNFCLCYFLFFADFCCLWCLFYFLCHFLF!LFYYFVGCYLYLYFV------
                          :   *  : :                          :*  *::   *   :: :

CR4ARF T. brucei          --LFICVFYFR!LWYWFYKMFF---------------
CR4ARF T. vivax           --LFICVLCF--------------------------
CR4ARF L. tarentolae 13U  FVLIVLVFCWLLFIFV-FRFFCMTFCILILIGLF!NL
CR4ARF L. amazonensis     --LFVCVLCFK---VG-YEFIFI--------------
                            *:: *: :
```

CR4 ARF Alignment with L. tarentolae 12U translation (Hypothetical Sequence)

```
CR4ARF T. brucei          -----IYCYLCVFIIILF!FWLCIFFFFIWCVCVLCTVYGIFYCCFVFCFCCLFVWVCFI
CR4ARF T. vivax           -----------------FFCLLCIFFFFIWCVCVLCIVCGIFCCCFFFCFFCLCVWVCFI
CR4ARF L. tarentolae 12U  QIH!NTYMYNCKSVV-VFGFFMY---------YFVC---CIFYFCFFFCLFDLCVMVYFI
CR4ARF L. amazonensis     ------DIKNIK!VI-FYYFYIFL---FTFVGWFLCIVVGIFWFYFFFYFCYL!FTVYFI
                                        :   :           .:*    **   *.* :  * . * **

CR4ARF T. brucei          CFVIVVCFFLLFWVVIFYWCFDCIVYFFCDVIILFIFFFYFVLCFLYCCFYLVCLVFFCI
CR4ARF T. vivax           CFVIVVCFFFLFWVVIFC!FIDCIVCFFCDVIILFIFFFCFVLCFLFCCFFLVCLVFFCI
CR4ARF L. tarentolae 12U  YIAFVCLFVLVCYVVILLLCFDCIVFFLFTVIIIFFFLFIFWVFIFLLLFWFVFGFLFFF
CR4ARF L. amazonensis     CIALFVCLFYVCYVVILLLCFDCIVFFLFIDIILFFFFFFYGLCLFF!LFLYLDLVFYFF
                           :.:.  :. : :***:   :**** *:   **:*:*:* : : ::   *  :  .:: :

CR4ARF T. brucei          FCCVL-CYFLIYFLCCFLFWVVFLFLFFFFVYVCFYLWLLLFC!FGCCCYLCIGLFICVF
CR4ARF T. vivax           FCYVL-CYFLICFLCCFLYWVVFLFLFFFFVYVYCFL!LLLFY!FGCCCYLCIGLFICVL
CR4ARF L. tarentolae 12U  FMFVFWFFILNFCLCCFLFFVDFCCLWCLFYFLYHFLF!LCWCFVGCYLYLCFDFFVWRF
CR4ARF L. amazonensis     F!FCVWFFILNFCLCYFLFFADFCCLWCLFYFLCHFLF!LFYYFVGCYLYLYFVLFVCVL
                          *   .  ::*   ** **::. *  *: :* ::  :*  * :  .**  ** : :*:  :

CR4ARF T. brucei          YFR!LWYWFYKMFF
CR4ARF T. vivax           CF------------
CR4ARF L. tarentolae 12U  VF!F!LDYFKI---
CR4ARF L. amazonensis     CFKVGYEFIFI---
                           *
```

C. ND3

ND3 ORF Alignment

```
ND3ORF T. brucei        --LLSLFLYLFLILWFIILFIGFFLCFLCFLLHFFDCFRCCLWFSCGLYDMNSRLVFYTL
ND3ORF T. vivax         --MLSIFLYLVLCLWFIILLLGFFLCFLCFLLHFFDCFRCCLWFSCGLYDMNSRLVFYTL
ND3ORF L. tarentolae    ------------------------FGRREKVLHFFDCFRCCLWFSCGLYDMNSRFVYVSI
ND3ORF L. amazonensis   MFFCSFYFNFVLIFCMLILSIGVLFYIFMFLLHFFDCFRCCLWFSCGLYDMNSRLCYIFI
                                              :     :************************: :  :

ND3ORF T. brucei        DLCFVSCLFFVLLNSIICVLLFVFVIVLFYFCYGFLFLWFLFFVVCIGFVWYFWDHVYLC
ND3ORF T. vivax         DLCFVSCLFFVLLNSVICVLLFVFVLVLFYFCYGFLFLWFLFFVLLLGFVWYFWDHVY--
ND3ORF L. tarentolae    DLCFAVLLCFVMFYSIIGLILFLIVVVLYFMCKL-FFVWFCFVFLL------FWSIV---
ND3ORF L. amazonensis   DLCFAILLCFVLFYSNFGLIIFVLVVLLYFMCKL-FFVWFLLLFFM------LWLMYLI-
                        ****.  * **:: * : :::*::*::*:::*   :*:** :...       :*

ND3ORF T. brucei        GVILFCLWCFLLYYTYYINKYIK----------
ND3ORF T. vivax         ---------------------------------
ND3ORF L. tarentolae    FNIWFCV---LFVFIFFLLILINSVFSYLILK!
ND3ORF L. amazonensis   FDSVYCL---FLFFFY!---------------
```

ND3 ARF Alignment

```
ND3ARF T. brucei        -SKNPRLFTLVCYHYFYICF-W-YCGLLFYL!VFFYVFYVFYCIFLIVFVVVCGFRVVCM
ND3ARF T. vivax         -----------CYQFFCIWF-C-VCGLLFYC!VFFYVFYVFCYIFLIVFVVVYDFHVVCM
ND3ARF L. tarentolae    HSKNSSI-Y-----------------NLYRSIGISLGGGKKCCIFLIVFVVVYDFRVVCM
ND3ARF L. amazonensis   TQKNSRFKFKICFFVHFILILC!FFVCWF!VLVYYFIFLCFCCIFLIVFVVVYDFHVVCM
                                              :              ********* .*:****

ND3ARF T. brucei        IWIHVWCFIHWIYVLLVVCFLYC!IPLFVFCCLFLW!CCFIFVMVFCFCGFCFLLYV!DL
ND3ARF T. vivax         IWIHVWCFIH!IYVLLVVCFLCYWILLFVFCYLYLCWYCFIFVMVFYFCGFCFLCCY-GL
ND3ARF L. tarentolae    TWIHDLFMFL!IYVSQFYYVLLCFIPLLV!FYF!!LWCCILCVNC-FLCGFVLYFCYFGV
ND3ARF L. amazonensis   TWIHVYVIFL!IYVSQFYCVLYCFIPILVWLFLY!WCCYILCVNY-FLCDFYCYFLCYGW
                         ***   ::  ***  .  .*   * ::*   :      :: *    :*.*       .

ND3ARF T. brucei        CGIFGITYI--CVV!FYFVYDVFCCIIHIILINI!---
ND3ARF T. vivax         CGIFEIMYI-----------------------------
ND3ARF L. tarentolae    L--YLIFDSVYCLFLFFF-Y!F!!IVYSLI!Y!NNK!I
ND3ARF L. amazonensis   CI!YLILCTV-CFC-FFF-INLINSVK--LSLD!NN--
                           : *
```

D. ND7 5' Editing Domain

ND7 5' Editing Domain ORF Alignment

```
ND7ORF T. brucei          -----------MLFLVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHR
ND7ORF T. vivax           ----------MFIFVVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHR
ND7ORF L. tarentolae      ILFSRLHDNYILYLLIVFLHLYRFTFGPQHPAAHGVLCCLLYLSGEFITYIDVIIGYLHR
ND7ORF P. serpens         -------IIFIFFIFVVFLHLYRFTFGPQHPAAHGVLCCLLYFSGEYITYIDVIIGYLHR
                                     : :.:*************************:.**:*.*** *******

ND7ORF T. brucei          GTEKLCE
ND7ORF T. vivax           GTEKLCE
ND7ORF L. tarentolae      GTEKLCE
ND7ORF P. serpens         GTEKLCE
                          *******
```

ND7 5' Editing Domain ARF Alignment

```
ND7ARF T. brucei          -IQKNMTTWYSIIVIFGSFFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFA
ND7ARF T. vivax           ------------VYFCCSFFAFVSFYIWPTASRSTWCFMLFIVFLWWIYCLYWLYYRLFA
ND7ARF L. tarentolae      LNFI!PTTR!LYFIFINCFFTLV!IYFRTPASSSPWRIMLFIISFWRIYNVYRCNYWVFT
ND7ARF P. serpens         -------SNNFYFFYFCCFFAFVSFYVWSTASSRTWCVMLFVIFFRWVYNIYWRNYRLLT
                                     .    .**::* :*.   **   * .***:: :  :* :*   * :::

ND7ARF T. brucei          SWYRKVMWI!
ND7ARF T. vivax           SRYRKVMWV!
ND7ARF L. tarentolae      SRYRKVMWI!
ND7ARF P. serpens         PWNGKIVWI!
                              *::*:*
```

E. ND9

ND9 ORF Alignment

```
ND9ORF T. brucei        MCIFLCLFRLCFCLILFYCLCCRWCFVCFVDCSFLFFYCFVSFFLFYCMFLFFNLWFLFL
ND9ORF T. vivax         MGLLLCLFRLCFCLILFYCLCCRWCFVCFVDCSFLFFYCFVSFFLFYCMFLFFNLWFLFL
ND9ORF L. tarentolae    MFLFLIMFRCVFVLLLFFCLCCRWVFLCFVDCSFVFFYLFVCFFLFFVMFLFFNLWFFLL
ND9ORF L. amazonensis   MFLFLIMFRCVFVLCLFFCLCCRWVFLCFVDCSFVFFYLFVCFFLFFVMFLFFNLWFFLL
                        * ::* :**  * * **:****** *:*******:*** **.****: *********::*

ND9ORF T. brucei        YCCDLLLIDFCGFCFCRFMLLYILFCLFLCVRLCFVLCCLFVFFGLCFSFSCFCYAFLLL
ND9ORF T. vivax         YCCDLLLIDFCGFCFCRFMLLYILFCLFLCFRLCFVLCCLFLFFGLCFSFSCFCYAFLLL
ND9ORF L. tarentolae    YCLDLFCIDFCGFCFVRFILIYVLFCLLLCFRVSFVLICFFLFFGLVFSLFFCSYALCIF
ND9ORF L. amazonensis   YCLDLFCIDFCGFCFVRFVLLYVLFCLILCFRVSFVLICFFLFFGLVFSLFFCSYALCIF
                        ** **: ******** **:*:*:****:**.*:.*** *:*:**** **:   .**: ::

ND9ORF T. brucei        ERECFDLFGFYFVGNDILHRLFVDWFFVGFFLLKCYPLFGLFVLLFCVLVEEIVCTFTML
ND9ORF T. vivax         ERECFDLFGFYFVGNDVLHRLFVDWFFVGFFLLKCYPLFGLFVLLFCVLVNEMICTFTMV
ND9ORF L. tarentolae    EREVFDLFGFVFCGNDCLHRFYVDWFFVGFFLCKVYPLFGLFMLNFCMLCEDIVVIATSC
ND9ORF L. amazonensis   ERECFDLFGFVFCGNDCLHRFYVDWFFVGFFLCKVYPLFGLFVLNFCMLVEEIIVYATCC
                        *** ****** * *** ***::********** * *******:* **:* ::::   *

ND9ORF T. brucei        FLLLHTNFYLHYFI!---------
ND9ORF T. vivax         FLLLHTNFYLHFFV!---------
ND9ORF L. tarentolae    FVLCFSNFAI!-------------
ND9ORF L. amazonensis   FVLVFPILHLFNLIYDNADLNIN!
                        *:* .  : :
```

ND9 ARF Alignment

```
ND9ARF T. brucei          NINLIFFINIILCVYFYVYFVYVFV!FYFIVCVVDGVLFVLLIVVFCFFIVLLVFFCFIV
ND9ARF T. vivax           ------------WVCCCVYFVCVFVWFCFIVCVVGGVLFVLLIVVFCFFIVLLVFFCFIV
ND9ARF L. tarentolae      -RYII!YLLSIICFYFWLCFVVCLCCCYFFVCVVDGFFYVLLIVVLFFFICLCVFFYFLW
ND9ARF L. amazonensis     IINCLY!NLIKLCFYF!LCFVVYLYYVYFFVYVVGEFFYVLLIVVLFFFICLCVFFYFLW
                                     .   : **  :    *:* **. .::******: *** * *** *:

ND9ARF T. brucei          CFYFLICDFCFYIVVICYWLIFVVFVFVVLCCCIFYFVCFCVFVCVLFCVVCLYFLDCVL
ND9ARF T. vivax           CFYFLIYGFCFCIVVIYCWLIFVVFVFVVLCYYIFCFVCFCVFVYVLFCVVCFCFLDCVL
ND9ARF L. tarentolae      CFYFLIYDFFYCIV!IYFV!IFAVFVLFVLFWYMFCFVYYYVFE!VLYWFVFFCFLVWFL
ND9ARF L. amazonensis     CFCFLIYDFFYCIVWICFV!IFAVFVLFDLFYYMFCFV!FCVFG!VLYWFVFFYFLVWFL
                          ** *** .* : ** *    **.***:. *   :* ** : **  **: .* : **  .*

ND9ARF T. brucei          VLVVFVMRFCCWNANVLICLVFILLVMIFYIVYLLIGFLLVFFCWSVIHYLVCLYCYFVC
ND9ARF T. vivax           VLVVFVMRFYC!NVSVLICLVFILLVMMFYTVYLLIDFLLVFFCWSVIHYLVCLYYCFVF
ND9ARF L. tarentolae      VYFFVVMRYVFLNVKFLICLVLFFVVMIVYIVFMLIDFLLVFFCVKFIHCLVCLCWIFVC
ND9ARF L. amazonensis     VYFFVVMRYVFLSENVLICLVLFFVVMIVYIVFMLIDFLLVFFYVKFIHCLVYLCWIFVC
                          * ...***:   . ..*****::::**:.* *::**.******  ..** ** *   **

ND9ARF T. brucei          WWRR!YVRLQCYFCCCIPIFICIILFNITVVILNFSL----
ND9ARF T. vivax           !LMRWYVRLQWFFCCCIPIFICIFLF---------------
ND9ARF L. tarentolae      YVKILLWLPLVVLCCVFPILQYSFYFIFV---MNLSIKLIY
ND9ARF L. amazonensis     WLKRLLFMLHVVLC!FFQFCIYLI!FMIM---LT!TLTKF-
                                     :*  : :    : *
```

F. RPS12

RPS12 ORF Alignment

```
RPS12ORF T. brucei        ----------MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWMLRCFF
RPS12ORF T. vivax         ----------MWFLYGCCLRFVLFVLCYYMSPRLPSSGNRRVLYAVFYLYNFVWLLRCFF
RPS12ORF L. tarentolae    --------MRVLFLYGLCVRFLYFCLVLYLSPRLPSSGNRRCLYAICYMFNILWFFCVF-
RPS12ORF L. amazonensis   TFLNLIYFVRVLYLYGLCVRFLFLCLVLYLSPRLPSSGNRRCLYAISIMFNILWYFLVF-
RPS12ORF P. serpens       -----MFFVRSYCLYGFCVRFCFVFLCIYVSPRLPSSGNRRVYVVCFNLYSFVIYCFLFG
RPS12ORF Perkinsela       ------------MLFGFLVRYGFIEFFFFVSPRLPSSGNRFCYELDMRFFFVCYDFVLLG
                                       *:*  :*:  . :  ::**********        :: .      :

RPS12ORF T. brucei        CC-FIGLVMSLFIIEGGGF---VDLP-GVKYYTRIVS!--
RPS12ORF T. vivax         CCVFFGLHLSLFIIEGGGF---VDLP-GIKYYTRMFIN!-
RPS12ORF L. tarentolae    CCVCF-LNHLLFIVEGGGF---IDLP-GVKYFSRFFLNA!
RPS12ORF L. amazonensis   CCFVF-VIFQLFIVEGGGF---IDLP-GVKYFSRFCNVS!
RPS12ORF P. serpens       CCVICYSQSFYFLCEGGGF---VDLP-CIKLYVRVPIA!-
RPS12ORF Perkinsela       FSV---LLSSLVFYEGFGFWLFMDVPFGLYYFSRG!----
                           .         .: ** **   :*:*  :  : *
```

RPS12 ARF Alignment

```
RPS12ARF T. brucei        NTLLITN!SK--------YILFFLRMWFCMVVVYVLFYLFYVIIWVRDCPVPVTDVYCMP
RPS12ARF T. vivax         -------------------------CDFCMVVVCVLFCLFYVIIWVPDCPVPVIDVCCMP
RPS12ARF L. tarentolae    ---NTYRPI!--------IIFILCVYYFCMVYVFVFYIFVWFYI!VHDYLVPVIDVVYMQ
RPS12ARF L. amazonensis   HK-YLFRPF!--------I!FILFVFYICMVYVFVFYFYVWFYI!VHDYQAPVIDVVYMQ
RPS12ARF P. serpens       -----LKPIF!LS!YFYLYLCFLFVVIVYMVFVYVFVLYFYVYMLVPVYPVQVIVVFMLF
RPS12ARF Perkinsela       ----------------------------CCLVFWFVMVL!SFFFLLALVCPVLVIGFVMSW
                                                      :*   *:    :. : .    . *  .

RPS12ARF T. brucei        YF--IYIILFGCCVVFFVVL-LV!LCHYLL!RVVVLLIYPV!SIIHVL!VRFRYKICF--
RPS12ARF T. vivax         CF--ICIILCGCCVVFFVVCFLVCICRYLL!RV-VLLIYPV!SIIHVCLLI---------
RPS12ARF L. tarentolae    YV--ICLIFYDFFVFFVV-FVFWIIC-CL!LKVVVLLICQE!SIFHVFFWMRKQ!VIIKI
RPS12ARF L. amazonensis   LV--LCLIFYDIFWFFAV-LFLWFFS-CL!LKVVVLLICQE!SIFRVFVMCRKFNYLYFY
RPS12ARF P. serpens       VL--ICIVLLFIVFYLVVVLFVILRVFIFYVRVVVLLIYHV!SYMSVCQ!PK!IIAS---
RPS12ARF Perkinsela       IWGFFLFVMILCCWVFPFCCQVWFFMKVLV---------------FGCLWMYRLDCIIFP
                              : :::      : .   .      :

RPS12ARF T. brucei        ----
RPS12ARF T. vivax         ----
RPS12ARF L. tarentolae    ILFR
RPS12ARF L. amazonensis   KN--
RPS12ARF P. serpens       ----
RPS12ARF Perkinsela       VV--
```

G. ND8 (Nondual-coding)

```
ND8 ORF Alignment
ND8ORF T. brucei        MFFFDFLFFFFVCFYMCFVCCVTICLPIELTIVSLLVRGNHFLRFYWCGLERCIACRLCD
ND8ORF T. vivax         MFFFDFLFFFFVCFYMCFVCCVTICLPIELTFCSLLTRGNHFLRFYWCGLERCIACRLCD
ND8ORF L. tarentolae    MFVYDFCFSFFVCFYMCFLCCVTLVLPLELTIVSICVRGNHFLRFYWCGLERCIACRLCD
ND8ORF L. amazonensis   MFCYDFVFSFFVCFYMCFLCCVTLILPCEITIVSICARGHHFLRFYWCGLERCIACRLCD
ND8ORF P. serpens       MFFFDFFFCFFCFVYMCFCCCVTIVVPCEVSLCSFLVRGTHFLRFYWCGLERCIACRMCD
                        ** :** * **  .**** ****: :* *::: *: .** *****************:**

ND8ORF T. brucei        LICPSLALDVRVGWSFGGHRFADWFTLSYRRCIYCGFCMHVCPTDAITHSLFVMCFCCLA
ND8ORF T. vivax         LICPSLALDVRVGWSFGGHRFADWFTLSYRRCIYCGFCMHVCPTDAITHSLFVMCFCCLA
ND8ORF L. tarentolae    FICPSLALDVRCVRSLCGYRFSDVFNISYRRCIYCGFCMHVCPTDAITHSCFLLFCCCIA
ND8ORF L. amazonensis   FICPSLAIDVRCIRSLCGYRYSDLFYISYRRCIYCGFCMHVCPTDAITHSCFLLFCCCIA
ND8ORF P. serpens       YICPSVAIDVRCGVSLIGHRFAHLFFISYRRCIYCGFCMHVCPTDAITHSFVVLFSVLLS
                         ****:*:***   *: *:*::. * :********************** .::    ::

ND8ORF T. brucei        MYLLAPKFLLFGCCFMLFDFYLCFV!
ND8ORF T. vivax         MYLLAPKFLLFGCCFMLFDFYLCFV!
ND8ORF L. tarentolae    MYLCAPKFVLFGCCFMLFDFYLCFV!
ND8ORF L. amazonensis   MYLCAPKFVLFGCCFMLFDFYLCFV!
ND8ORF P. serpens       SYLVAPKFILFGCCFMVFDLFLCFC!
                         ** ****:*******:**::***

ND8 ARF Alignment
ND8RF2 T. brucei        NLIILSFGWLLFFYFFIFVCFFLIFCFFFLFVFICVLFVVLLFVYPLS!PLLVYWFVVTI
ND8RF2 T. vivax         --------------------CFFLIFCFFFLFVFICVLFVVLPFVYPLN!HFVVCWPVVTI
ND8RF2 L. tarentolae    ---------KHI!CIRLKECLFMIFVFLFLFVFICVFYVVLLWFYHWSWPLLVFVFVVTI
ND8RF2 L. amazonensis   ---------NIIRSILIKICFVMILFFLFLFVFICVFYVVLLWFYHVRLPLLVFVLVVII
ND8RF2 P. serpens       ----------!I!CVNVIKCFFLIFFFVFFVLFICVFVVVLPLLFHVKYHCVVFWFVVLI
                                       *:.:*: *.*:.:****: ***  .:      :*   ** *

ND8RF2 T. brucei        FCVFIGVV!SVVLLVVYVI!FALV!HWMFVLGGVLVVIVLRIDLHWVIVVVFIVVFVCMF
ND8RF2 T. vivax         FCVFIDVV!NVVLLVVYVIWFVPV!HIMFVLGGVLVVIVLQIDLHWVTDVVFIVVFVCMF
ND8RF2 L. tarentolae    FCVFIDVV!NVVLPAVYVILYAQV!L!MFVVLEVYVVIGFPMCLILVIVVVFIVVFVCMF
ND8RF2 L. amazonensis   FCVFIDVV!NGVLPAVYVILYALVWPLMFVVLEVYVVIVIPIYFILVIVVVFIVVFVCMF
ND8RF2 P. serpens       FYVFIGVV!NVVLLVVCVIIFVLVLF!MFVVVLVWLVIVLHICFLLVIDVVFIVVFVCMF
                        * ***.***. ** .* ** :. *   ***:  * :** : : :  *  ***********

ND8RF2 T. brucei        ARQMPLRIHCLLCVFVV!PCIYWRPSFYCLVVVLCYLIFICVLCSYLFWVI------YCV
ND8RF2 T. vivax         VQQMPLHIHCLLCVFVV!PCIYWRPSFCYLVVVLCCLIFICVLC----------------
ND8RF2 L. tarentolae    VQQTPLRIHVFCYFVVVLPCIYAHLNLFYLVVVLCYLIFICVLFSLVYLEKY-IWLII!!
ND8RF2 L. amazonensis   VPPMLLRIHVFCYFVVVLPCIYAHLNLFYLVVVLCCLIFICVLFNL!FEYFFYIFYVVCS
ND8RF2 P. serpens       VQPMQLPIHLLFYLVCCYPVIWLHPNLFCLVVVLWFLIYFCVFVSCIV!FI!----LFVV
                        .    * ** :  ..   * *: : .:  *****  **::**:

ND8RF2 T. brucei        YDLKKFTVKLNFD!
ND8RF2 T. vivax         --------------
ND8RF2 L. tarentolae    INF-----------
ND8RF2 L. amazonensis   KNLVNV--------
ND8RF2 P. serpens       !!!!K!RTK!Y!--
```

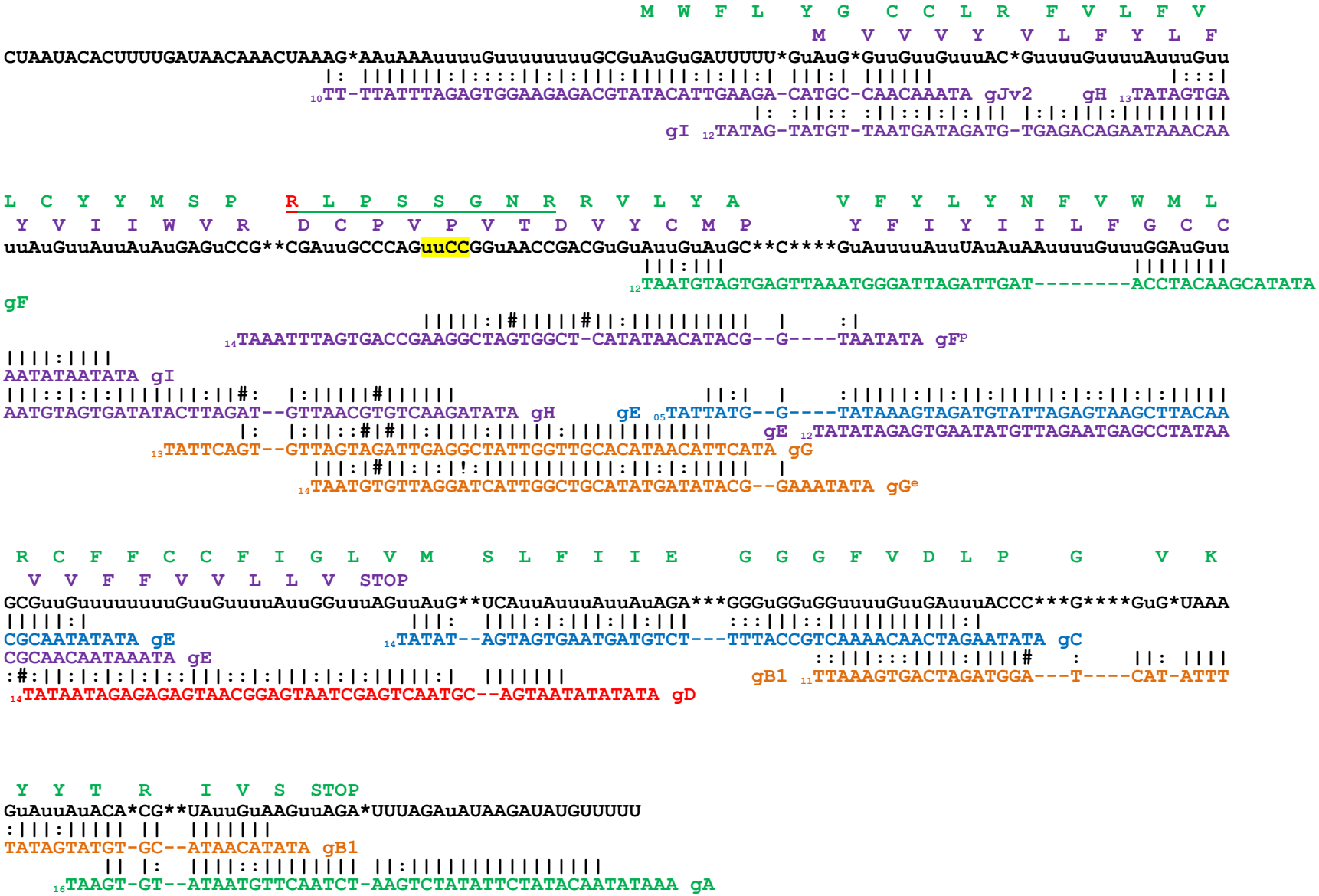
APPENDIX I. RPS12 gRNA Alignments for TREU 667 SDM79 (A) and EATRO 164 SDM79 cells (B), and all editing variants (C and D).

Amino acid translations are shown above the mRNA sequences. The cDNA sequence of the most abundant gRNA in its sequence class is shown aligned beneath the fully edited mRNA. Lowercase u's indicate uridines added by editing, asterisks indicate encoded uridines deleted during editing. Nucleotides and deletion sites in the fully edited mRNA were numbered starting from the 5' end (+1=0). gRNAs are colored based on transcript abundance as follows: Blue<100; Green<1,000; Purple<10,000; Orange<100,000; Red>100,000; Black=not quantified. Watson-Crick (|) and G:U base pairs (:) are indicated. Mismatches are indicated by an octothorpe (#). Highlighted sequence represents sequences were multiple CU configurations are possible.

A. TREU SDM79 gRNAs - Jv2, I, H, G, F, E, D, C, B1, A

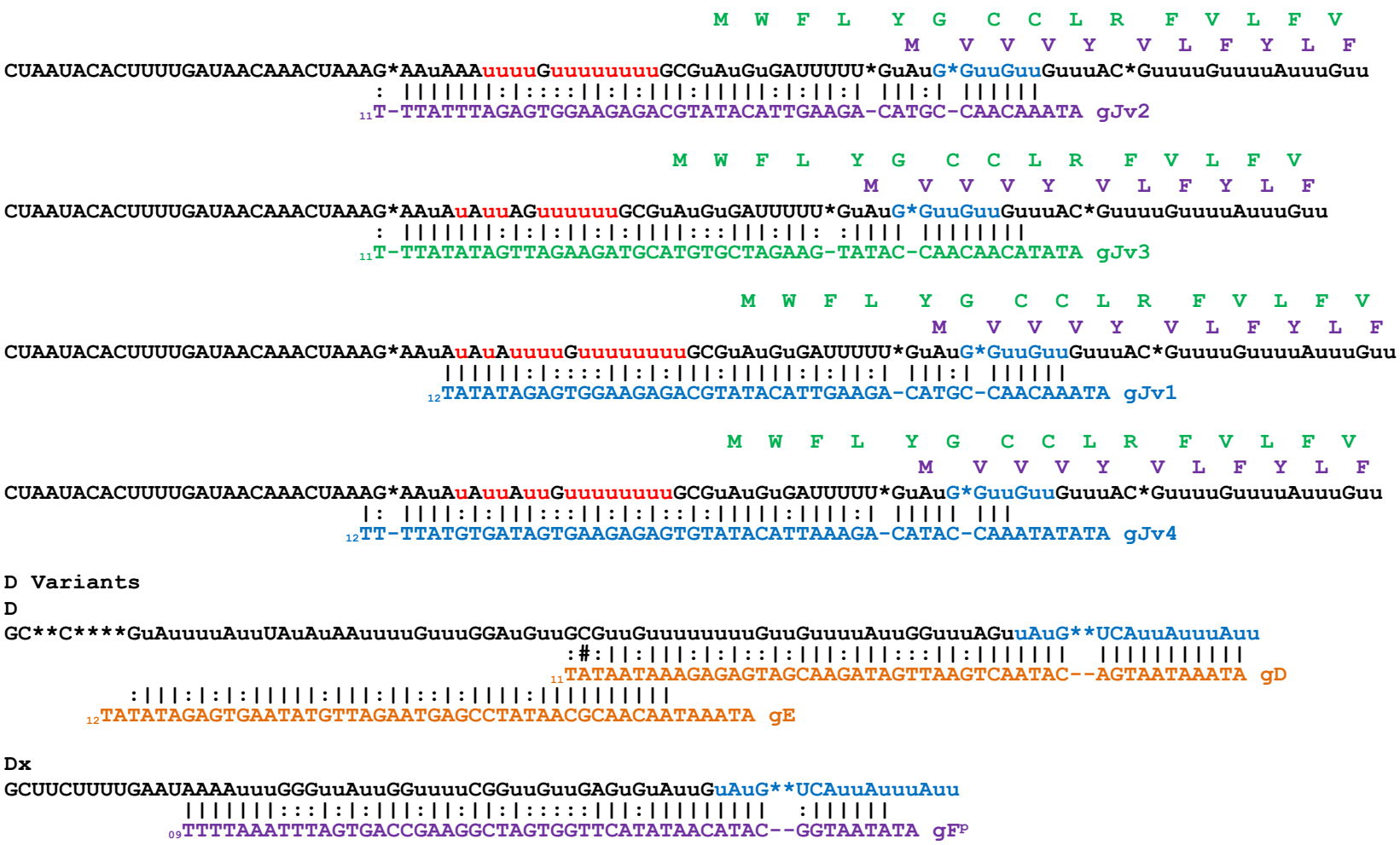
```
                                                                M  W  F  L   Y  G   C  C  L  R   F  V  L  F  V
                                                                   M   V  V  V  Y   V  L  F  Y  L  F
CUAAUACACUUUUGAUAACAAACUAAAG*AAuAAAuuuuGuuuuuuuuuGCGuAuGuGAUUUUU*GuAuG*GuuGuuGuuuAC*GuuuuGuuuuAuuuGuu
                             : |||||||:|:::::||:|:|||:|||||:|:||:| |||:| ||||||                    |:::|
                          11T-TTATTTAGAGTGGAAGAGACGTATACATTGAAGA-CATGC-CAACAAATA gJv2     gH 13TATAGTGA
                                                                  || :|||:: ||:::|:|:||| |||||:||:||:|||||
                                                              gI 14TAA-TATGT-CAGTGATAGATG-CAAAGTAAGATGAACAA

L  C  Y  Y  M  S  P   R  L  P  S  S  G  N  R  R  V  L  Y  A       V  F  Y  L  Y  N  F  V  W  M  L
 Y  V  I  I  W  V  R   D  C  P  V  P  V  T  D  V  Y  C  M  P       Y  F  I  Y  I  I  L  F  G  C  C
uuAuGuuAuuAuAuGAGuCCG**CGAuuGCCCAGuuCCGGuAACCGACGuGuAuuGuAuGC**C****GuAuuuuAuuUAuAuAAuuuuGuuuGGAuGuu
|||||:|||| |||:|||:|  |    :|||||:|||||:|:|||||||:|||||||||||
AATATAATATA gI                                  gF 18TCAATTAATATATG--G----TATAAGATAAGTGTATTAAGACAAACCTACAAAATATA
|||::|:|:|||||||:||#:  |:|||||#||||||                                  :|||:|:|:|||||:|||:||::|:||||:||
AATGTAGTGATATACTTAGAT--GTTAACGTGTCAAGATATA gH                       gE 12TATATAGAGTGAATATGTTAGAATGAGCCTATAA
                    |:  |:|||:#|#||:|||||:|||||:||||||||||||
          gG 13TATTCAGT--GTTAGTAGATTGAGGCTATTGGTTGCACATAACATTCATA gG
                                  ||||||:|#|||:|#||:||||||||||  |    :|
                       12TAAATTTAGTGACCGAAGGCTAGTGGTT-CATATAACATACG--G----TAATATA gFP

 R  C  F  F  C  C  F  I  G  L  V  M   S  L  F  I  I  E    G  G  G  F  V  D  L  P    G     V  K
  V  V  F  F  V  V  L  L  V  STOP
GCGuuGuuuuuuuuGuuGuuuuAuuGGuuuAGuuAuG**UCAuuAuuuAuuAuAGA***GGGuGGuGGuuuuGuuGAuuuACCC***G****GuG*UAAA
||||||||                          ||:  |||:|||:||:||:|||   ||:::#::|||||||||||||||
CGCAACAATAAATA gE                14TATTAT--AGTGATAGATGATGTCT---CCTGTAGTCAAAACAACTAAATATA gC
:#:|||:|||:|:|:::|:|||:|||:::||:|||||||  ||||||||||                  ::|||:::|||||:|||#   :    ||: ||||
11TATAATAAAGAGAGTAGCAAGATAGTTAAGTCAATAC--AGTAATAAATA gD          gB1 11TTAAAGTGACTAGATGGA---T----CAT-ATTT

 Y  Y  T   R    I  V  S  STOP
GuAuuAuACA*CG**UAuuGuAAGuuAGA*UUUAGAuAUAAGAUAUGUUUUU
:|||:|||||  ||  |||||||
TATAGTATGT-GC--ATAACATATA gB1
        ||  |:  ||||::|||||||| ||:|||||||||||||||
    16TAAGT-GT--ATAATGTTCAATCT-AAGTCTATATTCTATACAATATAAA gA
```

B. EATRO SDM79 gRNAs - Jv2, I, H, G, F, E, D, C, B1, A



C. TREU SDM79 Variants

J Variants



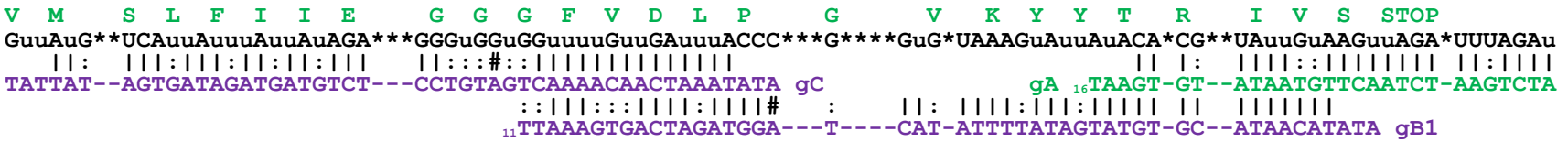
D Variants

D

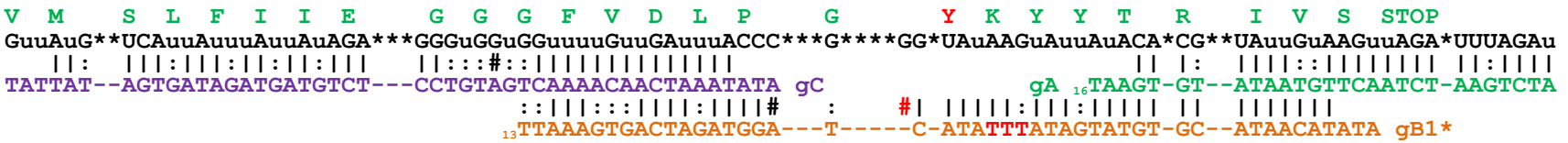
Dx

BC Variants

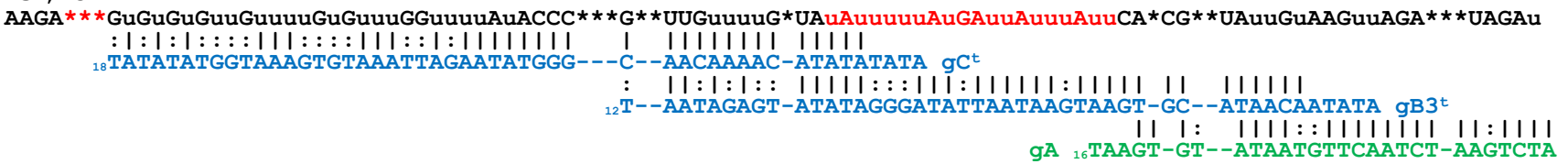
B1, C1



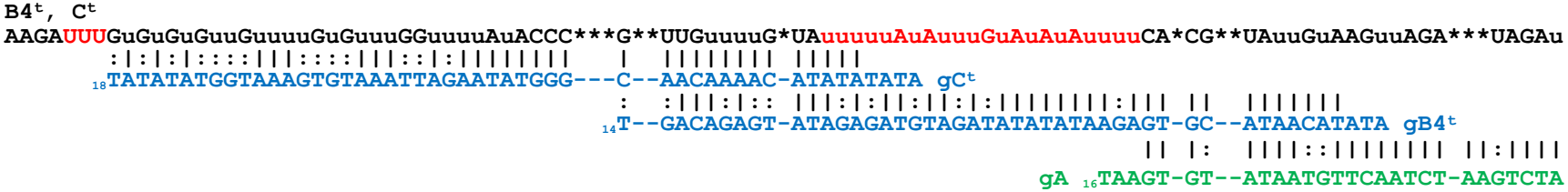
B1*, C1



B3^{t}, C^{t}

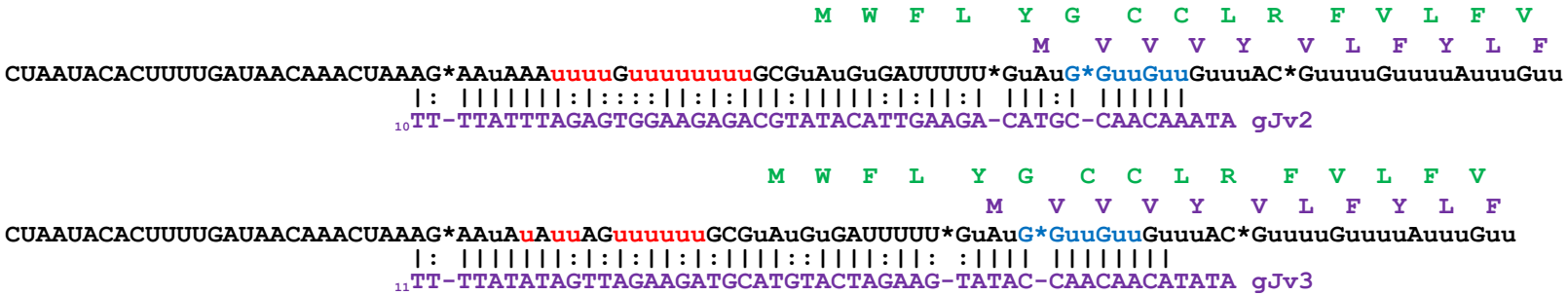


B4^{t}, C^{t}



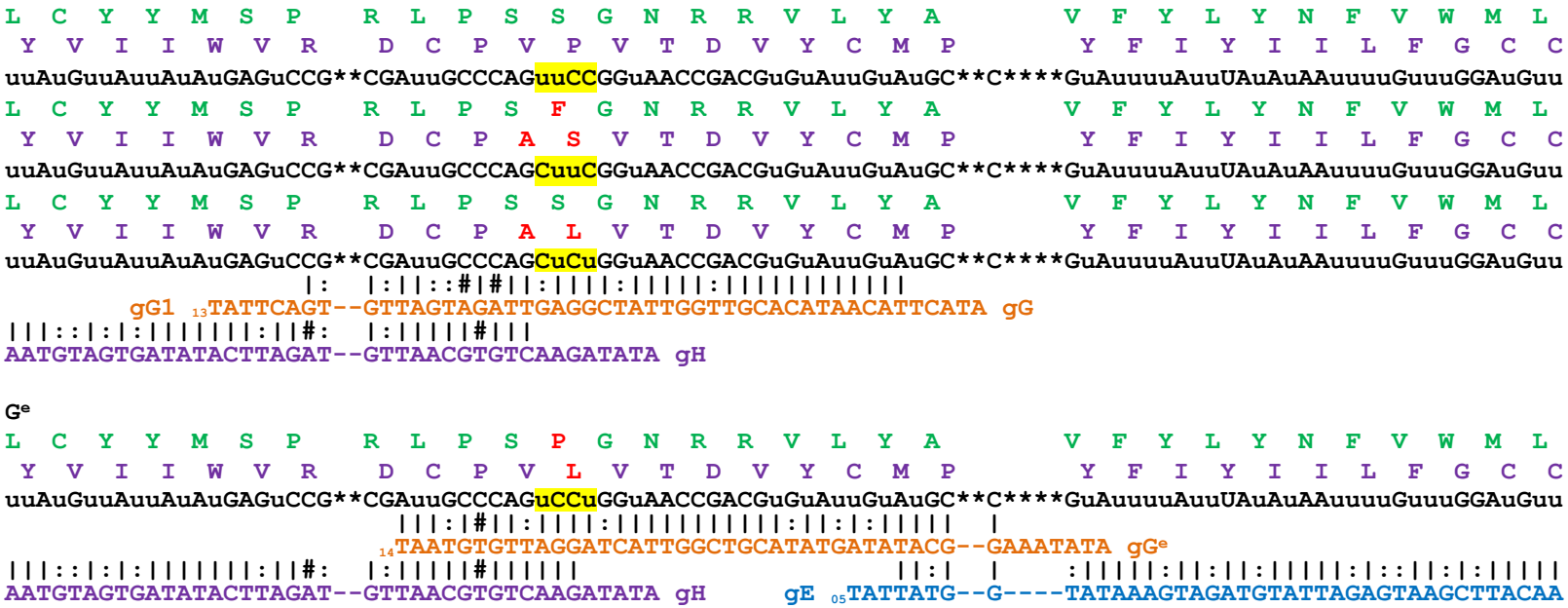
D. EATRO SDM79 Variants

J Variants



G Variants

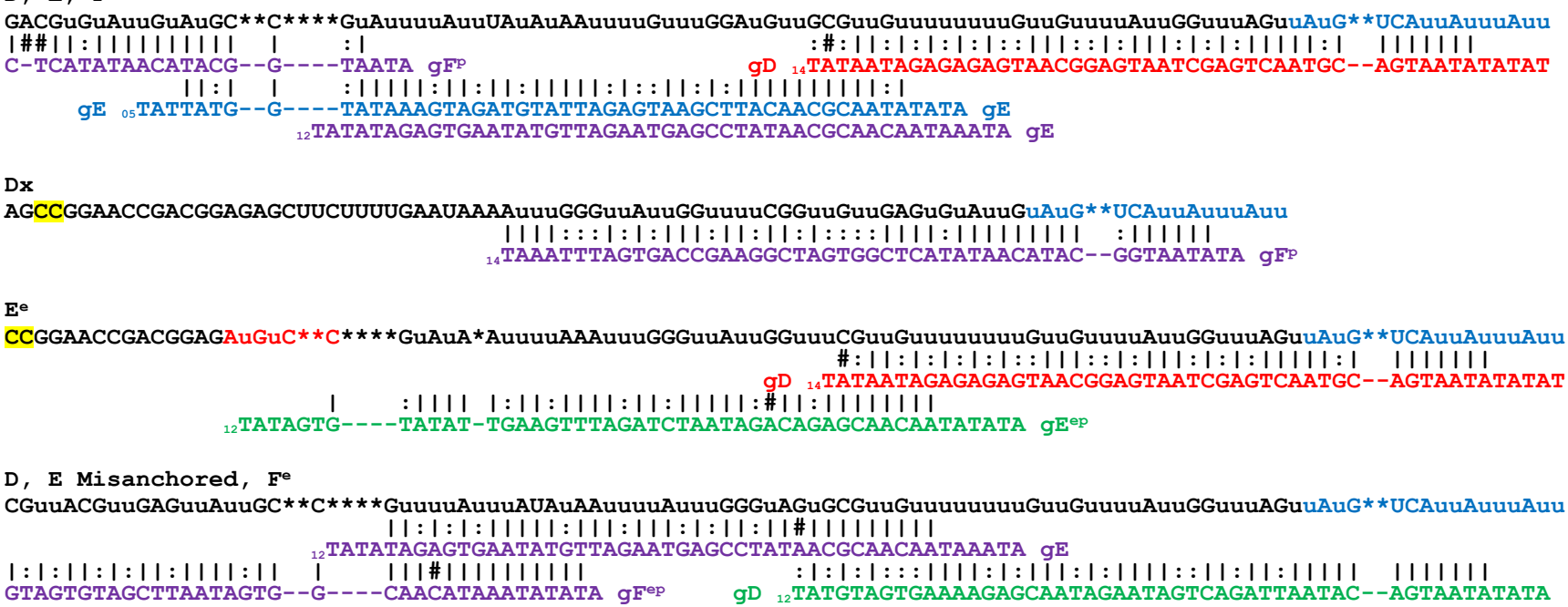
Multiple G sequences from gG



Ge

DEF Variants

D, E, F



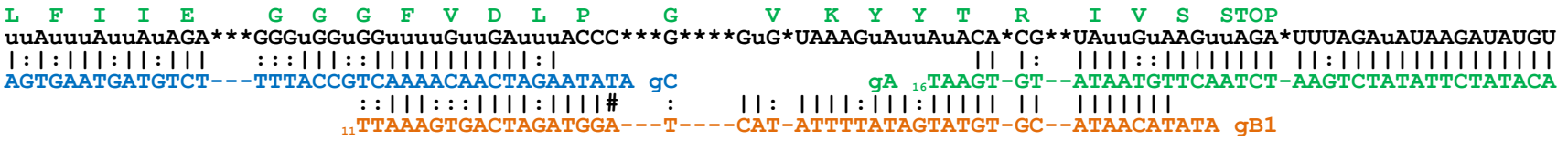
Dx

E^e

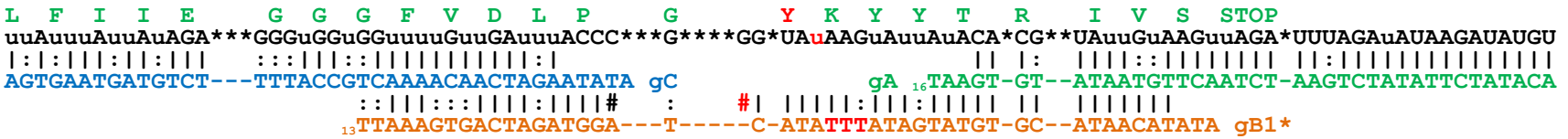
D, E Misanchored, F^e

B Variants

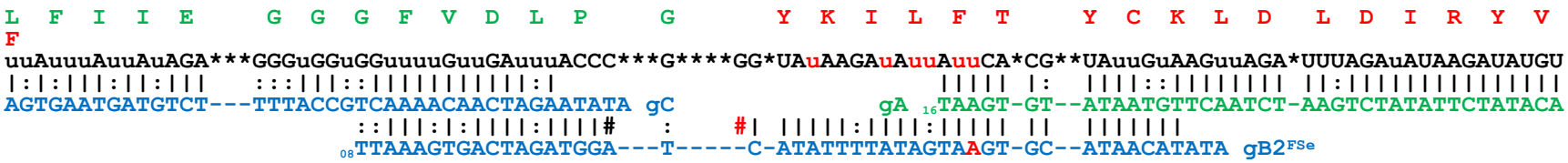
B1



B1′



B2^{FSe}



APPENDIX J. gRNAs identified to edit the RPS12 mRNAs of found in both TREU 667 and EATRO 164 gRNA transcriptomes.

Editing Region	Population	Sequence	Reads TREU 667	Reads EATRO 164
J	J variant 1	ATA AACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGATATATT TTTTTTATTT	1	
	J variant 2	ATA AACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGATTTATTTTTT TTTTTTT*	3,190	1,731
	J variant 2	ATA AACAACCGTACAGAAGTTACATATGCAGAGAAGGTGAGATTTAT ATTTTTTTTTTT	848	401
	J variant 3	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATAT TTTTTTTTTTTT	751	8,006
	J variant 3	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATATA TTTTTTTTTTT	628	3,974
	J variant 3	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATATAATTTTT TTTTTTTT		2,969
	J variant 3	ATAT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATAT TTTTTTTTTT	72	282
	J variant 3	TAT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATATAAT ATTTTTT		184
	J variant 3	AT ACAACAACCATATGAAGATCATGTACGTAGAAGATTGATATATAATTT ATTTT	28	144
	J variant 4	ATATATA AACCATACAGAAATTACATATGTGAGAGAAGTGATAGTGTATTT TTTTAATTTTT	8	
	J variant 4	ATATATA AACCATACAGAAATTACATATGTGAGAGAAGTGATAGTGT TTTTTT	6	
	J variant 4	ATATA AACCATACAGAAATTACATATGTGAGAGAAGTGATAGTGTAT ATTTTTTTT*	2	
I	I	AT ATAATATAAAACAAATAAGACAGAGTGTAGATAGTAATTGTATGA TATTTTTTTTTTTT	5,762	6,201
	I	ATAT ATATAAAACAAATAGAATAGAACGTAGATGAT TACTGTATAATTTTTTTTTTAT		2,718
	I	A TATATAATAACATAAAACAAATAGAACGAGATGTAAATGATA TCTATTTTTTTATTTTT	503	133
	I	AT ATAATATAAAACAAGTAGAATGAAACGTAGATAGTGACTGTATAA TTTTTTTTTTTTCT	6,937	
	I	ATAATATAAAACAAATGAAACGAAGCGTAGACAGTAATTATATGA TATTTTTTTTCTTTT	687	
	I	AT ATAATATAAAACAAGTAGAATGAAACGTAGATAG GGACTGTATAATTTTT	72	
H	H	ATATAGAACTGTGCAATTGTA GATTCATATAGTGATGTAAAGTGA TATTTTTTTTTTTTT*	12,736	4,786
	H	ATATAGAATTAGGCAATCA CGGATTTATATAGTAACGTAAAATGA TATTTTTTTTTTTTG		2,326
	H	ATATATAACTGGGCATCT CGGATTTGTATAGTGATATAAAGTGAATAA TTTTTTTTTTTTT		979
	H	ATATATAACTGGACAATCGTA GGCTTGTATGATGATATGAGATGAGTAAA TTTTTTTTTTTT		162
	H	ATATATAACTGGACAATCGTA GGCTTGTATGATGACATGAGATGAGTAAA TTTTTTTTTTTCTTT		90
	H	ATATATAACTGGACAATCGA TGGGCTTGTATAATGATATGAGATGAGTAAA TTTTTTTTTTTTTG	17,207	
	H	ATATATAACTGGACAATCGA TGGGCTTGTATAATGATATGAGATGAGTAA TTTTTTTTTTTT	1,140	
	H	ATATAAACTGTGCAATCGA TGGACTTATGTAGTGATATAAGATGA TAATTTTTTTTTTTT	828	
	H	TATATAACTGGACAATCGA TGGGCTTGTATAATGATATGAGATGAG GAAATTTTTTTTTTTT	505	
	H	ATATATAACTGGCAATAT CGGACTCATATAGTGATGTGAAGTAAATA TTTTTTTTTT	169	
G	G	ATACT TACAATACACGTTGGTTATCGGAGTT AGATGATTGTGACTTATTTTTTTTTTTTTC	791	15,732
	G^e	ATATAAA GGCATATAGTATACGTCGGTTACT AGGATTGTGTAATTTTTTTTTTTTT	94,870	90,998
	G^e	ATATAAAGC CATATAGTATACGTCGGTTACT AGGATTGTGTAATTTTTTTTTTT	196	142
F	F^p	AT ATAATGGCATACAATATACTCGGTGATCGGAAGCCAGTGATTTAAATTTT TTTTTTTTT	2,288	2,552
	F^p	AT ATAATGGCATACAATATACTCGGTGATCGGAAGCCAGTGATTT TTTTTTTTTT	599	346
	F^p	AT ATAATGGCATACAATATACTCGGTGATCGGAAGCCAGTGATTTAA TTTTTATTTTTTTTT	181	256
	F^p	T ATAATGGCATACAATATACTCGGTGATCGGAAGCCAGTGATTTAAATT AAATTTTTT	224	172
	F^p	AT ATAATGGCATACAATATACTCGGTGATCGGAAGCCAGTGATTTA TTTTTTTGTTTTTTTT*	87	120
	F^p	AT ATAATGGCATACAATATACTTGGTGATCGGAAGCCAGTGATTTAAATTT GTTTTTTTTTTT	156	
	F^p	AT ATAATGGCATACAATATACTTGGTGATCGGAAGCCAGTGATTTAAATTTTGTTT T	60	
	F	ATATAA AACATCCAAACAGAATTATGTGAATAGAATATGGTATATAAT TAACTTTTTTTTTTTTTTTTTT	63	
	F^{ep}	ATATATAAATAC AACGGTGATAATTCGATGTGATG ATATCTGTAATTTTTTTTTTTTTTAT		1,823

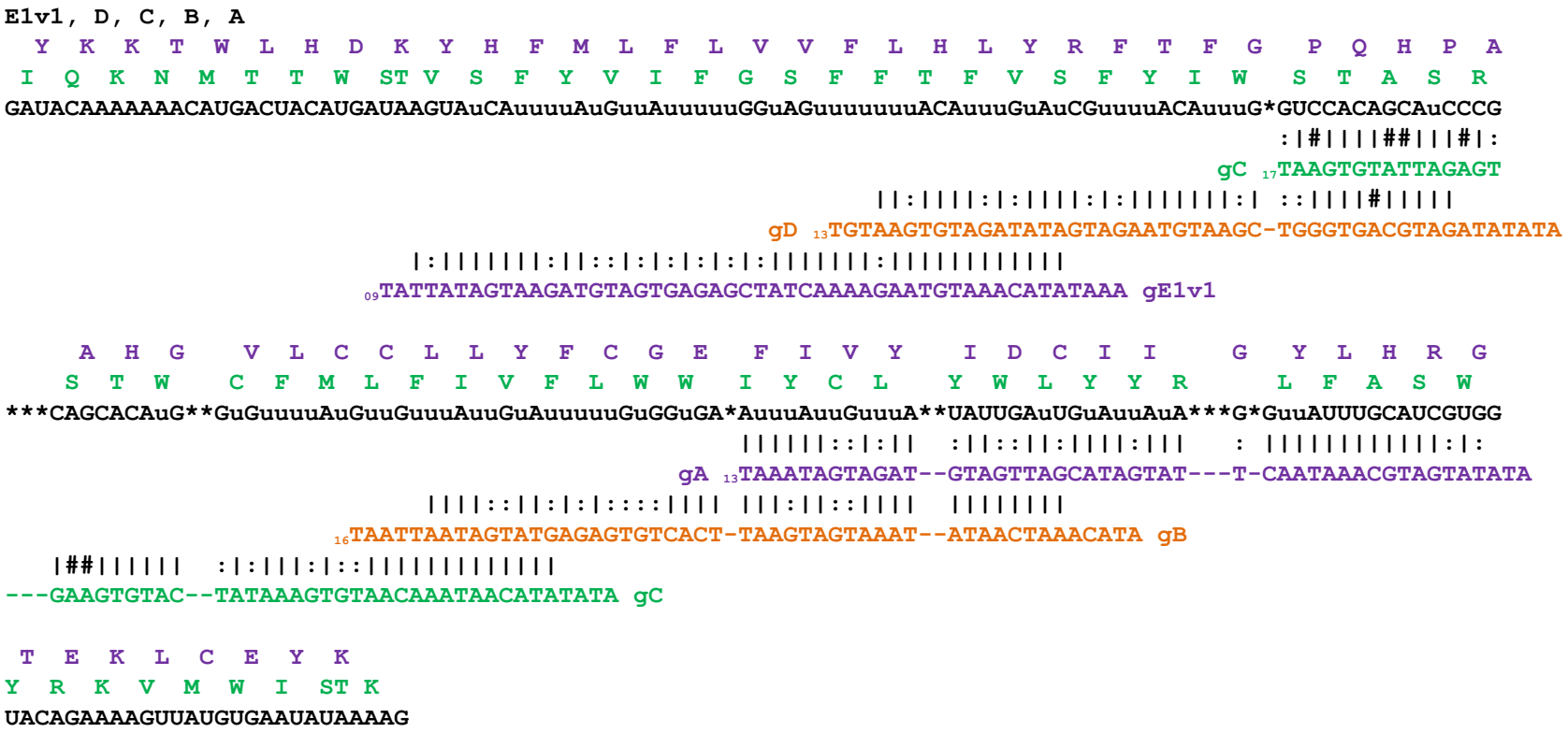
E	E	ATAAAT AACAACGCAATATCCGAGTAAGATTGTATAAGTGAGATAT ATTTTTTTTTTTT	10,751	4,860
	E	ATAT ATAACGCAACATTCGAATGAGATTATGTAGATGAAATATGGTAT TATTTTT	125	64
	E	ACAAATAACGCAACA TCAGATGAGATTATATAAGTGAGATATG ATATATTTTTTTTTTT		706
	E^{ep}	ATATATAACAACGAGACA GATAATCTAGATTTGAAGTTATATG TGATATTTTTTTTTTT	634	234
	E^{ep}	ATACATAACAACGAGACA GATAATCTAGATTTAGAGTTATATG TGATATTTTTT	113	
D	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTTTTTTT		344,244
	D	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGATGATGTAAT TTTTTTTTTTT	3,627	5,997
	D	ATATATAATGAC TAACTAAACTGATAAAGCAGTAGAAGAGATGATGTAATATTT TTTTTTTTTT	1,168	3,292
	D	ATATATATGACATAACTT GGCCAATGAGATAATGAAAGAGATGGTGTAAT TTTTTTTTTTTTCT	10,277	2,172
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATA TTTTTTTTTTTTT		1,088
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGAT TTTTTTTTTTTTTG		924
	D	ATATATATAATGACT TAACTGAGCTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTTTTT		809
	D	ATATATATAATGACGTAACTGAT CTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTTTTTT		747
	D	ATAAATTAATGACATAACTT GACTAATAGGACAGTGAAAGAGGCAGTGTAAT TCTTTTTTTTTT		558
	D	ATATAT ATAATGACGTAACTGAGCTAATGAAGCAATGAGAGAGATAAT ATTTTTTTTTTTTT		534
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGACAATGAGAGAGATAAT ATTTTTTTTTTTG		521
	D	ATATATAT ATGACGTAACTGAGCTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTT		512
	D	ATAT ATAATGACATAATTAGACTGATAAGATAACGAGAAAAGTGATGTATT TTTTTTTTT	2,340	382
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGGGATAAT ATTTTTTTTTTT		353
	D	ATATAT ATAATGACGTAACTGAACTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTTTT		308
	D	AT ATAAATAATGACATAACTAGGTTAGTAAAGTGACGAAGAAGATAAT ATTATTTT		261
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAAATAAT ATTTTTTTTTATTTTT		258
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGGGAGATAAT ATTTTTTTTTT		250
	D	ATATAT ATAATGACGTAACTGAGTTAATGAGGCAATGAGAGAGATAAT ATTTTTTTTATTTT		193
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAGGTAAT ATTTTTTGTTTTTTT		178
	D	ATATAT ATAATGACGTAACTGAGCTAATGAGGCAATGAGAGAAATA TTTTTTTTTGTTTTT		160
	D	ATAAATAATGACATAACTGAATTGATAGAACGATGAGAGAAATAAT ATTTTTTTTTTTC	39,218	
	D	ATAAAT TAATGACATAACTGAATCGATAGAATAATGAGAGAGATAAT ATTTTTTTTTTTG	23,327	
	D	ATAAATA AATGACATAACTGGATTAGTAAAGTGGTGAAAAAGATAAT ATTTTTTTTTTT	6,953	
	D	A AAATAATGACATAACTGAATTGATAGAACGATGAGAGAAATAAT TTTTTTTTTTCC	5,004	
	D	ATATATAATA ACGTAATTGGATCAGTGAGATAACGA TAGAAATGATATATTTTTTTTTTTTTGTTT	3,253	
	D	ATAT ATAATGACATAATTAGACTGATAAGATGACGAGAAAAGTGATGTA TTTTTTTTTTC	2,627	
	D	ATATATA AATGACATGACTAAACTAATAGGGCAGTGAAGAAGACAATG ATTTTTTTTTTTC	1,811	
	D	ATATATATC AATGACATAACTGAACTAGTAAGATAATGAGAGAAGT TTTTTTTTTTTAT	1,594	
	D	ATAT ATAAATAATGACGTAATTAAACTGGTAAGATGATAGAAAAAGT TTTTTTTTTTTTT	832	
	D	ATATATA AATGACATGACTAAACTGATAGGACAGTAAAGAGGACAATG ATTTTTTTTTTGTTT	815	
	D	ATATATATATC AATGACATAACTGAACTAGTAAGATAATGAGAGAA TTTTTTTTTCTTTT	756	
	D	ATAT ATAATGACATAACTGAACT TGTAAGATAGCGAGATTTTTTTTTTTTTT	636	
	D	ATATATATC AATGACATAACTGAACTAGTAAGATAATGAGAGAAGTGAT TTTTTTTTTTTT	220	
	D	ATAAATAATGACATAACTGAATTGATAGAACGATGAGAGAAAT TTTTTTTTTT	183	
	D	T TAAATAATGACATAACTGAATTGATAGAACGATGAGAGAAATAAT ATTAATTTTTTTTTTTTT	159	
	D	ATAAATAATGACATAACTAAATTGATAGAACGATGAGAGAAATAAT ATTTT	148	
	D	ATAT ATAAATAATGACGTAATTAAACTGGTAAGATGATAGAAAAAGTGAT TTTAAATTTTTTTTTT	129	

C	C	ATAC AAATCAACAAAACTACTACTCTCTATAGTGA TGATGATATGTATTTTTTGTTTTT	80	41
	C	ATATA AGATCAACAAAACTGCCATTTTCTGTAGTAAGTGATGATATAT TTTTTTTTATTTT		28
	C	ATATAAATCAACAAAACTGA TGTCCTCTGTAGTAGATAGTGATAT TATTTTTTTTTTCTTT	1,587	
	C	ATATAAATCAACAAAACTAG TGCCTTCTATAGTAGATGATGATATATGAT TTTTTTTTTTT	338	333
	C^{t}	ATAT ATATACAAAACAACGGGTATAAGATTAAATGTGAAATGGTATATAT TTTTTTTTTTTTTTTTT	8	
	C^{t}	ATAT ATATACAAAACAACGGGTATAAGATTAAATGTG TTTTTTTTT	1	
B	B1	ATA TACAATACGTGTATGATATTTTATACTAGGTAGATCAGTGAAATTTTTTTTTTT	1,699	17,696
	B1	ATA TACAATACGTGTATGATATTTTATACTGGGTAGATCAGTGAAATT TTTTTT		10
	B1*	ATA TACAATACGTGTATGATATTTATACT AGGTAGATCAGTGAAATTTTTTTTTTTTT	69,971	15,488
	$B2^{FSe}$	ATA TACAATACGTGAATGATATTTTATACT AGGTAGATCAGTGAAATTTTTTTT		4
	$B2^{FSe}$	ATA TACAATACGTGGATGATATTTTATACT AGGTAGATCAGTGAAATTTTTTTTTTTTT	1	2
	$B3^{t}$	ATATA ACAATACGTGAATGAATAATTATAGGGATATATGAGATAAT TTTTTTTTTTT	1,941	
	$B4^{t}$	ATA TACAATACGTGAGAATATATATAGATGTAGAGATATGAGACAGT TTTTTTTTTTTTT	3,453	
A	A	AAATAT AACATATCTTATATCTGAATCTAACTTGTAATATGTG AATTTTTTTTTTTTTTTT	758	158

APPENDIX K. ND7 gRNA Alignments for TREU 667 SDM79 (A) and EATRO 164 SDM79 cells (B), and all editing variants (C and D).

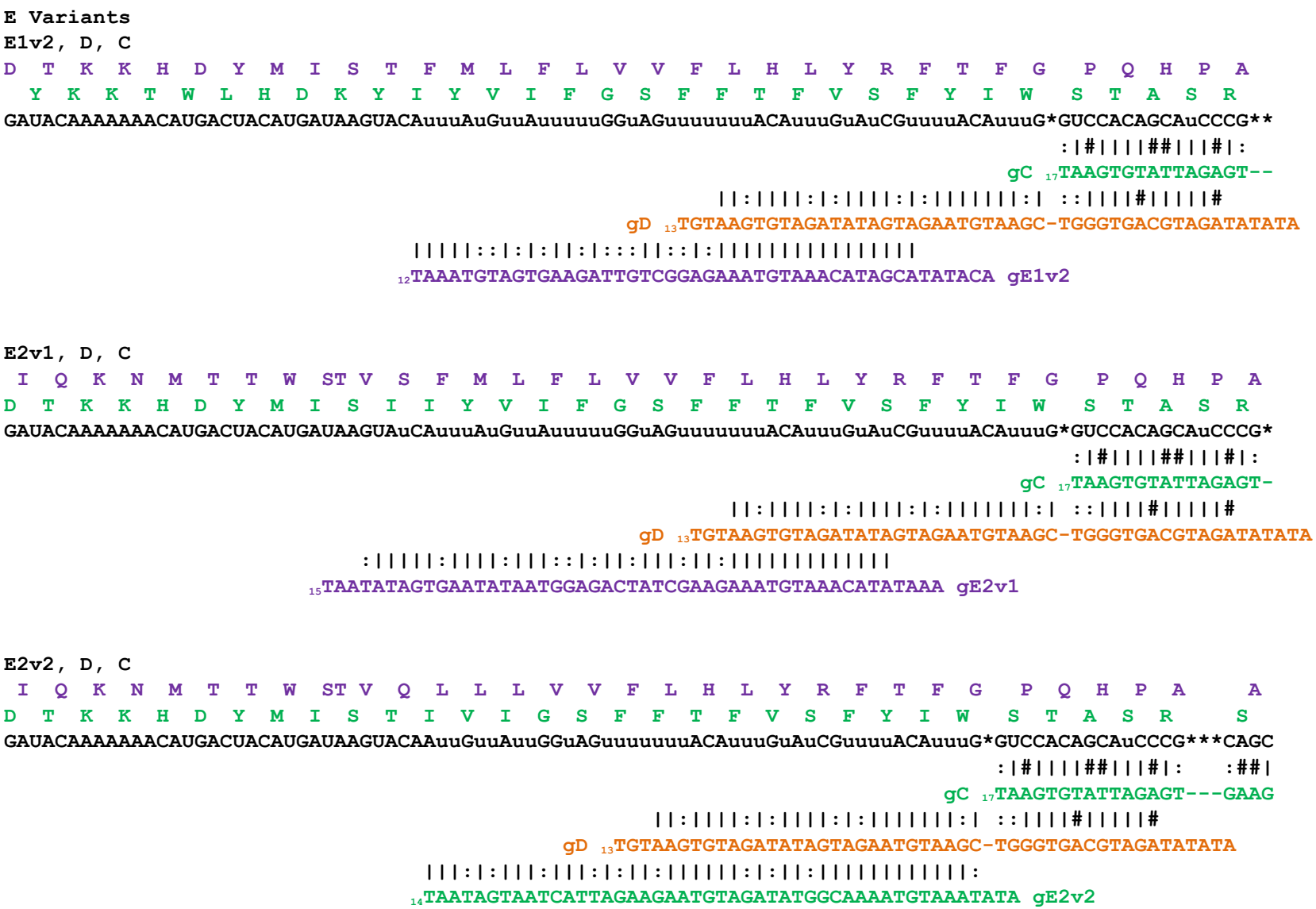
Amino acid translations are shown above the mRNA sequences. The cDNA sequence of the most abundant gRNA in its sequence class is shown aligned beneath the fully edited mRNA. Lowercase u's indicate uridines added by editing, asterisks indicate encoded uridines deleted during editing. Nucleotides and deletion sites in the fully edited mRNA were numbered starting from the 50 end (+1=0). gRNAs are colored based on transcript abundance as follows: Blue<100; Green<1,000; Purple<10,000; Orange<100,000; Red>100,000; Black=not quantified. Watson-Crick (|) and G:U base pairs (:) are indicated. Mismatches are indicated by an octothorpe (#). Highlighted sequence represents sequences were multiple CU configurations are possible.

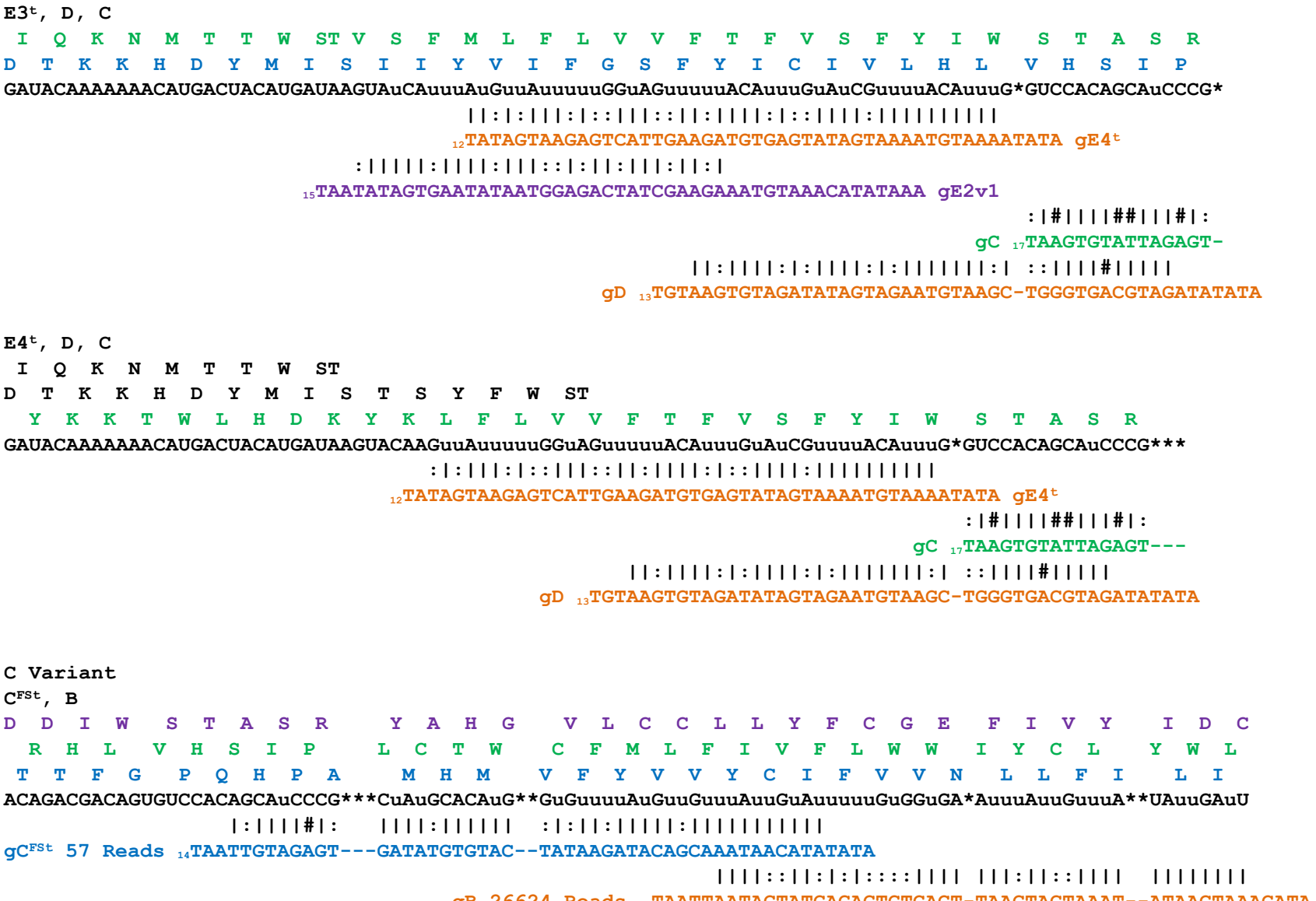
A. TREU 667 ND7



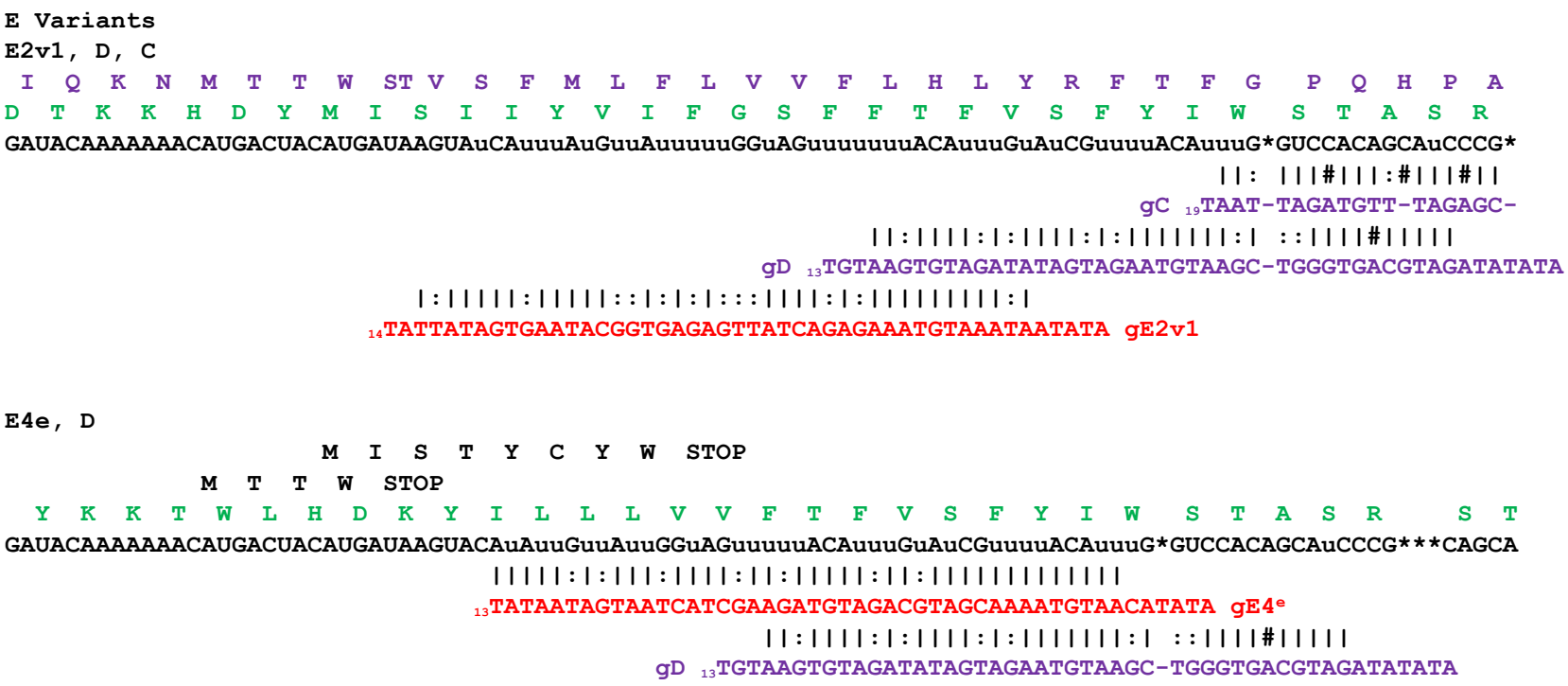
B. EATRO 164 ND7

C. TREU 667 ND7 Variants



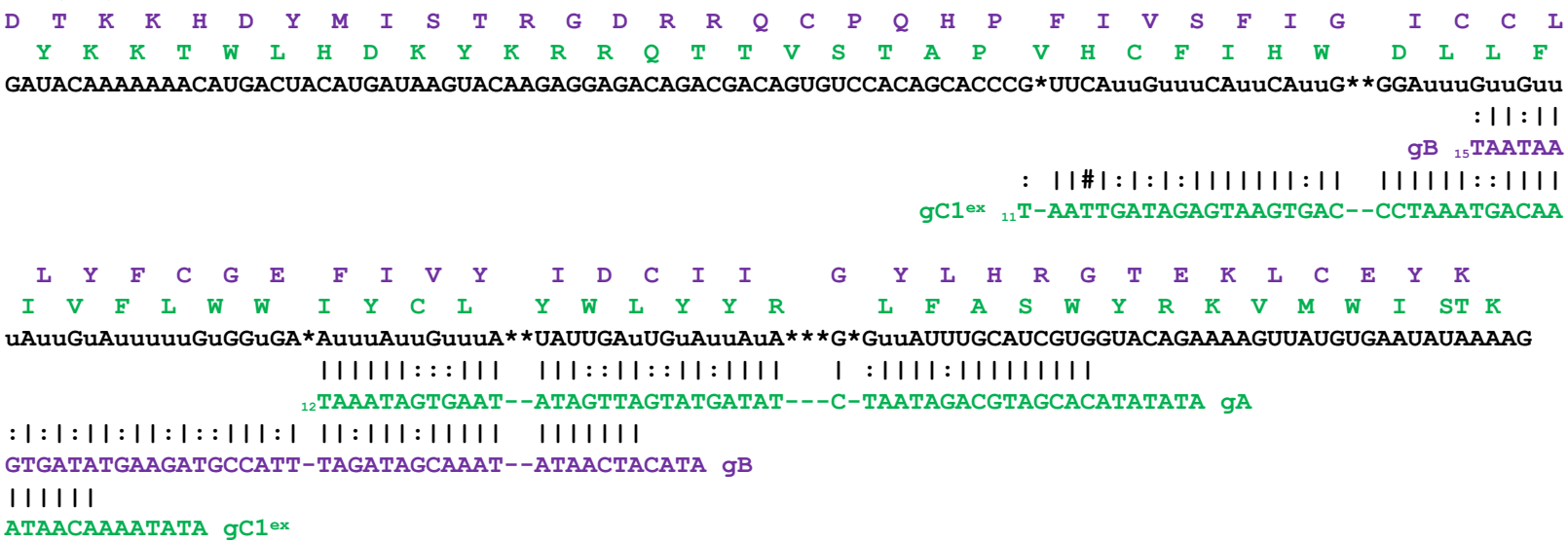

E3t, D, C
I Q K N M T T W ST V S F M L F L V V F T F V S F Y I W S T A S R
D T K K H D Y M I S I I Y V I F G S F Y I C I V L H L V H S I P
GAUACAAAAAAACAUGACUACAUGAUAAGUAuCAuuuAuGuuAuuuuuGGuAGuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAGCAuCCCG*
12TATAGTAAGAGTCATTGAAGATGTGAGTATAGTAAAATGTAAAATATA gE4t
15TAATATAGTGAATATAATGGAGACTATCGAAGAAATGTAAACATATAAA gE2v1
gC 17TAAGTGTATTAGAGT-
gD 13TGTAAGTGTAGATATAGTAGAATGTAAGC-TGGGTGACGTAGATATATA
E4t, D, C
I Q K N M T T W ST
D T K K H D Y M I S T S Y F W ST
Y K K T W L H D K Y K L F L V V F T F V S F Y I W S T A S R
GAUACAAAAAAACAUGACUACAUGAUAAGUACAAGuuAuuuuuGGuAGuuuuuACAuuuGuAuCGuuuuACAuuuG*GUCCACAGCAuCCCG***
12TATAGTAAGAGTCATTGAAGATGTGAGTATAGTAAAATGTAAAATATA gE4t
gC 17TAAGTGTATTAGAGT---
gD 13TGTAAGTGTAGATATAGTAGAATGTAAGC-TGGGTGACGTAGATATATA
C Variant
CFSt, B
D D I W S T A S R Y A H G V L C C L L Y F C G E F I V Y I D C
R H L V H S I P L C T W C F M L F I V F L W W I Y C L Y W L
T T F G P Q H P A M H M V F Y V V Y C I F V V N L L F I L I
ACAGACGACAGUGUCCACAGCAuCCCG***CuAuGCACAuG**GuGuuuuAuGuuGuuuAuuGuAuuuuuGuGGuGA*AuuuAuuGuuuA**UAuuGAuU
gCFSt 57 Reads 14TAATTGTAGAGT---GATATGTGTAC--TATAAGATACAGCAAATAACATATATA
gB 26624 Reads 16TAATTAATAGTATGAGAGTGTCACT-TAAGTAGTAAAT--ATAACTAAACATA

D. EATRO 164 ND7 Variants

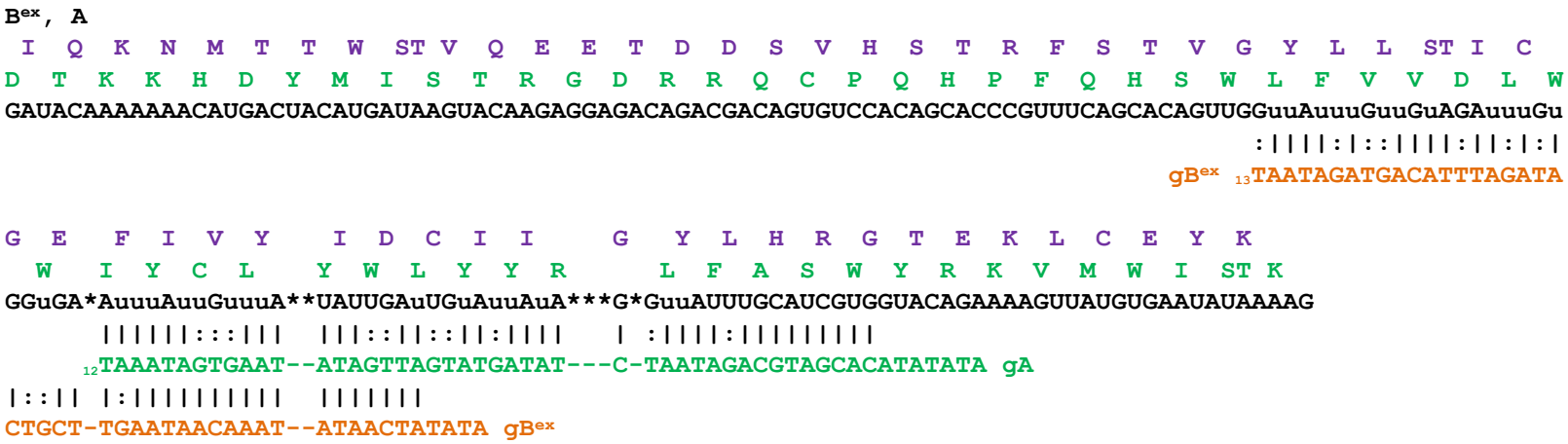


BC Variants

C1^{ex}, B, A



C2^{ex}, B, A


Bex, A
I Q K N M T T W ST V Q E E T D D S V H S T R F S T V G Y L L ST I C
D T K K H D Y M I S T R G D R R Q C P Q H P F Q H S W L F V V D L W
GAUACAAAAAAACAUGACUACAUGAUAAGUACAAGAGGAGACAGACGACAGUGUCCACAGCACCCGUUUCAGCACAGUUGGuuAuuuGuuGuAGAuuuGu
gBex 13TAATAGATGACATTTAGATA
G E F I V Y I D C I I G Y L H R G T E K L C E Y K
W I Y C L Y W L Y Y R L F A S W Y R K V M W I ST K
GGuGA*AuuuAuuGuuuA**UAUUGAuUGuAuuAuA***G*GuuAUUUGCAUCGUGGUACAGAAAAGUUAUGUGAAUAUAAAAG
12TAAATAGTGAAT--ATAGTTAGTATGATAT---C-TAATAGACGTAGCACATATATA gA
CTGCT-TGAATAACAAAT--ATAACTATATA gBex

APPENDIX L. gRNAs identified to edit the ND7 5' mRNAs of found in both TREU 667 and EATRO 164 gRNA transcriptomes.

Editing Region	Population	Sequences	Reads TREU 667	Reads EATRO 164
E	E1 version 1	AAAT ATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATATT ATTTTTTTTT	4,517	6
	E1 version 1	AAAT ATACAAGTGTAAGAAAACTATCGAGAGTGATGTAGAATGATATT ATTTTTTTTTTT		1
	E1 version 1	AAAT ATACAAATGTAAGAAAACTATCGAGAGTGATGTAGAATGATATTT TTTTTTTTTAT	803	
	E1 version 2	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGT TTTTATTTTTTTTTTTTAT		193
	E1 version 2	ACATAT ACGATACAAATGTAAAGAGGCTGTTAGAAGTGATGTAAAT TTTTTTTTTTG	7,073	
	E1 version 2	ATACATAT ACGATACAGATGTGAAGAAACTATTAGAGATAATGTAAAT TTTTCTTTTTT	213	
	E2 version 1	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATT ATTTTTTTTTTTTTT		148,603
	E2 version 1	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATTT TTTTTTTTTGT		6,174
	E2 version 1	ATATA ATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGAT TTTTTTTTTTCT		1,811
	E2 version 1	ATATA ATAAATGTAAAAAGACTATTGAGAGTGGCATAAGTGATATT ATTTTTTTTTGTTTTT		659
	E2 version 1	ATATAATAAA TAAAGAGACTATTGAGAGTGGCATAAGTGATATT ATAATGATTTTTTTTTTTTTT		593
	E2 version 1	AAAT ATACAAATGTAAAGAAGCTATCAGAGGTAATATAAGTGATAT AATTTTTTTGTTTTTTT	1,460	504
	E2 version 1	ATATAAT AATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATT ATAATGATATTTTTTTTT		241
	E2 version 1	AT ATATAAATGTAAAGAGACTATTGAGAGTGGCATAAGTGATATT ATTTTTT		203
	E2 version 1	ATATAC ACAAATGTAAAGAGACTATCGAGAGTGACATAAGTGATAT AATTTTTTTATTTT	192	
	E2 version 2	ATATA ATAAATGTAAAGAGACTATTGATAGTGG CATAAGTGATATTATAATGATTTTTTTTTTTTTTT		34
	E2 version 2	ATA TAAATGTAAAACGGTATAGATGTAAGAAGATTACTAATGATAATT TTTTTTTTTTT	259	
	E4[t]	ATATA ATAAATGTAAAGACTATTGAGAGTGGCATAAGTGATATT ATAATTTTTTTTTTTTT		49
	E4[t]	ATATA ATAAATGTAGAGACTATTGAGAGTGGCATAAGTGATATT ATAATTTTTTTTTTTTT		26
	E4[t]	ATATA AAATGTAAAATGATATGAGTGTAGAAGTTACTGAGAATGATAT TTTTTTTTTTC	10,349	
	E4[t]	ATATA AAATGTAAAATGATATGAGTGTAGAAGTTACTGAGAATGATATA TTTTTTTTTTC	2,680	
	E4[t]	ATATA AAATGTAAAATGATATGAGTGTAGAAGTTACTGAGAATGAT TTTTTTTTTTC	2,360	
	E4[e]	ATATAC AATGTAAAACGATGCAGATGTAGAAGCTACTAATGATAATAT TTTTTTTTTTTG		218,365
	E4[e]	ATATAC AATGTAAAACGATGCAGATGTAGAAGCTACTAATGATAAT TTTTTTTTTTC		18,562
	E4[e]	ATATACAA TAAAACGATGCAGATGTAGAAGCTACTAATGATAATAT TTTTTTATTTTTT		793
	E4[e]	ATATACAATGTAAAACGATT CAGATGTAGAAGCTACTAATGATAATAT TTTCTTTTTTTT		777
	E4[e]	ATATAC AATGTAAAACGATGCAGATGTAGAAGCTACTAATAATAATAT TTTTTTTAATTTTTTTTTTTTT		712
	E4[e]	ATATACAATGT AAACGATGCAGATGTAGAAGCTACTAATGATAATAT TTTTTCTTTTTTTT		439
	E4[e]	ATATAC AATGTAAAACGATACAGATGTAGAAGCTACTAATGATAATAT TTTCTTTTTTT		203
	E4[e]	ATATA AAATGTAAAATGATATGAGTGTAGAAGTTACTGATAATGATAT ATTTTTTTTTTTTT	16	
	E4[e]	ATATA AAATGTAAAATGATATGAGTGTAGAAGTTACTGATAATGAT TCTTTTTTTTTTTTT	5	
D	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTTTTTT TTTTTT	36,611	4,096
	D	ATATATAGATGCTGTA GATTAGATGTAGAGTGATATAAG CGTAAATTTTTTTTTTTTTG		2,136
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATTTTTTTT CTTTTTTT	3,590	797
	D	ATATATAA ATGCTGTGGATTAGATGTAGAATGATATGAGTGTGAAATTTTTTTT TTTTTTC	15,885	24
	D	ATATA AAATGTAAAATGATATGAGTGTAGAAG TTACTGAGAATGATATTTTTTTTTTTTC	12,929	
	D	ATATATAAATGCA GTGGATCAGATGTAAGATGGTATAAGTGTGAATATTTTTTT TTTTT	455	
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTT GTTTTTTTTTT	332	
	D	ATATATAA ATGCTGTGGATTAGATGTAGAATGATATGAGTGTGAATTTTTTTT TTTTTG	332	
	D	GC GGGTCGAATGTAAGATGATATAGATGTGAA TGTATTTTTTTTTTT	229	

	D	AGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTCTTTT TTTTTTT	211	
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTTTTT GTTTTTTT	203	
	D	TTATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATTTTTTTTGT TTT	195	
	D	ATATATAGATGCA GTGGGTCAAATGTAAGATGATATAGATGTGAA TGTTTTTTATTTT	158	
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTTTTTTGT TTTTT	154	
	D	ATATATAGATGCAGT GGTCGAATGTAAGATGATATAGATGTGAA TGTTTTTTTTT	153	
	D	ATATATAGATGCAT TGGGTCGAATGTAAGATGATATAGATGTGAA TGTTTGTTTTTTTTTTTT	128	
	D	GTGGGCCGAATGTAAGATGATATAGATGTGAA TGTTTTTTATTTTTTT	123	
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATGTTTTT GTTTTTTT	112	
	D	TATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTGAATTCTTTTT TTTTTTT	109	
	D	ATATATAAATGC TGGATTAGATGTAGAATGATATGAGTGTGAAA TTTTTTTTTTT	106	
	D	ATATATAGATGCA GTGGGTCGAATGTAAGATGATATAGATGTAAA TGTATTTTTTTTTTTTTT	101	
C	C	ATATAATAAACAACATAAAGTGCCATGT ACTCGAGATTTGTAGATTAATTTTTTTTTTTTTTTTTT		6,185
	C	ATAT ATACAATAAACAATGTGAAATATCATGTG AAGTGAGATTATGTGAATTTCTTTTTTTTTTTT	114	
	C1ex	ATATAAA ACAATAAACAGTAAATCCCAGTGAATGAGATAGT TAATTTTTTTTTTT	183	131
	C2ex	ATATATAAAC ACAATAGAACACACTGTAAGATAGT TATATGAGATATAATTTTTTT	22	103
	C2ex	ATATATAAAC ACAATAGAGCACACTGTAAGATAGT TATATGAGATATAATTTTTTTTTT	14	
	C2ex	ATATATATAAAC ACAATAGAACGCACTGTAAGATAGT TATATGAGATATTTTTTT	13	
	CFSt	ATATA ACAATAAACAACATAGAGCATTGTGTGTATAGTG AGATGTTACTTTTTCTTATTTTTTTGTATTTGTGTT	23	4
	CFSt	ATAT ATACAATAAACGACATAGAATATCATGTGTATAGTG AGATGTTAATTTTTTTTTGTTTT	57	
B	B	ATAC ATCAATATAAACGATAGATTTACCGTAGAAGTATAGTGAATAAT TTTTTTTTTTATTT	4,718	3,969
	B	ATATA CAATATAAACAGTAGATTCACTGCAGAAGTATGATAGATAAT TTTTTTTTTTG		3,557
	B	ATACAT ATATAAACAATGAATTCACTGTGAAGATACGATAGATGATA TATTTTTTTTTGTTTT		390
	B	ATACA AATCAATATAAATGATGAATTCACTGTGAGAGTATGATAA TTAATTTTTTTTTTTTTTCT	26,624	
	B	ATATA CAATATAAACAGTAGATTCACTGCAGAGATATGATAGATAATA TTTTTTTTTTT	3,050	
	B	ATATA CAATATAAACAGTAGATTCACTGCAGAGATATGATAGATAATAA TTTTTTTTTTTCTA	1,675	
	B	ATATA CAATATAAACAGTAGATTCACTGCAGAGATATGATAGATAATAAATTTT TTTTTTT	798	
	B	ATACAT ATATAAACAATGAATTCACTGTGAAGATACAATAGATGATA TATTTTTTTTTTTTT	579	
	B	ATATA CAATATAAACAGTAGATTCACTGCAGAGATATGATAGATAAT TTTTTTTTTT	327	
	B	ATACA AATCAATATAAATGATGAATTCACTGTGAGAGTATGAT TGTTTTTTTTTTT	110	
	B	ATATAT AACAATAAATTCATCATAGAGATATAGTAAGTGATATGAGACATT TTTTTT	30	
	Bex	ATAT ATCAATATAAACAATAAGTTCGTCATAGATTTACAGTAGATAATT TTTTTTTTTT		20,696
	Bex	ATAT ATCAATATAAACAATAAGTTCGTCATAGATTTACAGTAGATAATTA TTTTTTTTTT		1,024
	Bex	AT ATCAATATAAACAATAAGTTCGTCATAGATTTACAGTAGATAATTAA GTTTTTTTTTTT		714
	Bex	AT ATCAATATAAACAATAAGTTCGTCATAGATTTACAGTAGATAATTAATT TTTTTTTTT		693
	Bex	ATAT ATCAATATAAACGATGAATTTGTCATAGATTTACAGTAGATAATTA TTTTTTTTC	57	
	Bex	ATCAATATAAACGATGAATTTGTCATAGATTTACAGTAGATAATT TTTT	38	
	Bex	ATAT ATCAATATAAACAATGAGTTCATCATAGATTTACAGTAGATAATTA TTTTTTTT	25	
A	A	ATATATA CACGATGCAGATAATCTATAGTATGATTGATATAAGTGATAAATTT TTTTTTTT	993	535
	A	ATA TATGATGCAAATAACTTATGATACGATTGATGTAGATGATAAATTT TTTTTTTTTG	4,939	

APPENDIX M. Predicted ND7 protein sequences.

First start codons without premature termination codons were translated. Blue sequences are found in TREU cells only, orange sequences are found in EATRO cells only, and black sequences are found in both cell lines.

RF1

E1v1 MLFLVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYK
E1v2 MISTFMLFLVVFLHLYRFTFGPQHPAAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYK
E2v1FS MISIIYVIFGSFFTFVSFYIWSTASRYAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYK
E2v2FS MISTIVIGSFFTFVSFYIWSTASRYAHGVLCCLLYFCGEFIVYIDCIIGYLHRGTEKLCEYK

RF3

E2v1 MISIIYVIFGSFFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFASWYRKVMWI!
E2v2 MISTIVIGSFFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFASWYRKVMWI!

Minor RF2 variants

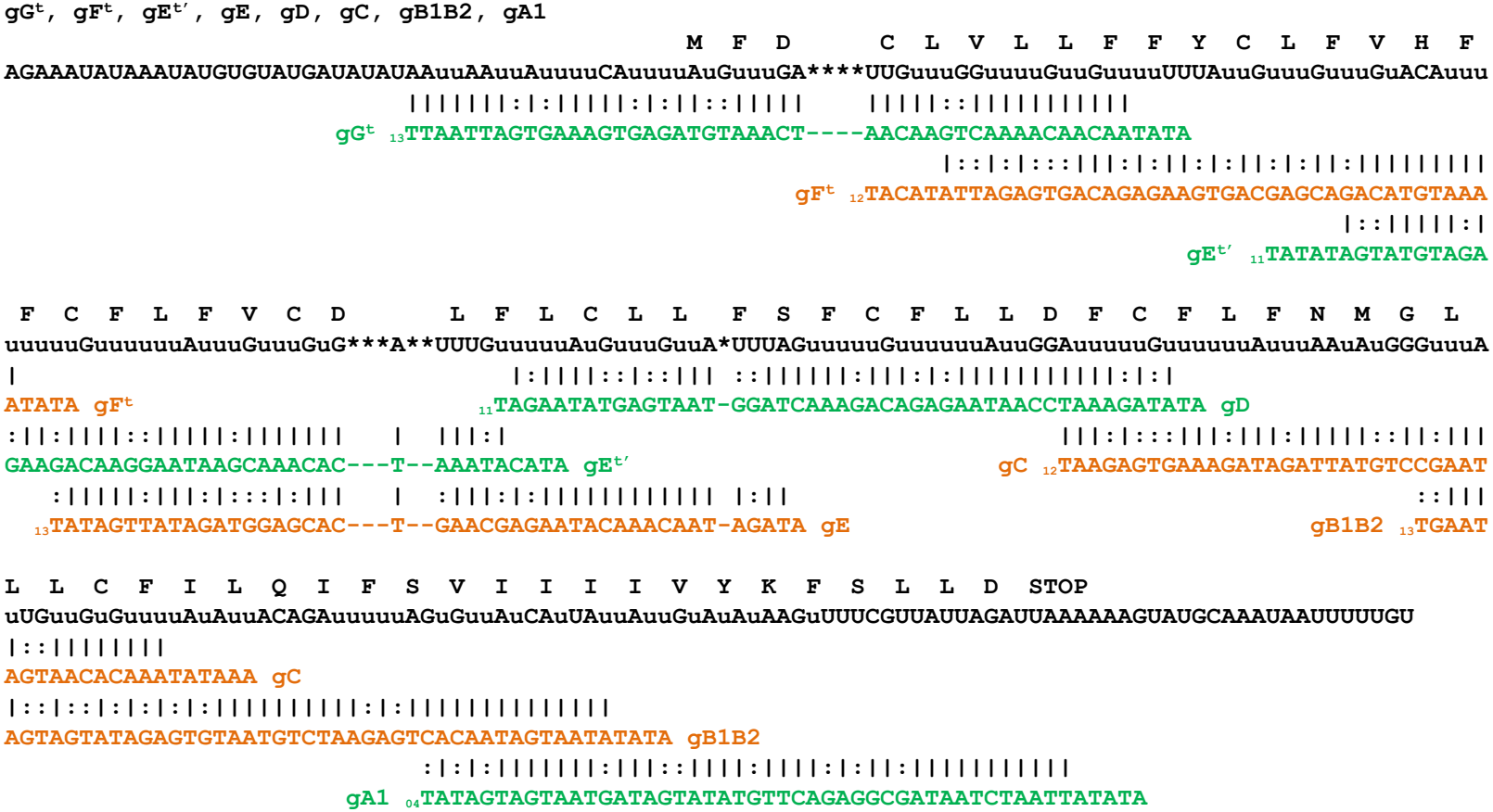
E3t MISIIYVIFGSFYICIVLHLVHSIPQHMVFYVVYCIFVVNLLFILIVL!
E1v1FS MLFLVVFLHLYRFTFGPQHPAMHMVFYVVYCIFVVNLLFILIVL!
E1v2FS MISTFMLFLVVFLHLYRFTFGPQHPAMHMVFYVVYCIFVVNLLFILIVL!

No AUG
E4tRF3 LVVFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFASWYRKVMWI!
E4eRF3 LLLVVFTFVSFYIWSTASRSTWCFMLFIVFLWWIYCLYWLYYRLFASWYRKVMWI!

APPENDIX N. CR3 gRNA Alignments for TREU 667 SDM79 (A), and all editing variants (B).

Amino acid translations are shown above the mRNA sequences. The cDNA sequence of the most abundant gRNA in its sequence class is shown aligned beneath the fully edited mRNA. Lowercase u's indicate uridines added by editing, asterisks indicate encoded uridines deleted during editing. Nucleotides and deletion sites in the fully edited mRNA were numbered starting from the 50 end (+1=0). gRNAs are colored based on transcript abundance as follows: Blue<100; Green<1,000; Purple<10,000; Orange<100,000; Red>100,000; Black=not quantified. Watson-Crick (|) and G:U base pairs (:) are indicated. Mismatches are indicated by an octothorpe (#). Highlighted sequence represents sequences were multiple CU configurations are possible.

A. TREU 667 CR3 gRNA alignment



B. TREU 667 Cell line variants

A Variants

A1

```
 L  L  C  F  I  L  Q  I  F  S  V  I  I  I  I  V  Y  K  F  S  L  L  D  ST K  S  M  Q  I  I  F
   C  C  V  L  Y  Y  R  F  L  V  L  S  L  L  L  Y  I  S  F  R  Y  ST I  K  K  V  C  K  ST F  L
I  V  V  F  Y  I  T  D  F  ST C  Y  H  Y  Y  C  I  ST V  F  V  I  R  L  K  K  Y  A  N  N  F  C
AuUGuuGuGuuuuAuAuuACAGAuuuuuAGuGuuAuCAuUAuuAuuGuAuAuAAGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU
||:::|::|:|:|:|:|||||||||||:|:|||||||||||||||
TAGTAGTATAGAGTGTAATGTCTAAGAGTCACAATAGTAATATATA gB1B2
                                  :|:|:|||||||:|||::||||:||||:|:||:||||||||||||
                                 04TATAGTAGTAATGATAGTATATGTTCAGAGGCGATAATCTAATTATATA gA1
```

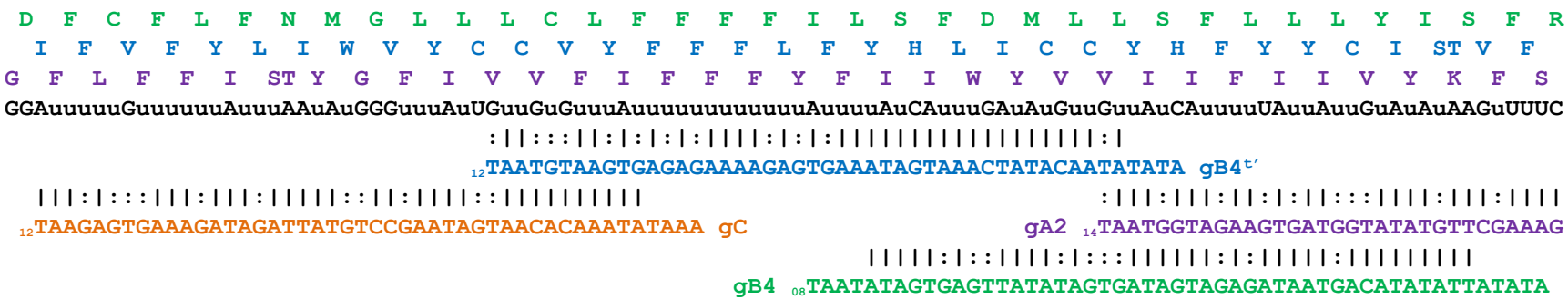
A2

```
 L  L  C  L  Y  Y  F  R  F  Y  G  I  I  F  I  I  V  Y  K  F  S  L  L  D  ST K  S  M  Q  I  I  F
   C  C  V  Y  I  I  S  D  F  M  V  S  F  L  L  L  Y  I  S  F  R  Y  ST I  K  K  V  C  K  ST F  L
I  V  V  F  I  L  F  Q  I  L  W  Y  H  F  Y  Y  C  I  ST V  F  V  I  R  L  K  K  Y  A  N  N  F  C
AuUGuuGuGuuuAuAuuAuuuCAGAuuuuAuGGuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU
                                    ||:|||:||:|:||:::|||||:|||:|||||||:||||||||
                                  14TAATGGTAGAAGTGATGGTATATGTTCGAAAGCAGTAATCTAAAATATA gA2
||:::|::|:|:||||:||:|:||:|||:|||||||||||||||||||
TAGTAGTATAGATATGATGAGGTTTAAGATACCATAGTAAAAATATATA gB3t
```

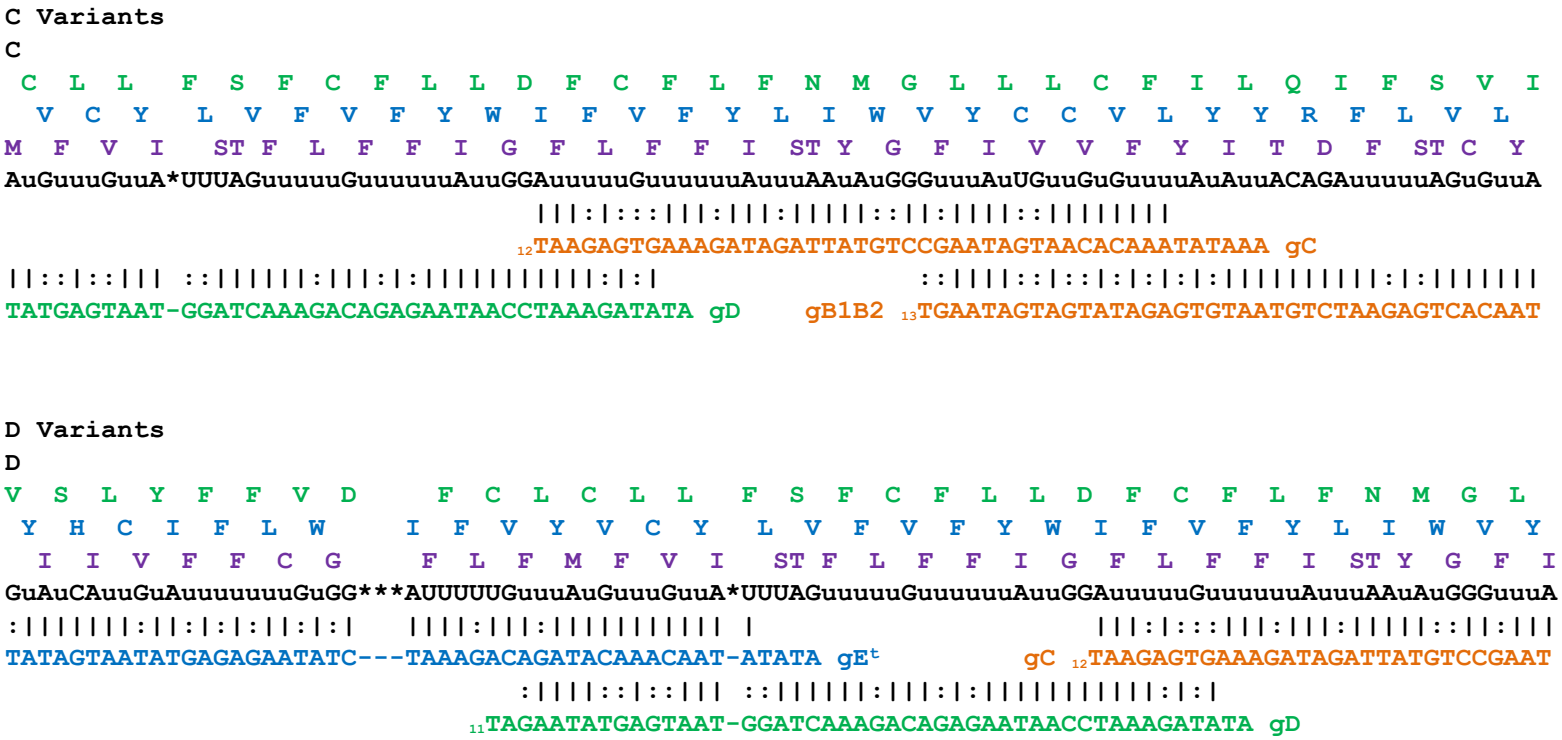
B Variants

B1B2

B4^{t}

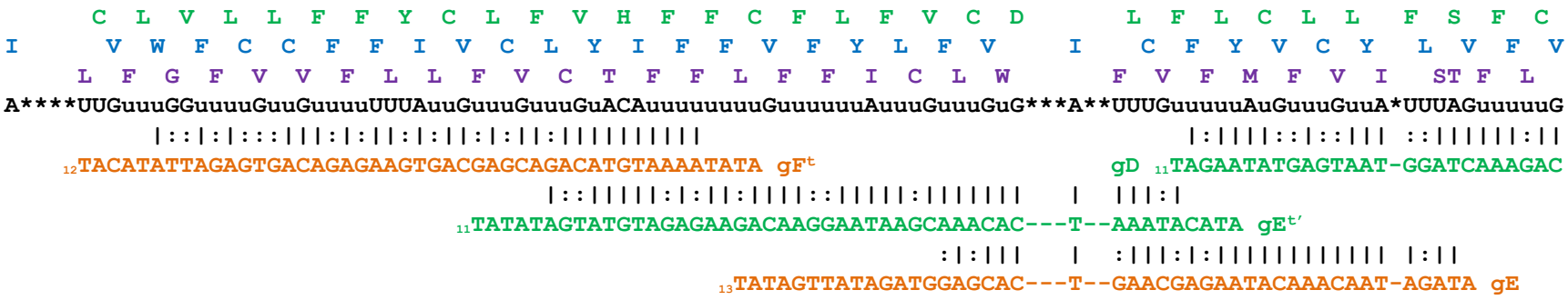


B3^{t}


C Variants
C
C L L F S F C F L L D F C F L F N M G L L L C F I L Q I F S V I
V C Y L V F V F Y W I F V F Y L I W V Y C C V L Y Y R F L V L
M F V I ST F L F F I G F L F F I ST Y G F I V V F Y I T D F ST C Y
AuGuuuGuuA*UUUAGuuuuuGuuuuuuAuuGGAuuuuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuuAuAuuACAGAuuuuuAGuGuuA
12TAAGAGTGAAAGATAGATTATGTCCGAATAGTAACACAAATATAAA gC
TATGAGTAAT-GGATCAAAGACAGAGAATAACCTAAAGATATA gD
gB1B2
13TGAATAGTAGTATAGAGTGTAATGTCTAAGAGTCACAAT
D Variants
D
V S L Y F F V D F C L C L L F S F C F L L D F C F L F N M G L
Y H C I F L W I F V Y V C Y L V F V F Y W I F V F Y L I W V Y
I I V F F C G F L F M F V I ST F L F F I G F L F F I ST Y G F I
GuAuCAuuGuAuuuuuuuGuGG***AUUUUUGuuuAuGuuuGuuA*UUUAGuuuuuGuuuuuuAuuGGAuuuuuGuuuuuuAuuuAAuAuGGGuuuA
TATAGTAATATGAGAGAATATC---TAAAGACAGATACAAACAAT-ATATA gE
t
gC
12TAAGAGTGAAAGATAGATTATGTCCGAAT
11TAGAATATGAGTAAT-GGATCAAAGACAGAGAATAACCTAAAGATATA gD

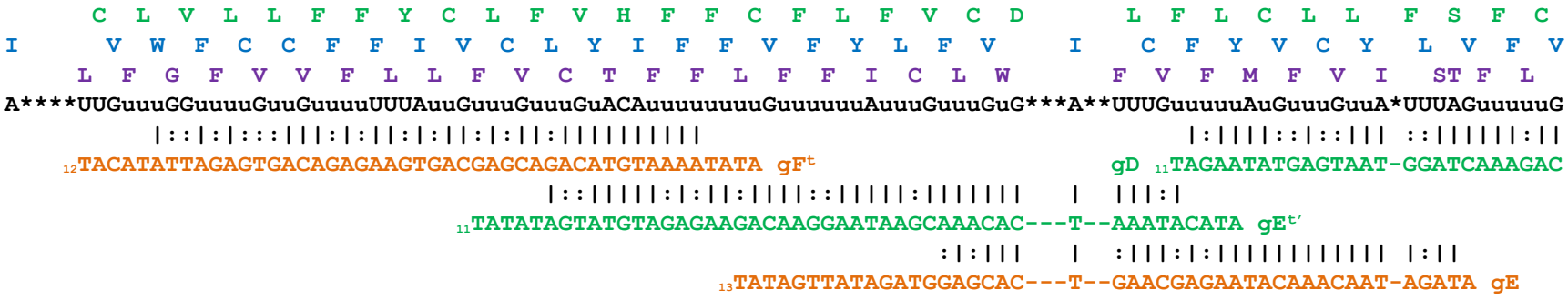
E Variants

E



E^t

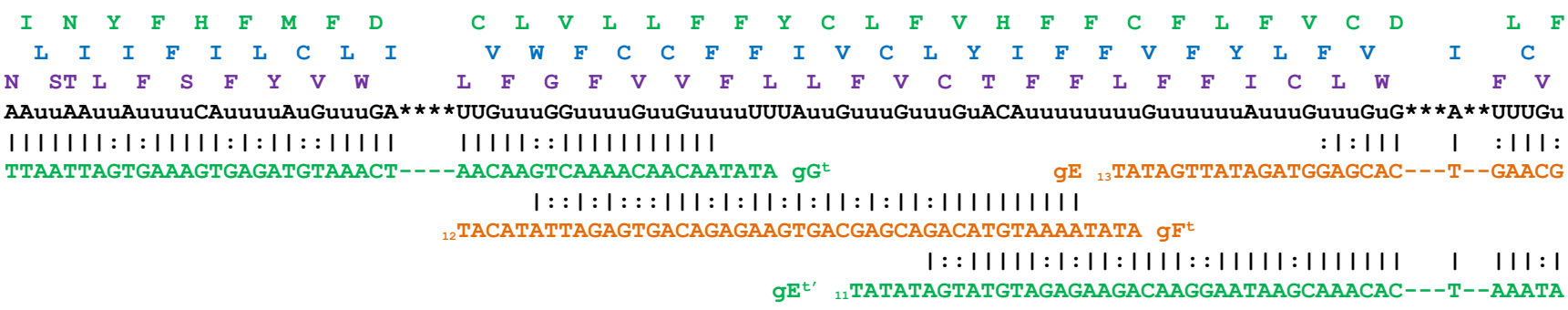
E^t'



FG Variants

FG^{tx}

F^{t}



FG^{t}


Gt
E I ST I C V W Y I I N Y F H F M F D C L V L L F F Y C L F V
K Y K Y V Y D I ST L I I F I L C L I V W F C C F F I V C L
R N I N M C M I Y N ST L F S F Y V W L F G F V V F L L F V C
AGAAAUAUAAAUAUGUGUAUGAUAUAUAAuuAAuuAuuuuCAuuuuAuGuuuGA****UUGuuuGGuuuuGuuGuuuuUUUAuuGuuuGuuuG
13TTAATTAGTGAAAGTGAGATGTAAACT----AACAAGTCAAAACAACAATATA gGt
gFt 12TACATATTAGAGTGACAGAGAAGTGACGAGCAGAC

APPENDIX O. CR3 gRNA Alignments for EATRO 164 (A), and all editing variants (B).

Amino acid translations are shown above the mRNA sequences. The cDNA sequence of the most abundant gRNA in its sequence class is shown aligned beneath the fully edited mRNA. Lowercase u's indicate uridines added by editing, asterisks indicate encoded uridines deleted during editing. Nucleotides and deletion sites in the fully edited mRNA were numbered starting from the 50 end (+1=0). gRNAs are colored based on transcript abundance as follows: Blue<100; Green<1,000; Purple<10,000; Orange<100,000; Red>100,000; Black=not quantified. Watson-Crick (|) and G:U base pairs (:) are indicated. Mismatches are indicated by an octothorpe (#). Highlighted sequence represents sequences were multiple CU configurations are possible.

A. EATRO 164 CR3 SDM79/SDM80

```
gFGep, gEep, gDep, gCe, gB5e, gB4, gA2
            M  C  M  I  Y  K  L  T  I  V  L        G  G     I  L  V  I  I  V  Y  L  V  V   M      S
AGAAAUAUAAAUAUGUGUAUGAUAUAUAAAuuAACAAuuGuGuuA******GGuGGG***AuuuuGGuGAuCAuuGuuuAuuuGGuuG*UUA****UGAG
                                |||||||||:::|:|||      |||:::   ||:|::|||||||||:|||
                              13TAATTGTTGGTATAAT------CCATTT---TAGAGTCACTAGTAGCAACATATA gFGep
                                                                      ||||||::||:|:||:::|| :||    ::||
                                                                  gEep 12TATAGTAGTAAGTGAATTGAC-GAT----GTTC
                                                                                  :|||: |:|    ::|:
                                                                               gDep 12TAAT-AGT----GTTT


 C  I  L  C  F  V  M  V  I  V  F  Y  L  I  W  V  Y  C  C  V  Y  I  T  Y  V  F  L  L  L  L  S  F  L
uuGuAUUUUAuGuuuuGuuAuGGuuAuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuAuCAuuuuUA
|::|||||||||||||||                                  |||:|::::||||:||:||||||:|||:|:||:||||||||||||
AGTATAAAATACAAAATATATA gEep                       12TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAATAGTATATA gB5e
:|||:|:|||||:|:|:::|:|||:||||||||||                                                |:|:::|||||:|:||
GACGTGAAATATAGAGTAGTACTAATAACAAAATATA gDep                              gB4 11TAATATAGTGAGTTATATAGTGATAGTAGAGAT
                        ::|||:::|||:|||:|||:||:|||||||::||||||||||||||                    :|||||::||:|
                      13TGATAGTGAAAGATAGATTGTATCCAAATAGTAACACAAATATAAA gCe          gA2 13TAATAGTGGAAGT


L  L  Y  I  S  F  R  Y  STOP
uuAuuGuAuAuAAGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU
|||:|||||||||
AATGACATATATTATATA gB4
|:||::||||:|||||:||||:||||||||
AGTAGTATATGTTCAAGAGCAGTAATCTAAAATATA gA2
```

B. EATRO 164 CR3 Variants

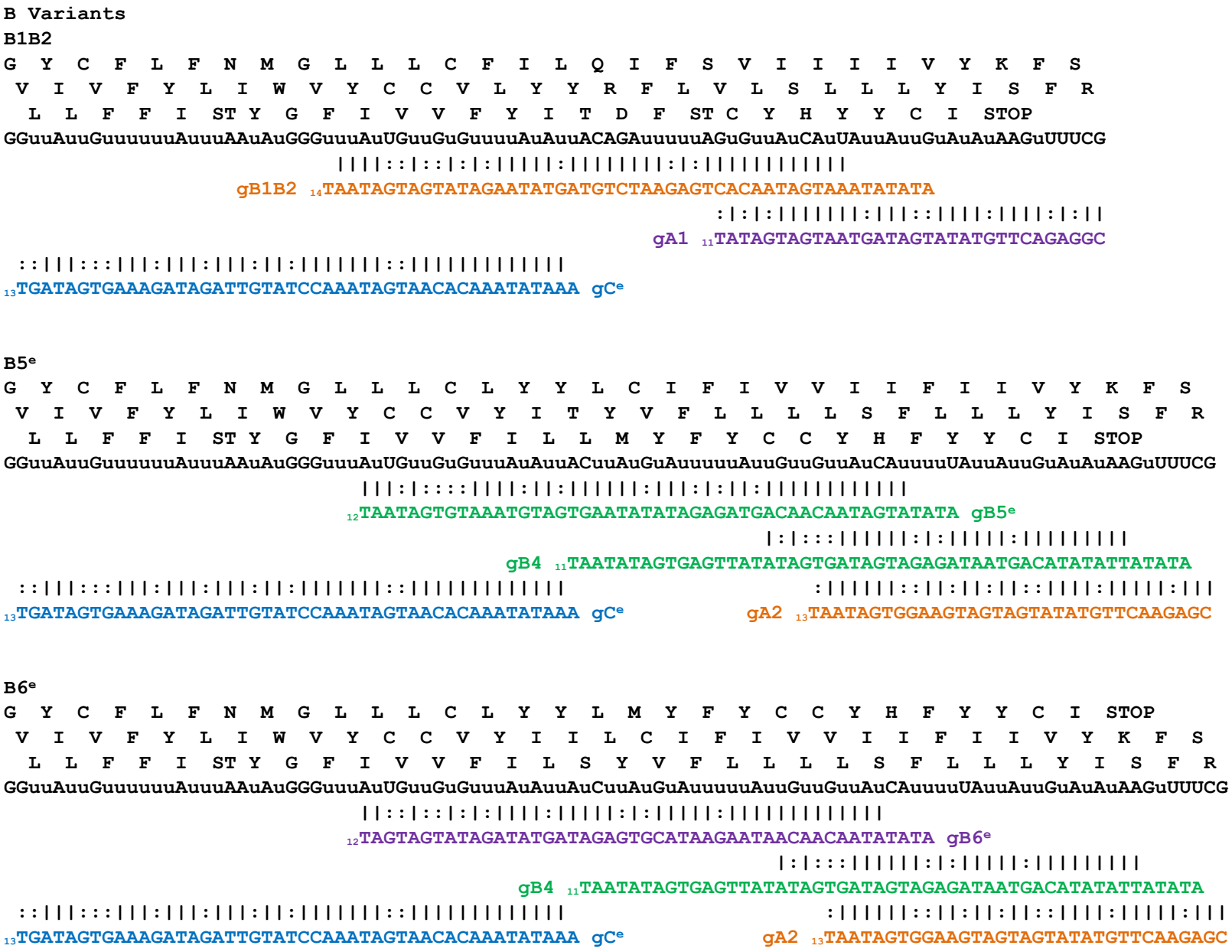
A Variants

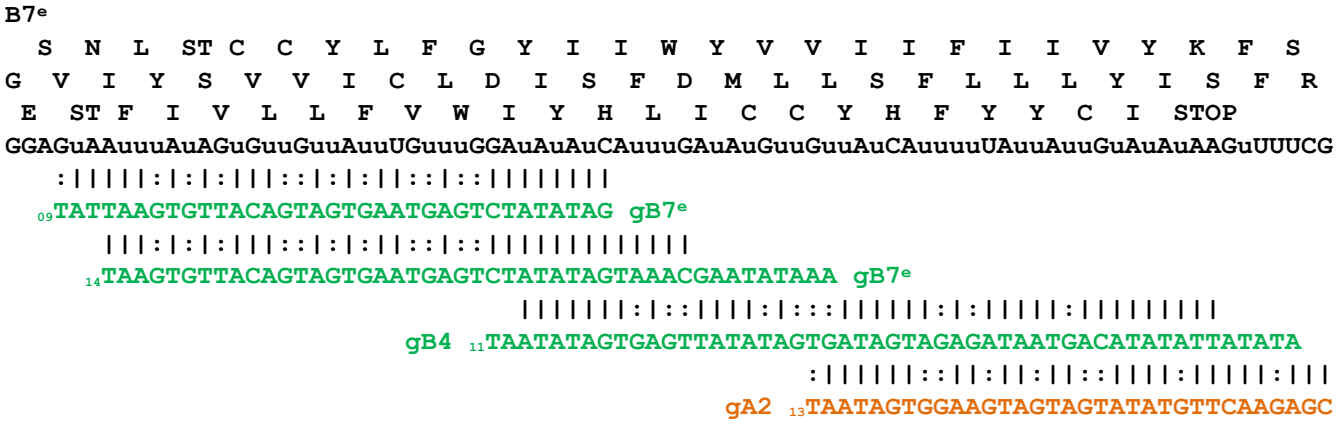
A1

```
 L  L  C  F  I  L  Q  I  F  S  V  I  I  I  I  V  Y  K  F  S  L  L  D  STOP
   C  C  V  L  Y  Y  R  F  L  V  L  S  L  L  L  Y  I  S  F  R  Y  STOP
I  V  V  F  Y  I  T  D  F  ST C  Y  H  Y  Y  C  I  STOP
AuUGuuGuGuuuuAuAuuACAGAuuuuuAGuGuuAuCAuUAuuAuuGuAuAuAAGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU
                                :|:|:|||||||:|||::||||:||||:|:|||||||||||||||
                               11TATAGTAGTAATGATAGTATATGTTCAGAGGCAATAATCTAATTATA gA1
||::|::|:|:|||||:|||||||||:|:||||||||||||
TAGTAGTATAGAATATGATGTCTAAGAGTCACAATAGTAAATATATA gB1B2
```

A2

```
 L  L  C  L  Y  Y  L  C  I  F  I  V  V  I  I  F  I  I  V  Y  K  F  S  L  L  D  STOP
   C  C  V  Y  I  T  Y  V  F  L  L  L  L  S  F  L  L  L  Y  I  S  F  R  Y  STOP
I  V  V  F  I  L  L  M  Y  F  Y  C  C  Y  H  F  Y  Y  C  I  STOP
AuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU
                                        :|||||||::||:||:||::||||:|||||:||||:||||||||
                                       13TAATAGTGGAAGTAGTAGTATATGTTCAAGAGCAGTAATCTAAAATATA gA2
                                     |:|:::|||||||:|:|||||:|||||||||
                        11TAATATAGTGAGTTATATAGTGATAGTAGAGATAATGACATATATTATATA gB4
|||:|::::||||:||:|||||||:|||:|:||:||||||||||||
TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAATAGTATATA gB5e
```


B Variants
B1B2
G Y C F L F N M G L L L C F I L Q I F S V I I I I V Y K F S
V I V F Y L I W V Y C C V L Y Y R F L V L S L L L Y I S F R
L L F F I STOP Y G F I V V F Y I T D F STOP C Y H Y Y C I
GGuuAuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuuAuAuuACAGAuuuuuAGuGuuAuCAuUAuuAuuGuAuAuAAGuUUUCG
gB1B2 14TAATAGTAGTATAGAATATGATGTCTAAGAGTCACAATAGTAAATATATA
gA1 11TATAGTAGTAATGATAGTATATGTTCAGAGGC
13TGATAGTGAAAGATAGATTGTATCCAAATAGTAACACAAATATAAA gCe
B5e
G Y C F L F N M G L L L C L Y Y L C I F I V V I I F I I V Y K F S
V I V F Y L I W V Y C C V Y I T Y V F L L L L S F L L L Y I S F R
L L F F I STOP Y G F I V V F I L L M Y F Y C C Y H F Y Y C I
GGuuAuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCG
12TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAATAGTATATA gB5e
gB4 11TAATATAGTGAGTTATATAGTGATAGTAGAGATAATGACATATATTATATA
13TGATAGTGAAAGATAGATTGTATCCAAATAGTAACACAAATATAAA gCe
gA2 13TAATAGTGGAAGTAGTAGTATATGTTCAAGAGC
B6e
G Y C F L F N M G L L L C L Y Y L M Y F Y C C Y H F Y Y C I STOP
V I V F Y L I W V Y C C V Y I I L C I F I V V I I F I I V Y K F S
L L F F I STOP Y G F I V V F I L S Y V F L L L L S F L L L Y I S F R
GGuuAuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuAuAuuAuCuuAuGuAuuuuuAuuGuuGuuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCG
12TAGTAGTATAGATATGATAGAGTGCATAAGAATAACAACAATATATA gB6e
gB4 11TAATATAGTGAGTTATATAGTGATAGTAGAGATAATGACATATATTATATA
13TGATAGTGAAAGATAGATTGTATCCAAATAGTAACACAAATATAAA gCe
gA2 13TAATAGTGGAAGTAGTAGTATATGTTCAAGAGC


B7e
S N L ST C C Y L F G Y I I W Y V V I I F I I V Y K F S
G V I Y S V V I C L D I S F D M L L S F L L L Y I S F R
E ST F I V L L F V W I Y H L I C C Y H F Y Y C I STOP
GGAGuAAuuuAuAGuGuuGuuAuuUGuuuGGAuAuAuCAuuuGAuAuGuuGuuAuCAuuuuUAuuAuuGuAuAuAAGuUUUCG
09TATTAAGTGTTACAGTAGTGAATGAGTCTATATAG gB7e
14TAAGTGTTACAGTAGTGAATGAGTCTATATAGTAAACGAATATAAA gB7e
gB4 11TAATATAGTGAGTTATATAGTGATAGTAGAGATAATGACATATATTATATA
gA2 13TAATAGTGGAAGTAGTAGTATATGTTCAAGAGC

C Variants

C

```
 C  L  L   F  S  F  C  F  L  L  D  F  C  F  L  F  N  M  G  L  L  L  C  L  Y  Y  L  C  I  F  I  V  V  I
  V  C  Y   L  V  F  V  F  Y  W  I  F  V  F  Y  L  I  W  V  Y  C  C  V  Y  I  T  Y  V  F  L  L  L  L
M  F  V  I   ST F  L  F  F  I  G  F  L  F  F  I  ST Y  G  F  I  V  V  F  I  L  L  M  Y  F  Y  C  C  Y
AuGuuuGuuA*UUUAGuuuuuGuuuuuuAuuGGAuuuuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuA
                            |||::||:||::|:|:|:|:|||||:|:|||||||||||||:|
                        gC 14TAATTTAGAAGTAGAGAGTGAATTATGCTCAAATAACAACATATATA
||::|::||| |:|||:||:|||:|:|||||||||||:|:|                    |||:|::::||||:||:|||||:|||:|:||:||||||||
TATGAGTAAT-AGATCGAAGACAGAGAATAACCTAAAGATA gD                gB5e 12TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAAT
```

Ce

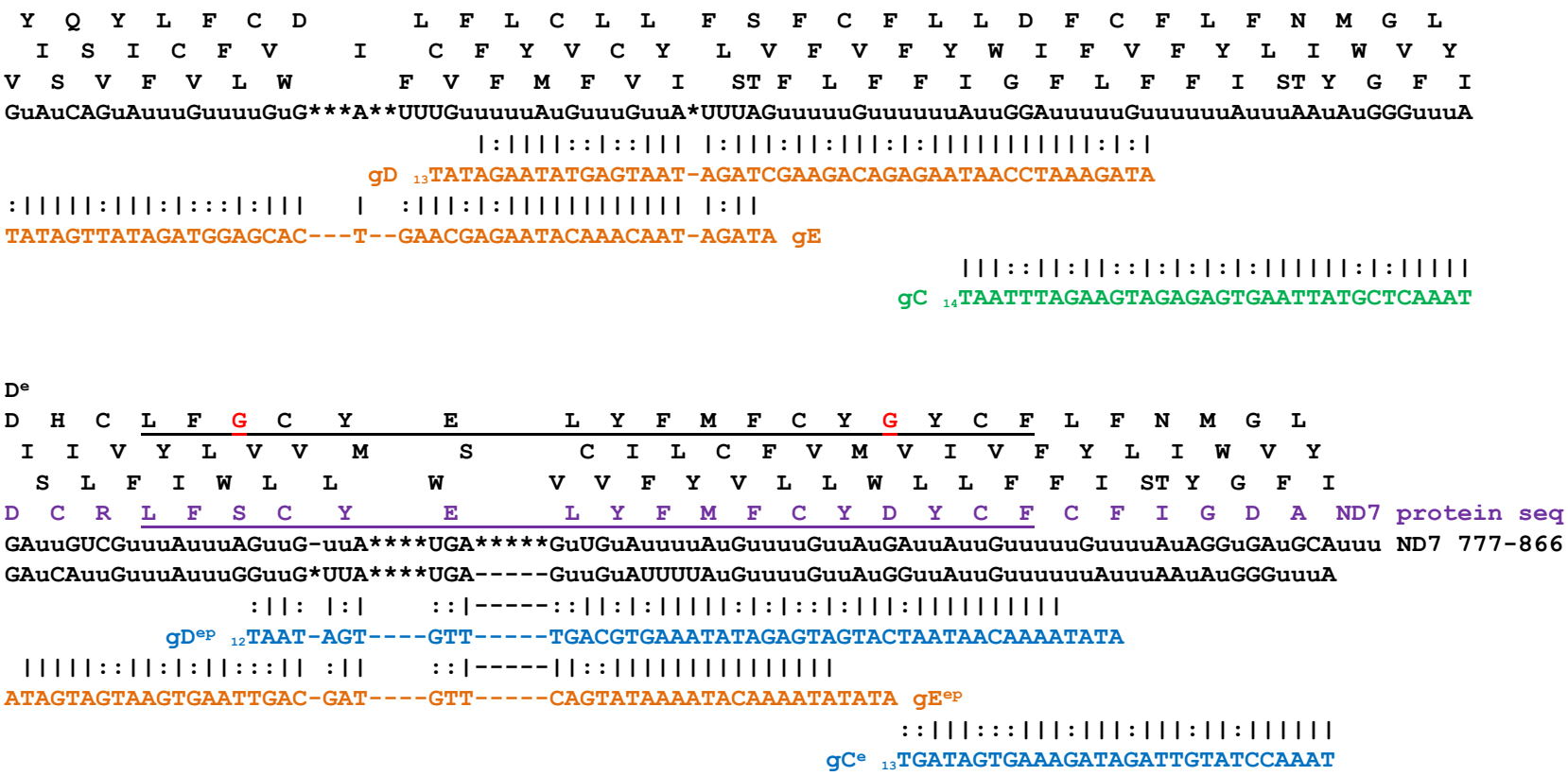
```
  L  Y  F  M  F  C  Y  G  Y  C  F  L  F  N  M  G  L  L  L  C  L  Y  Y  L  C  I  F  I  V  V  I
S  C  I  L  C  F  V  M  V  I  V  F  Y  L  I  W  V  Y  C  C  V  Y  I  T  Y  V  F  L  L  L  L
 V  V  F  Y  V  L  L  W  L  L  F  F  I  ST Y  G  F  I  V  V  F  I  L  L  M  Y  F  Y  C  C  Y
AGuuGuAUUUUAuGuuuuGuuAuGGuuAuuGuuuuuuAuuuAAuAuGGGuuuAuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuA
                          ::|||:::|||:|||:|||:||:|||||||::|||||||||||||
                        13TGATAGTGAAAGATAGATTGTATCCAAATAGTAACACAAATATAAA gCe
                                                          |||:|::::||||:||:|||||:|||:|:||:||||||||
                                                    gB5e 12TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAAT
|::||:|:|||||:|:|:::|:|||:||||||||||
TTGACGTGAAATATAGAGTAGTACTAATAACAAAATATA gDep
```

Ce80

```
S  C  I  L  C  F  V  M  F  C  M  I  I  L  ST C  D  L  L  C  L  Y  Y  L  C  I  F  I  V  V  I
 V  V  F  Y  V  L  L  C  F  V  W  L  F  Y  S  V  I  C  C  V  Y  I  T  Y  V  F  L  L  L  L
  L  Y  F  M  F  C  Y  V  L  Y  D  Y  F  I  V  W  F  V  V  F  I  L  L  M  Y  F  Y  C  C  Y
AGuuGuAUUUUAuGuuuuGuuAuGuuuuGuAuGAuuAuuuuAuAGuGuGAuuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuA
                    :|||||:|:|::|||:|:|||:|||:||||||||:||||||||||||||
                  14TAATATAGAGTATATTGATAGAATGTCACACTAGATAACACAAATATATATA gCe80
                                                          ||:|::::||||:||:|||||:|||:|:||:||||||||
                                                    gB5e 12TAATAGTGTAAATGTAGTGAATATATAGAGATGACAACAAT
|::||:|:|||||:|:|:::|:|||
TTGACGTGAAATATAGAGTAGTACTAATAACAAAATATA gDep
```

D Variants

D



De

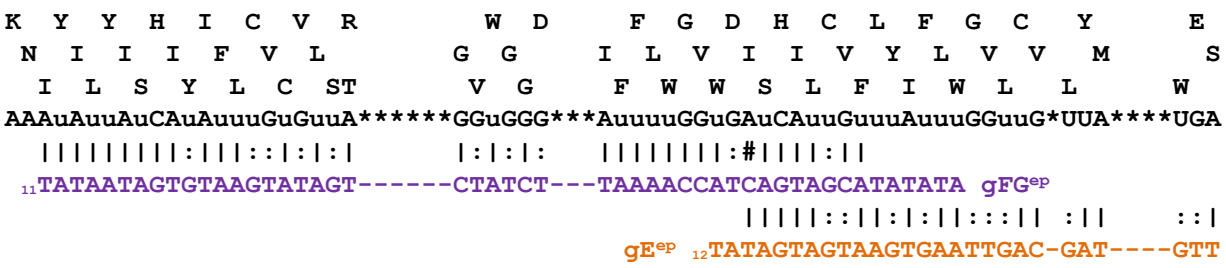
E Variants

E^e

E

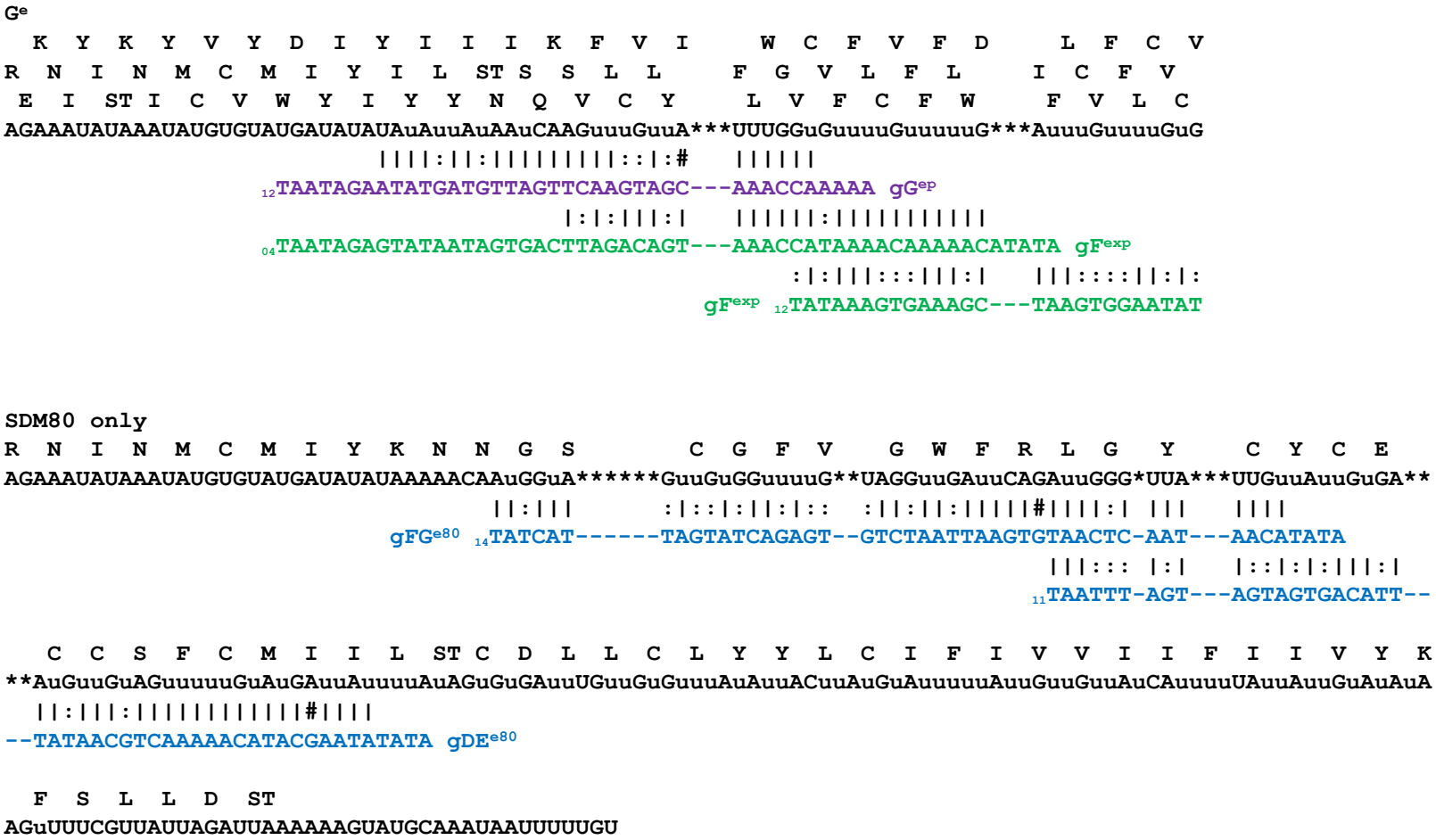
FG Variants

F^{e}



F^{e*}

F^{ex}


Ge
K Y K Y V Y D I Y I I I K F V I W C F V F D L F C V
R N I N M C M I Y I L ST S S L L F G V L F L I C F V
E I ST I C V W Y I Y Y N Q V C Y L V F C F W F V L C
AGAAAUAUAAAUAUGUGUAUGAUAUAUAuAuuAuAAuCAAGuuuGuuA***UUUGGuGuuuuGuuuuuG***AuuuGuuuuGuG
12TAATAGAATATGATGTTAGTTCAAGTAGC---AAACCAAAAA gGep
04TAATAGAGTATAATAGTGACTTAGACAGT---AAACCATAAAACAAAAACATATA gFexp
gFexp 12TATAAAGTGAAAGC---TAAGTGGAATAT
SDM80 only
R N I N M C M I Y K N N G S C G F V G W F R L G Y C Y C E
AGAAAUAUAAAUAUGUGUAUGAUAUAUAAAAACAAuGGuA******GuuGuGGuuuuG**UAGGuuGAuuCAGAuuGGG*UUA***UUGuuAuuGuGA**
gFGe80 14TATCAT------TAGTATCAGAGT--GTCTAATTAAGTGTAACTC-AAT---AACATATA
11TAATTT-AGT---AGTAGTGACATT--
C C S F C M I I L ST C D L L C L Y Y L C I F I V V I I F I I V Y K
**AuGuuGuAGuuuuuGuAuGAuuAuuuuAuAGuGuGAuuUGuuGuGuuuAuAuuACuuAuGuAuuuuuAuuGuuGuuAuCAuuuuUAuuAuuGuAuAuA
--TATAACGTCAAAAACATACGAATATATA gDEe80
F S L L D ST
AGuUUUCGUUAUUAGAUUAAAAAAGUAUGCAAAUAAUUUUUGU

APPENDIX P. gRNAs identified to edit the CR3 mRNAs of found in both TREU 667 and EATRO 164 gRNA transcriptomes.

Editing Region	Population	Sequence	Reads TREU 667	Reads EATRO 164
G	gG^{t}	ATATGT ACAACAAAACCGAGCAATCAGATATAGAGTGAAAGTGATTAATT TTTTTTTTTTTC		21,297
	gG^{t}	ATAT AACAACAAAACTGAACAATCAAATGTAGAGTGAAAGTGATTAATT TTTTTTTTTTTT	493	
	gG^{ep}	AAAAACCAAAC GATGAACTTGATTGTAGTATA AGATAATTTTTTTTTTTT	8,171	2,736
	gG^{ep}	AATACCGAGC GACAGATTTGATTGTAGTATA AGATAATATTTTTTTTTTTT		94
FG	gFG^{tp}	ATATAT AAAATGTACAACCAGAATTAAGATAAAGTGATGATGTATATATT TT	21	
	gFG^{txp}	ATATTAAATAC ATGATATACGCGATAGATTATTAGAGTTATGAGTTAAT TTTTTTTTTTTGTTT		2,772
	gFG^{txp}	ATAC ATGATATATACAGTGAACTATTAGAATTATAGGT AATGAGATTTATTTTTTTTTTTTT		354
	gFG^{txp}	ATATTAAATAC ATGATATGCGCAGTAGACTATTAAAGTTATGAGTTAAT TTTTTTTTTT		159
	gFG^{txp}	ATATTAAATAC ATGATATACACGGTAGATTATTAGAGCTATGAGTTAAT TTTTTTTTTTTT	44,119	
	gFG^{txp}	ATATTAAATAC ATGATATACACGGTAGATTATTAGAGCTATGAGTTA TTTTTTTTTTTCT	2,106	
	gFG^{txp}	ATATTAAATAC ATGATATACACGGTAGATTATTAGAGCTATGAGTT TTTTTTTTTT	936	
	gFG^{txp}	TATTAAATAC ATGATATACACGGTAGATTATTAGAGCTATGAGTTAA CTTTTTTTTTTTT	404	
	gFG^{txp}	ATATTAAATAC ATATACACGGTAGATTATTAGAGCTATGAGTTAAT TTTTT	133	
	gFG^{txp}	ATATTAAATAC ATGATATACACGGTAGATTATTAGAGCTATGA TTTAAAATTTTTTTTTTTTT	115	
	gFG^{e*p}	ATATAC AACGATGATCACTGAGATTTTACCTAATATGG ATGTTAATTTTTTTTTTTTTGGGGAACTGAA	53,389	53,459
	gFG^{e*p}	ATACA AAACGATGATTACCGAGATTTCA GTTAATATGATTGTCTAATTTTTTTTTTTT		717
	gFG^{e*p}	ATATACAACT ATGATCACTGAGATTTTACCTAATATGG ATGTTAATTTTTTTTTT	133	166
	gFG^{e*p}	ATATATACA AAACGATGATTACCGAGATTTCATTTAATATGATTGT CTAATTTT		143
	gFG^{e*p}	ATATAC AACGATGATCACTGAGATTTTACCTAATATG TATGTTAATTTTTTCTTTTT		97
	gFG^{e*p}	ATATAC AACGATGATCACTGAGATTTTACCTAATATGGTTGTTAATTT TAGATTTTTT	42	23
	gFG^{e*p}	ATATAC AACAATGATCACTGAGATTTTACCTAATATGGTTGTTAATTT TTTTTTTTTTT		1
	gFG^{e*p}	ATATAGAAACGATGAC TGCTAGAATTCTGCTTGATATGGATGT TAATTTTTTTTTTTTTT	12,250	
	gFG^{e*p}	ATATAC AACGATGATCACTGAGATTTTACCTAATACGA TGTTTAATTTTTTTTTTTTTTAT	1,418	
	gFG^{e*p}	ATATAC AACGATGATCACTGAGATTTTACCTAATATGGAT TTTTTTTTTTTTTTTAAAGTGCGGCCATAGTGGGTG	534	
	gFG^{ep}	ATATATACGATGAC TACCAAAATTCTATCTGATATGAATGTGATAATATTT TTTTTTTT		1,854
	gFG^{ep}	ATATATACGATGAC TGCCAAAATTCTATCTGATATGAATGTGATAATATTT TTTTTTTT	173	
	gFG^{ep}	ATATAC AACGATGATCACTGAGATTTTACCTAAT TTTTTTTTTTT	140	
	gFG^{e80}	ATATACAATAACTCAATG TGAATTAATCTGTGAGACTATGATTACT TTTTTTTGTTTTT	146	349
	gFG^{e80}	ATATACAATAACTCAATG TGAATTAATCTGTGAGACTATGATTACTATT TTTTTTTATTTT	112	239
F	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGAT ATTTTTTTTTTTTTC		138,710
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGATT TTTTTTTTTTTT		33,909
	gF^{t}	ATATAATT AAATGTACAGACAAATGATAGAGAGACGATGAGATTAAGT TATATTTTTTTTTTTT	16,892	3,738
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGAT TTTTTTTTGTTTTT		941
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATG TTTTTTTTTTT		932
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTAAAATTAGATGAT ATTTTTTTTTTTT		679
	gF^{t}	ATATATAAAATA TACAAACGGACAATGAGAGAACAGTGAAATTAGATGAT AATATATTTTT		603
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGA AATATTTTTTTTTTTTTT		459
	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATA TAATATTTTTTTTTTTA		263
	gF^{t}	ATATAT AAATGTACAAACGGACAATGAGAGAACAGTGAAATTAGATGAT ATTTTTATATTTTTTTTTT		234

	gF^{t}	ATATAT AAAATGTACAAACGGACAATGAGAGAACAGCGAAATTAGATGAT ATCTTTTTTT		208
	gF^{t}	ATAT AAAATGTACAGACGAGCAGTGAAGAGACAGTGAGATTA TACATTTTTTTTTTT	19,078	
	gF^{t}	ATATAA GTACAAACAGACAATGAAAAGATGGTGAGACTGAGTA TACATTTTTTTT	592	
	gF^{t}	AAATT AATGTACAAATAAACGATAGAGAGACAGTGAGATTA TGATTGTAATTCTTTTTTT	243	
	gF^{t}	ATAT AAAATGTACAGACAAGCAATGAAGAGACAGTGAGATTAGATAGTT TTTTTTTTC	210	
	gF^{exp}	ATACAT AATAAACAATGAGATATATAAGGTGAATCGAAAGTGAAATATT TTTTTTTTT		360
	gF^{exp}	ATATA CAAAAACAAAATACCAAATGACAGATT CAGTGATAATATGAGATAATTTT		235
	gF^{exp}	ATATAC AAAACAAAATACCAAATGACAGATT CAGTGATAATATGAGATAATTTTTTTTTTT		197
	gF^{exp}	ATACAT AATAAACAATGAGATATATAAGGTGAATCGAAAGTGAAATAT ATTTTTTTTT		179
E	gE	A TAGATAACAAACATAAGAGCAAGTCACGAGGTAGATATTGATATTTT TTTTTTTTC	10,598	15,900
	gE	A TAGATAACAAACATAAGAGCAAGTCACGAGGTAGATATTGATATTTTAA TTTTTTTATTTTT	867	1,342
	gE	A TAGATAACAAACATAAGAGCAAGTCACGAGGTAGATATTGATATTT GTTTTTTTTTTTT	100	199
	gE	TTAGATAACAAACATAAGAGCAAGTCACGAGGTAGATATTGATATTTTAA TTTTTT		102
	$gE^{t'}$	ATATAT AATCACAAACAAATAGAAAATGAGAGAGGTGTATGA TACTATTTTTTTTTCTTTTT		584
	$gE^{t'}$	ATAC ATAAATCACAAACGAATAAGGAACAGAAGAGATGTATGA TATATTTTTTTTTT	388	
	$gE^{t'}$	AT AATAAATCACAAACAGATAAAGAGCAAGAAAGGTGTATGA TTATATTTTTATTTTTT	86	
	$gE^{t'}$	ATAAAC AATCACAAACAGATAGAAGACAGAAGAGATGTATAGATA TTAAAATTTTTTTTT	80	
	gE^{t}	ATAT ATAACAAACATAGACAGAAATCTATAAGAGAGTATAATGATATGTGT TTTTTTTTTTTT	29	
	gE^{ep}	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTGAATGATGA TATTTTTTTTTTTT	2,493	27,823
	gE^{ep}	AAATA ATAACAAAACATGAGATATAACTTGTAGTGATTAGATGAATGAT TTTTTTTTTTT	12,105	5,923
	gE^{ep}	ATATAT TAAAATACAACTTATGATGACTAAGTGAATGATGATTGTCAA TTTTTTTTCTTTTT		5,559
	gE^{ep}	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTAAATGATGA TATTTTTTT		131
	gE^{ep}	ATATAT TAAAATACAACTTATGATGGCTAAGTGAATGATGATTGTCAA TTTTTTTTTTTTC	2,491	
	gE^{ep}	ATAT ATAAAACATAAAATATGACTTGTAGCAGTTAAGTGAATGATGATT TTTTTTTATTTTT	428	
	gE^{ep}	ATATT ATAAAACATAAGATATAACTCATAGTGATTGAATAAGTGAT AATTTTTTTTTGTTTT	157	
DE	gDE^{e80}	ATATATAAG CATACAAAAACTGCAATATTTACAGTGATGATGATTTAATTT TTTTTTTT	525	149
	gDE^{e80}	ATATATAAG CATACAAAAACTGCAGTATTTACAGTGATGATGATTTAATTT T		10
D	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATAAG ATATTTTTTTTTTTTT		27,320
	gD	ATATAT AAATCCAATAAGAAATGAAAGCTAGATAGTGAGTATAAGT TTTTTTTTTTTGTTT	535	331
	gD	GTAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATAAG ATATATTTTTTTTGTTTTTTTTTTTT		315
	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATAA TTTTTTTTTTTT		206
	gD	ATAT ACAAAAATCCAATGAAAAATAAAGACTGAGTGATGGATG CAATTTTCTTTTTTTTTT		142
	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTATA TTTTTTCTTTTTTTTTT		137
	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATGAGTAT TAGATATATTTTTTTTTTTT		90
	gD	ATAGAAATCCAATAAAAGACAGAAGCTAGATAATGAGTATAAG ATATATTTT		74
	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATCAGTATAAG ATATATTTTTTTAA		23
	gD	ATAGAAATCCAATAAGAGACAGAAGCTAGATAATTAGTATAAG ATATATATTTTTTTTTT		19
	gD	ATATAT AAACAAAAATTCAATAGAAAACGAGAACTGAGTAATTGATATAA TTTT		17
	gD	AT ATAGAAATCCAATAAGAGACAGAAACTAGGTAATGAGTATAAG ATTTTTTTTTTT	718	
	gD	ATAT AACAAAAATCCAATGAGAAATAGAGACTGAGTAATTGATATA TATTTTTATTTTTT	217	
	gD^{ep}	ATAT AAAACAATAATCATGATGAGATATAAAGTGCAGTTTGTGATAATT TTTTTGTTTT	12	68
	gD^{ep}	ATAT ATAATCATAACAAGATGTAGAGTACGATTTATAGTGATTAA TTTTTTTTTTTT	2	53
	gD^{ep}	ATAT AAAACAATAATCATGATGAGATATAAAGTGCAGTTTGTGATAATTAA TTTTT		31
	gD^{ep}	ATAT AAAACAATAATCATAATAAGATGTAAGGTACGATTTATGATAATT TTTTTTGTTTT	15	
	gD^{ep}	ATAT ATAATCATAACAGAGCATAGAATACAGTTTATAGTGATTAA TTTTTT	8	

	gD^{ep}	ATAT AAAACAATAATCATAATAAGATGTAAGGTACGATTTATGAT TTTTTTTTTATGCAACGTTGGATACTGGAG	1	
	gD^{ep}	ATAT ATAATCATAACAAGATATAGAATACGATTCATAGTGATTAA TTTTTTTTTTTTTT	1	
C	gC	ATAT ATACAACAATAAACTCGTATTAAGTGAGAGATGAAGATTTAAT TTTTTTTTATTTT		1,509
	gC	ATAT TAAACACAACGATAGATCTATATTAAGTAGAAGATAGAAATTTA TTTATTTTTTTTTTT		88
	gC	A AATATAAACACAATGATAAGCCTGTATTAGATAGAAAGTGAGAATTT TTTTTTTTTG	10,402	
	gC	TA AAACACAACAATAGATTCGTATTAAATAGAGAATAGAGATTTA TTTTTTTTTTT	209	
	gC	ATAT TAAACACAACAGTAGATCTATATTAAGTAGAAGATAGAAATTT TTTTTTTTT	170	
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATA ATTTTTTTTTTTTTT	214,918	187,465
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATAT TTTTTTTTTTTTN	180,589	170,393
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATATTTT TTTTTTTTTT	67,000	57,583
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGAT TTTTTTTTTTGTT	2,828	2,157
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATATTTTGT TTTTTT	553	932
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATATTT GTTTTTTTTT	621	871
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATA ATTTTTTTTTATTTTTTTT	1,014	829
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATAT AATTTTTTTT	330	250
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAAATAGTTATATGAGATATA ATTTTTCTTTTTTTT	358	242
	gC^{e80}	ATAT ATATAAACACAATAGATCGCACTGTAAGATAGTTATATGAGATAT TTTTTTTTTTTTG	33,434	
	gC^{e80}	ATAT ATATAAACACAATAGATCGCACTGTAAGATAGTTATATGAGATATA ATTTTTTTTTTTTTC	14,634	
	gC^{e80}	ATAT ATATAAACACAATAGATCGCACTGTAAGATAGTTATATGAGATATATTTT TTTTTTTTT	3,887	
	gC^{e80}	ATAT ATATAAACACAATAGATCGCACTGTAAGATAGTTATATGAGAT TTTTTTTTTTT	828	
	gC^{e80}	ATAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGGGATATA ATTTTTTTCTTTTTT	387	
	gC^{e80}	NTAT ATATAAACACAATAGATCACACTGTAAGATAGTTATATGAGATATACTTT TTTTT	374	
	gC^{e}	A AATATAAACACAATGATAAACCTATGTTAGATAGAAAGTGATAGTT TTTTTTTTTT		32
	gC^{e}	AAAA AATATAAACACAATGATAAGCCTGTATTAGATAGAAAGTGATAATT TTTT	3	
B	gB1B2	ATATATA AATGATAACACTGAGAATCTGTAGTATAAGATATGATGATAA TTTTTTTTTTTTTT		37,769
	gB1B2	ATATATA AATGATAACACTGAGAATCTGTAGTATAAGATATGATGATA TTTTTTTATTTT		330
	gB1B2	ATATATAAATA ATAACACTGAGAATCTGTAGTATAAGATATGATGATAA TTTGTTTTTTTTTTTTTT		180
	gB1B2	ATATATA AATGATAACACTGAGAATCTGTAGTATAAGATATGATAAT TTTTTTTTTT		98
	gB1B2	ATAT ATAATGATAACACTGAGAATCTGTAATGTGAGATATGATGATAAGTTT TTTTTTTTT	10,346	
	gB1B2	ATAT ATAATGATAACACTGAGAATCTGTAATGTGAGATATGATGATAA TTTTTTTTTTTT	2,337	
	gB1B2	ATATATA AATGATAACACTGAAGATCTGTAGTATAAGATATGATGATAA TTTTTTTCTTTTTT	139	
	gB1B2	ATAT ATAATGATAACACTGAGAATCTGTAATGTGAGATATGATGATA TTTTTTTTATTTT	124	
	gB3^{t}	ATAT ATAAAAATGATACCATAGAATTTGGAGTAGTATAGATATGATGATA TTTTTTTTTTT	1,183	
	gB3^{t}	ATAT ATAAAAATGATACCATAGAGTTTGAGGTGATGTAGATATGATGGTA TTTTT	474	
	gB3^{t}	ATAT ATAAAAATGATACCATAGAATTTGGAGTAGTATAGATATGATGAT TTTTTTTTTT	383	
	gB4	ATATA TTATATACAGTAATAGAGATGATAGTGATATATTGAGTGATATA ATTTTTTTTTT	260	154
	gB4	ATATACAGTAATAGAGATGATAGTGATATATTGAGTGATATAATTAATATGT GATTTTTAATTTTTTTTTTCTTT		8
	gB4	ATACAGTAATAGAGATGATAGTGATATATTGAGTGATATAATTAATATGTTTT T		6
	gB4	ATA TATATACAATAATAAGAGTGATAGCAGTG TATTAGTGATGATTAATTTTTTTTTTTTTTT		5
	gB4	ATA TATATACAATAATGAGAATGATGACAGTG TATTAGTGATGTAATATTTTTTTTTATTTT	367	
	gB4	ATATA TTATATACAGTAATAGAGATGATAGTGATATATTGAGTGATATAATTAATATGT GATTTTTAATATAATA	18	
	gB4$^{t'}$	ATAT ATAACATATCAAATGATAAAGTGAGAAAAGAGAGTGAATGTAAT TTTTTTCTTTTCTATCTATTATTACAT	13	
	gB5^{e}	ATAT ATGATAACAACAGTAGAGATATATAAGTGATGTAAATGTGATAAT TTTTTTTTTTT	3,708	859
	gB6^{e}	ATAT ATAACAACAATAAGAATACGTGAGATAGTATAGATATGATGAT TTTTTTTTTTTG		2,300
	gB6^{e}	ATATATAAT ACAATAAGAATACGTGAGATAGTATAGATATGATGAT TTTTTTTTTTA	222	

	gB7[e]	GATATATCTGAGTAAGTGATGACATTGTGAATTATTTTTTTT T		489
	gB7[e]	AAATATAAG CAAATGATATATCTGAGTAAGTGATGACATTGTGAATT TTTTTTTTTTT	3,042	250
	gB7[e]	ATGATATATCTGAGTAAGTGATGACATTGTGAATTAT GGTATATTTTTG		117
	gB7[e]	TGATATATCTGAGTAAGTGATGACATTGTGAATTACTTTTTT T		49
	gB7[e]	AATGATATATCTGAGTAAGTGATGACATTGTGAATTATTTTTTTT TTTTTT		14
	gB7[e]	AATATAAG CAAATGATATATCTGAGTAAGTGATGACATTGTGAATTATTTTTTTT TTTTTTTAAAAAAAAA	884	
	gB7[e]	AAATATAAG CAAATGATATATCTGAGTAAGTGATGACATTGTGAATTAT GGTATATAAAGTTAAATAATTTTATC	522	
A	gA1	ATA TTAATCTAATAACGGAGACTTGTATATGATAGTAATGATGATATT TTTTTTTTT		2,697
	gA1	ATATA TTAATCTAATAGCGGAGACTTGTATATGATAGTAATGATGATATT TT	104	
	gA2	ATATAA AATCTAATGACGAGAACTTGTATATGATGATGAAGGTGATA ATTTTTTTTTTTTTG		16,984
	gA2	ATA AATCTAATAACGAGAATTTATGTACGATAATGAAAGTGATAT ATTTTTTTTCTTTTT		540
	gA2	ATATAA AATCTAATGACGAAAGCTTGTATATGGTAATGAAGATGGTATT TTTTTTTTTT		540
	gA2	ATATAA AATCTAATGACGAAAGCTTGTATATGGTAATGAAGATGGTA ATTTTTTTTTCTTTTT		419
	gA2	ATATAA AATCTAATGACGAGAACTTGTATATGATGATGAAGGTGATATT TCTTTTTTTTT		331
	gA2	ATATA AATCTAATAACGAGAATTTATGTACGATAATGAAAGTGATATT TTTTTCTTTTTTTTT		106
	gA2	ATACAT ATCTAATAACGGAAGCTTATGTGTAGTAGTGAAGATGGTA TATTTTTTTTTTTTT		103
	gA2	ATATAA AATCTAATGACGAAAGCTTGTATATGGTAGTGAAGATGGTA ATTTTTTTTTTTTTT	1,234	
	gA2	ATATAA AATCTAATGACGAAAGCTTGTATATGGTAGTGAAGATGGTATT TTTTTTTTT	382	
	gA2	ATATA AATCTAATAACGGAAATTTGTATATGATGATAGAAGTGATAGT TTTTTTTTTTT	40	

REFERENCES

REFERENCES

1. Shapiro TA, Englund PT. The structure and replication of kinetoplast DNA. Annu Rev Microbiol. 1995;49: 117+.

2. Vickerman K. The evolutionary expansion of the trypanosomatid flagellates. Int J Parasitol. 1994;24: 1317–1331. doi:10.1016/0020-7519(94)90198-8

3. Simpson AGB, Stevens JR, Lukeš J. The evolution and diversity of kinetoplastid flagellates. Trends Parasitol. 2006;22: 168–174. doi:10.1016/j.pt.2006.02.006

4. Read LK, Lukeš J, Hashimi H. Trypanosome RNA editing: the complexity of getting U in and taking U out. Wiley Interdiscip Rev RNA. 2016;7: 33–51. doi:10.1002/wrna.1313

5. Priest JW, Hajduk SL. Developmental regulation of mitochondrial biogenesis inTrypanosoma brucei. J Bioenerg Biomembr. 1994;26: 179–191. doi:10.1007/BF00763067

6. Aphasizhev R, Aphasizheva I. Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer. Wiley Interdiscip Rev RNA. 2011;2: 669–685. doi:10.1002/wrna.82

7. Aphasizhev R, Aphasizheva I. Mitochondrial RNA editing in trypanosomes: Small RNAs in control. Biochimie. 2014;100: 125–131. doi:10.1016/j.biochi.2014.01.003

8. Hong M i. n., Simpson L. Genomic Organization of Trypanosoma brucei Kinetoplast DNA Minicircles. Protist. 2003;154: 265–279. doi:10.1078/143446103322166554

9. Koslowsky D, Sun Y, Hindenach J, Theisen T, Lucas J. The insect-phase gRNA transcriptome in Trypanosoma brucei. Nucleic Acids Res. 2014;42: 1873–1886. doi:10.1093/nar/gkt973

10. CDC - African Trypanosomiasis [Internet]. 2 May 2017 [cited 25 Jan 2019]. Available: https://www.cdc.gov/parasites/sleepingsickness/index.html

11. Programme Against African Trypanosomosis (PAAT) | Food and Agriculture Organization of the United Nations [Internet]. [cited 25 Jan 2019]. Available: http://www.fao.org/paat/en/

12. Horn D. Antigenic variation in African trypanosomes. Mol Biochem Parasitol. 2014;195: 123–129. doi:10.1016/j.molbiopara.2014.05.001

13. Sudarshi D, Lawrence S, Pickrell WO, Eligar V, Walters R, Quaderi S, et al. Human African Trypanosomiasis Presenting at Least 29 Years after Infection—What Can This Teach Us

about the Pathogenesis and Control of This Neglected Tropical Disease? PLOS Negl Trop Dis. 2014;8: e3349. doi:10.1371/journal.pntd.0003349

14. Matthews KR. Developments in the Differentiation of Trypanosoma brucei. Parasitol Today. 1999;15: 76–80. doi:10.1016/S0169-4758(98)01381-7

15. van Hellemond JJ, Bakker BM, Tielens AGM. Energy Metabolism and Its Compartmentation in Trypanosoma brucei. In: Poole RK, editor. Advances in Microbial Physiology. Academic Press; 2005. pp. 199–226. doi:10.1016/S0065-2911(05)50005-5

16. Hannaert V, Bringaud F, Opperdoes FR, Michels PA. Evolution of energy metabolism and its compartmentation in Kinetoplastida. Kinetoplastid Biol Dis. 2003;2: 11. doi:10.1186/1475-9292-2-11

17. Nolan DP, Voorheis HP. The mitochondrion in bloodstream forms of Trypanosoma brucei is energized by the electrogenic pumping of protons catalysed by the F1F0-ATPase. Eur J Biochem. 1992;209: 207–216. doi:10.1111/j.1432-1033.1992.tb17278.x

18. Vertommen D, Van Roy J, Szikora J-P, Rider MH, Michels PAM, Opperdoes FR. Differential expression of glycosomal and mitochondrial proteins in the two major life-cycle stages of Trypanosoma brucei. Mol Biochem Parasitol. 2008;158: 189–201. doi:10.1016/j.molbiopara.2007.12.008

19. Weelden SWH van, Fast B, Vogt A, Meer P van der, Saas J, Hellemond JJ van, et al. Procyclic Trypanosoma brucei Do Not Use Krebs Cycle Activity for Energy Generation. J Biol Chem. 2003;278: 12854–12863. doi:10.1074/jbc.M213190200

20. Weelden SWH van, Hellemond JJ van, Opperdoes FR, Tielens AGM. New Functions for Parts of the Krebs Cycle in Procyclic Trypanosoma brucei, a Cycle Not Operating as a Cycle. J Biol Chem. 2005;280: 12451–12460. doi:10.1074/jbc.M412447200

21. Bringaud F, Rivière L, Coustou V. Energy metabolism of trypanosomatids: Adaptation to available carbon sources. Mol Biochem Parasitol. 2006;149: 1–9. doi:10.1016/j.molbiopara.2006.03.017

22. Oberle M, Balmer O, Brun R, Roditi I. Bottlenecks and the Maintenance of Minor Genotypes during the Life Cycle of Trypanosoma brucei. PLOS Pathog. 2010;6: e1001023. doi:10.1371/journal.ppat.1001023

23. Abbeele JVD, Claes Y, Bockstaele DV, Ray DL, Coosemans M. Trypanosoma brucei spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis. Parasitology. 1999;118: 469–478.

24. Michelotti EF, Hajduk SL. Developmental regulation of trypanosome mitochondrial gene expression. J Biol Chem. 1987;262: 927–932.

25. Feagin JE, Jasmer DP, Stuart K. Developmentally regulated addition of nucleotides within apocytochrome b transcripts in Trypanosoma brucei. Cell. 1987;49: 337–345. doi:10.1016/0092-8674(87)90286-8

26. Read LK, Wilson KD, Myler PJ, Stuart K. Editing of Trypanosoma brucei maxicircle CR5 mRNA generates variable carboxy terminal predicted protein sequences. Nucleic Acids Res. 1994;22: 1489–1495. doi:10.1093/nar/22.8.1489

27. Koslowsky DJ, Bhat GJ, Perrollaz AL, Feagin JE, Stuart K. The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. Cell. 1990;62: 901–911. doi:10.1016/0092-8674(90)90265-G

28. Souza AE, Myler PJ, Stuart K. 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. 1992;12: 2100–2107. doi:10.1128/MCB.12.5.2100

29. Souza AE, Shu HH, Read LK, Myler PJ, Stuart KD. Extensive editing of CR2 maxicircle transcripts of Trypanosoma brucei predicts a protein with homology to a subunit of NADH dehydrogenase. Mol Cell Biol. 1993;13: 6832–6840. doi:10.1128/MCB.13.11.6832

30. Read LK, Myler PJ, Stuart K. Extensive editing of both processed and preprocessed maxicircle CR6 transcripts in Trypanosoma brucei. J Biol Chem. 1992;267: 1123–1128.

31. Bhat GJ, Koslowsky D, Feagin J, Smiley B, Stuart K. An Extensively Edited Mitochondrial Transcript in Kinetoplastids Encodes a Protein Homologous to ATPase Subunit 6. Cell. 1990;61: 885–894.

32. Feagin JE, Abraham JM, Stuart K. Extensive editing of the cytochrome c oxidase III transcript in Trypanosoma brucei. Cell. 1988;53: 413–422. doi:10.1016/0092-8674(88)90161-4

33. Blum B, Bakalara N, Simpson L. A model for RNA editing in kinetoplastid mitochondria: “Guide” RNA molecules transcribed from maxicircle DNA provide the edited information. Cell. 1990;60: 189–198. doi:10.1016/0092-8674(90)90735-W

34. Eperon IC, Janssen JWG, Hoeijmakers JHJ, Borst P. The major transcripts of the kinetoplast DNA of Trypanosoma brucei are very small ribosomal RNAs. Nucleic Acids Res. 1983;11: 105–125. doi:10.1093/nar/11.1.105

35. Adler BK, Harris ME, Bertrand KI, Hajduk SL. Modification of Trypanosoma brucei mitochondrial rRNA by posttranscriptional 3’ polyuridine tail formation. Mol Cell Biol. 1991;11: 5878–5884. doi:10.1128/MCB.11.12.5878

36. Aphasizheva I, Maslov D, Wang X, Huang L, Aphasizhev R. Pentatricopeptide Repeat Proteins Stimulate mRNA Adenylation/Uridylation to Activate Mitochondrial Translation in Trypanosomes. Mol Cell. 2011;42: 106–117. doi:10.1016/j.molcel.2011.02.021

37. Bhat GJ, Souza AE, Feagin JE, Stuart K. Transcript-specific developmental regulation of polyadenylation in Trypanosoma brucei mitochondria. Mol Biochem Parasitol. 1992;52: 231–240. doi:10.1016/0166-6851(92)90055-O

38. Hensgens LA, Brakenhoff J, De Vries BF, Sloof P, Tromp MC, Van Boom JH, et al. 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 mitochrondrial maxi-circle DNA. Nucleic Acids Res. 1984;12: 7327–7344.

39. Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell. 1986;46: 819–826. doi:10.1016/0092-8674(86)90063-2

40. Feagin JE, Stuart K. Differential expression of mitochondrial genes between life cycle stages of Trypanosoma brucei. Proc Natl Acad Sci. 1985;82: 3380–3384.

41. Payne M, Rothwell V, Jasmer DP, Feagin JE, Stuart K. Identification of mitochondrial genes in Trypanosoma brucei and homology to cytochrome c oxidase II in two different reading frames. Mol Biochem Parasitol. 1985;15: 159–170. doi:10.1016/0166-6851(85)90117-3

42. Kannan S, Burger G. Unassigned MURF1 of kinetoplastids codes for NADH dehydrogenase subunit 2. BMC Genomics. 2008;9: 455. doi:10.1186/1471-2164-9-455

43. Feagin JE, Jasmer DP, Stuart K. Apocytochrome b and other mitochondrial DNA sequences are differentially expressed during the life cycle of Trypanosoma brucei. Nucleic Acids Res. 1985;13: 4577–4596. doi:10.1093/nar/13.12.4577

44. Stuart K, Feagin JE, Jasmer DP. Regulation of Mitochondrial Gene Expression in Trypanosoma brucei. Sequence Specificity in Transcription and Translation. Alan R. Liss; 1985. pp. 621–631.

45. Jasmer DP, Feagin JE, Stuart K. Diverse patterns of expression of the cytochrome c oxidase subunit I gene and unassigned reading frames 4 and 5 during the life cycle of Trypanosoma brucei. Mol Cell Biol. 1985;5: 3041–3047. doi:10.1128/MCB.5.11.3041

46. Feagin JE, Stuart K. Developmental aspects of uridine addition within mitochondrial transcripts of Trypanosoma brucei. Mol Cell Biol. 1988;8: 1259–1265. doi:10.1128/MCB.8.3.1259

47. Stuart K. The RNA editing process in Trypanosoma brucei. Semin Cell Biol. 1993;4: 251–260. doi:10.1006/scel.1993.1030

48. Corell RA, Myler P, Stuart K. Trypanosoma brucei mitochondrial CR4 gene encodes an extensively edited mRNA with completely edited sequence only in bloodstream forms. Mol Biochem Parasitol. 1994;64: 65–74. doi:10.1016/0166-6851(94)90135-X

49. Hajduk SL, Adler BK, Madison S, McManus M, Sabatini R. Insertional and deletional RNA editing in trypanosome mitochondria. Nucleic Acids Symp Ser. 1996; 15–18.

50. Koslowsky DJ, Riley GR, Feagin JE, Stuart K. Guide RNAs for transcripts with developmentally regulated RNA editing are present in both life cycle stages of Trypanosoma brucei. Mol Cell Biol. 1992;12: 2043–2049. doi:10.1128/MCB.12.5.2043

51. Riley GR, Corell RA, Stuart K. Multiple guide RNAs for identical editing of Trypanosoma brucei apocytochrome b mRNA have an unusual minicircle location and are developmentally regulated. J Biol Chem. 1994;269: 6101–6108.

52. Greif G, Rodriguez M, Reyna-Bello A, Robello C, Alvarez-Valin F. Kinetoplast adaptations in American strains from Trypanosoma vivax. Mutat Res Mol Mech Mutagen. 2015;773: 69–82. doi:10.1016/j.mrfmmm.2015.01.008

53. Ooi C-P, Schuster S, Cren-Travaillé C, Bertiaux E, Cosson A, Goyard S, et al. The Cyclical Development of Trypanosoma vivax in the Tsetse Fly Involves an Asymmetric Division. Front Cell Infect Microbiol. 2016;6. doi:10.3389/fcimb.2016.00115

54. Ruvalcaba-Trejo LI, Sturm NR. The Trypanosoma cruzi Sylvio X10 strain maxicircle sequence: the third musketeer. BMC Genomics. 2011;12: 58. doi:10.1186/1471-2164-12-58

55. Cazzulo JJ. Protein and amino acid catabolism in Trypanosoma cruzi. Comp Biochem Physiol Part B Comp Biochem. 1984;79: 309–320. doi:10.1016/0305-0491(84)90381-X

56. Cazzulo JJ. Intermediate metabolism inTrypanosoma cruzi. J Bioenerg Biomembr. 1994;26: 157–165. doi:10.1007/BF00763064

57. Tyler KM, Engman DM. The life cycle of Trypanosoma cruzi revisited. Int J Parasitol. 2001;31: 472–481. doi:10.1016/S0020-7519(01)00153-9

58. Cannata JJB, Cazzulo JJ. The aerobic fermentation of glucose by Trypanosoma cruzi. Comp Biochem Physiol Part B Comp Biochem. 1984;79: 297–308. doi:10.1016/0305-0491(84)90380-8

59. Sanchez-moreno M, Fernandez-becerra MC, Castilla-calvente JJ, Osuna A. Metabolic studies by 1H NMR of different forms of Trypanosoma cruzi as obtained by “in vitro”

culture. FEMS Microbiol Lett. 1995;133: 119–125. doi:10.1111/j.1574-6968.1995.tb07871.x

60. Prevention C-C for DC and. CDC - Leishmaniasis [Internet]. 16 Oct 2018 [cited 25 Jan 2019]. Available: https://www.cdc.gov/parasites/leishmaniasis/index.html

61. McConville MJ, Naderer T. Metabolic Pathways Required for the Intracellular Survival of Leishmania. Annu Rev Microbiol. 2011;65: 543–561. doi:10.1146/annurev-micro-090110-102913

62. Saunders EC, Souza DPD, Naderer T, Sernee MF, Ralton JE, Doyle MA, et al. Central carbon metabolism of Leishmania parasites. Parasitology. 2010;137: 1303–1313. doi:10.1017/S0031182010000077

63. Simpson L, Thiemann OH, Savill NJ, Alfonzo JD, Maslov DA. Evolution of RNA editing in trypanosome mitochondria. Proc Natl Acad Sci. 2000;97: 6986–6993. doi:10.1073/pnas.97.13.6986

64. Porcel BM, Denoeud F, Opperdoes F, Noel B, Madoui M-A, Hammarton TC, et al. The Streamlined Genome of Phytomonas spp. Relative to Human Pathogenic Kinetoplastids Reveals a Parasite Tailored for Plants. PLOS Genet. 2014;10: e1004007. doi:10.1371/journal.pgen.1004007

65. Nawathean P, Maslov DA. 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. 2000;38: 95–103. doi:10.1007/s002940000135

66. Blum B, Simpson L. Guide RNAs in kinetoplastid mitochondria have a nonencoded 3’ oligo(U) tail involved in recognition of the preedited region. Cell. 1990;62: 391–397. doi:10.1016/0092-8674(90)90375-O

67. Ochsenreiter T, Hajduk SL. Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. EMBO Rep. 2006;7: 1128–1133. doi:10.1038/sj.embor.7400817

68. Ochsenreiter T, Anderson S, Wood ZA, Hajduk SL. Alternative RNA Editing Produces a Novel Protein Involved in Mitochondrial DNA Maintenance in Trypanosomes. Mol Cell Biol. 2008;28: 5595–5604. doi:10.1128/MCB.00637-08

69. Covello P, Gray M. On the evolution of RNA editing. Trends Genet. 1993;9: 265–268. doi:10.1016/0168-9525(93)90011-6

70. Gray MW. Evolutionary Origin of RNA Editing. Biochemistry (Mosc). 2012;51: 5235–5242. doi:10.1021/bi300419r

71. Gray MW, Lukeš J, Archibald JM, Keeling PJ, Doolittle WF. Irremediable Complexity? Science. 2010;330: 920–921. doi:10.1126/science.1198594

72. Stoltzfus A. On the Possibility of Constructive Neutral Evolution. J Mol Evol. 1999;49: 169–181. doi:10.1007/PL00006540

73. Stoltzfus A. Constructive neutral evolution: exploring evolutionary theory's curious disconnect. Biol Direct. 2012;7: 35. doi:10.1186/1745-6150-7-35

74. Leeder W-M, Hummel NFC, Göringer HU. Multiple G-quartet structures in pre-edited mRNAs suggest evolutionary driving force for RNA editing in trypanosomes. Sci Rep. 2016;6. doi:10.1038/srep29810

75. Speijer D. Evolutionary Aspects of RNA Editing. In: Göringer HU, editor. RNA Editing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. pp. 199–227. doi:10.1007/978-3-540-73787-2_10

76. Buhrman H, van der Gulik P, Severini S, Speijer D. A mathematical model of kinetoplastid mitochondrial gene scrambling advantage. ArXiv13071163 Q-Bio. 2013; Available: http://arxiv.org/abs/1307.1163

77. Thiemann OH, Maslov DA, Simpson L. Disruption of RNA editing in Leishmania tarentolae by the loss of minicircle-encoded guide RNA genes. EMBO J. 1994;13: 5689–5700.

78. Savill Nicholas J., Higgs Paul G. A theoretical study of random segregation of minicircles in trypanosomatids. Proc R Soc Lond B Biol Sci. 1999;266: 611–620. doi:10.1098/rspb.1999.0680

79. Lynch M, Bürger R, Butcher D, Gabriel W. The Mutational Meltdown in Asexual Populations. J Hered. 1993;84: 339–344. doi:10.1093/oxfordjournals.jhered.a111354

80. LaBar T, Adami C. Evolution of drift robustness in small populations. Nat Commun. 2017;8: 1012. doi:10.1038/s41467-017-01003-7

81. Muller HJ. The relation of recombination to mutational advance. Mutat Res Mol Mech Mutagen. 1964;1: 2–9. doi:10.1016/0027-5107(64)90047-8

82. Haigh J. The accumulation of deleterious genes in a population—Muller's Ratchet. Theor Popul Biol. 1978;14: 251–267. doi:10.1016/0040-5809(78)90027-8

83. Poon A, Otto SP. Compensating for Our Load of Mutations: Freezing the Meltdown of Small Populations. Evolution. 2000;54: 1467–1479. doi:10.1111/j.0014-3820.2000.tb00693.x

84. Whitlock MC. Fixation of New Alleles and the Extinction of Small Populations: Drift Load, Beneficial Alleles, and Sexual Selection. Evolution. 2000;54: 1855–1861. doi:10.1111/j.0014-3820.2000.tb01232.x

85. Normark S, Bergström S, Edlund T, Grundström T, Jaurin B, Lindberg FP, et al. Overlapping genes. Annu Rev Genet. 1983;17: 499–525.

86. Liang H. Decoding the dual-coding region: key factors influencing the translational potential of a two-ORF-containing transcript. Cell Res. 2010;20: 508–509. doi:10.1038/cr.2010.62

87. Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. 2007;17: 000–000. doi:10.1101/gr.6305707

88. Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. 2016;11: 26. doi:10.1186/s13062-016-0128-3

89. Mouilleron H, Delcourt V, Roucou X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. 2016;44: 14–23. doi:10.1093/nar/gkv1218

90. Liang H, Landweber LF. A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res. 2006;16: 190–196. doi:10.1101/gr.4246506

91. Ribrioux S, Brüngger A, Baumgarten B, Seuwen K, John MR. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genomics. 2008;9: 122. doi:10.1186/1471-2164-9-122

92. Pallejà A, Harrington ED, Bork P. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics. 2008;9: 335. doi:10.1186/1471-2164-9-335

93. Chung W-Y, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. A First Look at ARFome: Dual-Coding Genes in Mammalian Genomes. PLOS Comput Biol. 2007;3: e91. doi:10.1371/journal.pcbi.0030091

94. Klemke M. Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage. EMBO J. 2001;20: 3849–3860. doi:10.1093/emboj/20.14.3849

95. Yoshida H, Oku M, Suzuki M, Mori K. pXBP1(U) encoded in XBP1 pre-mRNA negatively regulates unfolded protein response activator pXBP1(S) in mammalian ER stress response. J Cell Biol. 2006;172: 565–575. doi:10.1083/jcb.200508145

96. Peleg O, Kirzhner V, Trifonov E, Bolshoy A. Overlapping Messages and Survivability. J Mol Evol. 2004;59: 520–527. doi:10.1007/s00239-004-2644-5

97. Sykes SE, Hajduk SL. Dual Functions of α-Ketoglutarate Dehydrogenase E2 in the Krebs Cycle and Mitochondrial DNA Inheritance in Trypanosoma brucei. Eukaryot Cell. 2013;12: 78–90. doi:10.1128/EC.00269-12

98. Sykes S, Szempruch A, Hajduk S. The Krebs Cycle Enzyme α-Ketoglutarate Decarboxylase Is an Essential Glycosomal Protein in Bloodstream African Trypanosomes. Eukaryot Cell. 2015;14: 206–215. doi:10.1128/EC.00214-14

99. Ochsenreiter T, Cipriano M, Hajduk SL. Alternative mRNA Editing in Trypanosomes Is Extensive and May Contribute to Mitochondrial Protein Diversity. PLoS ONE. 2008;3: e1566. doi:10.1371/journal.pone.0001566

100. Aphasizheva I, Maslov DA, Aphasizhev R. Kinetoplast DNA-encoded ribosomal protein S12. RNA Biol. 2013;10: 1679–1688. doi:10.4161/rna.26733

101. Stuart K, Gobright E, Jenni L, Milhausen M, Thomashow L, Agabian N. The Istar 1 Serodeme of Trypanosoma brucei: Development of a New Serodeme. J Parasitol. 1984;70: 747–754. doi:10.2307/3281757

102. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987;162: 156–159. doi:10.1016/0003-2697(87)90021-2

103. Agabian N, Thomashow L, Milhausen M, Stuart K. Structural Analysis of Variant and Invariant Genes in Trypanosomes. Am J Trop Med Hyg. 1980;29: 1043–1049.

104. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19: 92–105. doi:10.1101/gr.082701.108

105. Madina BR, Kumar V, Metz R, Mooers BHM, Bundschuh R, Cruz-Reyes J. Native mitochondrial RNA-binding complexes in kinetoplastid RNA editing differ in guide RNA composition. RNA. 2014; doi:10.1261/rna.044495.114

106. Clement SL, Mingler MK, Koslowsky DJ. An Intragenic Guide RNA Location Suggests a Complex Mechanism for Mitochondrial Gene Expression in Trypanosoma brucei. Eukaryot Cell. 2004;3: 862–869. doi:10.1128/EC.3.4.862-869.2004

107. Cristodero M, Seebeck T, Schneider A. Mitochondrial translation is essential in bloodstream forms of Trypanosoma brucei. Mol Microbiol. 2010;78: 757–769. doi:10.1111/j.1365-2958.2010.07368.x

108. MacLeod A, Turner CMR, Tait A. A high level of mixed Trypanosoma brucei infections in tsetse flies detected by three hypervariable minisatellites. Mol Biochem Parasitol. 1999;102: 237–248. doi:10.1016/S0166-6851(99)00101-2

109. Balmer O, Caccone A. Multiple-strain infections of Trypanosoma brucei across Africa. Acta Trop. 2008;107: 275–279. doi:10.1016/j.actatropica.2008.06.006

110. Szempruch AJ, Choudhury R, Wang Z, Hajduk SL. In vivo analysis of trypanosome mitochondrial RNA function by artificial site-specific RNA endonuclease-mediated knockdown. RNA. 2015; doi:10.1261/rna.052084.115

111. Surve S, Heestand M, Panicucci B, Schnaufer A, Parsons M. Enigmatic Presence of Mitochondrial Complex I in Trypanosoma brucei Bloodstream Forms. Eukaryot Cell. 2012;11: 183–193. doi:10.1128/EC.05282-11

112. Verner Z, Čermáková P, Škodová I, Kriegová E, Horváth A, Lukeš J. Complex I (NADH:ubiquinone oxidoreductase) is active in but non-essential for procyclic Trypanosoma brucei. Mol Biochem Parasitol. 2011;175: 196–200. doi:10.1016/j.molbiopara.2010.11.003

113. Speijer D. Is kinetoplastid pan-editing the result of an evolutionary balancing act? IUBMB Life. 2006;58: 91–96. doi:10.1080/15216540600551355

114. Hudson K m., Taylor AE r., Elce B j. Antigenic changes in Trypanosoma brucei on transmission by tsetse fly. Parasite Immunol. 1980;2: 57–69. doi:10.1111/j.1365-3024.1980.tb00043.x

115. Gibson W. The origins of the trypanosome genome strains Trypanosoma brucei brucei TREU 927, T. b. gambiense DAL 972, T. vivax Y486 and T. congolense IL3000. Parasit Vectors. 2012;5: 71.

116. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2014;7: 539–539. doi:10.1038/msb.2011.75

117. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science. 1992;256: 1443–1445.

118. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825−2830.

119. Kirby LE, Sun Y, Judah D, Nowak S, Koslowsky D. Analysis of the Trypanosoma brucei EATRO 164 Bloodstream Guide RNA Transcriptome. PLOS Negl Trop Dis. 2016;10: e0004793. doi:10.1371/journal.pntd.0004793

120. Duarte M, Tomás AM. The mitochondrial complex I of trypanosomatids - an overview of current knowledge. J Bioenerg Biomembr. 2014;46: 299–311. doi:10.1007/s10863-014-9556-x

121. Simpson L, Neckelmann N, Cruz VF de la, Simpson AM, Feagin JE, Jasmer DP, et al. Comparison of the maxicircle (mitochondrial) genomes of Leishmania tarentolae and Trypanosoma brucei at the level of nucleotide sequence. J Biol Chem. 1987;262: 6182–6196.

122. Hanada K, Shiu S-H, Li W-H. The Nonsynonymous/Synonymous Substitution Rate Ratio versus the Radical/Conservative Replacement Rate Ratio in the Evolution of Mammalian Genes. Mol Biol Evol. 2007;24: 2235–2241. doi:10.1093/molbev/msm152

123. Firth AE, Brown CM. Detecting overlapping coding sequences with pairwise alignments. Bioinformatics. 2005;21: 282–292. doi:10.1093/bioinformatics/bti007

124. Firth AE, Brown CM. Detecting overlapping coding sequences in virus genomes. BMC Bioinformatics. 2006;7: 75. doi:10.1186/1471-2105-7-75

125. Landweber LF, Gilbert W. RNA editing as a source of genetic variation. Nat Lond. 1993;363: 179.

126. Tielens AGM, van Hellemond JJ. Surprising variety in energy metabolism within Trypanosomatidae. Trends Parasitol. 2009;25: 482–490. doi:10.1016/j.pt.2009.07.007

127. Verner Z, Čermáková P, Škodová I, Kováčová B, Lukeš J, Horváth A. Comparative analysis of respiratory chain and oxidative phosphorylation in Leishmania tarentolae, Crithidia fasciculata, Phytomonas serpens and procyclic stage of Trypanosoma brucei. Mol Biochem Parasitol. 2014;193: 55–65. doi:10.1016/j.molbiopara.2014.02.003

128. Jackson AP, Berry A, Aslett M, Allison HC, Burton P, Vavrova-Anderson J, et al. Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. Proc Natl Acad Sci U S A. 2012;109: 3416–3421. doi:10.1073/pnas.1117313109

129. Morrison LJ, Vezza L, Rowan T, Hope JC. Animal African Trypanosomiasis: Time to Increase Focus on Clinically Relevant Parasite and Host Species. Trends Parasitol. 2016;32: 599–607. doi:10.1016/j.pt.2016.04.012

130. Maslov DA, Hollar L, Haghighat P, Nawathean P. Demonstration of mRNA editing and localization of guide RNA genes in kinetoplast–mitochondria of the plant trypanosomatid Phytomonas serpens1Note: Nucleotide sequences from P. serpens 1G reported in this work were deposited in GenBank™ database with the following accession numbers: AF034624 (Sau3AI-cut minicircle), AF034625 (HindIII-cut minicircle), AF034626 (fully edited sequence of RPS12 mRNA), AF034627 (genomic sequence of RPS12 cryptogene).1. Mol Biochem Parasitol. 1998;93: 225–236. doi:10.1016/S0166-6851(98)00028-0

131. David V, Flegontov P, Gerasimov E, Tanifuji G, Hashimi H, Logacheva MD, et al. Gene Loss and Error-Prone RNA Editing in the Mitochondrion of Perkinsela, an Endosymbiotic Kinetoplastid. mBio. 2015;6: e01498-15. doi:10.1128/mBio.01498-15

132. Maslov DA. Complete set of mitochondrial pan-edited mRNAs in Leishmania mexicana amazonensis LV78. Mol Biochem Parasitol. 2010;173: 107–114. doi:10.1016/j.molbiopara.2010.05.013

133. Käll L, Krogh A, Sonnhammer ELL. A Combined Transmembrane Topology and Signal Peptide Prediction Method. J Mol Biol. 2004;338: 1027–1036. doi:10.1016/j.jmb.2004.03.016

134. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic Acids Res. 2007;35: W429–W432. doi:10.1093/nar/gkm256

135. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10: 845–858. doi:10.1038/nprot.2015.053

136. Xu Y, Tao Y, Cheung LS, Fan C, Chen L-Q, Xu S, et al. Structures of bacterial homologues of SWEET transporters in two distinct conformations. Nature. 2014;515: 448–452. doi:10.1038/nature13670

137. Bender T, Pena G, Martinou J-C. Regulation of mitochondrial pyruvate uptake by alternative pyruvate carrier complexes. EMBO J. 2015;34: 911–924. doi:10.15252/embj.201490197

138. Štáfková J, Mach J, Biran M, Verner Z, Bringaud F, Tachezy J. Mitochondrial pyruvate carrier in Trypanosoma brucei. Mol Microbiol. 2016;100: 442–456. doi:10.1111/mmi.13325

139. Saunders EC, Ng WW, Kloehn J, Chambers JM, Ng M, McConville MJ. Induction of a Stringent Metabolic Response in Intracellular Stages of Leishmania mexicana Leads to Increased Dependence on Mitochondrial Metabolism. PLOS Pathog. 2014;10: e1003888. doi:10.1371/journal.ppat.1003888

140. Aphasizheva I, Aphasizhev R. U-Insertion/Deletion mRNA-Editing Holoenzyme: Definition in Sight. Trends Parasitol. 2016;32: 144–156. doi:10.1016/j.pt.2015.10.004

141. Kirby LE, Koslowsky D. Mitochondrial dual-coding genes in Trypanosoma brucei. PLoS Negl Trop Dis. 2017;11: e0005989. doi:10.1371/journal.pntd.0005989

142. Lamour N, Rivière L, Coustou V, Coombs GH, Barrett MP, Bringaud F. Proline Metabolism in Procyclic Trypanosoma brucei Is Down-regulated in the Presence of Glucose. J Biol Chem. 2005;280: 11902–11910. doi:10.1074/jbc.M414274200

143. Coustou V, Biran M, Breton M, Guegan F, Rivière L, Plazolles N, et al. Glucose-induced Remodeling of Intermediary and Energy Metabolism in Procyclic Trypanosoma brucei. J Biol Chem. 2008;283: 16342–16354. doi:10.1074/jbc.M709592200

144. Bochud-Allemann N, Schneider A. Mitochondrial Substrate Level Phosphorylation Is Essential for Growth of Procyclic Trypanosoma brucei. J Biol Chem. 2002;277: 32849–32854. doi:10.1074/jbc.M205776200

145. Horváth A, Horáková E, Dunajčíková P, Verner Z, Pravdová E, Šlapetová I, et al. Downregulation of the nuclear-encoded subunits of the complexes III and IV disrupts their respective complexes but not complex I in procyclic Trypanosoma brucei. Mol Microbiol. 2005;58: 116–130. doi:10.1111/j.1365-2958.2005.04813.x

146. Gnipová A, Panicucci B, Paris Z, Verner Z, Horváth A, Lukeš J, et al. Disparate phenotypic effects from the knockdown of various Trypanosoma brucei cytochrome c oxidase subunits. Mol Biochem Parasitol. 2012;184: 90–98. doi:10.1016/j.molbiopara.2012.04.013

147. Kuile BH ter. Adaptation of metabolic enzyme activities of Trypanosoma brucei promastigotes to growth rate and carbon regimen. J Bacteriol. 1997;179: 4699–4705. doi:10.1128/jb.179.15.4699-4705.1997

148. Simpson RM, Bruno AE, Bard JE, Buck MJ, Read LK. High-throughput sequencing of partially edited trypanosome mRNAs reveals barriers to editing progression and evidence for alternative editing. RNA. 2016;22: 677–695. doi:10.1261/rna.055160.115

149. Carnes J, McDermott S, Anupama A, Oliver BG, Sather DN, Stuart K. In vivo cleavage specificity of Trypanosoma brucei editosome endonucleases. Nucleic Acids Res. 2017;45: 4667–4686. doi:10.1093/nar/gkx116

150. Otaka E, Hashimoto T, Mizuta K. The ribosomal proteins. I: An introduction to a compilation of the protein species equivalents from various organisms by a universal code system. Protein Seq Data Anal. 1993;5: 285–300.

151. Lawson SD, Igo RP, Salavati R, Stuart KD. The specificity of nucleotide removal during RNA editing in Trypanosoma brucei. RNA. 2001;7: 1793–1802.

152. Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. Crystal structure of the entire respiratory complex I. Nature. 2013;494: 443–448. doi:10.1038/nature11871

153. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. Nat Protoc. 2012;7: 1511–1522. doi:10.1038/nprot.2012.085

154. Peng J, Xu J. A multiple-template approach to protein threading. Proteins Struct Funct Bioinforma. 2011;79: 1930–1939. doi:10.1002/prot.23016

155. Peng J, Xu J. Raptorx: Exploiting structure information for protein alignment by statistical inference. Proteins Struct Funct Bioinforma. 2011;79: 161–171. doi:10.1002/prot.23175

156. Sturm NR, Maslov DA, Blum B, Simpson L. Generation of unexpected editing patterns in Leishmania tarentolae mitochondrial mRNAs: Misediting produced by misguiding. Cell. 1992;70: 469–476. doi:10.1016/0092-8674(92)90171-8

157. Maslov DA, Thiemann O, Simpson L. Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of Leishmania tarentolae. Mol Biochem Parasitol. 1994;68: 155–159. doi:10.1016/0166-6851(94)00160-X

158. Alatortsev VS, Cruz-Reyes J, Zhelonkina AG, Sollner-Webb B. Trypanosoma brucei RNA Editing: Coupled Cycles of U Deletion Reveal Processive Activity of the Editing Complex. Mol Cell Biol. 2008;28: 2437–2445. doi:10.1128/MCB.01886-07

159. Necas D, Ohtamaa M, Määttä E, Haapala A. python-Levenshtein 0.12.0.

160. Zimmer SL, Simpson RM, Read LK. High throughput sequencing revolution reveals conserved fundamentals of U-indel editing. Wiley Interdiscip Rev RNA. 2018;9: e1487. doi:10.1002/wrna.1487

161. Simpson RM, Bruno AE, Chen R, Lott K, Tylec BL, Bard JE, et al. Trypanosome RNA Editing Mediator Complex proteins have distinct functions in gRNA utilization. Nucleic Acids Res. 2017;45: 7965–7983. doi:10.1093/nar/gkx458

162. Koslowsky DJ, Jayarama Bhat G, Read LK, Stuart K. Cycles of progressive realignment of gRNA with mRNA in RNA editing. Cell. 1991;67: 537–546. doi:10.1016/0092-8674(91)90528-7

163. Abraham JM, Feagin JE, Stuart K. Characterization of cytochrome c oxidase III transcripts that are edited only in the 3’ region. Cell. 1988;55: 267–272. doi:10.1016/0092-8674(88)90049-9

164. Sturm NR, Simpson L. Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from leishmania tarentolae mitochondria: RNA editing intermediates. Cell. 1990;61: 871–878. doi:10.1016/0092-8674(90)90197-M

165. Decker CJ, Sollner-Webb B. RNA editing involves indiscriminate U changes throughout precisely defined editing domains. Cell. 1990;61: 1001–1011. doi:10.1016/0092-8674(90)90065-M

166. Ammerman ML, Presnyak V, Fisk JC, Foda BM, Read LK. TbRGG2 facilitates kinetoplastid RNA editing initiation and progression past intrinsic pause sites. RNA. 2010;16: 2239–2251. doi:10.1261/rna.2285510

167. Wang Z, Drew ME, Morris JC, Englund PT. Asymmetrical division of the kinetoplast DNA network of the trypanosome. EMBO J. 2002;21: 4998–5005. doi:10.1093/emboj/cdf482

168. Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. Mol Biochem Parasitol. 2014;195: 115–122. doi:10.1016/j.molbiopara.2014.05.007