CONTENTS

INTRODUCTION

BACKGROUND

TriTryp Genomes

TriTryp refers to genomes T. cruzi [1] , T. brucei [2] and Leishmania major [3] published in 2005. Here we use term TriTryp to refer to trypanosomatid parasites. Trypanosomatids are unicellular flagellates parasites that include Trypanosoma and Leishmania and belong to the phylum Kinetoplastida. these parasites can infect both plant and animals and cause milions of deaths annually. Figure 1 shows a summarized consensus on Kinetoplastid phylogeny [4]

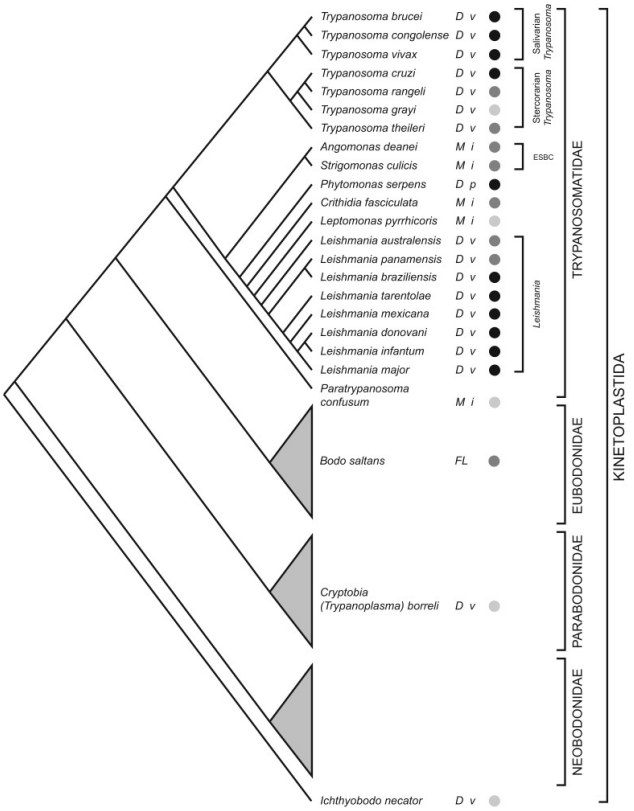

Figure 1: Kinetoplastid phylogeny

tRNA identity detereminant visualization

There are 21 (excluding initiators) class of tRNA genes, which code for 21 class of amino acids. We reperesent set of 22 single letter amino acid code as $Y = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y\}$. tRNAs with different anticodon that accept same amino acid are called isoacceptors. Here we use IUPAC one-letter code of the amino acid to label tRNAs. letter X is used for labeling initiators. tRNA identity refers to the specific amino acid it codes for. Enzymes called aminoacyl-tRNA synthetases (aaRSs) bind the right amino acid with the right tRNA. each aaRS recognizes the right tRNA based on its sequence and structural features which are called tRNA

identity determinants. Bellow I describe four identity determinants visualization tools we use in out research.

Function Logo

Function logos [5] are a generalization of sequence logos [6] made from a set of aligned tRNA sequences with length L and are defined as the Cartesian product of a state $x \in X$ where $X = \{A, C, G, U, -\}$ and $l \in L$ where $1 \leqslant l \leqslant L$. Every specific state x at position l is called feature x_l. Functional information $I_l(Y|x)$ that a state x confers about the frequencies of different classes Y at position l is calculated for each feature x_l as $I_l(Y|x) = H(Y) - e(n_l(x)) - H_l(Y|x))$. Where $H_l(Y|x) = -\sum_{y \in Y} p_l(y|x) \log_2(p_l(y|x))$ is the class entropy calculated based on the frequency of sequences that carry state x at position l, $H(Y)$ is the background class entropy which is calculated as $-\sum_{y \in Y} p(y) \log_2(p(y))$, and $e(n_l(x))$ is a correction factor to correct for biases caused by small sample size.
Function logo is a symbol-stacked-bar graph with positions on x axis and informations on y axis where each symbol within a bar is one of the functional classes of tRNAs in set Y and symbols are sorted based on their height. Height of each symbol y ($y \in Y$) for feature x_l is proportional to the frequency of sequences of class y with that feature and is calculated as $(\sum_{w \in Y} p_l(w|x)/p(w))I_l(Y|x)$. Figure 2 shows an example of function logo for state C generated by tsfm. Information at each bar are shown in bits and the maximum information of each stack is 4.2 bits. Postition 22 is an example of an indentity determinant which is fixed for functional class S. This means that sequences with C at position 22 are very likely to belong to functional class Serine.

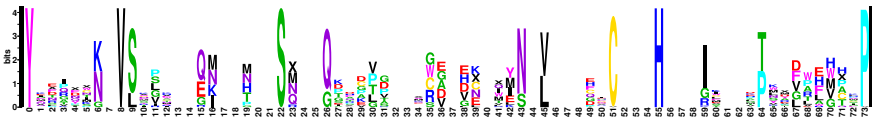

Figure 2: Function logo of one specific state generated by tsfm

Information Difference (ID) logo

ID logos [7] visualize the evolutionary gain or loss of functional information between two genomes refered as foreground and background. ID logos look like function logos except that for the height of each stack at each feature we calculate The functional information difference of two genomes refered as foreground and background as $\Delta I(Y|X_l = x) = I^F(Y|X_l = x) - I^B(Y|X_l = x)$ and the height of each symbol within a bar is proportional to $\frac{p^F(y|x_l)/p^F(y)}{p^B(y|x_l)/p^B(y)}$

KullbackeLeibler divergence difference (KLD) logos

KLD logos [7] complement the information difference measure to visualize the changes in the functional associations of features of foreground genome against background genome. KLD logos look like function logos except that the height of each stack is calculated based on sum of KLDs calculated for probability distribution of functional class y for each feature xl of two genomes as $D_{KL}(Y|X_l = x) = D_{KL}(P^F(y|x_l)||p^B(y|x_l)) =$

$\sum_{y \in Y} p^F(y|x_l) \log_2(\frac{p^F(y|x_l)}{p^B(y|x_l)})$ and the height of each symbol within a bar is proportional to $\frac{p^F(y|x_l)/p^F(y)}{p^B(y|x_l)/p^B(y)}$.

SPECIFIC AIMS

Developing a Eukaryotic tRNA identity classifier

1. Predict and annotate tRNA gene models from TriTryp genomes.
2. Create a tRNA identity classifier based on Bayesian Networks that accepts phylogenetically structured data as input and outputs posterior probabilities of functional identities for query tRNA sequences.
3. Investigate anticodon shift/functional conversion events in tRNA genes of TryTrypDB, fly, yeast, worm, etc.

Reconstructing ancestral rearrangements of tRNA gene clusters in eukaryotic genomes

Develop algorithm(s) to reconstruct ancestral rearrangements of tRNA gene clusters in eukaryotic genomes, including functional conversions and genic sequence conversion events, and apply these to the above-named eukaryotic datasets to discover functionally converting genes in these datasets and predict boundaries of gene conversion events occurring in them.

Developing a machine learning framework to model the evolution of tRNA genes on an input phylogenetic tree

Develop a machine learning framework to model the evolution of (either or both: consensus structure, structure-function map) tRNA genes on an input phylogenetic tree, and use this framework to improve alignment and gene-finding of evolutionarily diverse tRNA gene-sets.

APPROACH

Developing a Eukaryotic tRNA identity classifier

Work to Date

Predicting and annotating tRNA gene models

From tritrypdb website, we downloaded the version 41 of 46 TryTryp genomes released on 2018-12-05.Genomes are compared based on number of sequence fragments relative to their length as shown in figure 3 on the following page. later, We annotated tRNA genes for the sequenced TryTryp genomes using two computational methods for tRNA prediction, tRNAscan-SE(ref) and Aragorn(ref). We integrated the result of both genefinders by keeping the union of tRNA gene predictions generated by tRNAscan-SE v2.0 using defult options (Lowe and Eddy 1997) and Aragorn v1.2.38 using options -i116 -t -br -seq -w -e -l -d (Laslett and Canback 2004). Genes with overlapped coordinate were considered as one gene. However, the identity and exact coordinate of both genefinders were saved seperately to be anal-

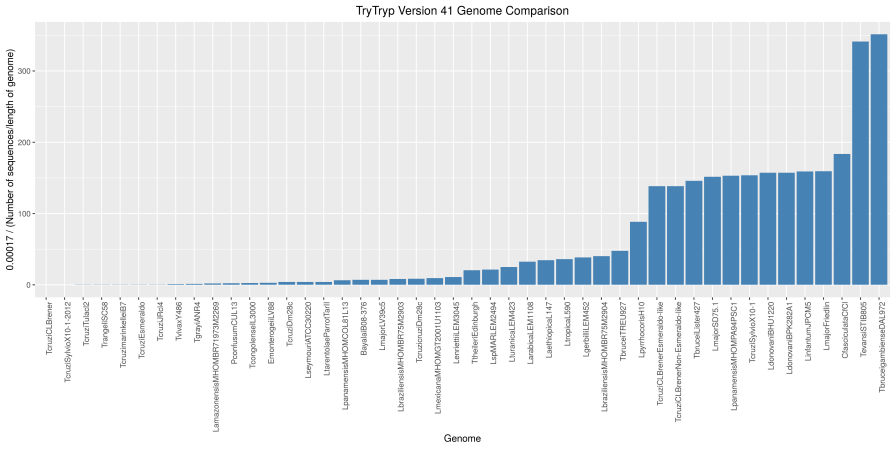

Figure 3: Comparing genomes based on formula $\left(\frac{f(x)}{\max(f(x))}\right)^{-1}$. f(x) = Number of sequences in genome devided by the length of genome. The length of the bars shows how good the genomes are sequenced.

ysed later. Since these genefinders cannot predict initiators, we predicted the initiator tRNAs for the genes with anticodon 'CAT' from intersection of both tRNAscan (TSE) and Aragorn (ARA) Based on Conserved positions of initiators in Eukarya from the study by CHRISTIAN MARCK and HENRI GROSJEAN (ref?).

Summary of predicted TryTryp tRNA genes

To investigate and compare tRNA genes predicted by two gene finders TSE and ARA, we made four sets of genes. Set one, TSE and ARA intersection, Set two, TSE and ARA union, Set three, genes found by ARA and Set four, genes found by TSE . For the intersection set, we dismissed genes which had different identity by ARA and TSE. for union set, for the porpuse of only making a summary of our genes, we picked TSE identity over ARA for overlapped genes. Table 1 shows a summary of these four sets. Further, to compare the coordinates of genes annotated by ARA and TSE we made a heatmap shown in figure 4. We see from this figure that the coordinates of same genes annotated by ARA and TSE do not always match. We then Analysed the reason for each set of displacement. Some of the main results of this analysis are: 1. Genes with same identity found by both genefindrs, usually have same reported structure, except for those with insertion or possible introns in their anticodon loop (ref). 2. ARA was seen to report up to 3 extra bases at the 3 prime end with many of them fallowing the pattern acc or ?cc, however, TSE reports only up to position 73. 4. few of the genes reported by Aragorn had Amino Acid arm one base longer than what TSE reports which caused a displacement at 5 prime end. Further, to inspect whether our genefinders annotated tRNA genes of all 22 functional classes for each genome, we visualized the number of genes annotated for each genome in Figure 5 and tRNA functional classes by both TSE and ARA for each genome in Figure 6.

Table 1: summary of the predicted genes by TSE and ARA. We marked pseudo genes as $, initiators as X, stop as #, sup as "?", sec as Z and pyl as O

GeneSet	# tRNA	# nucleotides	N/T	gene length	%G	%C	%T	%A	%intron	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	X	Z	$	?	#	O
TSE2	3631	270355	74.46	50-164	31.99	26.11	23.22	18.68	2.616	214	64	105	163	110	234	80	179	190	338	108	126	201	162	350	238	219	241	52	94	76	78	28	3	0	0
ARA	4347	372539	85.70	70-215	32.64	26.92	22.87	17.57	14.677	257	86	124	193	125	339	129	213	194	393	101	153	228	175	420	362	248	282	60	90	76	82	0	0	2	4
UNION	4381	377734	86.22	50-215	32.81	26.66	22.87	17.65	15.339	259	86	119	194	130	344	129	220	197	380	112	143	229	175	421	369	249	282	57	106	76	82	28	3	2	2
INTERSECTION	3562	265160	74.44	68-89	32.01	26.13	23.22	18.64	2.330	212	64	105	162	105	229	80	172	187	338	97	125	200	162	349	230	218	241	52	78	76	78	6	0	0	0

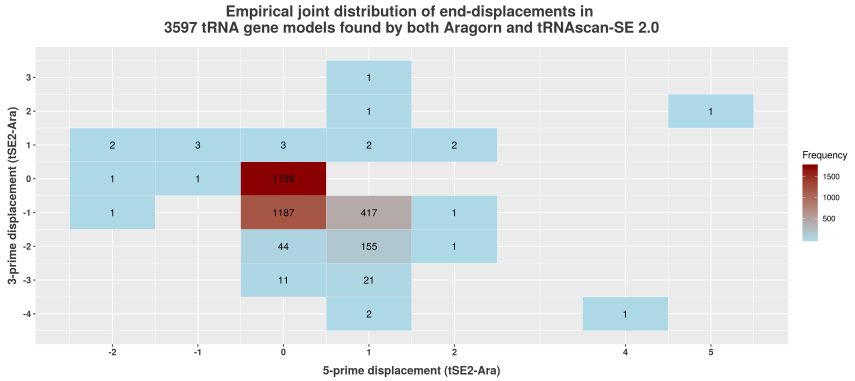

Figure 4: Empirical joint distribution of end-displacements in ? tRNA gene models found by both Aragorn and tRNAscan-SE 2.0.

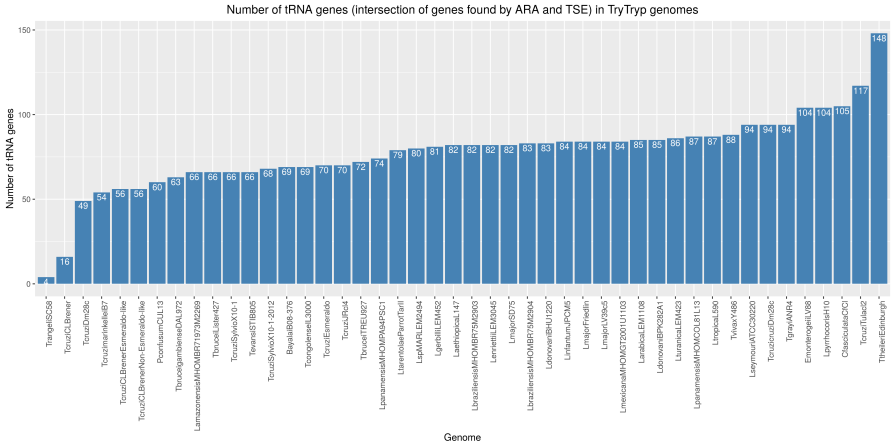

Figure 5: Number of genes annotated by both TSE and ARA for each TryTryp genome.

Pipline for alignment of TriTryp tRNA gene models

to be able to compare tRNA gene sequences, we need to build a consensus tRNA gene model. To do so, I wrote a pipline which will accept our integrated gene file as input and returns a fasta file with genes aligned to together and a consensus structure which describes their folding pattern. The first step in this pipeline is to remove the varibale arms, introns and and other nucleotides in non-conserved positions using the secondary structures reported by our gene-finders. Second, we run covea v2.4.2 (Sean Eddy 1994) for the structural alignment of our genes based on the Eukaryotic model. Third, removing sites with more than 99% gap, genes with more than 8 gaps in their aligned sequence, and genes with letter N in their sequence. at the end, we map the consensus structure to the standard numbering system (Sprinzl et al. 1991) (ref?).

Identity classifier for TriTryp tRNA gene models

0.using the intersection set of our aligned gene files as training set 1.split the gene models based on their functional class, 2.for each functional class, find the outlier using OD-seq and dismiss the outliers from training set 3.make profiles of each model class and score each sequence of our train-

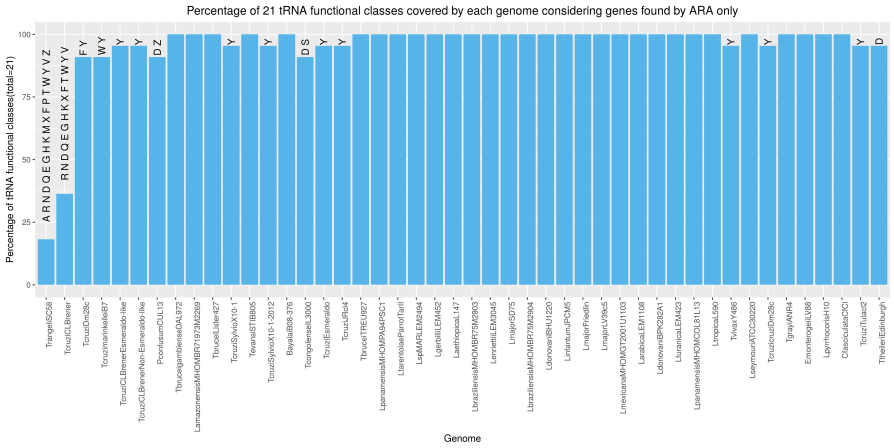

Figure 6: Percentage of 22 tRNA types annotated by both TSE and ARA for each TryTryp genomes. The label on top of each bar shows which tRNA classes are not annotated for the genome.

ing set according to the log-odds of belonging to a specific functional class. 3.inspecting the distribution of our gene file. we would like to normalize the distribution of scores against each model to be able to use z-score of sequences scores against all profile models, and assign the best fit model it them. 4. after visualizing the distribution of score in the first round and inspecting the alignment of each model using an alignment viewer seaview(ref?), we saw that in each model tRNAs of specific genomes are better aligned to eachother than others. So, to investigate sub populations of our gene population in each model class... ? what ?

Potential Anti-codon shifts

To find the potential anti-codon shifts, we calculated the z-score of the outliers from each model class against all the other models. for outlier with a relatively better z-score in another model class called ?, we compared the reported anti-codon for the genes to the possible anticodons of model ? to find the most possible shifts(shifts with the minimum numebr of substitution). later, we would like to verify these anti-codon shifts by using the synteny of gene clusters across closly related genomes (ref?).

TriTryp-specific tRNA identity determinants in compare to Human tRNA genes

We visualized differences in tRNA identity determinants between TryTryp and Human, and across TryTryp genomes, using four different Logos:

1 Function Logos to estimate the potential identity determinants for each genome
2 Information Difference logos (ID logos), to show the evolutionary gain or loss of functional information between Human and TryTryp genomes
3 KullbackeLeibler divergence Difference logos (KLD logos) to show changes in the functional associations of features between Human and TryTryp genomes
4 Using Three Logos mentioned above, we made bubble plots to show gains and shifts in functions of tRNAs in Trypanosomea contrasted against human tRNAs.

Using phylogenetic trees of Trypanosoma from these works [8–11], we grouped TryTryp genomes as table **??**. We excluded genome PconfusumCUL13 from the study, until we find a well sequenced version of this genome for which we can annotate all 22 functional classes of tRNAs. The Logo data for TryTryp genomes and Human can be found here. you can find the bubble plots for each cluster here. we have 11 pages and each page has 21 models for all tRNA classes in a cluster.

Analysis of the output ... Do I need a picture of the logos here ?

SIGNIFICANCE

tRNA identity classifier

Annotation of query tRNA genes

Determining the compete set of identity determinants and creating a structure function map of tRNA gene models is an open question which varies among different species. Gene finders such as tRNAscan-SE [12] and Aragorn [13], assign identities based on only anticodon sequence and they may not always agree specialty when there are insertions or introns within the anticodon loop [14] as we have seen in comparison of their predicted tRNA gene models for TriTryp genomes. Having a tRNA classifier which can predict the identity determinants of a group of closely related species based on the structure and sequence of their tRNA gene models, and assigns a functional class to a gene model based on those features will be robus to both anticodon prediction errors and sequence errors. TFAM [15], one major tRNA classifier which classifies the function of tRNA genes using sequence profile models, only provides bacterial tRNA identity models and models for identifying only initiator in eukaryote and archaeal. Also, tfam builds its models based on tRNA sequences of one group of species. However, identity determinants of gene models of a same functional class are different across different taxa. Building a classifier that takes a phylogenetically structured data as input, and builds profile models for classification of tRNA gene queries according to each taxa can produce more accurate information to annotate tRNA genes and resolve disagreements between two genefinders.

Detecting possible anticodon shifts

Anticodon is one of the major identity determinant in tRNAs. It is possible for a tRNA gene to switch to a different functional class after one or more mutations in its anticodon sequence and it has been shown to be possible in vitro as well (Schulman and Pelka 1989; Pallanck and Schulman 1991). mutations of bases in anticodon sequence resulting in change of tRNA's amino acid charging is called alloaccepter shifts. relative number of anticodon shifts in a taxa can be the can suggest that tRNA gene redundancy is likely the driving factor (ref?) also, by detecting the antocodon shifts, we can find sites with high covariation with anticodon shifts as potential determinants of tRNAs.
Previous works on anticodon shifts in eukaryotic tRNA genes (ref) has been done using a synteny-conservation-based method which looks for different anticodons within ortholog tRNAs as potential anticodon shifts. although this method is .... Mappings of flanking regions for each tRNA agains all other flanking regions and and ortholog set compilation may not be compu-

tationally efficient. By building a tRNA classifer we can score all the query tRNAs against all the models with time complexity of length of query. further, tRNA gene models with mismatch identity assigned by the classifier and the identity assigned by genefinders based on anti-codon, can be used as the potential anticodon shifts. these quences can be further be used in flank-mapping method for the perpuse of verification.

Detecting differences of indentity determinants between Trypanosoma and Human tRNA genes

The trypanosomatids are flagellated protozoan parasite that include species Trypanosoma brucei, Trypanosoma cruzi and Leishmania major also known as TriTryp genomes. Three major diseases caused by these species are African trypanosomiasis, South American trypanosomiasis and leishmaniasis in order, which can cause permanent disability or death in humans. Trypanosomatids are eukaryotic single cells and it is a challenge to develop drugs which can selectively target their pathogen and not affecting human host. Available treatments for Leishmanianis used for many years has been found to evolve drug resistance (3 ref). It is important to find targets that has been diverged significantly from human with less chance of developing resistance. Previous works has shown that targeting aaRSs via its interaction with tRNA can be a great target for parasites(22). Here, we would like to reannotate and classify tRNA genes in Trypanosomatids, study evolution of these tRNAs across TriTyp genomes and find differences of tRNA identity determinants between TriTryps and humans to be used in future as a target to develop non-toxic drugs for human.

REFERENCES

[1] Najib M. El-Sayed and et al. Myler. The genome sequence of trypanosoma cruzi, etiologic agent of chagas disease. *Science*, 309(5733):409–415, 2005.

[2] Matthew Berriman and et al. Ghedin. The genome of the african trypanosome trypanosoma brucei. *Science*, 309(5733):416–422, 2005.

[3] Alasdair C. Ivens and et al. Peacock. The genome of the kinetoplastid parasite, leishmania major. *Science*, 309(5733):436–442, 2005.

[4] ANDREW P. JACKSON. Genome evolution in trypanosomatid parasites. *Parasitology*, 142(S1):S40–S56, 2015.

[5] Eva Freyhult, Vincent Moulton, and David H. Ardell. Visualizing bacterial trna identity determinants and antideterminants using function logos and inverse function logos. *Nucleic Acids Research*, 34:905 – 916, 2006.

[6] Thomas Schneider and R Michael Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18, 01 2002.

[7] Eva Freyhult, Yuanyuan Cui, Olle Nilsson, and David H. Ardell. New computational methods reveal trna identity element divergence between proteobacteria and cyanobacteria. *Biochimie*, 89(10):1276 – 1288, 2007. Functional diversity of RNA.

[8] Denise Andréa Silva. et al. de Souza. Evolutionary analyses of myosin genes in trypanosomatids show a history of expansion, secondary losses and neofunctionalization. *Scientific Reports*, 8, 2018.

[9] Helen Hughes, Austin L. Piontkivska. Molecular phylogenetics of trypanosomatidae: contrasting results from 18s rrna and protein phylogenies. *PubMed*, 2, 2003.

[10] Chaiwarith R Jariyapan N Wannasan A Siriyasatien P et al. Pothirat T, Tantiworawit A. First isolation of leishmania from northern thailand: Case report, identification as leishmania martiniquensis and phylogenetic position within the leishmania enriettii complex. *PLoS Negl Trop Dis*, 12, 2014.

[11] Steven. et al. Kelly. An alternative strategy for trypanosome survival in the mammalian bloodstream revealed through genome and transcriptome analysis of the ubiquitous bovine parasite trypanosoma (megatrypanum) theileri. *Genome biology and evolution*, 9, 2017.

[12] Todd Lowe and S R Eddy. trnascan-se: A program for improved detection of transfer rna genes in genomic sequence. *Nucleic Acids Res*, 25, 01 1997.

[13] Dean Laslett and Björn Canbäck. Aragorn, a program to detect trna genes and tmrna genes in nucleotide sequences. *Nucleic acids research*, 32:11–6, 02 2004.

[14] Norma E. Padilla-Mejía, Luis E. Florencio-Martínez, Elisa E. Figueroa-Angulo, Rebeca G. Manning-Cela, Rosaura Hernández-Rivas, Peter J. Myler, and Santiago Martínez-Calvillo. Gene organization and sequence analyses of transfer rna genes in trypanosomatid parasites. *BMC Genomics*, 10(1):232, May 2009.

[15] Helena Taquist, Yuanyuan Cui, and David Ardell. Tfam 1.0: an online trna function classifier. *Nucleic acids research*, 35:W350–3, 08 2007.