1. INTRODUCTION

Trypanosomatids are unicellular flagellates parasites that include species Trypanosoma brucei, Trypanosoma cruzi and Leishmania major also known as TriTryp genomes. They belong to the phylum Kinetoplastida parasites and can infect both plant and animals and cause milions of deaths annually. 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 without affecting human host. Available treatments for Leishmanianis used for many years has been found to evolve drug resistance [1, 2]. 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 [3]. Here, we would like to re-annotate and classify tRNA genes in Trypanosomatids, study their evolution across TriTryp 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.

2. BACKGROUND

2.1 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 [4] are a generalization of sequence logos [5] 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 1 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.

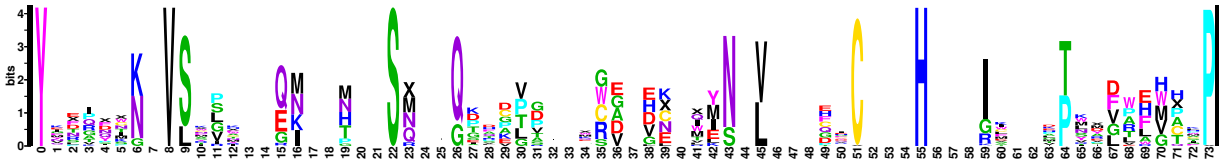


Figure 1: Function logo of one specific state generated by tsfm

Information Difference (ID) logo

ID logos [6] 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 [6] 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 $\frac{p^F(y|x_l)/p^F(y)}{p^B(y|x_l)/p^B(y)}$. Figures 3,4,7,8 show an example of function logo, ID logo and KLD logo of state C for Human and Leishmania. For example in figure 7 and 8 at position 0 you can see the functional differences associated to feature C0.

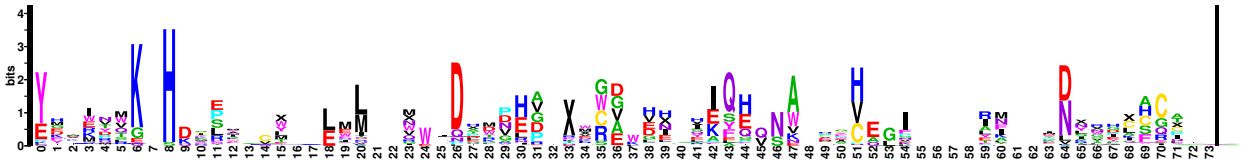


Figure (3) Function logo of state C for Human

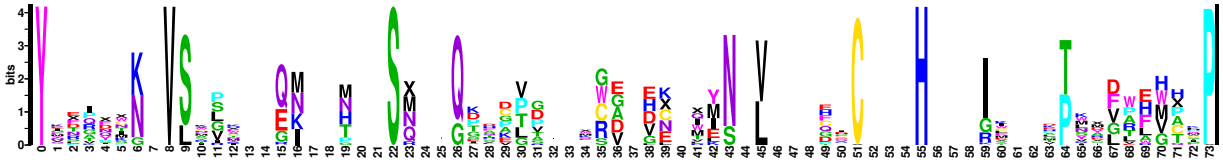


Figure (4) Function logo of state C for Leishmania

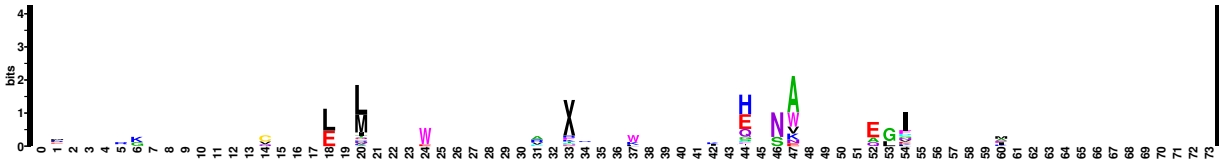


Figure (5) Id logo of state C showing features (and the functions associated to them) in Human tDNAs with excess functional information when contrasted against those features in Leishmania tDNAs

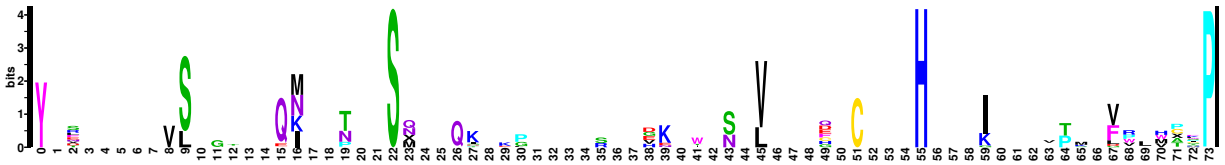


Figure (6) Id logo of state C showing features (and the functions associated to them) in Leishmania tDNAs with excess functional information when contrasted against those features in Human tDNAs

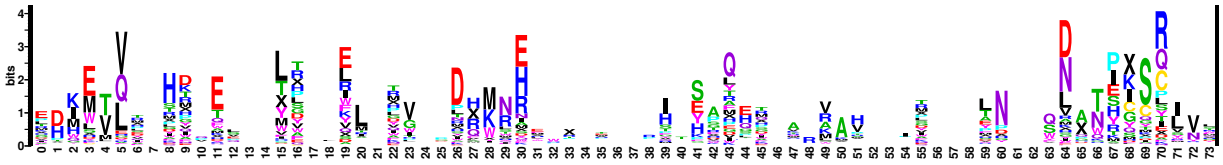


Figure (7) KLD logo of state C showing which functions are excessively associated to features in Leishmania tDNAs when compared to Human tDNAs

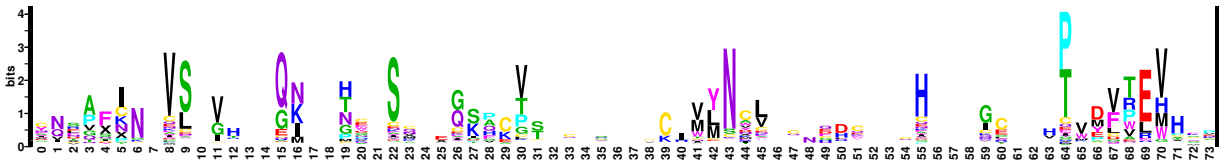


Figure (8) KLD logo of state C showing which functions are excessively associated to features in Human tDNAs when compared to Leishmania tDNAs

3. SPECIFIC AIMS

3.1 Developing a Eukaryotic tRNA identity classifier

- Predict and annotate tRNA gene models from TriTryp genomes.
- 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.
- Investigate anticodon shift/functional conversion events in tRNA genes of TryTrypDB, fly, yeast, worm, etc.

3.2 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.

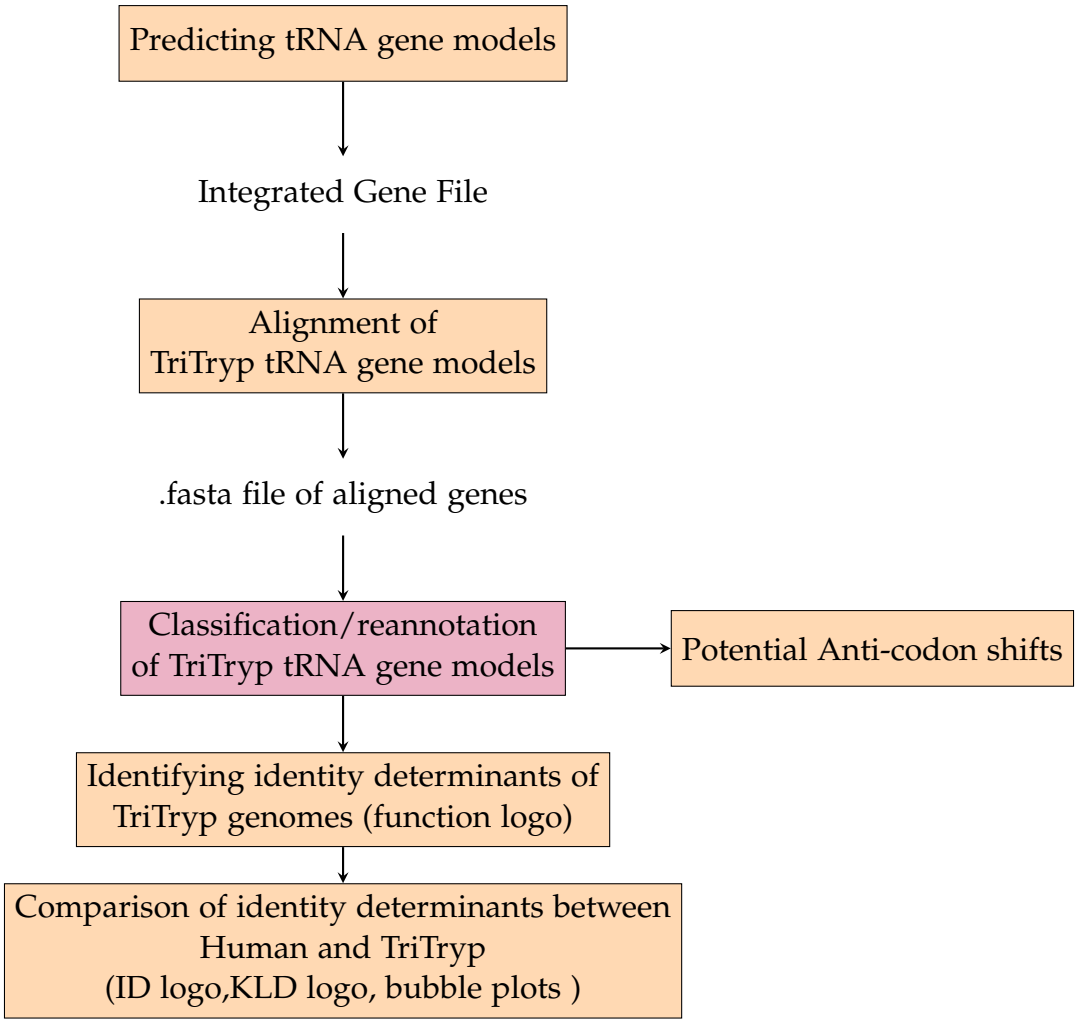
3.3 Developing a machine learning framework to model the evolution of tRNA genes on an input phylogenetic tree

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.

4. METHOD

4.1 Developing a Eukaryotic tRNA identity classifier

The fallowing flowchart illustrates the workflow of building a classifier and its use in comparison of TriTryp genomes together and versus Human and detecting potential Anticodon shifts.



4.1.1 Predicting and annotating tRNA gene models

From Tritrypdb [7], 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 fig. 8 on the following page. Later, We annotated tRNA genes for the sequenced TryTryp genomes using two computational methods for tRNA prediction, tRNAscan-SE (TSE) [8] and Aragorn (ARAx) [9]. We integrated the result of both gene finders 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 analysed later. Since these genefinders cannot predict initiators, we predicted the initiator tRNAs for the genes with anticodon 'CAT' from intersection of both genefinders Based on Conserved positions of initiators in Eukarya from the study by Christian Marck and Henri Grosjean [10].

4.1.2 Summary of predicted TriTryp 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.

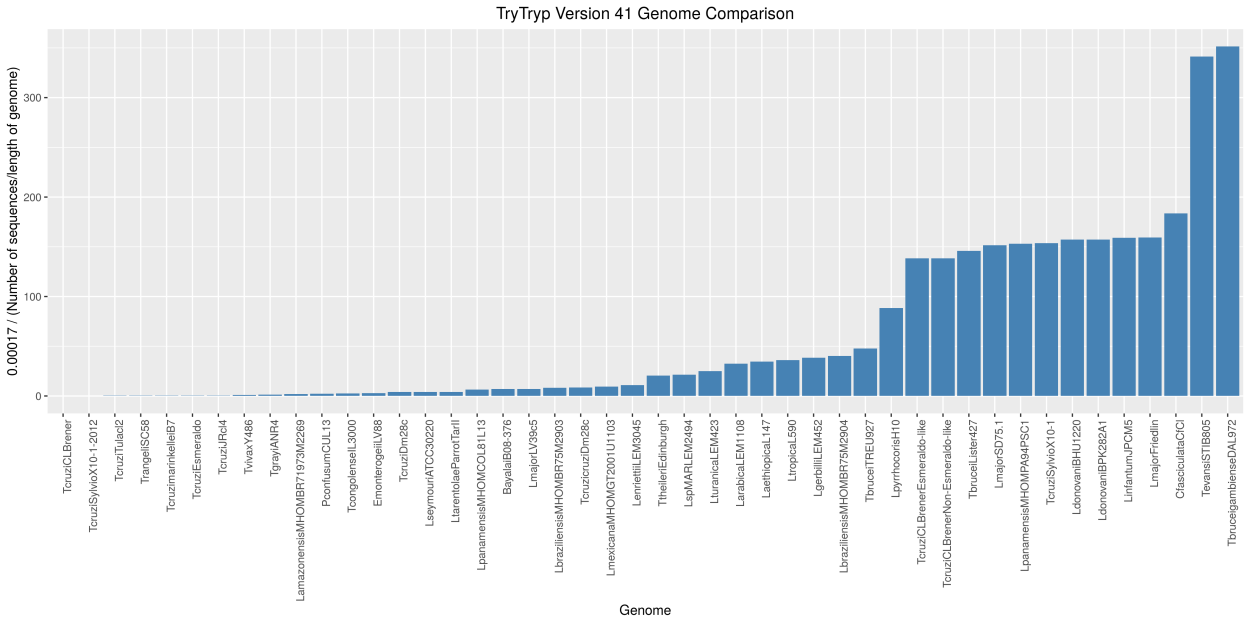


Figure 8: Comparing genomes based on formula $(\frac{f(x)}{\max(f(x))})^{-1}$. f(x) = Number of sequences in genome devided by the length of genome. The length of the bars shows quality of genome sequencing.

For union set, for the purpose of making a summary of our gene model's identity, 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 heat-map shown in figure 9. We see from this figure that the coordinates of same genes annotated by ARA and TSE do not always match. We then investigated 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, often have same reported structure, except for those with insertion or introns in their anticodon loop [11]. 2) ARA reported genes up to 76 bases with 3 extra bases at the 3 prime end, however, TSE reported genes up to position 73. 3) Few of the genes had Amino Acid arm one base longer in ARA output in compare to TSE output 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 10 and tRNA functional classes by both TSE and ARA for each genome in Figure 11.

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

4.1.3 Alignment of TriTryp tRNA gene models

The secondary structure of most tRNAs is made up of three stem-loop and one stem with a cloverleaf structure. Nucleotides at each position in different tRNAs with canonical structure [12] have a comparable function. Building a consensus secondary structure for tRNAs based on a Eukaryotic tRNA gene model, is a necessity for detection of conserved positions for a subset of genes that determines their function. To build a consensus tRNA gene model we wrote a pipeline that accepts our integrated gene file as input and

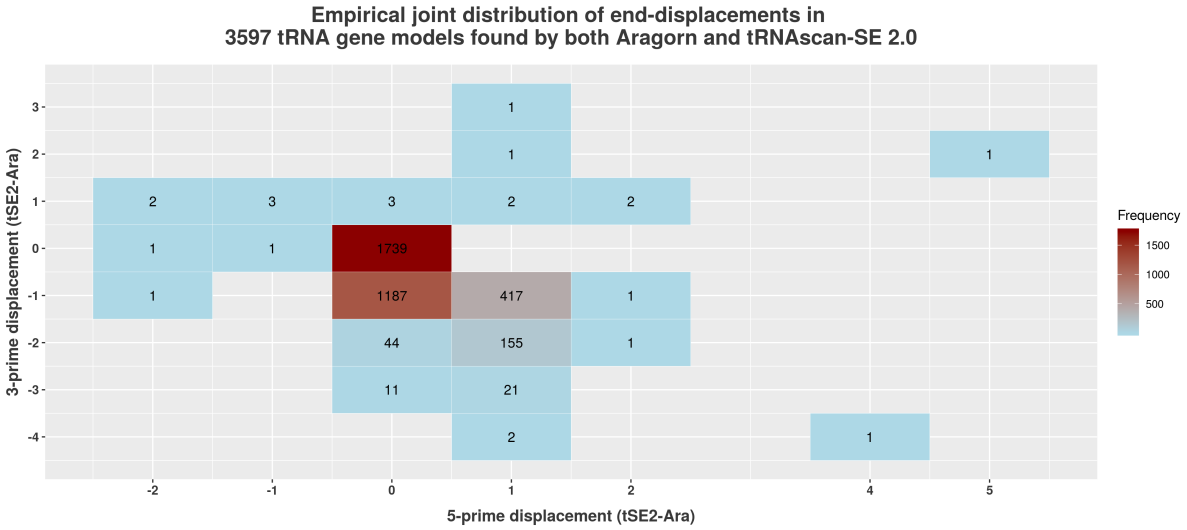


Figure 9: Empirical joint distribution of end-displacements in ? tRNA gene models found by both Aragorn and tRNAscan-SE 2.0.

returns a fasta file of aligned gene models along with a text file of consensus structure which describes their folding pattern. The pipeline starts with removing the variable arms, introns and and other nucleotides in non-conserved positions using the secondary structures reported by our gene-finders. Later, it calls covea v2.4.2 (Sean Eddy 1994) for the structural alignment of our genes based on the Eukaryotic model. Then, it removes 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) [12].

4.1.4 Classification of TriTryp tRNA gene models

A tRNA gene classifier built from tRNA genes of closely related genomes can take one or more novel tRNA genes, align them to its consensus secondary structure model and return their most possible functional class based on similarity of nucleotides at each position to the training set. Here we used the intersection set of our aligned genefile as training set to build a classifier for TriTryp tRNA genes and will extend our classifier later to be applied to other Eukaryotes. We used a package called OD-seq [13] to find the outliers of tRNA genes of each functional class from our training set. Then we made profiles of each class as a 5 * L matrix where L is the length of aligned tRNA genes and each row is for one nucleotide symbol and gap. The details of creating profiles are similar to profiles created in tfam as described here [14]. We score each sequence of our training set against each profile according to the log-odds of belonging to a each functional class. Later, to compare the scores against each profile we standardized the scores by first inspecting the distribution of scores for each class and later deconvolving the distributions to make subpopulations of scores with normal distributions.

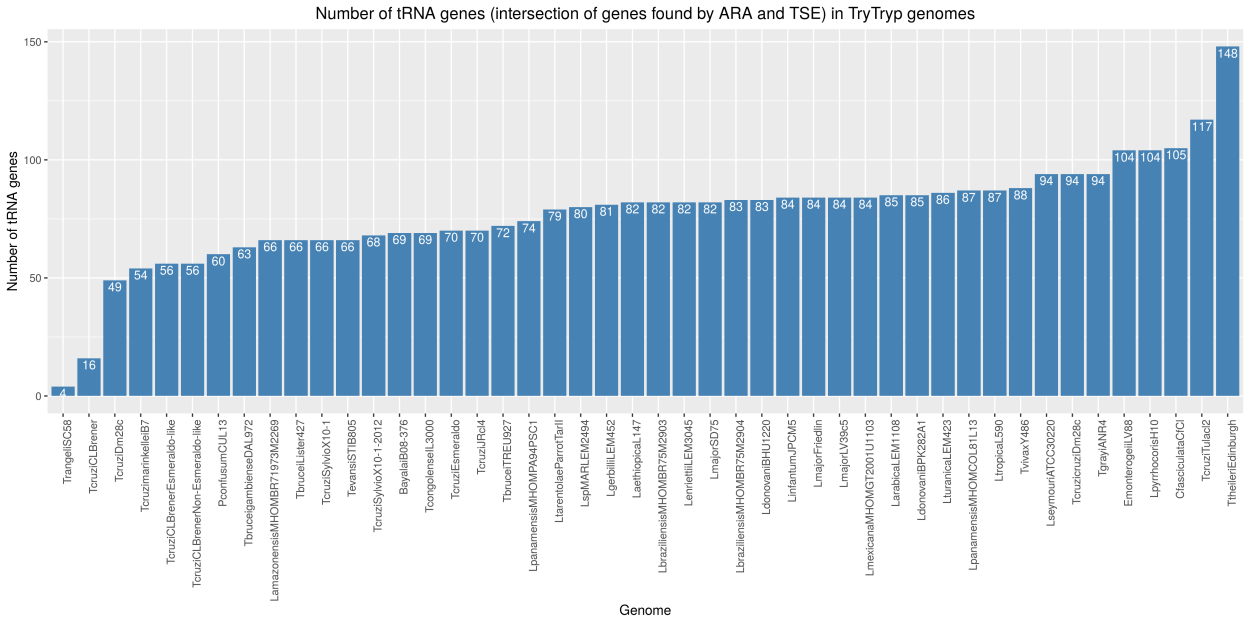


Figure 10: Number of genes annotated by both TSE and ARA for each TryTryp genome.

4.1.5 Potential Anti-codon shifts

Anticodon is one of the major identity determinant in tRNAs. It is shown in vitro that it is possible for a tRNA gene to switch to a different functional class after one or more mutations in its anticodon sequence [15, 16]. Mutation in anticodon resulting in change of tRNA's amino acid charging is called alloaccepter shift. To find the potential anti-codon shifts, we score the outliers against each profile model and assign a functional class to each sequence against which it has maximum score. If the functional class assigned by the classifier did not match the class assigned by genefinders, we consider that as a potential anticodon shift. In order to verify the shifts we will define clusters of tRNA genes for TriTryp genomes as a set of genes located on a same sequence within the distance of K (number of gene clusters can change according to k). Later we will find clusters that have potential anticodon shifts. Then, using the synteny of gene clusters across closely related genomes we will look for ortholog clusters of these clusters. Within each two ortholog clusters we will look for an ortholog gene (lets call it x') for our tRNA gene with potential anticodon shift (lets call it x). We will mark a shift in gene x verified If the functional class assigned by classifier to tRNA gene x matches the functional class assigned by genefinders to gene x'.

4.1.6 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 Logo, to estimate the potential identity determinants for each genome, 2) ID logo, to show the evolutionary gain or loss of functional information between Human and TryTryp genomes, 3 Later using phylogenetic trees of Trypanosoma from these works [17–20], we grouped TryTryp genomes as two clusters of Trypanosoma and Leishmania, created the bubble plots for each of them against human to explore their differnces in each group. Analysis of the output ... continues LOL! Do I need to put a picture of the logos and bubble plots here ?

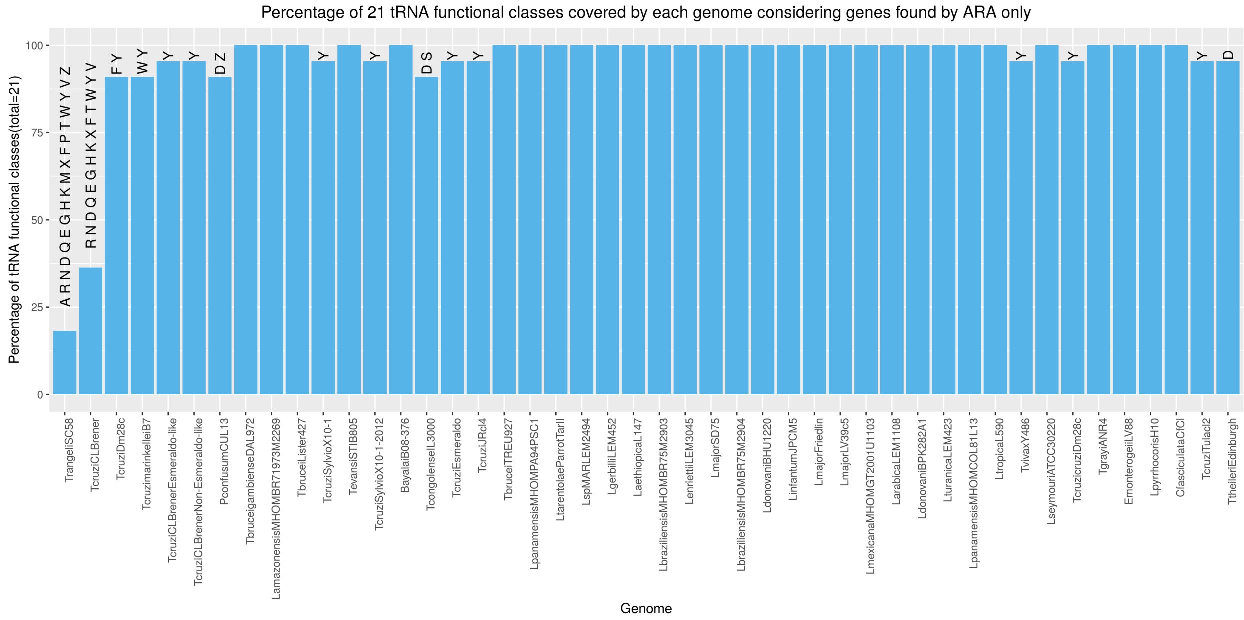


Figure 11: Percentage of 22 tRNA types annotated by both TSE and ARA for each TryTryp genomes. The label on top of each bar shows which tRNA classes are not annotated for the genome.

5. SIGNIFICANCE

5.1 tRNA identity classifier

5.1.1 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 [8] and Aragorn [9], assign identities based on only anticodon sequence and they may not always agree specialty when there are insertions or introns within the anticodon loop [11] 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 anti-codon prediction errors and sequence errors. TFAM [21], a 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.

5.1.2 tRNA anticodon shifts

Previous work on detecting anticodon mutations in eukaryotic genomes suggests that tRNA redundancy can be a reason for relative number of anticodon shifts in a taxa [22]. A synteny-conservation-based method is used in this work which looks for different anti-

codons within ortholog tRNAs as potential anticodon shifts. Although the restriction of a tRNA gene having at least one ortholog to be considered active is important, mapping of flanking regions for each tRNA against all other flanking regions to find ortholog sets may not be computationally efficient. A tRNA classifier can score all the query tRNAs against all the profile models of a classifier with time complexity of length of queries. tRNA gene models with mismatch identity assigned by the classifier and the identity assigned by gene finders based on anticodon, can be used as the potential anticodon shifts. Later it can be used in flank-mapping method for the purpose of verification. Further, by predicting anticodon shifts, we can find sites that covary with these anticodon mutations as potential determinants of tRNAs.

REFERENCES

[1] Simon L. Croft, Shyam Sundar, and Alan H. Fairlamb. Drug resistance in leishmaniasis. *Clinical Microbiology Reviews*, 19(1):111–126, 2006.

[2] Alicia Ponte-Sucre, Francisco Gamarro, Jean-Claude Dujardin, Michael P. Barrett, Rogelio López-Vélez, Raquel García-Hernández, Andrew W. Pountain, Roy Mwenechanya, and Barbara Papadopoulou. Drug resistance and treatment failure in leishmaniasis: A 21st century challenge. *PLOS Neglected Tropical Diseases*, 11:1–24, 12 2017.

[3] Paul Schimmel, JS Tao, and Jason Hill. Aminoacyl trna synthetases as targets for new anti-infectives. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 12:1599–609, 01 1999.

[4] 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.

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

[6] 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.

[7] et al. Aslett M, Aurrecoechea C. Tritrypdb: a functional genomic resource for the trypanosomatidae. In *Nucleic Acids Research*, 2010.

[8] 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.

[9] 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.

[10] Christian Marck and Henri Grosjean. trnomics: Analysis of trna genes from 50 genomes of eukarya, archaea, and bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA (New York, N.Y.)*, 8:1189–232, 11 2002.

[11] 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.

[12] Anatoli Ioudovitch, Sergey Steinberg, Carsten Horn, Melissa Brown, and Mathias Sprinzl. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 26(1):148–153, 01 1998.

[13] Peter Jehl, Fabian Sievers, and Desmond G. Higgins. Od-seq: outlier detection in multiple sequence alignments. *BMC Bioinformatics*, 16(1):269, Aug 2015.

[14] David H. Ardell and Siv G. E. Andersson. Tfam detects co-evolution of trna identity rules with lateral transfer of histidyl-trna synthetase. *Nucleic Acids Research*, 34:893 – 904, 2006.

[15] L H Schulman and H Pelka. The anticodon contains a major element of the identity of arginine transfer rnas. *Science (New York, N.Y.)*, 246:1595–7, 01 1990.

[16] L Pallanck and L H Schulman. Anticodon-dependent aminoacylation of a noncognate trna with isoleucine, valine, and phenylalanine in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 88:3872–6, 06 1991.

[17] 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.

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

[19] 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.

[20] 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.

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

[22] Hubert H. Rogers and Sam Griffiths-Jones. trna anticodon shifts in eukaryotic genomes. *RNA*, 20 3:269–81, 2014.