0. BACKGROUND

TriTryp Phylogenetic Tree

Phylogenetic trees of Trypanosomatids from these works [1–4] are used for classification of genomes in this work. figure 1 shows a simplified classification of Trypanosoma and Leishmania genomes.

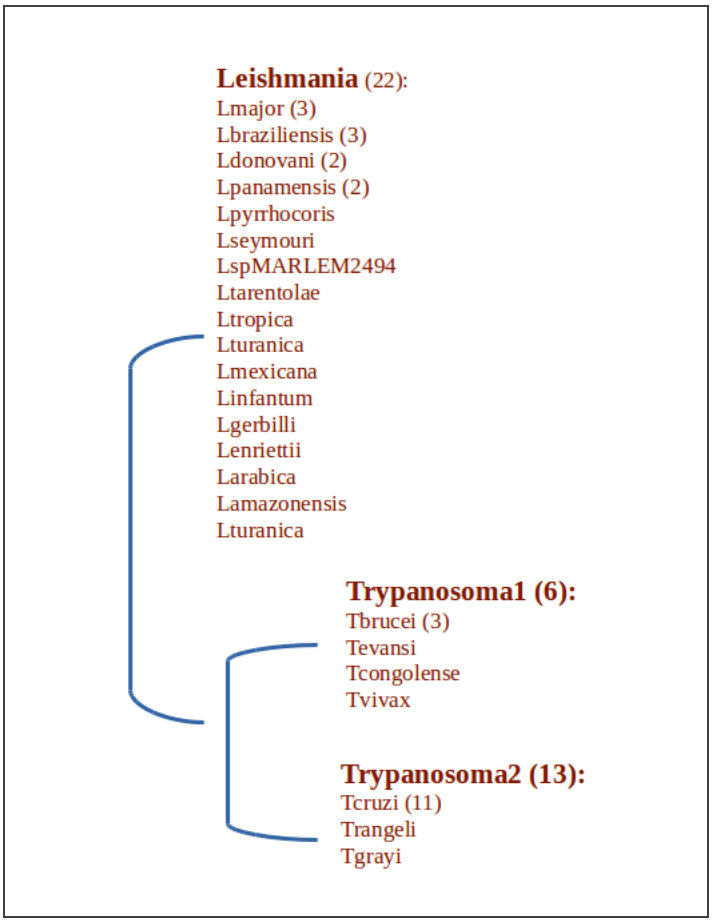

Figure 1: TriTryp genomes clustered into three classes of Leishmania(L), Trypanosoma1(T1), and Trypanosoma2(T2). other TriTryp genomes within this tree which does not fit into these categories are labeled as Others(o)

Gene Cluster

We define clusters as a group of two or more genes found within a genome located within a thousand base pairs of each other. For example, cluster "IVQRLTRKGW" is a sequence of genes shown with their one letter amino acid code which are located on same double stranded sequence of a genome and are within a 1000bp of each other. Also the orientation of a cluster is shown as a sequence of $+$s and $-$s which refer to the strand of each gene within the cluster, in order. for instance the orientation of cluster "IVQRLTRKGW" can be shown as $"+-----++-+"$.

1. PREDICTING AND ANNOTATING TRNA GENE MODELS

From Tritrypdb [5], we downloaded the version 41 of 46 TryTryp genomes released on 2018-12-05. The quality of sequenced genomes are compared based on number of sequence fragments relative to their length as shown in figure 2. Later, We annotated tRNA genes for the sequenced TryTryp genomes using two computational methods for tRNA prediction, tRNAscan-SE (TSE) [6] and Aragorn (ARA) [7]. 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. 4381 genes were predicted from which 3597 genes were predicted by both gene finders, 750 genes predicted by ARA only and 34 genes predicted by TSE only. Since Aragorn cannot predict the initiators, we predicted the initiator tRNAs for the genes with anticodon 'CAT' from union of both genefinders Based on Conserved positions of initiators in Eukarya from the study by Christian Marck and Henri Grosjean [8].

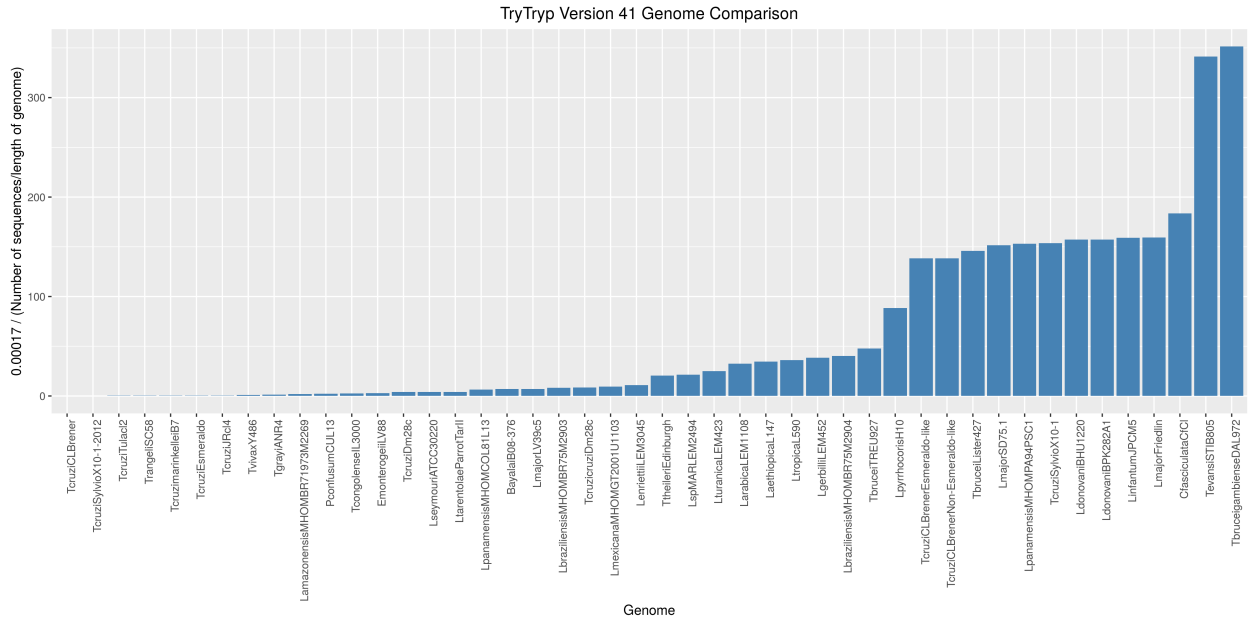

Figure 2: Comparing genomes based on formula $\left(\frac{f(x)}{\max(f(x))}\right)^{-1}$. f(x) = Number of sequences in genome devided by the length of genome. The length of the bars shows quality of genome sequencing.

1.2 Initiator tRNA prediction

We predicted the initiator tRNAs for the genes with anticodon 'CAT' from union of both tRNAscan (TSE) and Aragorn (ARA) Based on Conserved positions of initiators in Eukarya from the study by CHRISTIAN MARCK and HENRI GROSJEAN. Based on this study we have the fallowing citeria for initiators:

1. In all eukaryotic tDNA-iMet, positions 11–24 are occupied by C-G, However, eukaryotic elongators also prefer C-G at these positions.

2. Initiator tDNAs from Eukarya use A54 and A60. Some eukaryotic elongators also use either A54 or A60 but none (with only one exception) uses both
3. Initiator tDNA-iMet (CAT) from all domains display the GGG sequence (Mandal et al+, 1996) or, very seldom, the AGG sequence at positions 29 to 31, pairing with the complementary CCC or CCT sequences at positions 39 to 41
4. Another domain-specific feature in all eukaryotic initiators is the systematic nonoccupancy of all optional positions of the D-loop (17, 17a, 20a, and 20b) whereas in elongators, only position 17a is always unoccupied.
5. At position 20, A is strictly conserved in all eukaryotic initiators

To investigate all these features first we used CAT genes from the gene set. We had 188 genes with anticodon CAT from which 4 genes were found by only ARA and the rest were found by both genefinders. Later, we aligned all the genes from the intersection of ARA and TSE along with 4 CAT genes found only by ARA. in the result gene set, we had 185 CAT genes left (3 of them which were found by ARA were removed during the alignment due to unusual structure. these genes were also low score ARA genes with ARA score of less than 102). Later we clustered these 185 CAT genes using Levenshtein (edit) distance between gene sequences and Ward.D2 method to measure the dissimilarity between each two clusters. We ended up with three clusters. Table 1 investigates each of these features in each column. from this table we see that only tRNA genes in cluster 1 have almost all the conserved features for eukaryotic initiators. So, we marked these genes as initiators represented with letter X in our gene file. The only CAT gene found by ARA was is included in cluster 3. Further, to verify these 76 initiator, we compared them to the CAT genes marked as initiators (iMet) by TSE. There were 76 genes marked as initiator by TSE which matched with all the CAT genes we marked as initiator.

Table 1: Table of CAT clusters to show how many tRNA genes in each cluster satisfy each feature

Clusters	# tRNAs	11–24(C-G)	54-60(A-A)(T-T)	1-72(A-T)	29-31(GGG)	39-41(CCC/CCT)	# posisInDloop	20A	distanceRange
Cluster1	76	76	76	76	76	76	7	75	
Cluster2	95	95	2	0	0	0	8	0	
Cluster3	14	2	13	0	0	0	1	0	

Detailed Gene Annotation

The fallowing is a list of ambiguities in predicted genes by two gene finders which needs to be resolved in order to integrate and filter out the untrusted genes.

1. Genes found by Aragorn only
2. Genes found by tRNAScan-SE only
3. Genes marked as pseudo or truncated by TSE
4. Genes with unidentified identity by both gene finders
5. Genes with different identities between two gene finders

1.3 Pseudo Genes

There are 43 genes marked as pseudo or truncated by TSE. Table 2 shows the number of pseudo and truncated genes from our initial gene set. Among these, 4 genes found only by TSE with score of between 20-30 (genes which barely passed the score threshold (20) by TSE) were not assigned any identity or anticodon. From 39 remain genes, 15 genes were found by both gene finders and the rest (24 genes) were predicted by only TSE. In order to verify these 39 genes with assigned identity, we analyzed them among clusters. Table 3 shows the pseudo/truncated genes within clusters along with singletons. Pseudo genes are marked with lower letter characters. Table 4 shows the sets of clusters which are similar to the pseudo containing clusters with possible variations including deletion, duplicated, insertion and inversion. From these tables we can see that all the pseudo containing clusters except for cluster "Eye" are similar to other non-pseudo gene containing clusters, which means that these genes have potentially lost some functionalities due to structural variations in chunks of genomes including inversions and duplication. Also, 21 of these pseudo genes appeared as singletons from which 8 genes were found by both gene finders. Also, one of the the pseudo genes found by both gene finders has been assigned with two different anticodon/identity, I (by TSE) and D (by ARA).

Table 2: Genes labeled as pseudo, truncated or both by TSE

pseudo	truncated	truncated,pseudo
28	11	4

1.4 Other Unidentified Genes

There were **15 genes** from our initial gene set with unassigned identity both genefinders, 9 found by TSE (with score < 30), 4 found by ARA (with scores between 100-101) and 2 found by both (with ARA score of 116,112 and TSE score of 54,51). Three of the 9 genes found by TSE are marked as pseudo and one as truncated. TSE has not been able to assign any anticodon to these genes. However ARA assigned anticodon of length 2 or 4 to these genes. These genes seems to have unusual arm length including unpaired base and/or insertion in their arms. **(These 15 genes are excluded in annotation of genes in the fallowing sections which leaves the total number of 4366 genes: 3595 found by both TSE and ARA, 30 genes by TSE only and 741 genes by ARA only)**
(note! we can find the identity of these two genes using the profiles! and look at them within clusters for potential shifts or inversions).

Also, there were 4 genes with identities A,K,Z,E. 1 found by only TSE and 3 found by both genefinders which has 1- 3 letters of N. with TSE score of 60-11 and ARA score of 107-117.

1.5 Genes found by only one gene finder

1.5.1 ARA only genes

741 genes of our total initial gene set were found by only ARA. Table 5 shows the ARA-only genes within clusters along with ARA-only singletons. ARA -only genes are shown

Table 3: Table of pseudo-gene containing clusters along with singleton pseudo genes. Column genes foundby shows which gene finder predicted each gene within the cluster, in order. Column cluster foundby shows whether the whole cluster has been found by both gene finders or TSE only. Cluster i/d refers to genes found by both genefinders with different assigned identity. TSE identity of I and ARA identity of D.

cluster	frequency	genes foundby	cluster foundby
DSa	1	both-both-both	both
Eye	1	both-tse-both	tse
f	4	both,tse	both,tse
fIQ	1	tse-both-both	tse
FIQa	1	both-both-both-tse	tse
FIQFIQafIQ	1	both-both-both-both-both-both-tse-tse-both-both	tse
i	1	both	both
i/d	2	both	both
IQl	1	both-both-both	both
k	3	both	both
lS	1	both-both	both
Nk	1	both-tse	tse
p	1	tse	tse
PTn	1	both-both-tse	tse
r	1	both	both
s	7	tse	tse
sK	1	both-both	both
sKHSk	1	both-both-both-both-tse	tse
t	1	tse	tse
YTyTTy	2	both-both-tse-both-both-tse	tse
z	1	both	both

Table 4: Sets of pseudo containing clusters with potential variations between clusters of each set including duplication, deletion, inversion. There are three genome classes T, L and O which refer to Trypanosoma, Leishmania and Other genomes, in order. Clusters with same set number are considered similar. Three last column shows the orientation of genes within clusters for each class of genomes.

clusters	set	Genomeclass	Tdirs	Ldirs	Odirs
Eye	0	T	++-		
Nk	1	T	–		
KN	1	L		+-	
ENKRRA	1	O			–++-+
DSa	2	T	-++		
ASD	2	LT	–+	–+	
AS	2	O			+-
DSA	2	OLT	-++	-++	-++
ASDD	2	T	–++		
DS	2	T	-+		
DSV	2	T	-++		
YTyTTy	3	L		—-++	
YTTY	3	L		–++	
YTY	3	L		–+	
YT	3	LT	+-	–	
YTRD	3	OL		—-	—-
TY	3	OLT	+-	++	++
lS	4	T	–		
MEYSL	4	O			—-+
LS	4	OLT	–	–	–
SL	4	OLT	–|++	++	++
LSM	4	T	-+-		
LSMEMSMYV	4	T	-+-++-+-+		
LSMEMYV	4	T	-+-++-+|-+–+-+		
MEMSL	4	T	–+-+|-++-+		
VYMEMSL	4	T	-+-++-+|-+–+-+		

clusters	set	Genomeclass	Tdirs	Ldirs	Odirs
FIQa	5	T	—+		
fIQ	5	T	—		
FIQFIQafIQ	5	T	——+—		
QFIQ	5	T	—-		
IQl	6	T	-++		
ILQ	6	T	+–		
ILQI	6	T	+-++		
ILQQI	6	T	+–++		
IQ	6	T	-+		
IQL	6	T	-++		
IQQL	6	T	–++		
IQQLI	6	T	–++-		
QLI	6	LT	-+-	+-+	
QLIIIII	6	O			+-++–+
PTn	7	T	-++		
NPTYN	7	L		–+++	
YTN	7	O			—
NYTPN	7	OL		—++	—++
TN	7	OT	++		++
NT	7	T	–		
NTP	7	T	–+		
NY	7	T	-+		
PTN	7	T	-++		
PTYN	7	T	-+-+		
sK	8	T	–		
sKHSk	8	T	—-		

written in lower case letters. Among these genes 20 are within clusters and the rest appeared as singletons. Column #similar shows number of times each of these ARA-only-gene containing cluster sequences has appeared as subsequence of other clusters made from non-ara-only genes. Last column shows the number of each ara-only gene within the clusters (or singletons) with score of > 106. Based on these two columns, we can see that ara-only-genes which are complementing already seen clusters (clusters that have occurred many times shown in column similar), have higher scores than most of the ara-only singletons genes. However, the ara-only gene containing clusters which have not been seen in other clusters (their #similar is 0 or 1), have relatively low scores (lower than 107). This shows that using a cutoff score of 107 for ARA genes will keep 36 genes which are potentially most of the reliable ara-only genes figure 3a. Table 6 shows few sample of these similar clusters. *(note: the sequences of these 20 genes are potentially same as already found genes by TSE and ARA, which were missed by TSE. this should be checked!)*

1.5.2 TSE only genes

TSE predicted **30 genes** which were not predicted by ARA. 29 of these genes are very low scored TSE genes scored from 20 to 48. and one gene with score 58. Table 7 shows the frequency of these genes as three sets of pseudo, truncated, and truncated,pseudo. the last three sets has already been annotated. Table 8 shows a summary of TSE-only genes as clusters. Genes appeared as singletons have very low score, however genes appered within the clusters have relatively higher scores. cluster VYIMLs which contains gene S found by only TSE has been seen before as exact same clusters with genes found by both genefinders. Also, cluster QLIiIII which contains gene i, has also appeared many times as form of QLI, ILQ and other similar clusters which can be validify the existence of this TSE-only gene. Figure 3b shows the TSE scores of genes found by TSE.

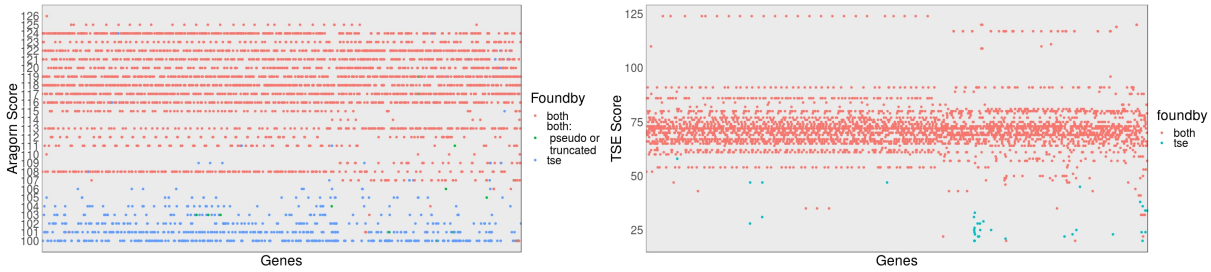

Figure 3: a) ARA score of genes found by both genefinders TSE and ARA, and genes found by only ARA. b) TSE score of genes found by both genefinders and genes found by only TSE

Table 5: Table of ara-only-gene containing clusters and singletons. column #similar shows number of times cluster sequence has appeared as subsequence of other clusters made from the non-ara-only genes from our initial gene set. Last column shows the number of occurances of the ara-only gene within the clusters (or singletons) with score of > 106. this column will be used later for assigning a cutoff score for ara genes. $\star$ in the last columns shows an except. Cluster sl has been found as substring of 22 clusters, however its genes have scores lower than 107. the score for these two genes are 106,105 in order which is close to the threshold.

cluster	frequency	# similar	# genes with score > 106
dSA	1	37	1
fRA	1	4	1
rh	1	25	1
gT	1	29	1
lA	2	24	2
lI	1	26	1
sl	1	22	$\star$
tt	1	14	1
aE	1	1	0
pppp	1	0	0
Ek	1	0	0
Za	1	0	0
Zr	1	0	0
#	2	-	0
a	43	-	2
c	22	-	1
d	13	-	0
e	31	-	1
f	19	-	1
g	109	-	2
h	48	-	3
i	41	-	1
k	6	-	1
l	38	-	3
m	4	-	0
n	17	-	0
o	2	-	0
p	24	-	0
q	13	-	0
r	69	-	1
s	130	-	7
t	28	-	2
v	41	-	0
w	5	-	1
y	12	-	0
z	4	-	0

Table 6: Sets of ara-only-gene containing clusters with potential variations between clusters of each set including duplication, deletion, inversion. There are three genome classes T, L and O which refer to Trypanosoma, Leishmania and Other genomes, in order. Clusters with same set number are considered similar. Three last column shows the orientation of genes within clusters for each class of genomes.

clusters	set	Genomeclass	Tdirs	Ldirs	Odirs
RRAE	1	O			–++
EARR	1	L		–++	
Ek	1	L		–	
dSA	4	T	-++		
DSA	4	OLT	-++	-++	-++
DSV	4	T	-++		
ASD	4	LT	–+	–+	
ASDD	4	T	–++		
DS	4	T	-+		
fRA	1	T	-++		
FRA	1	T	-++		
ARF	1	T	–+		
ARFARF	1	T	–+–+		
gT	2	T	-+		
GT	2	OLT	-+	++	++
GTGP	2	L		+–+|++-+	
GTGPVK	2	L		+–+++	
GTP	2	L		+-+	
PGTTG	2	L		-++–	
TG	2	OT	-+		–
TGP	2	OL		–+	–+
PTG	2	L		—	

clusters	set	Genomeclass	Tdirs	Ldirs	Odirs
pppp	2	T	—-		
PP	2	O			-+
rh	7	T	–		
EVRH	7	OL		–++	–++
FHVRH	7	L		+–++	
RHLP	7	T	++++		
HRVE	7	L		–++	
sl	5	T	–		
SL	5	OLT	++	++	++
SLS	5	T	+–|++-		
SLMIV	2	L		-+–+	
MEMSL	5	T	–+-+|-++-+		
MEYSL	5	O			—-+
VYMEMSL	5	T	-+-++-+|-+–+-+		
LS	5	OLT	–	–	–
LSM	5	T	-+-		
LSMEMSMYV	5	T	-+-++-+-+		
LSMEMYV	5	T	-+-++-+|-+–+-+		
LSSS	5	O			—-
LSMIVY	2	O			-+—+
Za	13	L		–	
Zr	1	O			–

Table 7: TSE-only genes

others	pseudo	truncated	truncated,pseudo
6	17	5	2
28-58	20-38	28-47	22-25

Table 8: TSE-only genes

cluster	frequency	TSE score
g	2	28
i	1	33
k	1	24
QLIiIII	1	58
VYIMLs	1	47

1.6 Deviations between two gene finders

Deviation between two gene finders for the predicted genes is defined in form of fallowing question: Assuming two genes: A1 (predicted by TSE with identity I1) and B (predicted by ARA with identity I2) which I1 can be equal to I2 or not. The coordinates of A and B overlaps, However, there is a small displacement at the 3′ and/or 5′ ends of these genes. We say that these two genes are one gene (lets call it gene A) with different predicted structure by two gene finders which had caused the displacement and potentially the identity disagreement. The question is which coordinate and which identity best describe gene A? We will approach this problem in two cases described in next two sections for TriTryp genes:

a. Genes with unmatched Identity

b. End displacement between two genefinders

1.6.1 Genes with unmatched Identity

From **3595 genes** found by both gene finders there are 35 genes with mismatched identity by TSE and ARA. Table 10 shows the frequency of clusters which contains each of these ambiguities. From this table we see that for clusters of length > 2, which contain an ambiguity, have been always appeared with this ambiguity across different genomes.

Table 9: 35 genes with unmatch identity/anticodon reported by ARA and TSE. first row shows the ambiguities with the format <ara identity| tse identity>

Ambiguity	(D|I)	(L|?)	(L|E)	(L|M)	(N|Y)	(O|M)	(S|R)	(W|G)
Frequency	5	3	1	9	11	2	1	3

Table 10: Table of clusters and singletons which contain the Undetermined-indentity genes. column #similar shows number of times cluster sequence has appeared as subsequence of other clusters made from the non-ara-only genes from our initial gene set.

TSE Cluster	TSE Cluster Freq	Ambiguity	ARA Cluster	ARA Cluster Freq	arascore	tsescore	#ARAsimilar	#TSEsimilar
e	1	(L|E)	l	1	113	48	327	150
gV	2	(W|G)	wV	2	113	54,66	0	0
i	5	(D|I)	d	5	115,101	49,20⋆	103	161
LSMEMSMyV	1	(N|Y)	LSMEMSMnV	1	108	72	0	0
LSMEMyV	5	(N|Y)	LSMEMnV	5	108	72	0	0
m	2	(O|M)	o	2	112	47	0	74
m	9	(L|M)	l	9	112	50	327	74
rV	1	(S|R)	sV	1	113	43	7	3
?V	3	(L|?)	lV	3	113	43	1	215
Vg	1	(W|G)	Vw	1	113	54	0	2
Vy	1	(N|Y)	Vn	1	101	60	0	6
VyMEMSL	3	(N|Y)	VnMEMSL	3	108	72	0	0
yV	1	(N|Y)	nV	1	108	72	0	1

continues...

1.6.2 End displacement between two genefinders

continues...

2 SUMMERY OF FINAL GENE SET ANNOTATION

We made the final gene set data by removing genes with TSE score of less than 50 and genes with ARA score of less than 107. The cutoff score is set based on the ARA and TSE score distribution shown in figure 3. This cutoff score as shown in section 1.5 will keep the most trusted genes found by only one of the gene finders (Genes appeared within common clusters across genomes which also complement their clusters to look like other clusters). It will only remove 43 of our genes from intersection set from which 13 are pseudo and/or truncated genes, 10 genes have different identities by two genefinders and 20 other genes removed are shown in table12. .

Table 11: Summary of Gene Set. last three columns are each for one gene set. Sets are defined as: a) Intersection Of two genes finders. Two genes are labeled as same if their coordinate overlaps. Displacement of overlapped genes between ARA and TSE does not pass 4bp. b) Union of two gene finders. c) Genes found by only ARA. Genes marked as ?? include: pseudo|truncated genes, genes with unmatched identity, genes with unassigned identity|anticodon by any of genefinders, and genes with letter N in their sequence(we had 4 of these genes).

Sets	Intersection	ARAonly	Union
#tRNA	3554	36	3591
#N/#G	74	98	75
Gene Length	68-88	71-206	68-206
%intron	2	28	3
%G	32	33	32
%C	26	26	26
%T	23	23	23
%A	19	18	19
A	210	2	212
C	63	1	64
D	105	1	106
E	160	1	161
F	104	2	106
G	228	3	231
H	79	4	83
I	171	1	172
K	183	1	184
L	334	6	340
M	97	0	97
N	125	0	125
P	200	0	200
Q	161	0	161
R	348	2	350
S	227	7	234
T	218	4	222
V	235	0	235
W	52	1	53
X	76	0	76
Y	78	0	78
Z	71	0	71
??	29	0	30

Table 12: non-pseuso|truncated|unmatched-identity genes removed from the gene set after the applying score cutoff

Functional Class	?	A	C	E	G	H	I	L	M	Q	R	S	V	Y	Z
Frequency	3	1	1	1	1	1	3	2	2	1	1	1	6	1	5

2.1 Criteria for filtering out untrusted genes

For the purpose of creating CIFs, genes marked with "??" were removed and the remain were aligned using the alignment pipeline described in section 3. From GtRNAdb, we downloaded the last version of Homo sapiens High confidence tRNA sequences which include 430 genes with functional class distribution shown in table13 and length size distribution shown in table 14.

Table 13

Functional Class	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
Frequency	38	29	13	16	10	28	9	23	27	32	11	25	20	19	28	25	20	28	7	9	14	1

Table 14

Gene Length	70	71	72	73	74	75
Frequency	1	20	151	118	60	1

2.1.1 summery of final gene set

...
...

3. ALIGNMENT PIPELINE

...
...

4. CIF RESULTS

...
...

REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.