CONTENTS

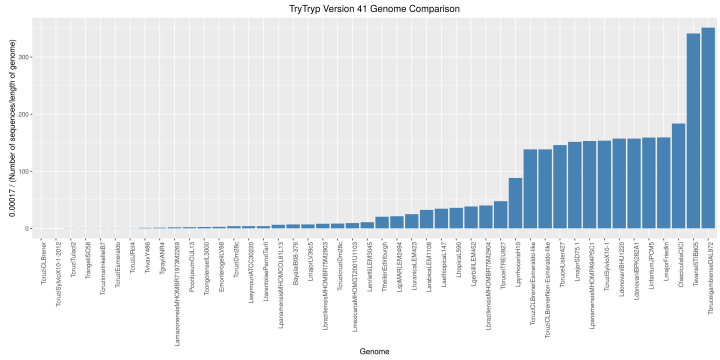

Figure 1: Comparing genomes based on formula $\left(\frac{f(x)}{\max(f(x))}\right)^{-1}$. f(x) = Number of sequences in genome devided by the length of genome. The length of the bars shows how good the genomes are sequenced.

PREPARING DATA

TryTryp genome data. 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 1. All the results, scripts for TryTryp version 41 can be found here.

tRNA gene annotation

tRNA gene prediction

In order to annotate tRNA genes for the sequenced TryTryp genomes, we used two computational methods for tRNA prediction, tRNAscan-SE and Aragorn. 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 we saved seperately to be analysed later.

Initiator tRNA prediction

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. 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 we clustered CAT tRNA 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.

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	0-6
Cluster2	95	95	2	0	95	95	8	0	0-8
Cluster3	2	2	2	0	0	0	8/9	0	0-22

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 intersection set, we dismissed genes which had different identity by ARA and TSE. for union set, we picked TSE identity over ARA. Table 2 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 2. We see from this figure that the coordinates of same genes annotated by ARA and TSE do not always match. We Analysed the reason for each set of displacementin this figure as fallow:

1. Genes with 0 displacement in both ends: These genes have same identity for both TSE and ARA except for 33 genes with Anticodon loop of more than 8bp. Also, both ARA and TSE reported genes up to base 73.
2. Genes with 0 displacement at 5 prime end and 1 displacement at 3 prime end: Identity of these genes matches between ARA and TSE. They all have anticodon loop of length 7. The reason for the displacement in 3 prime end is that ARA reports up to position 74, however, TSE reports only up to position 73.
3. Genes with one base displacement at both ends: In this case ARA reports one extra base at both ends which is because these two bases pair together and in most cases this makes the Amino Acid arm one base longer than what TSE reports. Although in a few cases, the AminoAcid arm will stay 7bp, but we see insertions in this arm.
4. Genes with two base displacement at 3 prime end and one at 5 prime end: In this case ARA is reporting 2 extra bases at one end and 1 for the other end which usually leads to a longer AminoAcid arm. Also, the extra 2 bases reported by ARA at 3 prime end are mostly 'ac'.
5. Genes with two base displacement at 3 prime end and 0 displacement at 5 prime end: In this case TSE reports up to position 73 as always,

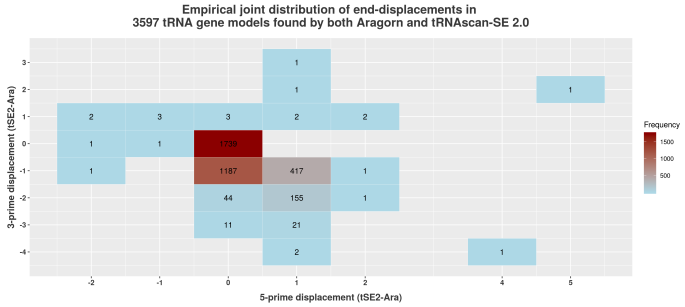

Figure 2: Empirical joint distribution of end-displacements in ? tRNA gene models found by both Aragorn and tRNAscan-SE 2.0.

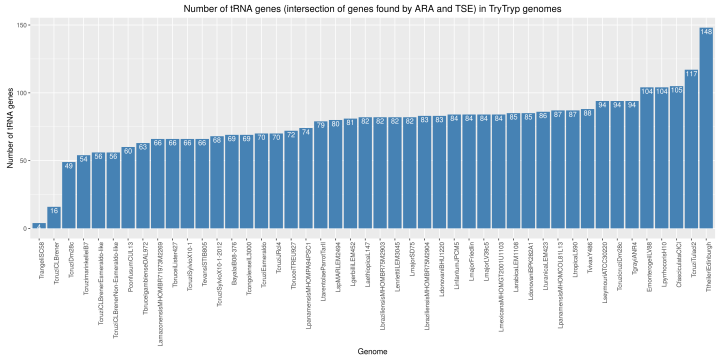

Figure 3: Number of genes annotated by both TSE and ARA for each TryTryp genome.

but ARA is reporting genes up to position 75. the last three bases of ARA are mainly 'acc'

6. Genes with three base displacement at 3 prime end and 0 displacement at 5 prime end. In this case ARA is reporting 3 extra bases at 3 prime end and these three bases are '?cc'.

We expect our genefinders to annotate tRNA genes of all 22 functional classes for each genome. To investigate this we visualized the number of genes annotated for each genome in Figure 3 and tRNA functional classes annotated by both TSE and ARA for each genome in Figure 4.

Table 2: 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
TSEs	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	50-215	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

from figure 3 we see that few of the 22 tRNA classes are not annotaded for all the genomes. To improve the annotation we included 33 gene annotated by both gene finders, with mismatched identity. you can see a summary of these genes in table 3. To determind the identity of these genes we built a structural alignment of all the TryTryp tRNA genes from our intersection set and these 33 genes (We used TSE reported sequences over ARA for alignment, because TSE reports genes up to position 73, however ARA can report afew bases after 73. We used removed introns and other nucleotides in non-conserved positions, and variable arms prior to the alignment. We used covea v2.4.2 (Sean Eddy 1994) for the structural alignment and edited the alignment by removing sites with more than 99% gap, genes with more

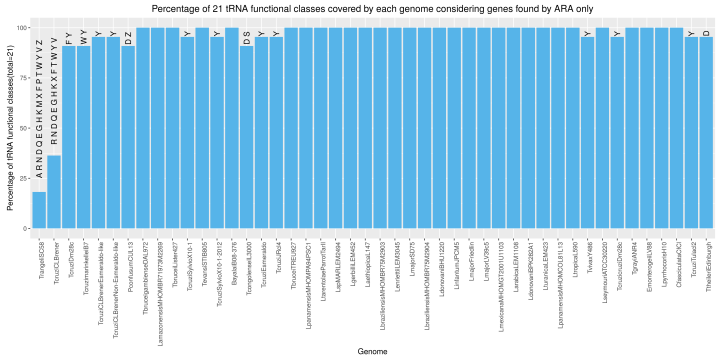

Figure 4: Percentage of 22 tRNA types annotated by both TSE and ARA for each TryTryp genomes. The label on top of each bar shows which tRNA classes are not annotated for the genome.

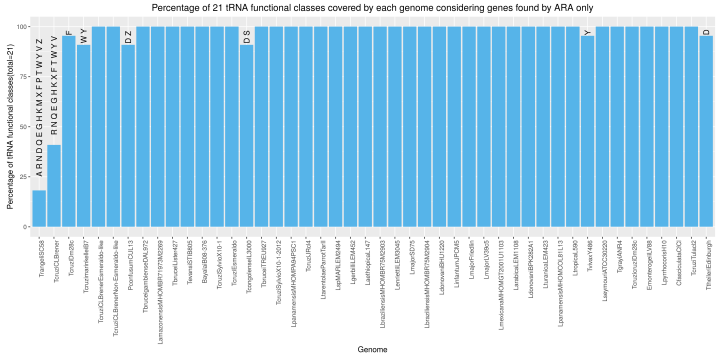

Figure 5: Percentage of 22 tRNA types annotated by both TSE and ARA for each TryTryp genomes after including 32 genes to intersection set. The label on top of each bar shows which tRNA classes are not annotated for the genome.

than 8 gaps in their aligned sequence, and genes with N in their sequence.). Later, using only the intersection aligned genes we made a profile covariance model for 22 functional classes. We calculated the score of 33 genes for each of these 22 models. For each gene, we compared the score of two models made for the indentities reported by TSE and ARA and picked the one with higher score. you can see the result in this file. We were able to include 32 of these genes to our intersection set which improved the annotation of genomes as you can see in figure 5

Table 3: Table of 33 genes annotated by both TSE and ARA with mismatched identity

ARA|TSE	D|I	L|?	L|E	L|M	N|Y	O|M	S|R	W|G
Number of genes	5	3	1	9	11	2	1	3

REFERENCES