# ARTICLE TITLE

FATEMEH HADI NEZHAD[1]

2019

## CONTENTS



* *Department of Quantitative System Biology, University of California, Merced, United States*

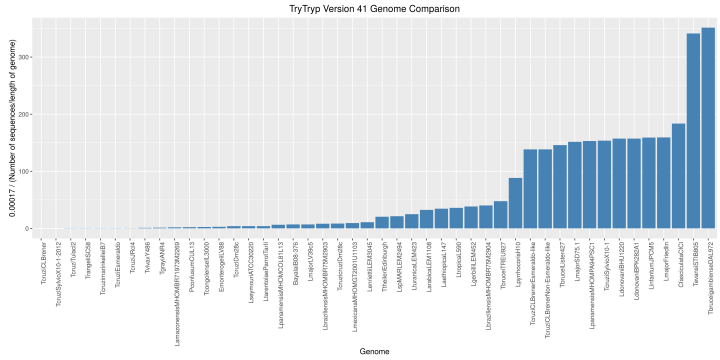


**Figure 1:** Comparing genomes based on formula $\left(\frac{f(x)}{\max(f(x))}\right)^{-1}$ which f(x) = Number of sequences in genome devided by the length of genome. the higher the bar is the better genome is sequenced.

## PREPARING DATA

**TryTryp genome data**. From tritrypdb website, we downloaded the version 41 of 46 TryTryp genomes released on 2018-12-05 (Script1). Genomes are compared based on number of sequence fragments relative to their length as shown in figure 1.

### tRNA gene annotation

*tRNA gene prediction*

In order to annotate the 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 one gene. However, the identity and exact coordinate of both genefinders we saved seperately to be compared 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 the fallowing citeria:

1. In all eukaryotic tDNA-iMet, positions 1124 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,

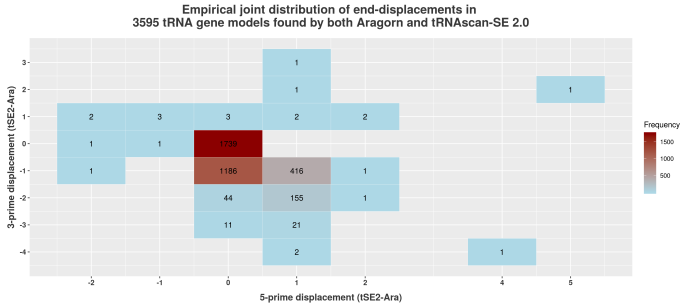


**Figure 2:** Empirical joint distribution of end-displacements in ? tRNA gene models found by both Aragorn and tRNAscan-SE 2.0.

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.

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

| Clusters | # tRNAs | 1124(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 ...

Figure 3 shows the number of genes annotated for each genome, and Figure 4 shows which tRNA functional classes were not annotated by both TSE and ARA for each genome.

**Table 2:** summary of the predicted genes by TSE and ARA. pseudo genes are marked 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 | 3609 | 370211 | 74.46 | 59-184 | 31.99 | 26.11 | 23.22 | 18.68 | 2.618 | 214 | 64 | 104 | 162 | 110 | 234 | 80 | 179 | 190 | 338 | 108 | 126 | 201 | 162 | 350 | 238 | 219 | 241 | 52 | 94 | 76 | 78 | 28 | 3 | 0 | 0 |
| ARA | 4345 | 372392 | 85.71 | 70-215 | 32.64 | 26.92 | 22.87 | 17.57 | 14.684 | 257 | 86 | 123 | 192 | 125 | 339 | 129 | 213 | 194 | 393 | 101 | 153 | 228 | 175 | 420 | 362 | 248 | 282 | 60 | 90 | 76 | 82 | 0 | 0 | 2 | 4 |
| UNION | 4379 | 377587 | 86.23 | 59-215 | 32.81 | 26.66 | 22.87 | 17.65 | 15.346 | 259 | 86 | 118 | 193 | 130 | 344 | 129 | 220 | 197 | 380 | 112 | 143 | 229 | 175 | 421 | 369 | 249 | 282 | 57 | 106 | 76 | 82 | 28 | 3 | 2 | 2 |
| INTERSECTION | 3560 | 265016 | 74.44 | 68-89 | 32.01 | 26.13 | 23.22 | 18.64 | 2.331 | 212 | 64 | 104 | 161 | 105 | 229 | 80 | 172 | 187 | 338 | 97 | 125 | 200 | 162 | 349 | 230 | 218 | 241 | 52 | 78 | 76 | 78 | 6 | 0 | 0 | 0 |

....

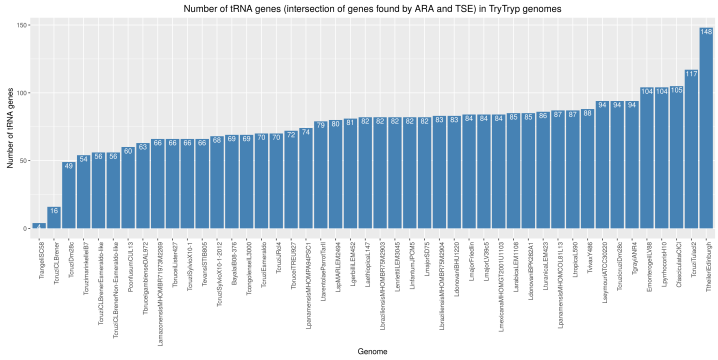


**Figure 3:** Number of genes annotated by both TSE and ARA for each TryTryp genome.

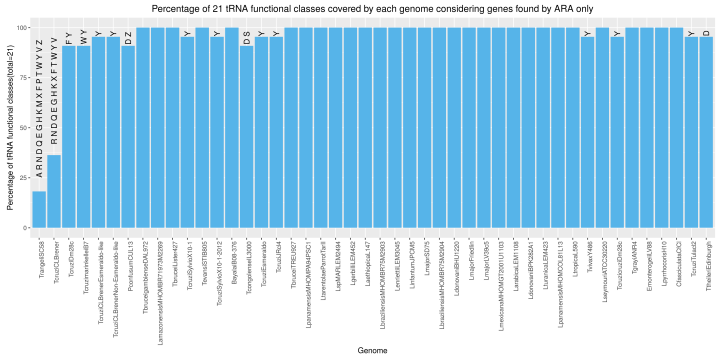


**Figure 4:** Percentage of 22 tRNA types annotated by both TSE and ARA for each TryTryp genomes. The labels on top of each bar shows which tRNA classes are not annotated for the genome.

## REFERENCES