Custom Annotation of tRNA Genes in TryTrypDB Genomes

We obtained 4381 unified gene records from the output of two tRNA gene-finders, Aragorn and tRNAscan-SE v.2.0, to TryTrypDB v.41. Of these, 3597 were found by both gene-finders, 750 were found by Aragorn only, and 34 were found by tRNAscan-SE 2.0 only. We identified the same 76 genes as initiator tRNA genes, using either tRNAscan-SE 2.0's profile-based predictions or our own edit-distance-based clustering approach, in which sequence variation at base-pairs at positions 1:72, 29:41, 30:40, and 31:39 identified a unique cluster of initiator tRNA-containing genes according to the rules described in (39).

To further annotate all genes, we identified tRNA gene clusters in TryTrypDB genomes using a maximum intergenic distance criterion of 1000 bp on either strand. Doubling this distance criterion did not substantially increase cluster number or size. We found that 77% of tRNA gene records occur in clusters of size two or greater. The largest clusters we identified were of size ten, accounting for 9% of gene. We used gene content and organization, including strandedness, to identify tRNA gene clusters with similar gene content, and could thereby identify putatively orthologous tRNA gene clusters conserved within each of the *Leishmania* and *Trypanosoma* genera, and sometimes both genera, with substantial evidence of evolution in gene organization through duplication, divergence, inversion and other changes (Supplementary Tables 1-3).

We used both gene cluster similarities across genomes and score statistics to finalize our gene set annotation. Using the similarities of tRNA gene clusters across genomes, we could find putative homologs for some of the 43 genes marked as pseudogenes or

truncated genes by tRNAscan-SE 2.0. We were also able to find putative homologs through similar clusters of genes detected only by Aragorn, which were detected by both gene-finders. With these results in mind, we plotted the densities of gene scores according to whether they were found by both gene-finders or only one (Fig. 2).

_____ added:

This section needs to be finalized. Based on this evidence, we retained genes that had either or both of an Aragorn bit-score of at least 107 and a tRNAscan-SE 2.0 bit-score of at least 50, retaining 3579 genes found by both gene-finders, 36 genes found by Aragorn only, and 1 gene found by only tRNAscan-SE (Table 1). These score cutoffs separated Aragorn-only genes within conserved gene clusters from singletons, which had lower scores (evidence not shown). We labeled 45 genes with identity # including 2 genes with unassigned identity by both genefinders, 6 genes marked as truncated or pseudo by tRNAscan-SE, 4 genes containing letter N, and 33 genes with different predicted structure and identity by genefinders (Table 9 from the Latex document shows ambiguous cases for 33 genes). Later, we were able to resolve the ambiguity of 10 genes predicted as Tyr by tRNAscan-SE and Asn by Aragorn. These 10 genes predicted for 10 genomes of T.Cruzi clade had same TSE structure and same ARA folding, with relatively high TSE score (72) and low ARA score (108). They have been marked as intron containing genes by both genefinders with different reported intron locations within the anticodon loop. tRNAscan-SE reported an intron of length 13 between sites 37 and 38, however Aragorn reported an intron of length 13 between sites 35-36. It has been shown in previous works (ref:

https://www.ncbi.nlm.nih.gov/pubmed/19450263) that tRNA-Tyr genes of T. cruzi contain introns of length 13 between positions 37-38. Further, we observed that these genes appeared in similar common clusters (all having the same ambiguity) of clade T.cruzi which lacked class Tyr after annotation. So, based on these evidence from genes' predicted structure and scores, and completeness of functional types of 9 genomes we annotated these 10 genes as Tyr.

Sets	Intersection	ARAonly	Union
#tRNA	3579	36	3616
#N/#G	74	98	75
Min Gene Length	68	71	68
Max Gene Length	89	206	206
%intron	2	28	3
%G	32	33	32
%C	26	26	26
A	210	2	212
C	64	1	65
D	105	1	106
E	160	1	161
F	104	2	106
G	228	3	231
H	80	4	84
I	171	1	172
K	183	1	184
L	335	6	341
M	97	0	97
N	125	0	125
P	200	0	200
Q	161	0	161
R	348	2	350
S	228	7	235
T	218	4	222
V	236	0	236
W	52	1	53
X	76	0	76
Y	88	0	88
Z	76	0	76
#	34	0	35

A breakdown of statistics of our finalized gene annotation by individual genome is available in Supplementary Table 4. The median number of genes per genome was 83. Two apparently incomplete genomes, belonging to *T. cruzi* Brener and *T. rangeli* SC58, were annotated with only six or 18 genes respectively, and so were excluded from further analysis. Table 1 below will be updated and replaced with a normal Table rather than a figure.

organism	A%	T%	C%	G%	GC%	Seq#	Gene#
TbruceigambienseDAL972	26	26	24	24	47	11	64
TevansiSTIB805	27	27	23	23	47	13	67
CfasciculataCfCl	21	22	29	28	57	31	105
TbruceiLister427	26	26	22	23	45	32	67
LpanamensisMHOMPA94PSC1	21	21	28	28	56	35	74
LdonovaniBHU1220	19	20	29	29	57	36	84
LdonovaniBPK282A1	19	20	29	29	57	36	85
LinfantumJPCM5	20	20	30	30	60	36	84
LmajorFriedlin	20	20	30	30	60	36	84
LmajorSD75.1	20	20	30	30	59	36	82
TcruziCLBrenerEsmeraldo-like	20	20	20	20	40	41	57
TcruziCLBrenerNon-Esmeraldo-like	21	21	22	22	43	41	57
TcruziSylvioX10-1	24	23	25	25	50	47	69
LpyrrhocorisH10	21	22	28	28	56	60	104
TbruceiTREU927	27	27	23	23	45	131	73
LbraziliensisMHOMBR75M2904	21	21	29	29	58	139	83
LgerbilliLEM452	20	20	30	29	59	142	81
LaethiopicaL147	19	20	30	29	59	160	83
LtropicaL590	19	19	29	28	57	160	87
LarabicaLEM1108	20	20	29	29	58	168	85
LturanicaLEM423	19	20	30	29	59	219	86
LspMARLEM2494	20	20	30	29	59	251	80
TtheileriEdinburgh	26	26	17	17	35	253	159
LenriettiiLEM3045	20	20	29	29	59	495	82
BayalaiB08-376	22	22	27	27	55	546	69
LmexicanaMHOMGT2001U1103	20	20	30	30	60	588	84
LbraziliensisMHOMBR75M2903	19	20	27	27	53	745	86
LmajorLV39c5	20	20	29	29	59	809	84
LpanamensisMHOMCOL81L13	21	21	29	28	57	856	88
TcruzicruziDm28c	24	24	26	26	52	1029	97
TcruziDm28c	25	25	26	25	50	1210	51
LseymouriATCC30220	22	22	28	28	55	1222	94
LtarentolaeParrotTarII	21	21	27	27	55	1351	79
EmonterogeiiLV88	23	23	26	26	52	1961	104
PconfusumCUL13	18	18	28	28	57	2188	61
LamazonensisMHOMBR71973M2269	20	20	30	30	59	2627	66
TcongolenseIL3000	21	21	20	20	40	2839	72
TgrayiANR4	23	23	27	27	54	2871	95
TrangeliSC58	24	23	27	26	53	7433	6
TvivaxY486	21	21	23	23	46	8290	82
TcruziJRcl4	24	23	26	24	50	15312	74
TcruziEsmeraldo	23	22	24	23	47	15803	74
TcruzimarinkelleiB7	22	22	23	23	45	16783	57
TcruziSylvioX10-1-2012	24	24	26	26	51	27019	72
TcruziCLBrener	23	23	27	27	53	29407	18
TcruziTulacl2	22	21	23	23	46	45711	121