Diaz-Miranda *et al,* Additional File 2 (top)

Additional File 2: PDF file containing supplementary figures S1-S6. Legends are included at the beginning of Additional file 2.

Supplementary Figure S1. Minimizing false discovery based on intron abundance (ISRPM) and excision efficiency (ISRPM/FPKM). For all 59,755 introns considered in this study, table cells represent the number of annotated (**A**) or unannotated (**B**) introns whose maximum abundance (in any sample) exceeds the specified levels for ISRPM (rows) and splicing efficiency ISRPM/FPKM (columns). Red lines indicate cut-offs selected to minimize false discovery (MaxISRPM ≥2.8, ISRPM/FPKM ≥1.8), based on panel (**C**), where table cells sum the estimated number of true positives (annotated introns above the cut-off) and true negatives (unannotated introns below the cut-off) from **A** & **B**.

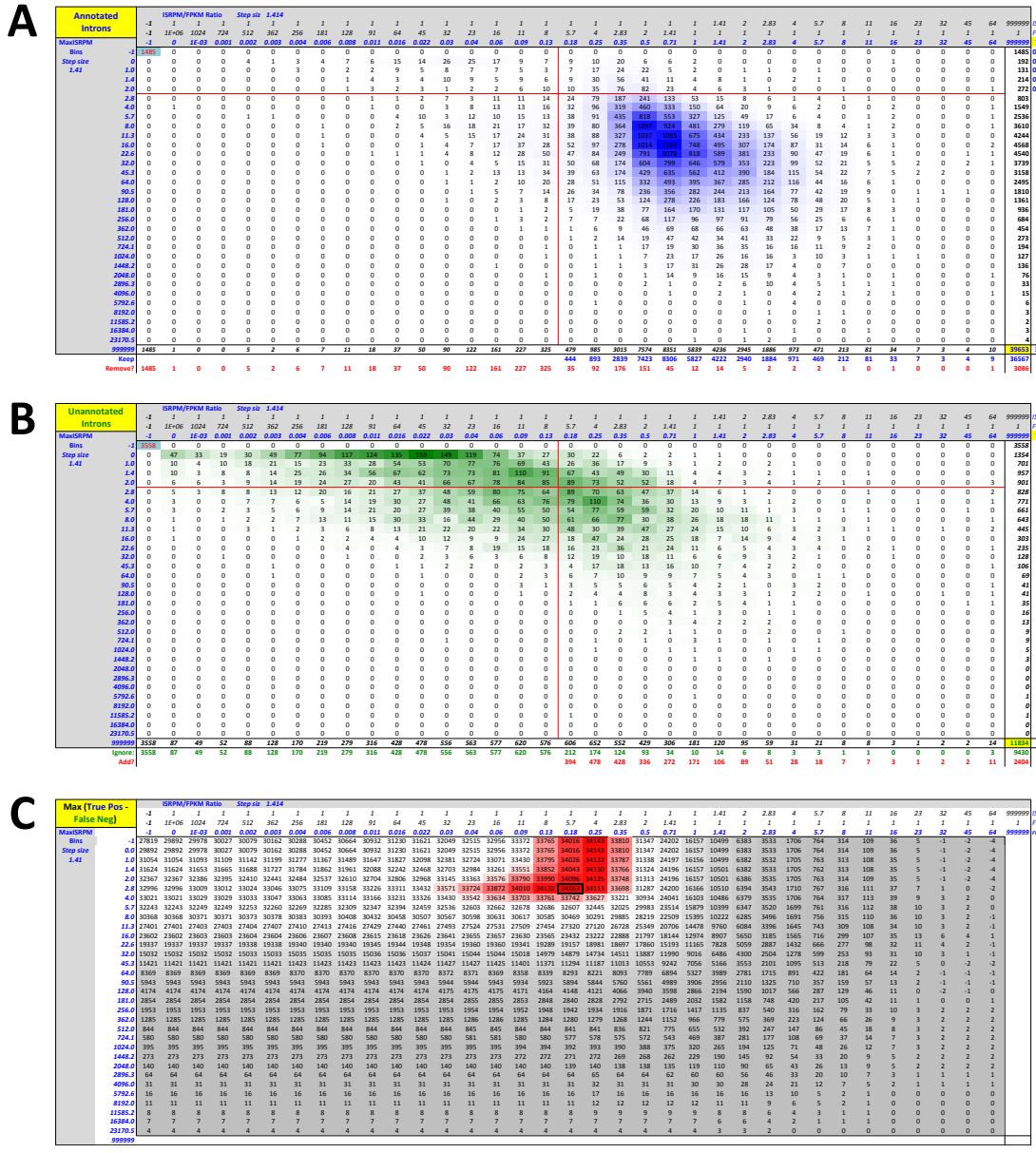
Supplementary Figure S2. *Tg*ME49_234450 coding potential and RPS15A multiple sequence alignment.

Supplementary Figure S3. Multiple sequence alignment for isocitrate dehydrogenase.

Supplementary Figure S4. *Tg*ME49_278830 coding potential and multiple sequence alignment.

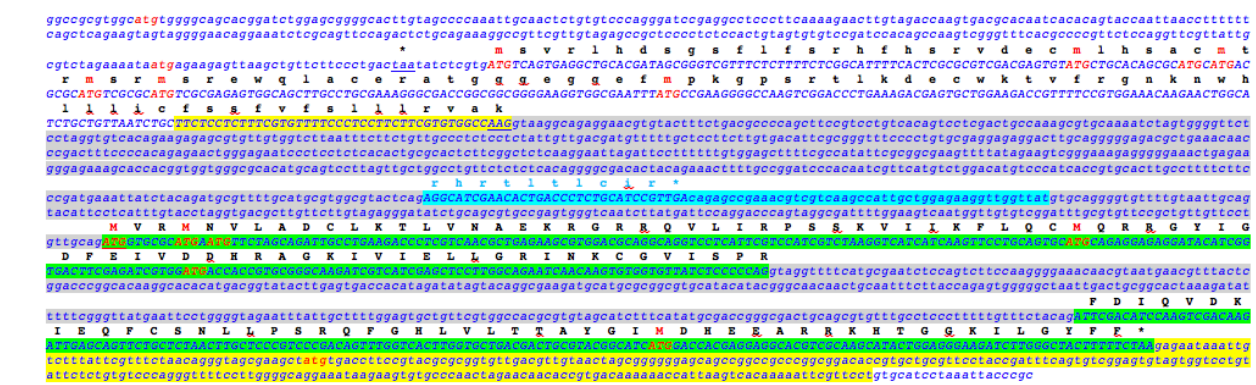
Supplementary Figure S5. Intron abundance (ISR & ISRPM) and excision efficiency (ISRPM/FPKM) for G6PDH introns. Intron abundance across the 20 samples selected for this study are presented for the annotated and the maximal alternative introns (unannotated) of the gene G6PDH. The ratio of unannotated to annotated ISRs is colored to indicate the samples in which the alternative unannotated intron is most highly expressed (darker red).

Diaz-Miranda *et al,* Supplementary Figure S1 (top)

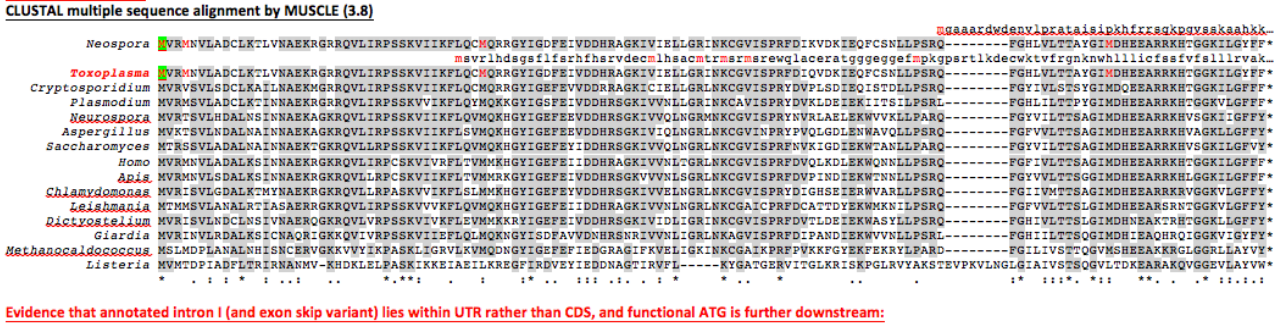


Supplemental Figure S1. Minimizing false discovery based on intron abundance (ISRPM) and excision efficiency (ISRPM/FPKM). For all 59,755 introns considered in this study, table cells represent the number of annotated (**A**) or unannotated (**B**) introns whose maximum abundance (in any sample) exceeds the specified levels for ISRPM (rows) and splicing efficiency ISRPM/FPKM (columns). Red lines indicate cut-offs selected to minimize false discovery (MaxISRPM ≥2.8, ISRPM/FPKM ≥1.8), based on panel (**C**), where table cells sum the estimated number of true positives (annotated introns above the cut-off) and true negatives (unannotated introns below the cut-off) from **A** & **B**.

Diaz-Miranda *et al,* Supplementary Figure S2 (top)



Blue italic nucleic acid sequence reflects current mRNA annotation; exons shaded white, introns shaded gray; UTRs in lower case, CDS in upper case.
Yellow shading indicates revised UTRs based on RNAseq; turquoise shading shows exon-retention variant.
Black amino acid text reflects current protein annotation; upper case indicates revised protein annotation based on alternative ATG (underlining indicates appropriate sequence context); turquoise AAs reflect premature termination in exon retention variant.



Evidence that annotated intron I (and exon skip variant) lies within UTR rather than CDS, and functional ATG is further downstream:

- *Annotated intron I is relatively inefficiently excised (ave ISRPM ~50% of the level observed for annotated intron II), and a cassette exon (exon skip variant) is also observed within this intron. If present in the coding sequence, this would yield a frame-shift mutation and premature termination. Note: this exon skip variant appears to be stage specific: ~30% in tachyzoites, ~60% in bradyzoites, 10% in gametocytes; likely attributable in part to alternative stage-specific promoters, including internal promoters that preclude excision..*
- *The annotated ATG, and all of the first nine(!) ATGs (including six in-frame) lack the appropriate sequence context for efficient translation & stability in T.gondii (aaaATG), but the 1st ATG downstream of the splice acceptor (precisely at predicted ATG) displays appropriate sequence context.*
- *Orthologs in other species and protein-level sequence alignment support the proposed ATG, permitting alignment of the entire 130 AA protein … but not the long (101 AA) N-terminal extension*
- *(which cannot be aligned even with closely-related Neospora parasites)*
- *Proteomics datasets provide abundant support for the body of the protein, but no evidence for translation of a long N-terminal extension.*

TgME49_234450 (Rps15a)		Exon I (long)	Exon I (short)	*Intron I*	Exon II	*Intron II*	*Intron I-II*	Exon III	*Intron III*	Exon IV					
	current annotation:	+	–	–	–	–	+	+	+	+					
CZ-H3 Tachyzoites (Hehl)	coverage depth (log2):	1	13	*8*	12	*8*		13	*5*	13					
	% maximum:	<1%	100%		50%			100%		100%					
	intron-spanning reads (per million):			*142.1*		*184.5*	*395.8*		*592.5*						
	excision efficiency (ISRPM/FPKM):			*28%*		*36%*	*78%*		*117%*						
	% maximum:			*24%*		*31%*	*67%*		*100%*						
Z-H3 Gametocytes, d7 *in vivo* (Hehl)	coverage depth (log2):	5	15	*8*	13	*8*		16	*7*	16					
	% maximum:	<1%	50%		12.5%			100%		100%					
	intron-spanning reads (per million):			*97.6*		*157.3*	*912.3*		*1401.6*						
	excision efficiency (ISRPM/FPKM):			*8%*		*13%*	*75%*		*115%*						
	% maximum:			*7%*		*11%*	*65%*		*100%*						
M4 Bradyzoites, In vivo (Buchholz)	coverage depth (log2):	*0*	*11*	*6*	11	*6*		12	*2*	12					
	% maximum:	<1%	50%		50%			100%		100%					
	intron-spanning reads (per million):			*206.5*		*268.5*	*147.9*		*805.9*						
	excision efficiency (ISRPM/FPKM):			*32%*		*42%*	*23%*		*126%*						
	% maximum:			*26%*		*33%*	*18%*		*100%*						
	revised annotation:	–	+	–	–	–	+	+	+	+					

TgME49_200320 (HXGPRT)		Exon I (long)	Exon I (short)	*Intron I*	Exon II	*Intron II*	Exon III	*Intron III*	*Intron II-III*	Exon IV	Intron IV	Exon V	Intron V	Exon VI (sho	Exon VI (long
	current annotation:														
Tachyzoites (Hehl)	coverage depth (log2):	3	9.5	*6*	9.5	*4*	8	*4*		9.5	2	9.5	2	9.5	5
	% maximum:	<1%	100%		100%		33%			100%		100%		100%	4%
	intron-spanning reads (per million):			*48.86*		*9.88*		*14.31*	*30.84*		40.21		51.46		
	excision efficiency (ISRPM/FPKM):			*75%*		*16%*		*23%*	*49%*		64%		82%		
	% maximum:			*91%*		*19%*		*28%*	*60%*		78%		100%		
d7 Gametocytes (Hehl)	coverage depth (log2):	3	15	*11*	15	*9*	13	*10*		15	7	15	7	15	8
	% maximum:	<1%	100%		100%		25%			100%		100%		100%	
	intron-spanning reads (per million):			*938.56*		*161.28*		*196.31*	*566.33*		815.66		912.7		
	excision efficiency (ISRPM/FPKM):			*74%*		*13%*		*15%*	*44%*		64%		72%		
	% maximum:			*100%*		*17%*		*21%*	*60%*		87%		97%		
In vivo Bradyzoites (Buchholz)	coverage depth (log2):	3	8	*6*	8	*0*	5	*4*		7	0	7	0	8	2
	% maximum:	<1%	100%		100%		12.5%			50%		50%		100%	
	intron-spanning reads (per million):			*16.24*		*0*		*2.65*	*9.95*		31.16		12.93		
	excision efficiency (ISRPM/FPKM):			*39%*		*0%*		*6%*	*24%*		75%		31%		
	% maximum:			*52%*		*0%*		*9%*	*32%*		10000%		42%		
	revised annotation:														

Diaz-Miranda *et al,* Supplementary Figure S3 (top)

TgME49_266760 (Isocitrate Dehydrogenase) -- OG5_127057

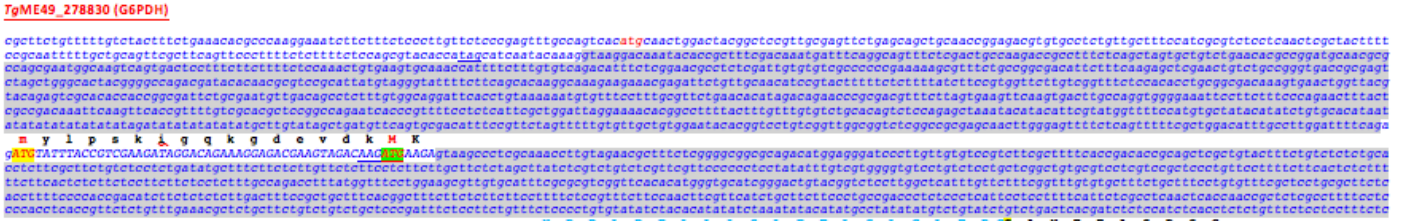
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Evidence that annotated intron I lies within UTR rather than CDS, and functional ATG is further downstream:

- *Annotated intron I is largely absent in gametocytes, which use a stronger downstream promoter that lies within this annotated intron and would preclude translation from the annotated ATG*
- *Orthologs in other species and protein-level sequence alignment support the proposed ATG, but not the long N-terminal extension or upstream ATG usage, and proteomics datasets provide abundant support for the body of the protein, but no evidence for translation of a long N-terminal extension.*

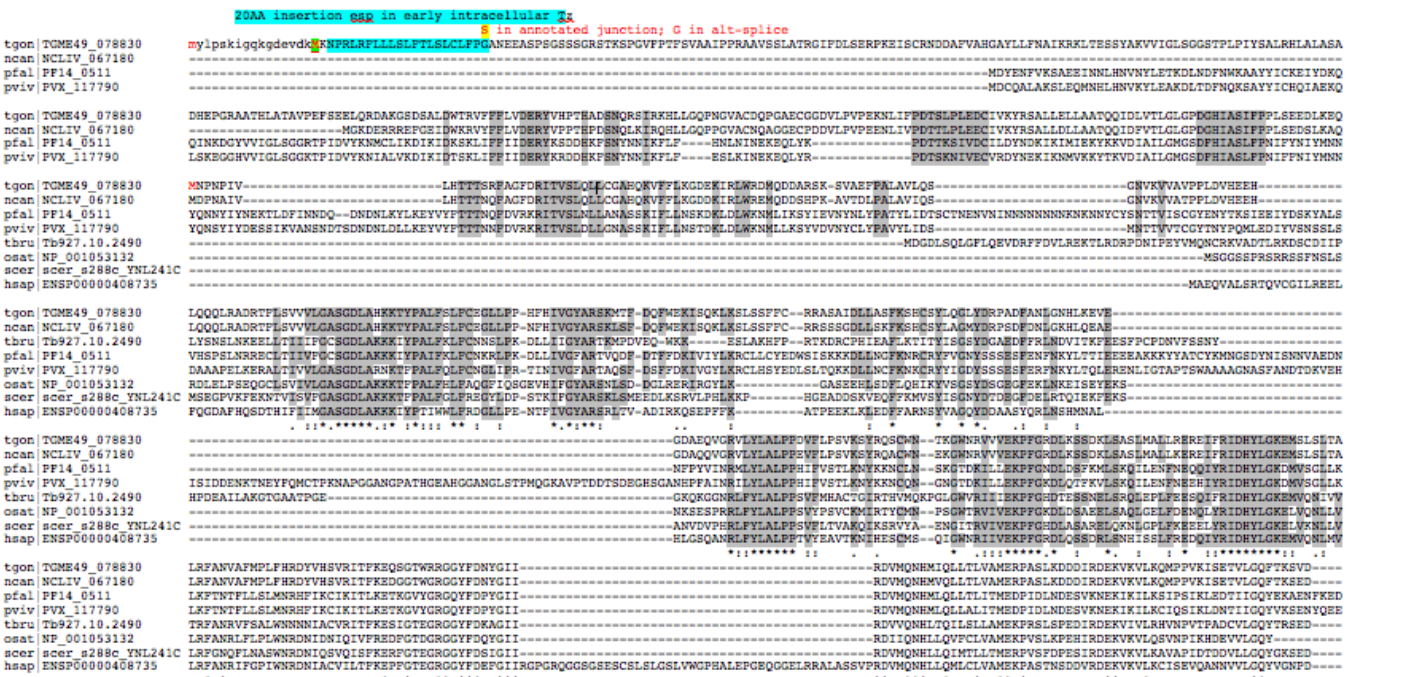
Diaz-Miranda *et al,* Supplementary Figure S4 (top)

*Tg*ME49_278830 (G6PDH)



OG5_127121

CLUSTAL multiple sequence alignment by MUSCLE (3.8)



Evidence that the stage-specific splice acceptor variant is part of the CDS rather than UTR

- *Based on sequence context, second ATG is more plausible than the first (annotated) ATG, but this lies in-frame, and still upstream of the additional sequence introduced by the splice acceptor variant. While this region is not conserved in other apicomplexan parasites, no other ATGs are found upstream of conserved protein sequences.*
- *All apicomplexans share a long N-terminal extension relative to other species, and proteomics data show ample evidence for expression of this protein.*

Diaz-Miranda *et al,* Supplementary Figure S5 (top)

TgME49_278830 (G6PDH) intron 2		Annotated: TGME49_chrXII:5834356..5835329					Unannotated (Alternative): TGME49_chrXII:5834356..5835269					
Experiment	Sample	Unique ISRs	ISRPM	Nonunique	ISRPM/FPKM	% Sample	Unique ISRs	ISRPM	Nonunique	ISRPM/FPKM	% Sample	Unann/Ann
Bradyzoite *in vitro* Transcriptome (ME49)	bradyzoite	79	44.43	0	355.8674	8.0	58	32.62	0	261.2697	5.9	0.73
Bradyzoite *in vivo* transcriptome (M4)	bradyzoite	94	31.16	0	0.7113	23.3	5	1.66	0	0.0378	1.2	0.05
Oocyst Time Series (M4)	day 0	11	14.91	0	3.9817	100.0						0.00
	day 4	362	290.43	0	320.1447	19.4	1	0.80	0	0.8844	0.1	0.00
	day 10	527	444.63	0	503.7622	33.1	12	10.12	0	11.4709	0.8	0.02
Tachyzoite Transcriptome 3 and 4 days post-infection (VEG NcLIV)	day 3	91	43.38	0	0.3115	30.2	34	16.21	1	0.1164	11.3	0.37
	day 4	65	25.40	0	0.1987	19.6	33	12.89	1	0.1009	10.0	0.51
Tachyzoite Transcriptome Time Series (GT1)	hour 2						1	24.66	0	0.2527	10.0	
	hour 4	1	74.37	0	0.7019	20.0						
	hour 8	3	44.09	0	0.5758	13.0						
	hour 16	56	18.60	0	0.2019	5.9	12	3.99	0	0.0433	1.3	0.21
Tachyzoite Transcriptome Time Series (ME49)	hour 2	1	4.50	0	0.0598	2.1	6	26.98	0	0.3589	12.5	6.00
	hour 4	5	12.59	0	0.1405	5.7	12	30.22	0	0.3371	13.8	2.40
	hour 8						2	5.07	0	0.0726	4.9	
	hour 16	5	15.48	0	0.1952	9.1	2	6.19	0	0.0781	3.6	0.40
	hour 36	37	26.35	0	0.1992	4.3	34	24.21	0	0.1830	3.9	0.92
	hour 44	32	23.44	0	0.1368	2.8	61	44.69	0	0.2609	5.3	1.91
Tachyzoite Transcriptome Time Series (RH)	hour 2						2	95.00	0	0.4833	11.1	
	hour 22	23	17.83	0	0.0888	2.1	10	7.75	0	0.0386	0.9	0.43
	hour 36	44	24.87	0	0.1046	2.4	51	28.83	0	0.1212	2.8	1.16
Tachyzoite Transcriptome Time Series (VEG)	hour 2	14	16.77	0	0.1389	5.1	26	31.15	0	0.2580	9.4	1.86
	hour 4	10	19.55	0	0.1659	7.0	13	25.42	0	0.2156	9.2	1.30
	hour 8	4	6.42	0	0.0633	2.9	6	9.63	0	0.0950	4.3	1.50
	hour 16	7	13.17	0	0.1146	5.7	10	18.81	0	0.1637	8.1	1.43
	hour 36	123	75.04	0	0.4352	9.0	58	35.39	0	0.2052	4.3	0.47
	hour 44	91	81.75	0	0.4922	12.3	26	23.36	0	0.1406	3.5	0.29
Transcriptome during acute or chronic infection in mouse brain	chronic infection 2	2	24.37	0	0.4041	50.0						0.00
Transcriptomes of 29 strains during murine macrophage infection	ARI infected	20	14.39	0	0.1950	19.8	12	8.63	0	0.1170	11.9	0.60
	B41 infected	6	16.81	0	0.2514	19.4	2	5.60	0	0.0838	6.5	0.33
	B73 infected	15	47.07	0	0.5829	51.7	1	3.14	0	0.0389	3.4	0.07
	BOF infected	7	16.73	0	0.2393	20.0	3	7.17	0	0.1026	8.6	0.43
	CAST infected	9	15.77	0	0.2740	25.7	1	1.75	0	0.0304	2.9	0.11
	CASTELLS infected	5	9.61	0	0.1582	15.6						0.00
	CEP delta HXGPRT	12	23.90	0	0.2915	30.8	3	5.98	0	0.0729	7.7	0.25
	COUGAR infected	15	34.42	0	0.3643	30.6	1	2.29	0	0.0243	2.0	0.07
	DEG infected	12	18.78	0	0.3205	30.0	3	4.69	0	0.0801	7.5	0.25
	FOU infected	4	17.63	0	0.3622	26.7						
							1	4.96	0	0.0935	8.3	
	GT1 infected	7	9.02	0	0.1553	14.6	3	3.87	0	0.0665	6.3	0.43
	GUYDOS infected	7	14.09	0	0.1230	7.1	1	2.01	0	0.0176	1.0	0.14
	GUYKOE infected	8	7.90	0	0.0842	7.8	6	5.93	0	0.0632	5.8	0.75
	GUYMAT infected	5	11.68	0	0.1543	12.5						0.00
	MAS infected	7	15.57	0	0.2186	21.2	4	8.90	0	0.1249	12.1	0.57
	ME49 infected	18	36.53	0	0.4548	32.7	2	4.06	0	0.0505	3.6	0.11
	P89 infected	1	1.25	0	0.0236	2.6	3	3.74	0	0.0707	7.9	3.00
	PRU delta HXGPRT	20	21.19	0	0.2149	22.2	7	7.42	0	0.0752	7.8	0.35
	RAY infected	13	27.01	0	0.3716	35.1	1	2.08	0	0.0286	2.7	0.08
	RH delta HXGPRT i	16	11.49	0	0.1713	13.0	1	0.72	0	0.0107	0.8	0.06
	ROD infected	18	16.00	0	0.2013	19.4	4	3.56	0	0.0447	4.3	0.22
	RUB infected	1	1.21	0	0.0312	3.1	4	4.86	0	0.1249	12.5	4.00
	TgCATBr44 infecte	1	1.20	0	0.0250	2.9	1	1.20	0	0.0250	2.9	1.00
	TgCATBr5 infected	7	25.45	0	0.3037	28.0						0.00
	TgCATBr9 infected	17	14.58	0	0.3265	37.8	4	3.43	0	0.0768	8.9	0.24
	VAND infected	6	4.11	0	0.0779	8.2	1	0.69	0	0.0130	1.4	0.17
	VEG infected	8	17.27	0	0.2188	24.2	4	8.63	0	0.1094	12.1	0.50
	WTD3 infected	13	17.30	0	0.2165	22.8	6	7.98	0	0.0999	10.5	0.46
Transcriptomes of Cat Enteroepithelial Stages (CZ-H3)	Tachyzoite	172	29.31	0	0.2524	27.6	168	28.63	0	0.2465	27.0	0.98
	day 3	21	11.48	0	0.4996	56.8	1	0.55	0	0.0238	2.7	0.05
	day 5	217	17.64	1	0.4330	56.8	2	0.16	0	0.0040	0.5	0.01
	day 7	310	15.72	0	0.4282	56.1	1	0.05	0	0.0014	0.2	0.00