Genome-Wide Evolutionary Analysis of the Non-Coding RNA Genes and Non-Coding DNA of *Paramecium tetraurelia*

Chun-Long CHEN[1, 2, 3, 4], Hui ZHOU[3], Jian-You LIAO[3]

Liang-Hu QU[3] and Laurence AMAR[1, 2]

[1] Université Paris Sud, Orsay 91405, France.

[2] CNRS, UMR 8080, Orsay 91405, France.

[3] Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Zhongshan University, Guangzhou, 510275, People's Republic of China.

[4] Present address: CNRS, UPR 2167, CGM, Gif sur Yvette, 91198, France

Corresponding authors: Laurence AMAR and Liang-Hu QU

Mail : laurence.amar@u-psud.fr, lssqlh@mail.sysu.edu.cn

Supplemental Data

NcRNA gene identification.

SnoRNA genes.

Profile-based programs methods identified 35 box H/ACA snoRNA and 89 box C/D snoRNA putative genes with guiding antisequences complementary to snRNA and/or rRNA sequences. The computationally identified snoRNA genes grouped into 19 box H/ACA and 41 box C/D OHSets, 13 (70%) and 38 (95%) of which respectively had ≥2 members.

Experimental screenings of box C/D snoRNAs were performed using size selection and box C/D-anchored primers, identifying 26 of the 38 (68%) OHSets of size ≥2, and 1 of the 3 OHSets of size 1. Stage-specific transcription patterns or low transcript levels might have accounted for the lack of experimental identification of the 12 OHSets with sizes ≥2 (60% of the experimentally identified snoRNA species were represented by ⩽2 clones, see Table SI). In addition, experimental screenings identified 6 non-predicted OHSets of size ≥2 and 1 of size 1, the members of which displayed rRNA or snRNA antisequences. Computation analysis of box C/D snoRNA genes therefore predicted 27 out of the 34 (80%) experimentally recovered rRNA- or snRNA-guiding box C/D OHSets. Experimental screenings also identified 1 orphan OHSet of size ≥2 and 3 of size 1, the members of which displayed 3-4-nt stems, boxes C and D but no rRNA or snRNA antisequences (these could not be predicted, see Methods).

Computational, comparative genomics and experimental approaches altogether identified 35 box H/ACA and 107 box C/D snoRNA genes that respectively grouped into 19 and 52 OHSets (Table I). Within the 58 OHSets with sizes ≥2, the ohnologous genes harbor identical guiding sequences in 47 OHSets, and mutations of C→T or A→G that allow identical pairing in 9 OHSets. Members of OHSet PtCD03 display one base difference, three positions upstream of box D (Figure S6). Mismatches at this position have already been characterized within functional rRNA:snoRNA duplexes (Chen et al.,

2007). Members of OHSets PtCD06 harbor one base difference, four positions upstream of box D. No functional snoRNA-rRNA duplex with mismatched bases at this position could be characterized in a survey of 415 duplexes (Chen et al., 2007).

snRNA genes and Pol II signature.

We identified 4U1, 4 U2, 2 U4, 3 U5 and 3 U6 snRNA putative genes (See Table I). Analysis of the DNA ohnologous blocks for local duplication events and genome screening for snRNA-related sequences identified a U6-related snRNA sequence with a deletion of 11 bp in the region that was complementary to the U4 snRNA. This sequence was part of the OHSet that also included the 3 U6 snRNA genes.

An Upstream Sequence Element (USE) with the consensus sequence 5'-AAACCCATAA(A/T)A-3' and a TATA-like box flanked these genes (Figure 2A & B). A distance of 18 bp separated the USE and the TATA-like box that flanked the U1, U2, U4 and U5 snRNA genes while distances of 21-22 bp separated those flanking the U6 snRNA genes. In metazoan and plants, transcription studies demonstrated that the former were transcribed by Pol II and the U6 snRNA genes, mainly by Pol III. The U1, U2, U4 and U5 genes *of P. tetraurelia* were also flanked on their 3' sides by the conserved Downstream Element (5'- TTAAACAAACGAAAT-3') and on their 5' sides, by an Initiation Element (IE) consisting of a pyrimidine trinucleotide (IE). The USE, TATA-like box, IE and DE, respectively separated by 18 bp, 22-26 bp, and 50-300 bp, were used as Pol II signature to screen the *P. tetraurelia* genome.

SRP RNA genes, telomerase RNA gene and Pol III signature.

We identified 2 SRP RNA putative genes that grouped into 1 OHSet. A related sequence mapped 3 kb distant from one of the former (see below). Our bioinformatics screening of the genome identified the previously characterized telomerase RNA sequence but failed to identify any telomerase RNA-related sequence (McCormick-Graham & Romero, 1996).

In most if not all organisms, the SRP RNA, U6 snRNA and telomerase RNA genes are transcribed by Pol III. In the *P. tetraurelia* genome, the 2 SRP RNA genes, the telomerase RNA gene and the 3 U6 snRNA genes displayed a Pol III signature, being flanked on their 3' ends with a polythymidine stretch of 5-6bp, and on their 5' ends with a USE (consensus sequence 5'-AAA/TCCCATAAA(A/T)-3') and a TATA-like box separated by 21-22 bp. The SRP-related sequence lacked any flanking consensus motif while the U6-related snRNA sequence displayed mutations in the flanking TATA-like box and polythymidine stretch suggesting they were non-functional. The USE, TATA-like box and polythymidine stretch, respectively separated by 21-22 bp and 50-300 bp were used as Pol III signature to screen the *P. tetraurelia* genome.

5S RNA genes.

The *P. tetraurelia* genome harbors 26 5S RNA genes that grouped into 2, 2, 3 and 3 OHSets respectively displaying 1, 3, 2 and 4 members.

tRNA genes and codon usage

We identified 216 tRNA putative genes that grouped into 106 OHSets, 62 of which had sizes ≥2. The programs that we used also identified 13 tRNA-related sequences that were expected non-functional (see Methods). Sparring strategies allow an organism to read the genetic code information with a limited repertoire of 46 anticodons. The 216 tRNA genes harbored 44 anticodons complementary to the 61 universal sense codons and 2 anticodons for the specific decoding of the 5'-TAG-3' and 5'-TAA-3' codons (see below).

The *P. tetraurelia* anticodon repertoire harbored a single guanine-ending anticodon for pairing with the pyrimidine-ending codons encoding the Asp, Asn, Cys, His, Phe, Ser and Tyr residues (Figure S7). There were no adenine-ending anticodons for the decoding of the Asn, His, Phe and Tyr

residues, and the 5'-ATC-3' and 5'-ACA-3' anticodons are likely to define tRNA potential pseudogenes rather than Asp and Cys residues, respectively. If not, A and its modified nucleoside I that is able to pair with A would lead to mistranslation (Osawa et al., 1992). As in most eukaryote genomes, two tRNAs with pyrimidine-ending anticodons paired with the two purine-ending codons encoding the Arg, Glu, Gln, Leu or Lys residues, the three 5'-(A/G/T)AT-3' anticodons paired with the three 5'-AT(T/C/A)-3' codons encoding the Ile residue, the four 5'-(A/C/G/T)CC-3' anticodons paired with the four synonymous codons encoding the Gly residue while three 5'-(A/C/T)XX-3' anticodons (with X being A, T, G or C depending on the decoded codon) paired with the four synonymous codons encoding the Ala Arg, Leu, Pro, Ser, Thr or Val residues. Finally the *P. tetraurelia* tRNA genes also harbored two 5'-(C/T)TA-3' anticodons that paired with the 5'-TA(A/G)-3' codons encoding a Glu residue instead of stopping translation.

tRNA genes could be grouped into functional sets of 6-20 members accordingly to the anticodon sequence (Supplemental Figure S7). Within each functional set, gene number correlated with *P. tetraurelia* codon usage that we calculated by using the whole proteome. The thymidine-ending codons displayed higher frequencies than the cytosine-ending codons for the Asp, Asn, Cys, His, Ser and Tyr residues in the *P. tetraurelia* genome as in other eukaryote genomes. However the Phe-encoding codons provided an exception with an over-abundance of the 5'-TTC-codons relatively to the 5'-TTT-3' codons.

Conserved sequence identification.

Sequence conservation was analyzed with CDS coverage values of 55%, 65% or 75% (see Figure S2). With a CDS coverage of 55%, about 13%, 19%, 31% and 31% of the macronuclear genome were conserved between the DNA ohnologous blocks arisen from the AD, OWG, IWGD and RWGD, respectively. With a CDS coverage of 75%, about 18%, 28%, 44% and 41% of the macronuclear genome were conserved between the same DNA ohnologous blocks. Finally, with the retained CDS coverage of 65%, about 23%, 37% and 36% of the macronuclear genome were conserved between the

OWG, IWGD and RWGD DNA ohnologous blocks. The three genomic subsets are mainly included into each other, but not totally, so that altogether the conserved sequences identified in that case account for 47% of the *P. tetraurelia* genome. While the percentages of the conserved sequences between the DNA ohnologous blocks displayed some variation depending upon the coverage values, the relative abundance of the different genetic elements were only slightly sensitive: conserved sequences distributed into 0.1% of RNA gene sequences, 3.5-4.6% of intron sequences, 4.9-10.9% of intergenic DNA and 84.4% to 89.5% of CDS in the four genomic sets. CDS coverage of 65% drew out intergenic DNA, intron and ncRNA gene coverages of 20%, 71% and 88%, respectively (Figure S2) The high intron coverage was a consequence of the short intron size of 25 bp on average and minimal conservation length of 100 bp with which the Phastcons program worked. This led introns to be enclosed within larger fragments.

Conserved structure identification.

Using higher cutoffs, less candidates with a higher specificity (a parameter that estimates the percent of the true candidates) and a lower sensitivity (the percent of snRNA and tRNA recovered amongst those present in the different subsets) were recovered. The RNAz program was also run onto conserved sequences of the OWGD genomic subset and two additional candidates were recovered. The relative paucity in recovering RNA structures from this subset did not appear to be due to the incapacity of the RNAz program to run on sets of >6 sequences. Although OWGD subsets may theoretically harbor up to 8 homologous sequences, none of the ncDNA subsets did it. The relative paucity in recovering RNA candidates from the OWGD genomic subset should rather reflect a lack of such sequences in the *P. tetraurelia* genome.

Supplemental Table Legend

Table SI. ncRNA putative genes and evolutionary conserved RNA putative secondary structures. The 5S RNA, snRNA, snoRNA, SRP RNA, telomerase RNA and tRNA putatives genes are listed with their genome positions on the corresponding pages. The singleton or cluster organization, the number of experimental clones recovered, the targets and the existence of orthologous targets within mammals (M), plants (P) or yeasts (Y) are indicated in the case of the snoRNA genes. The structures identified by the RNAz program, adjacent to protein-coding genes within putative or experimentally characterized UTRs, and those mapping further apart are listed with their genome positions on pages RNAz-1 and RNAz-2, respectively.

Supplemental Figure Legends

Figure S1. Sequence identity. Identity of aligned (left) and conserved (right) sequences evolved from recent duplicates

Figure S2. Identification of conserved sequences. (A) Aligned and **(B**) conserved sequences evolved from recent duplicates were split into CDSs, introns, ncRNA genes and intergenic DNA. While the aligned sequences distributed into 0.06% of ncRNA genes, 3.7% of introns, 17.5% of intergenic DNA and 78.8% of CDSs, the Phastcons sequences distributed into 0.08% of ncRNA genes, 4.6% of introns, 6.4% of intergenic DNA and 89.3% of CDSs. **(C)** Coverage of 65% for CDSs drew out coverage of 72% for intron sequences, 88% for ncRNA gene sequences and 28% for intergenic DNA sequences.

Figure S3. Sequences flanking ncRNA genes of *Tetrahymena thermophila*. (A) Pol II-transcribed genes**.** USEs, TATA-boxes and IEs map upstream of the U1, U2, U4 and U5 snRNA gene 5' regions, DEs map downstream of the gene 3' regions. The USE, TATA-box, IE and DE consensus sequences and inter-element distances are shown. **(B)** Pol III-transcribed genes. USEs and TATA-boxes map upstream of the U6 snRNA, SRP RNA and telomerase RNA gene 5' regions, polythymidine stretches map at the gene 3' regions. The USE, TATA-box and polythymidine stretch consensus sequences and inter-element distances are shown.

Figure S4: ncRNA gene characterization. Graphic representation of the gene candidates recovered with different profile-based programs**. (A)** snoRNA genes, (B) tRNA genes.

Figure S5. Screen shot of conservation tracks. The *P. tetraurelia* genome has three ohnologous genes encoding U6 snRNAs, one of which maps on scaffold 52. Conservation tracks were established

for ohnologous sequences arisen from the RWGD (upper track), and for ohnologous sequences arisen from the RWGD and IWGD (lower track). Red arrows point to sequence conservation of USEs (see Figure 2). Upstream of the USE of the U6_snRNA.1 gene lies one protein-coding gene (GSPATG00017346001) with three paralogs in the genome. The lack of any sequence conservation downstream of the U6_snRNA.1 gene indicates that deletions occurred within the ohnologous regions.

Figure S6. SnoRNA gene OHSets. Box H/ACA snoRNA genes grouped into 19 OHSets and box C/D snoRNA genes, into 52 OHSets. 18 box C/D snoRNA OHSets were further organized into clusters. Clusters 1-1 to 1-4 are inserted within scaffolds 91, 92, 134 or 15. **(A)** Cluster organization. Clusters 1_1 and 1_2 harbors 8 snoRNA genes, cluster 1_3, 6 genes and "cluster" 1_4, 1 gene. Gene sizes are indicated (upper lanes), as well as distances separating boxes C and D of adjacent genes (lower lanes). A TATA box and a USE lie upstream of ohnologs PtCD01_1, PtCD01_2 and PtCD01_3, and a polythymidine stretch, immediately downstream of ohnologs PtCD06_1, PtCD06_2 and PtCD06_3. Conserved nucleotide motifs are indicated by hatched boxes. Corresponding sequences are aligned below. "Cluster 1_4" only harbored one gene and a polythymidine stretch. **(B)** Cluster conservation. SnoRNA sequences that form duplexes with RNA sequences for the guiding of nucleotide modifications are shaded. Boxes C and D are boxed. Positions identical between ohnologs are marked by stars.

Figure S7. tRNA genes. The 216 tRNA genes grouped into 50 functional sets of 6-20 members accordingly to the anticodon sequences. Codon and anticodon frequencies are indicated so that the codon or anticodon sums for a given amino-acid are 100%. Codon usage frequency was calculated by using the whole proteome.

Table SII . Conserved elements amongst aligned sequences. The four genomic sets were analyzed so as to recover 55%, 65% or 75% of the CDSs amongst the aligned sequences.

Genomic Set	Method	Target coverage	Expected length	Conserved elements(bp)	CDS coverage	tRNA coverage	PIT	Lmin
RWGD	55%	0.600	55	22,279,398	56.04	86.40	9.82	81.61
	65%	0.825	185	25,938,353	65.05	85.96	9.81	100.90
	75%	/	/	/	/	/	/	/
IWGD	55%	0.460	27	21,864,333	54.91	91.67	9.82	21.76
	65%	0.666	55	26,219,934	65.18	92.14	9.80	25.50
	75%	0.896	180	31,300,132	76.51	95.46	9.81	31.74
OWGD	55%	0.470	29	13,850,317	54.73	94.82	9.86	10.58
	65%	0.670	50	16,659,965	65.02	94.88	9.81	12.03
	75%	0.852	104	19,661,506	75.46	99.88	9.81	14.22
AD	55%	0.450	25	9,318,383	57.65	92.27	/	/
	65%	0.640	50	10,617,768	65.24	94.60	/	/
	75%	0.808	50	12,609,434	76.09	99.37	/	/

Figure S1

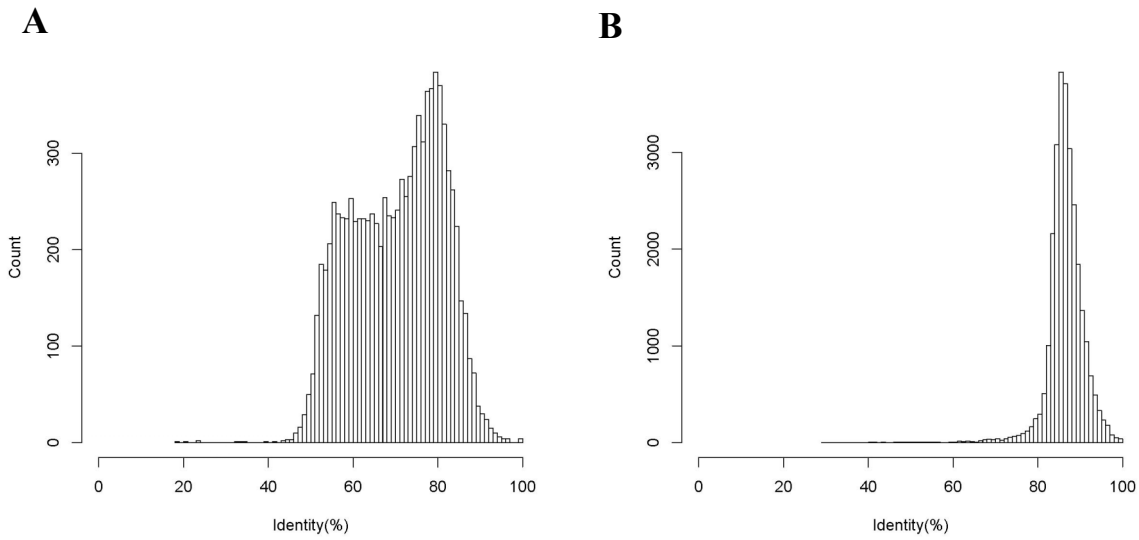

A
Count
0
100
200
300
0
20
40
60
80
100
Identity(%)
B
Count
0
1000
2000
3000
0
20
40
60
80
100
Identity(%)

Figure S2

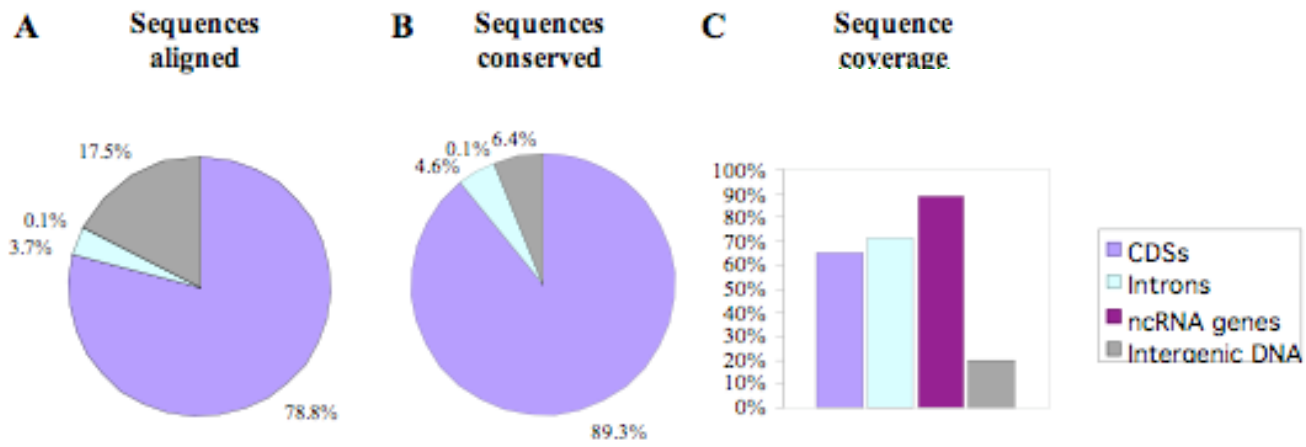
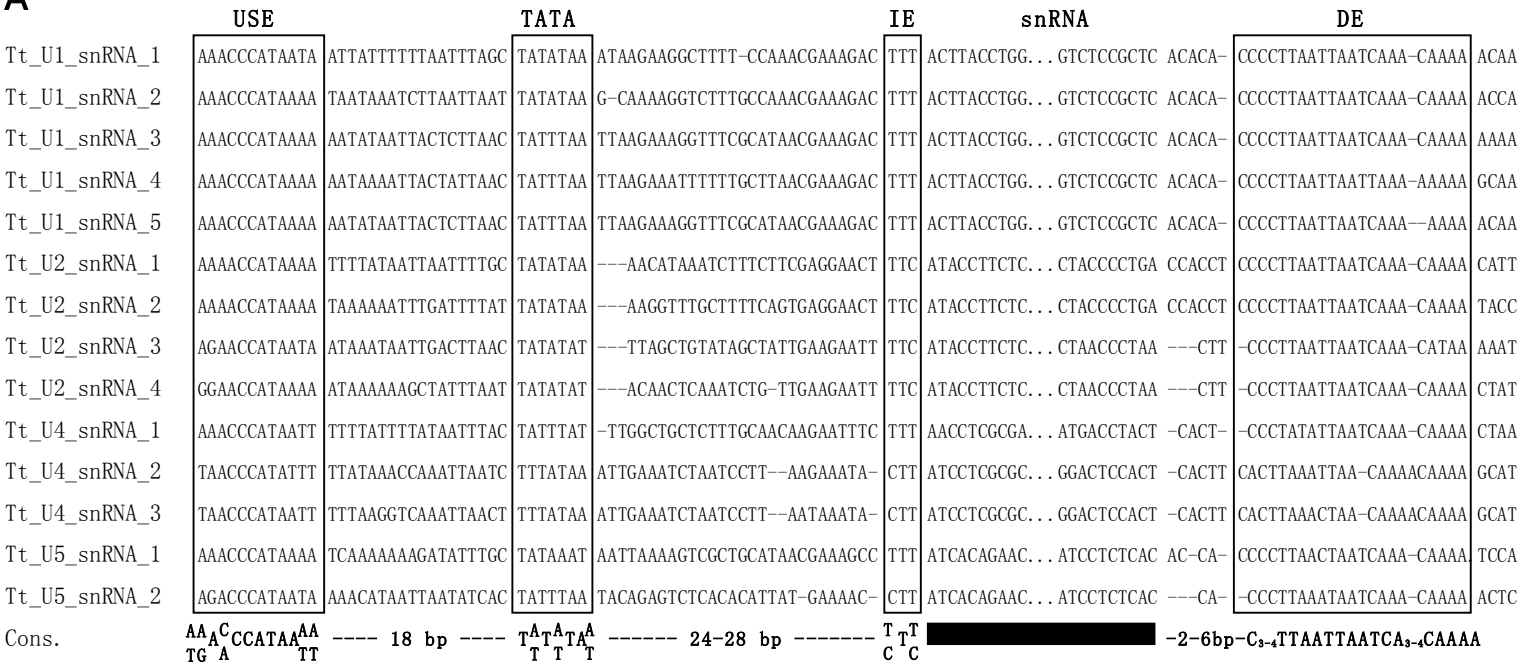

A
Sequences aligned
17.5%
0.1%
3.7%
78.8%
B
Sequences conserved
0.1%
6.4%
4.6%
89.3%
C
Sequence coverage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
CDSs
Introns
ncRNA genes
Intergenic DNA

Figure S3

A


USE
TATA
IE
snRNA
DE
Tt_U1_snRNA_1 AAACCCATAATA ATTATTTTTTAATTTAGC TATATAA ATAAGAAGGCTTTT-CCAAACGAAAGAC TTT ACTTACCTGG...GTCTCCGCTC ACACA- CCCCTTAATTAATCAAA-CAAAA ACAA
Tt_U1_snRNA_2 AAACCCATAAAA TAATAAATCTTAATTAAT TATATAA G-CAAAAGGTCTTTGCCAAACGAAAGAC TTT ACTTACCTGG...GTCTCCGCTC ACACA- CCCCTTAATTAATCAAA-CAAAA ACCA
Tt_U1_snRNA_3 AAACCCATAAAA AATATAATTACTCTTAAC TATTTAA TTAAGAAAGGTTTCGCATAACGAAAGAC TTT ACTTACCTGG...GTCTCCGCTC ACACA- CCCCTTAATTAATCAAA-CAAAA AAAA
Tt_U1_snRNA_4 AAACCCATAAAA AATAAAATTACTATTAAC TATTTAA TTAAGAAATTTTTTGCTTAACGAAAGAC TTT ACTTACCTGG...GTCTCCGCTC ACACA- CCCCTTAATTAATTAAA-AAAAA GCAA
Tt_U1_snRNA_5 AAACCCATAAAA AATATAATTACTCTTAAC TATTTAA TTAAGAAAGGTTTCGCATAACGAAAGAC TTT ACTTACCTGG...GTCTCCGCTC ACACA- CCCCTTAATTAATCAAA--AAAA ACAA
Tt_U2_snRNA_1 AAAACCATAAAA TTTTATAATTAATTTTGC TATATAA ---AACATAAATCTTTCTTCGAGGAACT TTC ATACCTTCTC...CTACCCCTGA CCACCT CCCCTTAATTAATCAAA-CAAAA CATT
Tt_U2_snRNA_2 AAAACCATAAAA TAAAAAATTTGATTTTAT TATATAA ---AAGGTTTGCTTTTCAGTGAGGAACT TTC ATACCTTCTC...CTACCCCTGA CCACCT CCCCTTAATTAATCAAA-CAAAA TACC
Tt_U2_snRNA_3 AGAACCATAATA ATAAATAATTGACTTAAC TATATAT ---TTAGCTGTATAGCTATTGAAGAATT TTC ATACCTTCTC...CTAACCCTAA ---CTT -CCCTTAATTAATCAAA-CATAA AAAT
Tt_U2_snRNA_4 GGAACCATAAAA ATAAAAAAGCTATTTAAT TATATAT ---ACAACTCAAATCTG-TTGAAGAATT TTC ATACCTTCTC...CTAACCCTAA ---CTT -CCCTTAATTAATCAAA-CAAAA CTAT
Tt_U4_snRNA_1 AAACCCATAATT TTTTATTTTATAATTTAC TATTTAT -TTGGCTGCTCTTTGCAACAAGAATTTC TTT AACCTCGCGA...ATGACCTACT -CACT- -CCCTATATTAATCAAA-CAAAA CTAA
Tt_U4_snRNA_2 TAACCCATATTT TTATAAACCAAATTAATC TTTATAA ATTGAAATCTAATCCTT--AAGAAATA- CTT ATCCTCGCGC...GGACTCCACT -CACTT CACTTAAATTAA-CAAAACAAAA GCAT
Tt_U4_snRNA_3 TAACCCATAATT TTTAAGGTCAAATTAACT TTTATAA ATTGAAATCTAATCCTT--AATAAATA- CTT ATCCTCGCGC...GGACTCCACT -CACTT CACTTAAACTAA-CAAAACAAAA GCAT
Tt_U5_snRNA_1 AAACCCATAAAA TCAAAAAAAGATATTTGC TATAAAT AATTAAAAGTCGCTGCATAACGAAAGCC TTT ATCACAGAAC...ATCCTCTCAC AC-CA- CCCCTTAACTAATCAAA-CAAAA TCCA
Tt_U5_snRNA_2 AGACCCATAATA AAACATAATTAATATCAC TATTTAA TACAGAGTCTCACACATTAT-GAAAAC- CTT ATCACAGAAC...ATCCTCTCAC ---CA- -CCCTTAAATAATCAAA-CAAAA ACTC
Cons. AA/TG A C/A CCATAA AA/TT ---- 18 bp ---- T A/T T A/T T A A/T ------ 24-28 bp ------- T/C T T/C -2-6bp-C3-4TTAATTAATCA3-4CAAAA

B

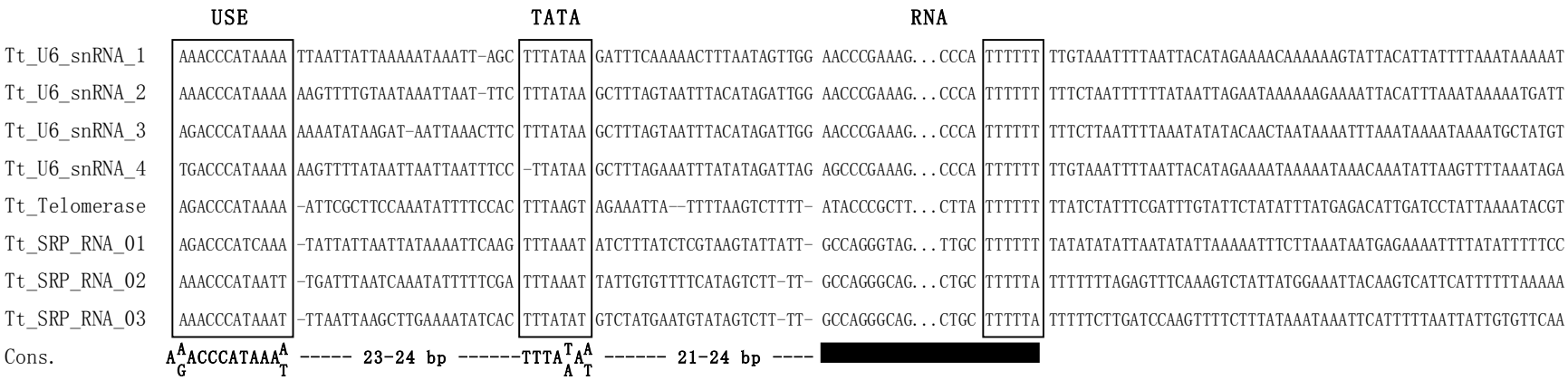
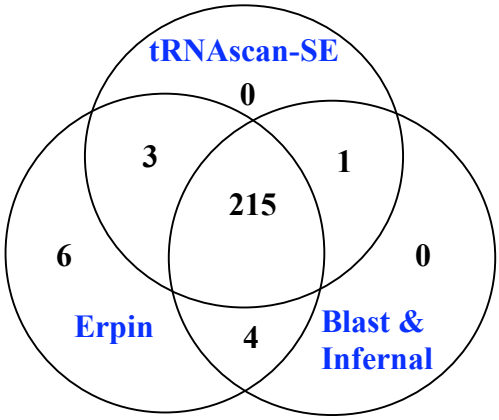

USE
TATA
RNA
Tt_U6_snRNA_1 AAACCCATAAAA TTAATTATTAAAAATAAATT-AGC TTTATAA GATTTCAAAAACTTTAATAGTTGG AACCCGAAAG...CCCA TTTTTT TTGTAAATTTTAATTACATAGAAAACAAAAAAGTATTACATTATTTTAAATAAAAAT
Tt_U6_snRNA_2 AAACCCATAAAA AAGTTTTGTAATAAATTAAT-TTC TTTATAA GCTTTAGTAATTTACATAGATTGG AACCCGAAAG...CCCA TTTTTT TTTCTAATTTTTTATAATTAGAATAAAAAAGAAAATTACATTTAAATAAAAATGATT
Tt_U6_snRNA_3 AGACCCATAAAA AAAATATAAGAT-AATTAAACTTC TTTATAA GCTTTAGTAATTTACATAGATTGG AACCCGAAAG...CCCA TTTTTT TTTCTTAATTTTAAATATATACAACTAATAAAATTTAAATAAAATAAAATGCTATGT
Tt_U6_snRNA_4 TGACCCATAAAA AAGTTTTATAATTAATTAATTTCC -TTATAA GCTTTAGAAATTTATATAGATTAG AGCCCGAAAG...CCCA TTTTTT TTGTAAATTTTAATTACATAGAAAATAAAAATAAACAAATATTAAGTTTTAAATAGA
Tt_Telomerase AGACCCATAAAA -ATTCGCTTCCAAATATTTTCCAC TTTAAGT AGAAATTA--TTTTAAGTCTTTT- ATACCCGCTT...CTTA TTTTTT TTATCTATTTCGATTTGTATTCTATATTTATGAGACATTGATCCTATTAAAATACGT
Tt_SRP_RNA_01 AGACCCATCAAA -TATTATTAATTATAAAATTCAAG TTTAAAT ATCTTTATCTCGTAAGTATTATT- GCCAGGGTAG...TTGC TTTTTT TATATATATTAATATATTAAAAATTTCTTAAATAATGAGAAAATTTTATATTTTTCC
Tt_SRP_RNA_02 AAACCCATAATT -TGATTTAATCAAATATTTTTCGA TTTAAAT TATTGTGTTTTCATAGTCTT-TT- GCCAGGGCAG...CTGC TTTTTA TTTTTTTAGAGTTTCAAAGTCTATTATGGAAATTACAAGTCATTCATTTTTTAAAAA
Tt_SRP_RNA_03 AAACCCATAAAT -TTAATTAAGCTTGAAAATATCAC TTTATAT GTCTATGAATGTATAGTCTT-TT- GCCAGGGCAG...CTGC TTTTTA TTTTTCTTGATCCAAGTTTTCTTTATAAATAAATTCATTTTTAATTATTGTGTTCAA
Cons. A A/G ACCCATAAA A/T ----- 23-24 bp ------TTTA T/A A A/T ------ 21-24 bp ----

Figure S4

A. tRNA gene identification



B. ncRNA gene identification

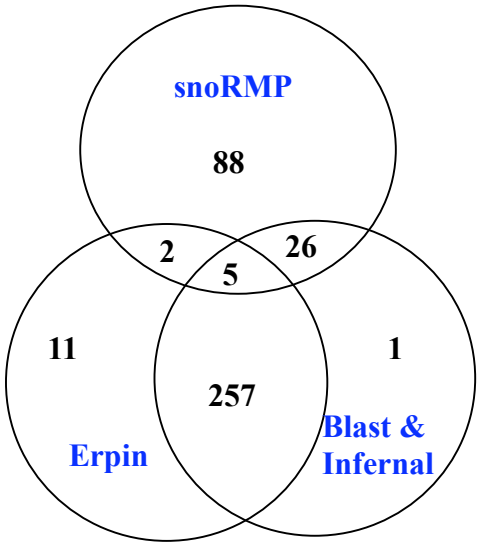


Figure S5

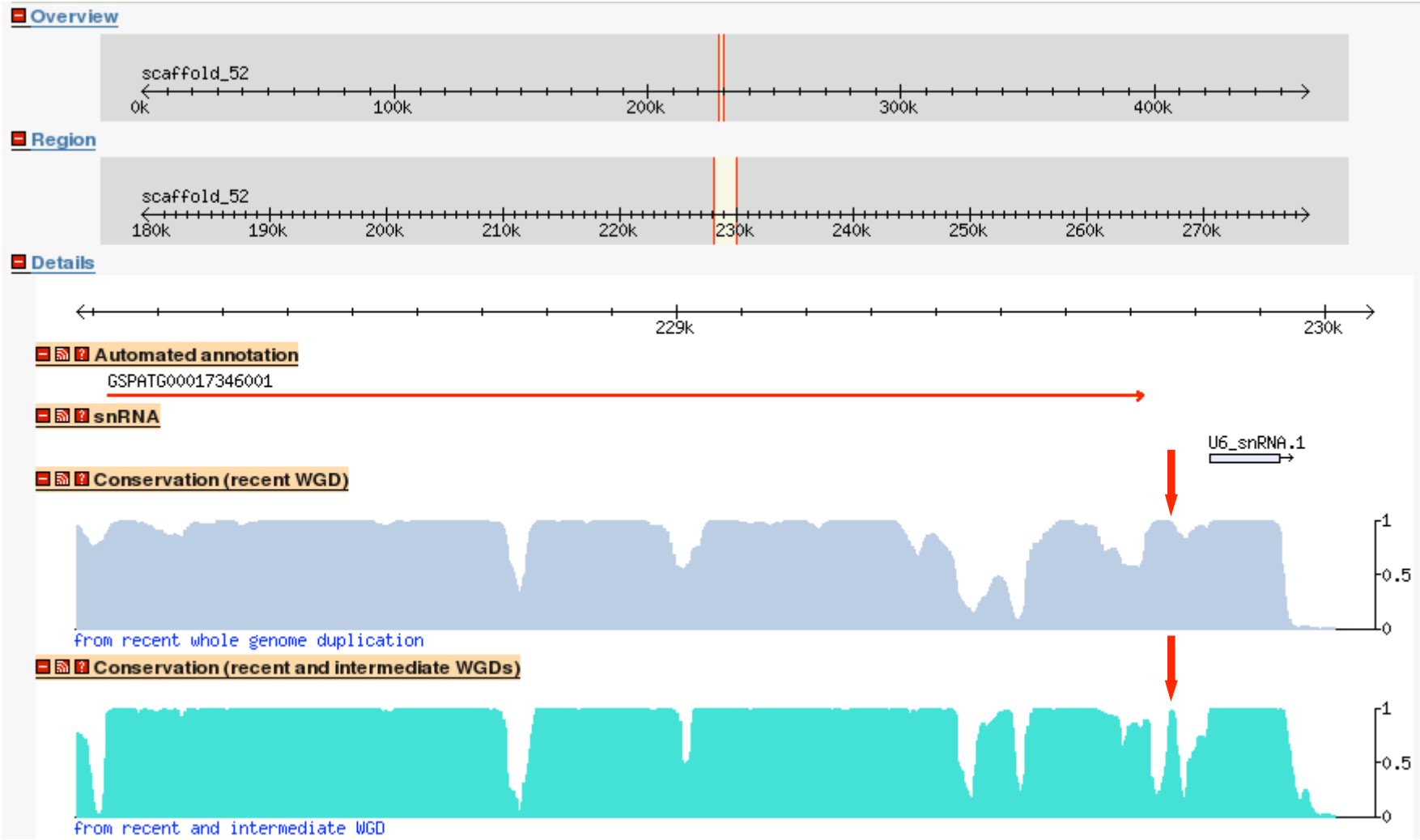
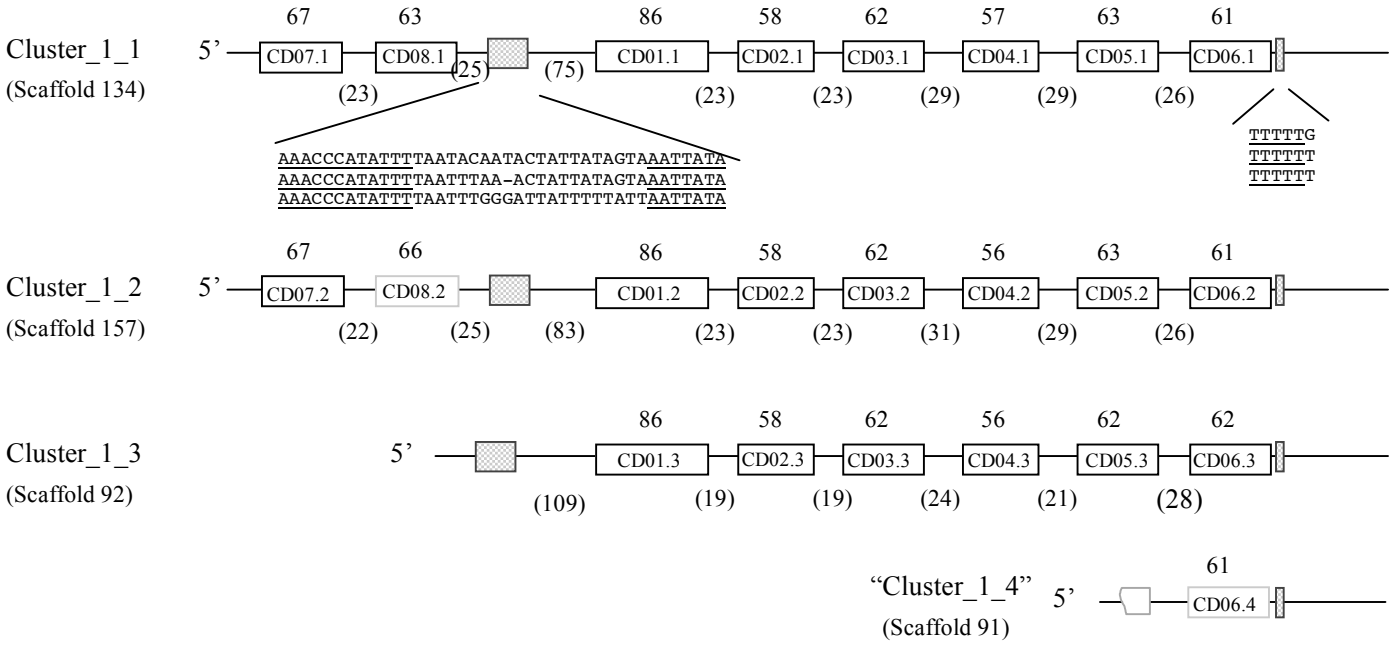

Overview
scaffold_52
0k
100k
200k
300k
400k
Region
scaffold_52
180k
190k
200k
210k
220k
230k
240k
250k
260k
270k
Details
229k
230k
Automated annotation
GSPATG00017346001
snRNA
U6_snRNA.1
Conservation (recent WGD)
1
0.5
0
from recent whole genome duplication
Conservation (recent and intermediate WGDs)
1
0.5
0
from recent and intermediate WGD

Figure S6

A


Cluster_1_1
(Scaffold 134)
5’
67
63
86
58
62
57
63
61
CD07.1
CD08.1
CD01.1
CD02.1
CD03.1
CD04.1
CD05.1
CD06.1
(23)
(25)
(75)
(23)
(23)
(29)
(29)
(26)
AAACCCATATTTTAATACAATACTATTATAGTAAATTATA
AAACCCATATTTTAATTTAA-ACTATTATAGTAAATTATA
AAACCCATATTTTAATTTGGGATTATTTTTATTAATTATA
TTTTTG
TTTTTT
TTTTTT
Cluster_1_2
(Scaffold 157)
5’
67
66
86
58
62
56
63
61
CD07.2
CD08.2
CD01.2
CD02.2
CD03.2
CD04.2
CD05.2
CD06.2
(22)
(25)
(83)
(23)
(23)
(31)
(29)
(26)
Cluster_1_3
(Scaffold 92)
5’
86
58
62
56
62
62
CD01.3
CD02.3
CD03.3
CD04.3
CD05.3
CD06.3
(109)
(19)
(19)
(24)
(21)
(28)
“Cluster_1_4”
(Scaffold 91)
5’
61
CD06.4

B

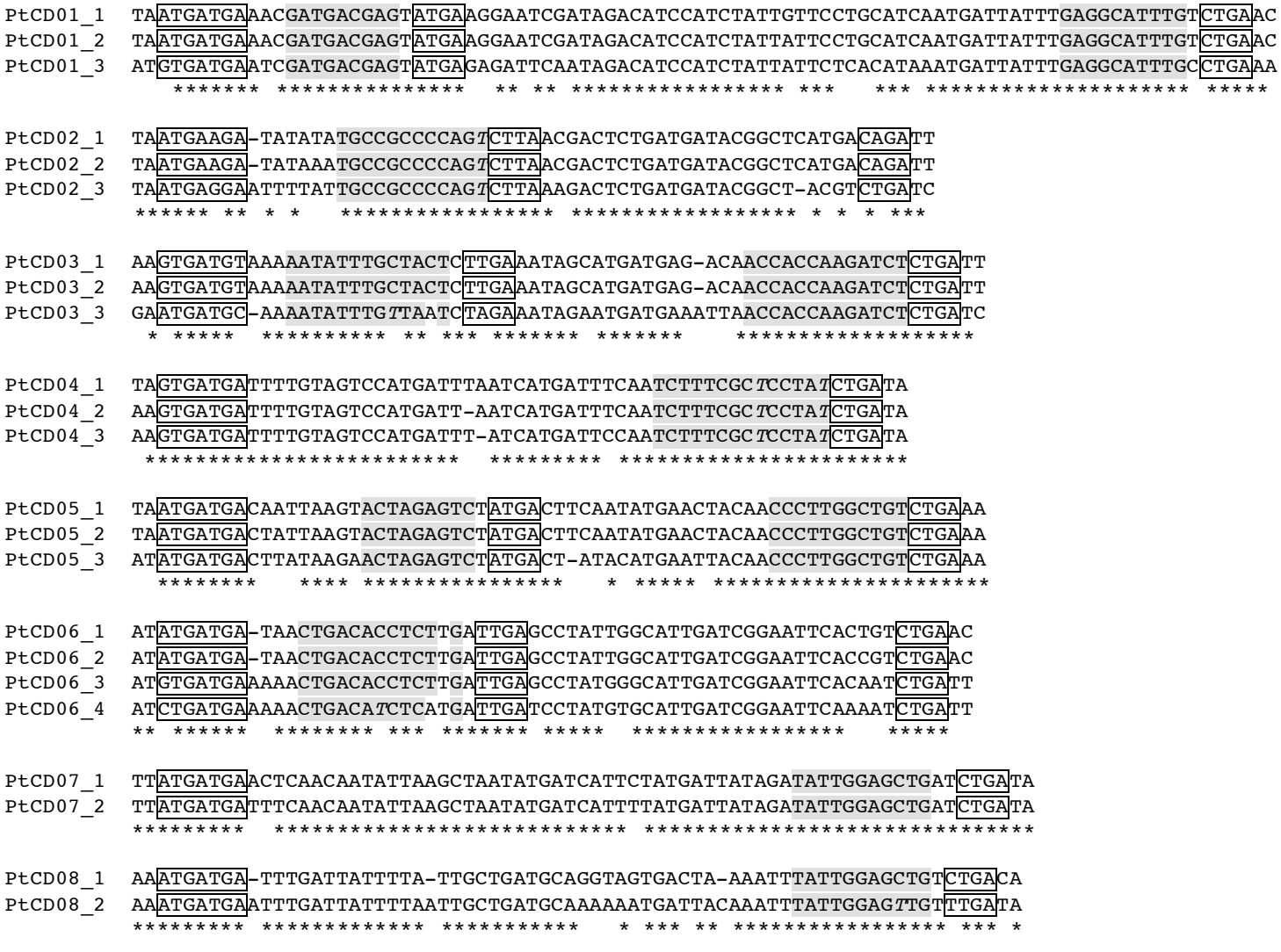
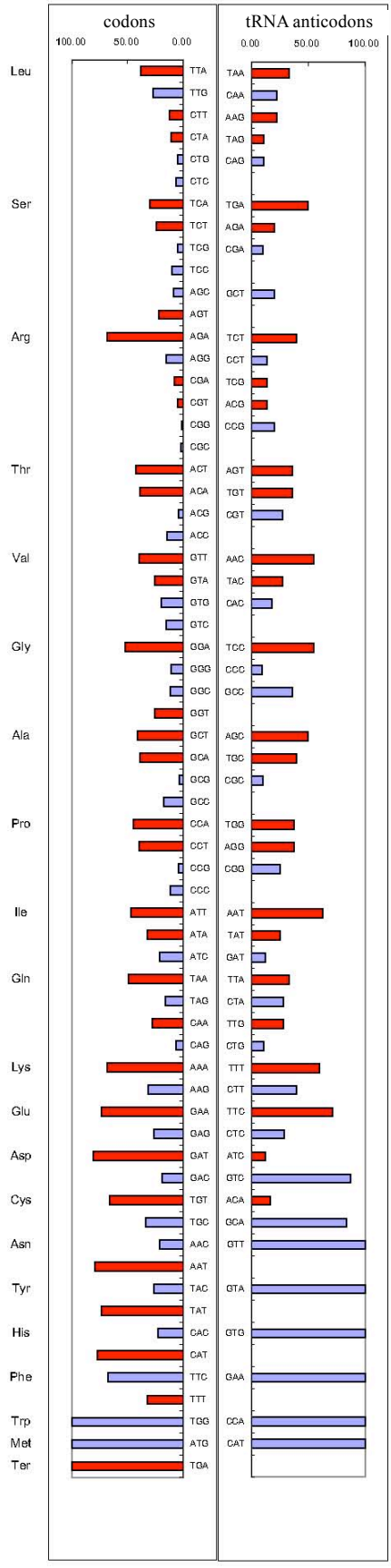

PtCD01_1
PtCD01_2
PtCD01_3
PtCD02_1
PtCD02_2
PtCD02_3
PtCD03_1
PtCD03_2
PtCD03_3
PtCD04_1
PtCD04_2
PtCD04_3
PtCD05_1
PtCD05_2
PtCD05_3
PtCD06_1
PtCD06_2
PtCD06_3
PtCD06_4
PtCD07_1
PtCD07_2
PtCD08_1
PtCD08_2

Figure S7


codons
100.00
50.00
0.00
tRNA anticodons
0.00
50.00
100.00
Leu
Ser
Arg
Thr
Val
Gly
Ala
Pro
Ile
Gln
Lys
Glu
Asp
Cys
Asn
Tyr
His
Phe
Trp
Met
Ter
TTA
TTG
CTT
CTA
CTG
CTC
TCA
TCT
TCG
TCC
AGC
AGT
AGA
AGG
CGA
CGT
CGG
CGC
ACT
ACA
ACG
ACC
GTT
GTA
GTG
GTC
GGA
GGG
GGC
GGT
GCT
GCA
GCG
GCC
CCA
CCT
CCG
CCC
ATT
ATA
ATC
TAA
TAG
CAA
CAG
AAA
AAG
GAA
GAG
GAT
GAC
TGT
TGC
AAC
AAT
TAC
TAT
CAC
CAT
TTC
TTT
TGG
ATG
TGA
TAA
CAA
AAG
TAG
CAG
TGA
AGA
CGA
GCT
TCT
CCT
TCG
ACG
CCG
AGT
TGT
CGT
AAC
TAC
CAC
TCC
CCC
GCC
AGC
TGC
CGC
TGG
AGG
CGG
AAT
TAT
GAT
TTA
CTA
TTG
CTG
TTT
CTT
TTC
CTC
ATC
GTC
ACA
GCA
GTT
GTA
GTG
GAA
CCA
CAT