SUPPLEMENTARY MATERIAL

Different sequence signatures in the upstream regions of plant and animal tRNA genes shape distinct modes of regulation

Gong Zhang[1], Radoslaw Lukoszek[1,2], Bernd Mueller-Roeber[2], Zoya Ignatova[1]

[1]Biochemistry, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany

[2]Molecular Biology, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany

Corresponding authors: Zoya Ignatova, Biochemistry, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14467 Potsdam, Germany; Tel: +49 331 977 5130, Fax: +49 331 977 5128; e-mail: ignatova@uni-potsdam.de

Bernd Mueller-Roeber, Molecular Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14467 Potsdam, Germany; Tel: +49 331 977 2810, Fax: +49 331 977 2512; e-mail: bmr@uni-potsdam.de

Table S1. Animal tRNAs encoding the same amino acid share common motif in their upstream regions. tRNAs from each organism were grouped dependent on the amino acid they decode and the tRNAs with at least one conserved motif in their upstream region are marked with +. Upstream segments without any conserved motif are marked with −.

Amino acid	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Anopheles gambiae*	*Gallus gallus*	*Mus musculus*	*Rattus norvegicus*	*Homo sapiens*
Ala	+	−	−	+	+	+	+
Arg	+	+	+	−	−	−	−
Asn	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	+
Asp	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.
Cys	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.
Gln	+	−	+	+	+	+	+
Glu	+	+	+	+	+	+	+
Gly	+	+	-	+	+	−	+
His	+	s.a.	s.a.	+	s.a.	s.a.	s.a.
Ile	+	−	+	+	+	+	+
Leu	+	−	−	−	+	+	+
Lys	−	+	+	−	+	−	−
Met	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.
Phe	s.a.	s.a.	s.a.	+	s.a.	+	s.a.
Pro	−	−	−	+	−	−	+
Ser	−	+	−	+	−	+	−
Thr	−	−	−	−	−	−	−
Trp	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.
Tyr	s.a.	s.a.	s.a.	s.a.	s.a.	s.a.	+
Val	−	−	−	−	+	+	+

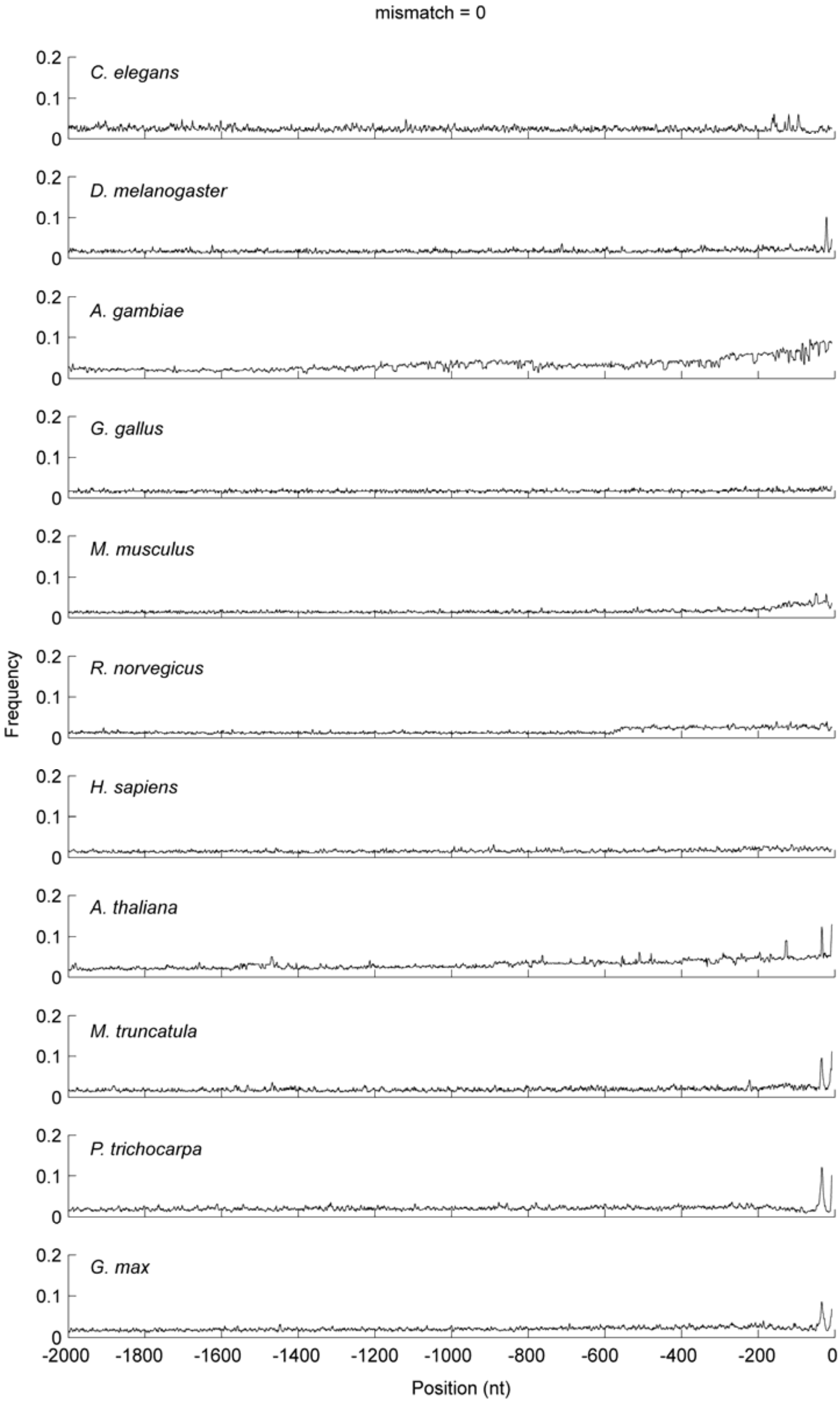
s.a. denotes that all tRNAs decoding this amino acid bear the same anticodon.

Table S2. Human Brf1 and Bdp1 splicing variants. Data were retrieved from the ECgene alternative splicing database [http://genome.ewha.ac.kr/ECgene/].

Gene	ID	mRNA length (bp)	CDS length (bp)
Brf1	H14C11627.1	3511	579
	H14C11627.2	2874	609
	H14C11627.3	913	402
	H14C11627.4	496	259
	H14C11627.5	967	483
	H14C11627.6	1673	402
	H14C11627.7	4169	1419
	H14C11627.8	2861	1317
	H14C11627.9	3538	2031 *
	H14C11627.10	3603	1950
	H14C11627.11	3948	2031 *
	H14C11627.12	1906	450
	H14C11627.13	1492	339
	H14C11627.14	574	297
	H14C11627.15	1452	624
Bdp1	H5C7152.1	740	476
	H5C7152.2	4447	4116
	H5C7152.3	7265	6762
	H5C7152.4	10600	7872
	H5C7152.5	5031	1392
	H5C7152.6	5013	1797

* mRNA splice variants H14C11627.9 and H14C11627.11 are identical in protein sequence but differ in the non-coding flanking regions.

A


mismatch = 0
C. elegans
D. melanogaster
A. gambiae
G. gallus
M. musculus
R. norvegicus
H. sapiens
A. thaliana
M. truncatula
P. trichocarpa
G. max
0.2
0.1
0
Frequency
-2000
-1800
-1600
-1400
-1200
-1000
-800
-600
-400
-200
0
Position (nt)

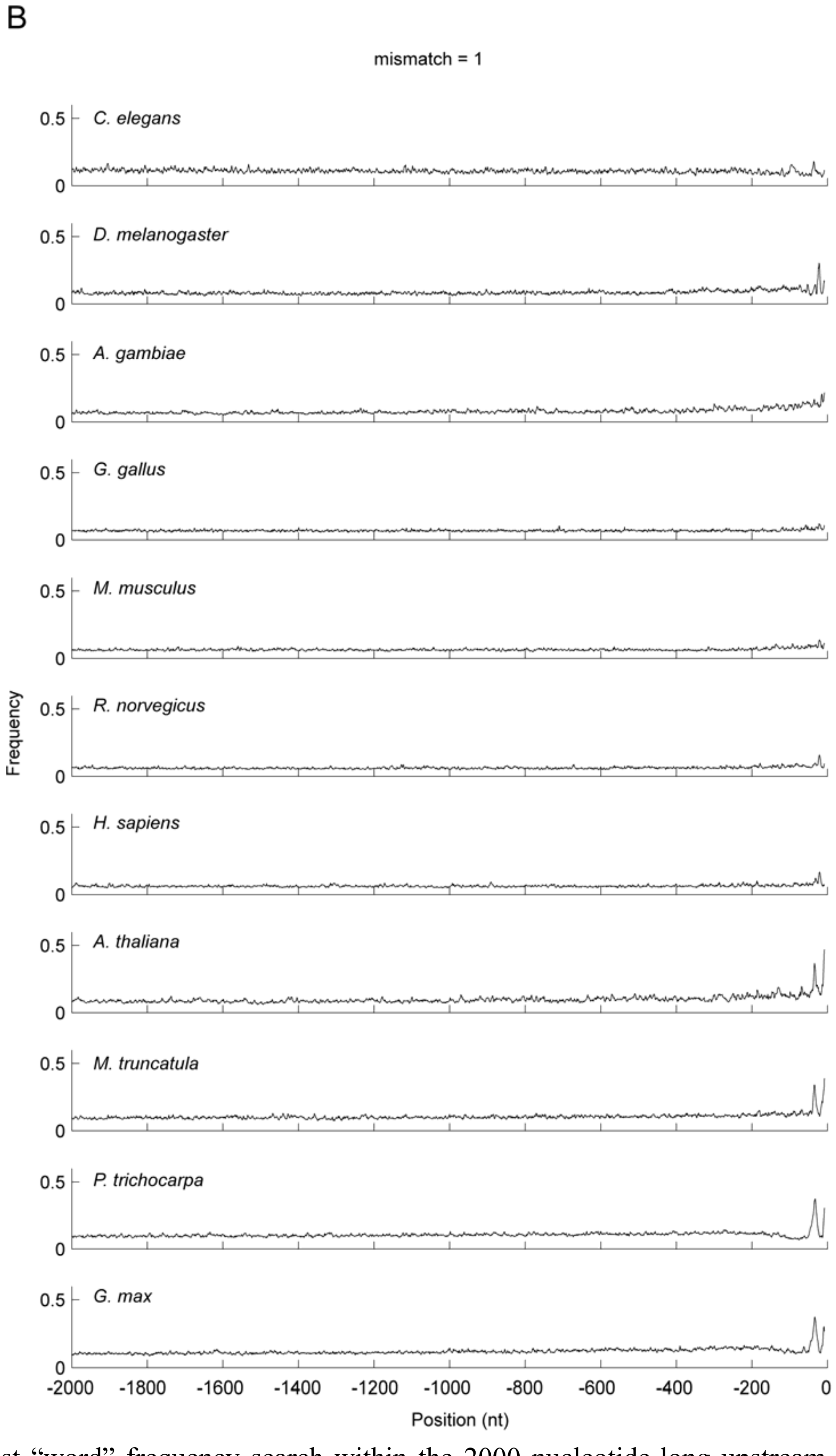


Figure S1. Best "word" frequency search within the 2000 nucleotide-long upstream regions of the mature tRNA sequences revealed no additional significant motif that could serve as a putative recognition sites for TFIIIB. The occurrence of the most frequent six-letter "word" at each position (x-axis) is plotted on the y-axis allowing a search with none (A) or one mismatch (B). Position 0 is determined by the first 5'-nucleotide of the mature tRNA. The significant peaks within the first 100 nucleotides upstream of the5'-initial nucleotide of the mature tRNA are identical to those identified in Figure 1.

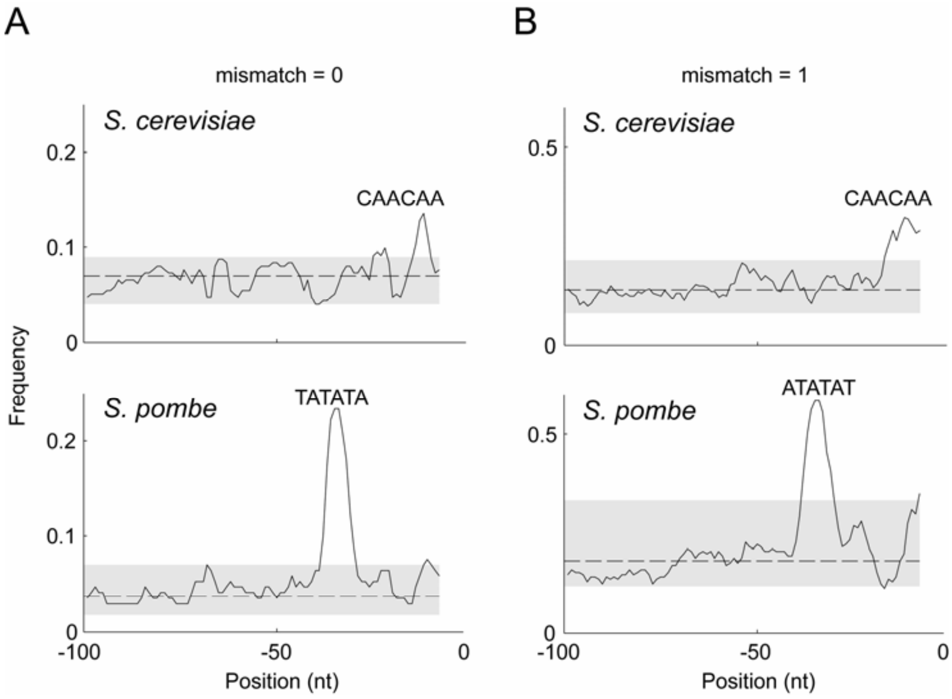


Figure S2. The best frequency "word" search in two yeast genomes, *Saccharomyces cerevisiae* (SacCer1, Oct. 2003) and *Schizosaccharomyces pombe* (build 1.1., Apr. 2007), revealed different conserved motifs. The search was performed in the near upstream region (up to -100 nucleotides) of the tRNAs genes allowing none (A) or one mismatch (B) (for details see the legend to Figure 1). Position 0 is determined by the first nucleotide of the mature tRNA. The dashed line indicates the average frequency value obtained from randomized sequences. Maximum and minimum frequency values determine the boundaries of the shadowed area and were also derived from the randomized dataset.

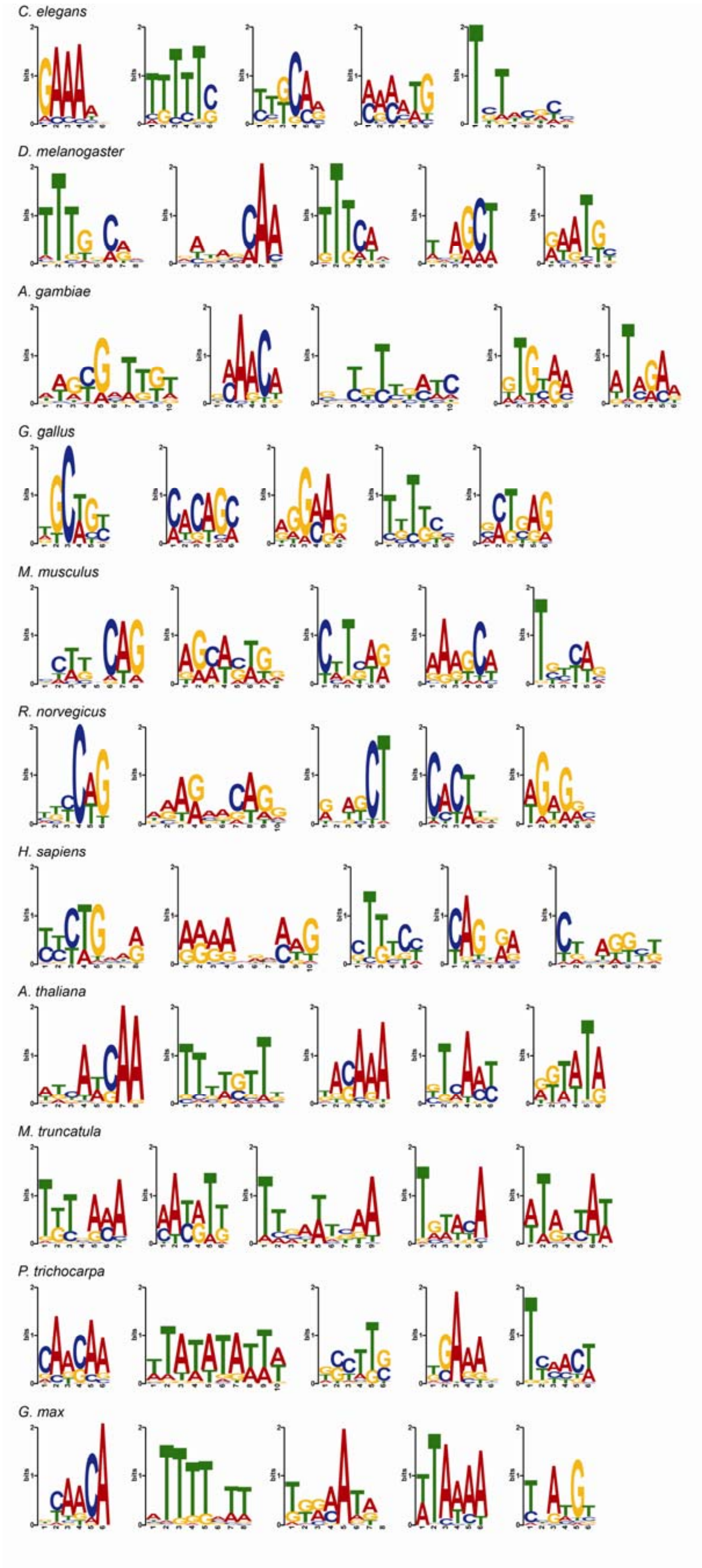
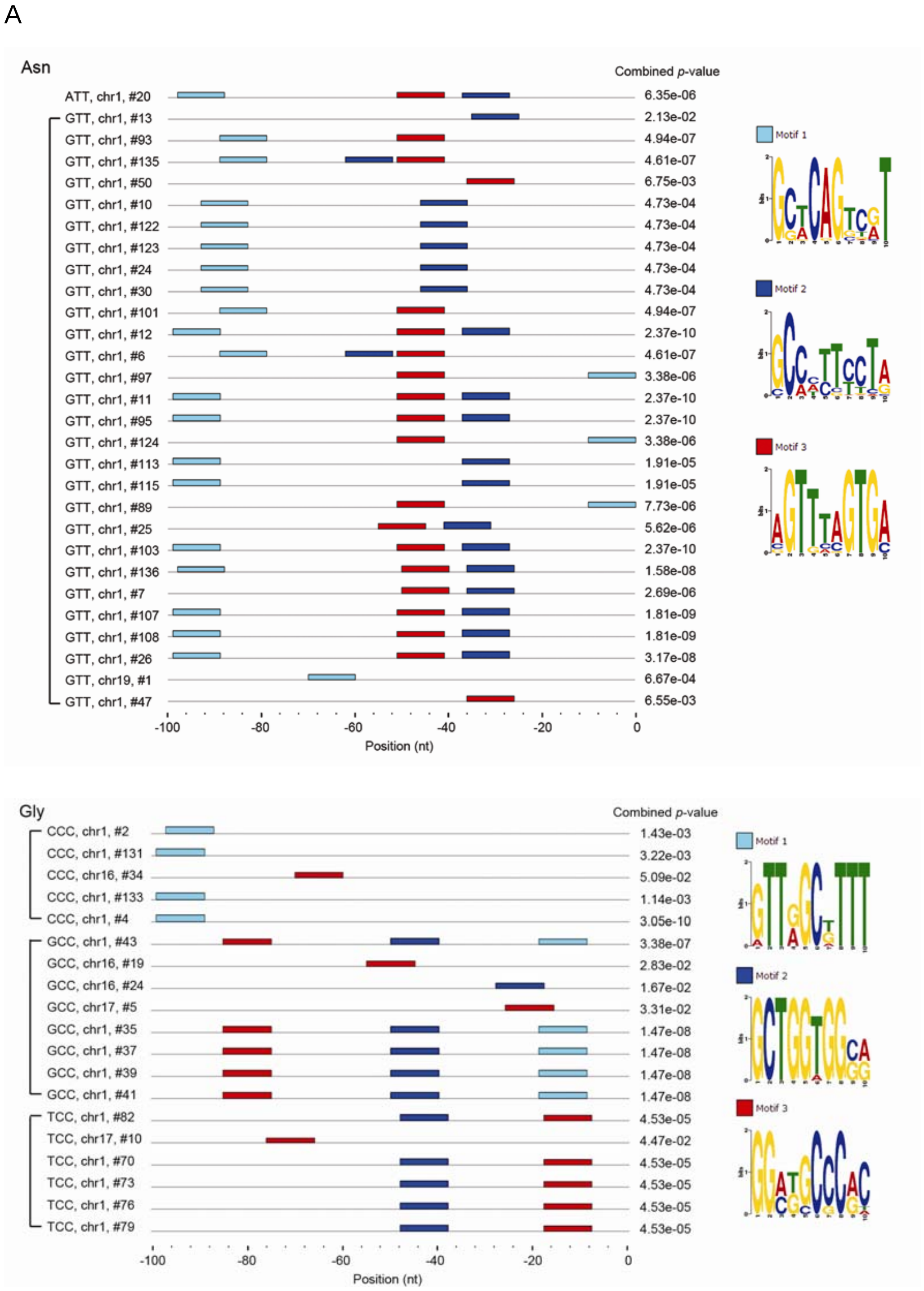
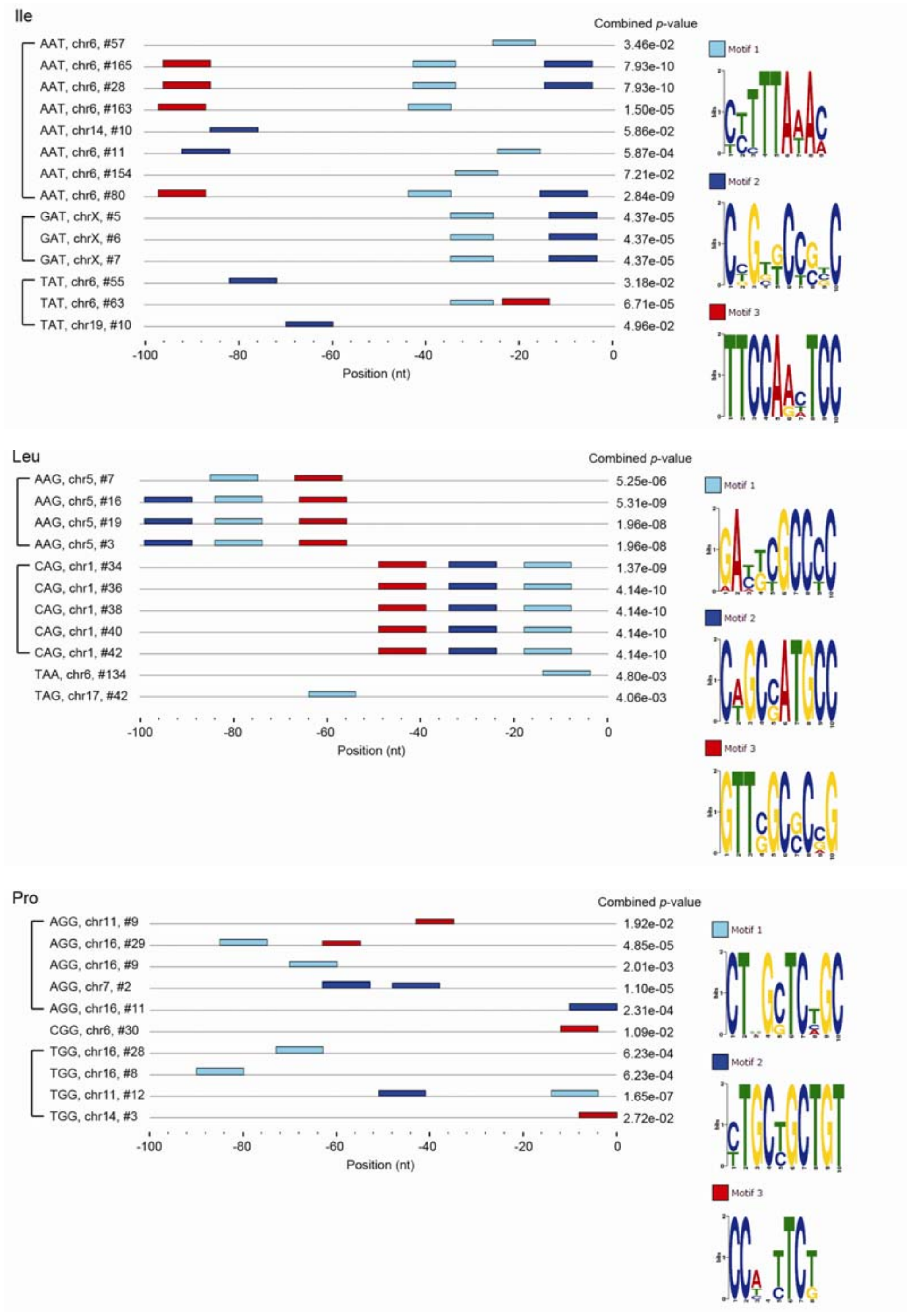
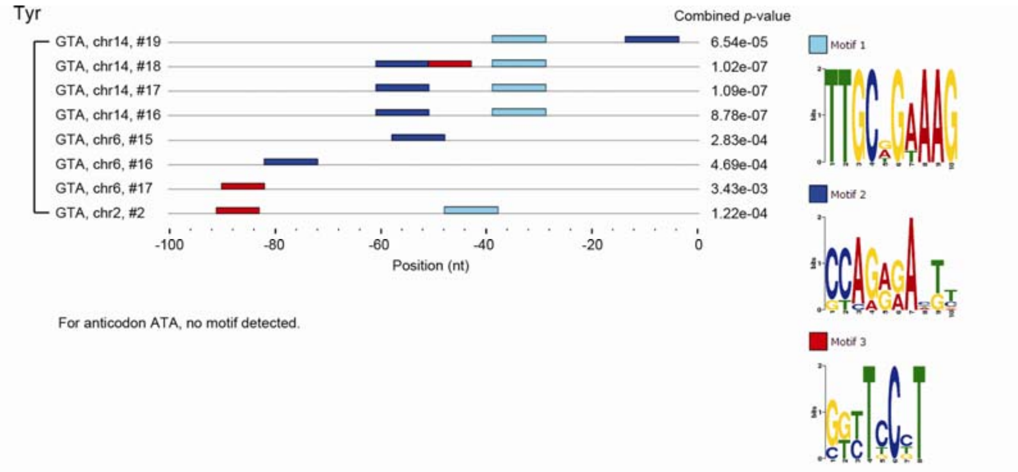
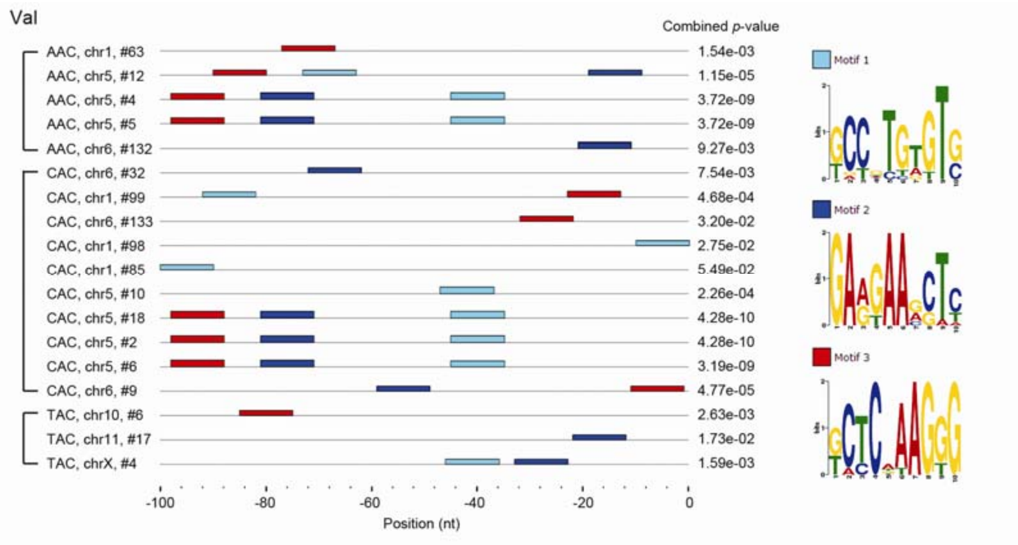


Figure S3. Motifs identified within the 100 nucleotides upstream of the 5’-start of the mature tRNAs. The top five most significant motifs for each genome are shown. TA-rich motifs are detected in every plant genome, but are completely lacking in the animal genomes. The motif search was performed using MEME version 4.4.0 (http://meme.sdsc.edu/meme4_4_0/cgi-bin/meme.cgi).

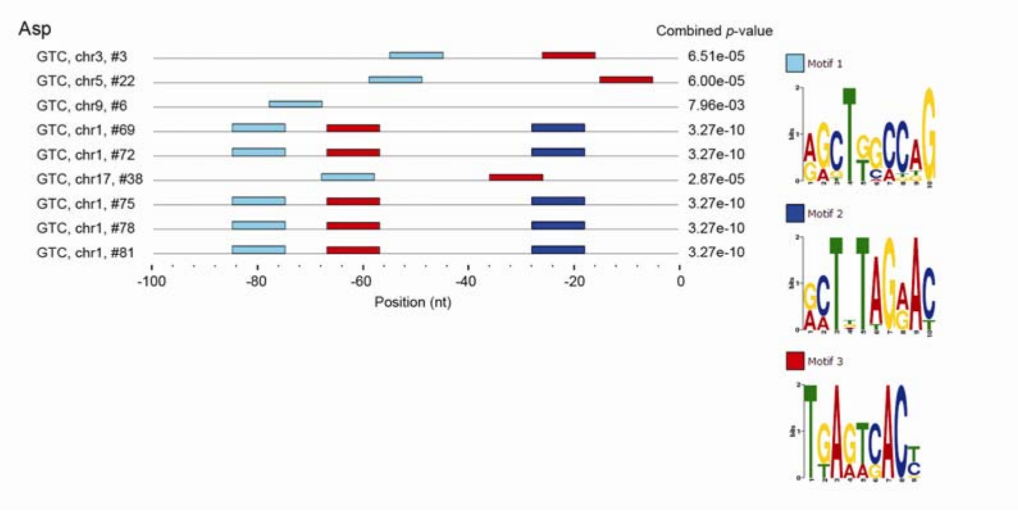

A
Asn
Combined p-value
ATT, chr1, #20 6.35e-06
GTT, chr1, #13 2.13e-02
GTT, chr1, #93 4.94e-07
GTT, chr1, #135 4.61e-07
GTT, chr1, #50 6.75e-03
GTT, chr1, #10 4.73e-04
GTT, chr1, #122 4.73e-04
GTT, chr1, #123 4.73e-04
GTT, chr1, #24 4.73e-04
GTT, chr1, #30 4.73e-04
GTT, chr1, #101 4.94e-07
GTT, chr1, #12 2.37e-10
GTT, chr1, #6 4.61e-07
GTT, chr1, #97 3.38e-06
GTT, chr1, #11 2.37e-10
GTT, chr1, #95 2.37e-10
GTT, chr1, #124 3.38e-06
GTT, chr1, #113 1.91e-05
GTT, chr1, #115 1.91e-05
GTT, chr1, #89 7.73e-06
GTT, chr1, #25 5.62e-06
GTT, chr1, #103 2.37e-10
GTT, chr1, #136 1.58e-08
GTT, chr1, #7 2.69e-06
GTT, chr1, #107 1.81e-09
GTT, chr1, #108 1.81e-09
GTT, chr1, #26 3.17e-08
GTT, chr19, #1 6.67e-04
GTT, chr1, #47 6.55e-03
-100
-80
-60
-40
-20
0
Position (nt)
Motif 1
Motif 2
Motif 3
Gly
Combined p-value
CCC, chr1, #2 1.43e-03
CCC, chr1, #131 3.22e-03
CCC, chr16, #34 5.09e-02
CCC, chr1, #133 1.14e-03
CCC, chr1, #4 3.05e-10
GCC, chr1, #43 3.38e-07
GCC, chr16, #19 2.83e-02
GCC, chr16, #24 1.67e-02
GCC, chr17, #5 3.31e-02
GCC, chr1, #35 1.47e-08
GCC, chr1, #37 1.47e-08
GCC, chr1, #39 1.47e-08
GCC, chr1, #41 1.47e-08
TCC, chr1, #82 4.53e-05
TCC, chr17, #10 4.47e-02
TCC, chr1, #70 4.53e-05
TCC, chr1, #73 4.53e-05
TCC, chr1, #76 4.53e-05
TCC, chr1, #79 4.53e-05
-100
-80
-60
-40
-20
0
Position (nt)
Motif 1
Motif 2
Motif 3

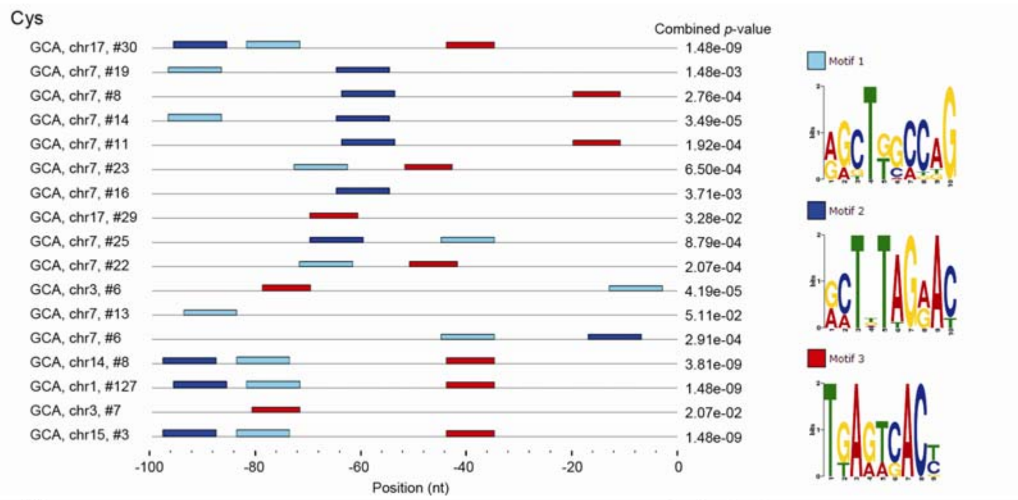

Ile
Combined p-value
AAT, chr6, #57 3.46e-02
AAT, chr6, #165 7.93e-10
AAT, chr6, #28 7.93e-10
AAT, chr6, #163 1.50e-05
AAT, chr14, #10 5.86e-02
AAT, chr6, #11 5.87e-04
AAT, chr6, #154 7.21e-02
AAT, chr6, #80 2.84e-09
GAT, chrX, #5 4.37e-05
GAT, chrX, #6 4.37e-05
GAT, chrX, #7 4.37e-05
TAT, chr6, #55 3.18e-02
TAT, chr6, #63 6.71e-05
TAT, chr19, #10 4.96e-02
-100
-80
-60
-40
-20
0
Position (nt)
Motif 1
Motif 2
Motif 3
Leu
Combined p-value
AAG, chr5, #7 5.25e-06
AAG, chr5, #16 5.31e-09
AAG, chr5, #19 1.96e-08
AAG, chr5, #3 1.96e-08
CAG, chr1, #34 1.37e-09
CAG, chr1, #36 4.14e-10
CAG, chr1, #38 4.14e-10
CAG, chr1, #40 4.14e-10
CAG, chr1, #42 4.14e-10
TAA, chr6, #134 4.80e-03
TAG, chr17, #42 4.06e-03
-100
-80
-60
-40
-20
0
Position (nt)
Motif 1
Motif 2
Motif 3
Pro
Combined p-value
AGG, chr11, #9 1.92e-02
AGG, chr16, #29 4.85e-05
AGG, chr16, #9 2.01e-03
AGG, chr7, #2 1.10e-05
AGG, chr16, #11 2.31e-04
CGG, chr6, #30 1.09e-02
TGG, chr16, #28 6.23e-04
TGG, chr16, #8 6.23e-04
TGG, chr11, #12 1.65e-07
TGG, chr14, #3 2.72e-02
-100
-80
-60
-40
-20
0
Position (nt)
Motif 1
Motif 2
Motif 3

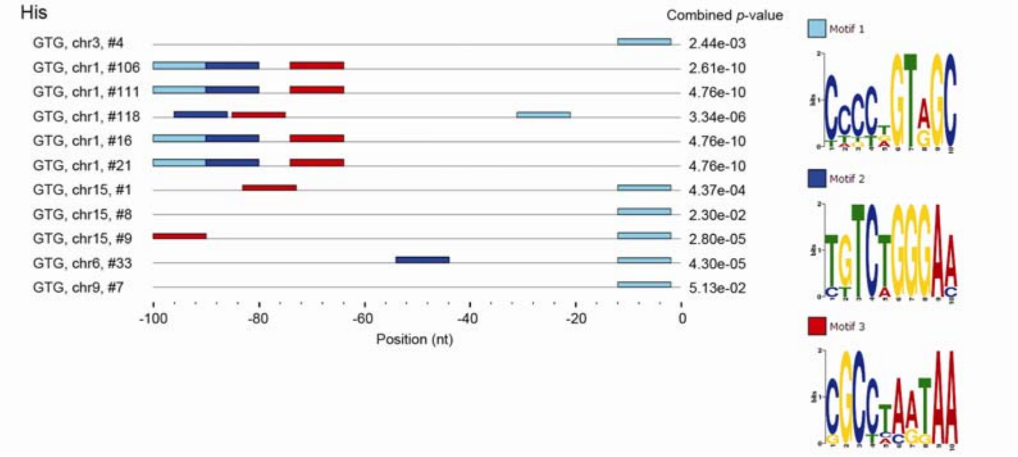

Tyr
Combined p-value
GTA, chr14, #19 6.54e-05
GTA, chr14, #18 1.02e-07
GTA, chr14, #17 1.09e-07
GTA, chr14, #16 8.78e-07
GTA, chr6, #15 2.83e-04
GTA, chr6, #16 4.69e-04
GTA, chr6, #17 3.43e-03
GTA, chr2, #2 1.22e-04
-100 -80 -60 -40 -20 0
Position (nt)
For anticodon ATA, no motif detected.
Motif 1
Motif 2
Motif 3

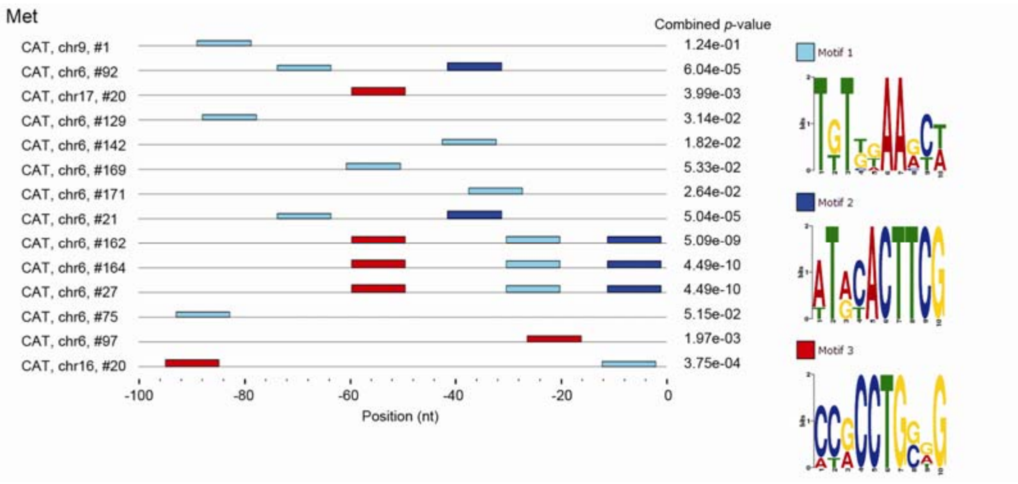

Val
Combined p-value
AAC, chr1, #63 1.54e-03
AAC, chr5, #12 1.15e-05
AAC, chr5, #4 3.72e-09
AAC, chr5, #5 3.72e-09
AAC, chr6, #132 9.27e-03
CAC, chr6, #32 7.54e-03
CAC, chr1, #99 4.68e-04
CAC, chr6, #133 3.20e-02
CAC, chr1, #98 2.75e-02
CAC, chr1, #85 5.49e-02
CAC, chr5, #10 2.26e-04
CAC, chr5, #18 4.28e-10
CAC, chr5, #2 4.28e-10
CAC, chr5, #6 3.19e-09
CAC, chr6, #9 4.77e-05
TAC, chr10, #6 2.63e-03
TAC, chr11, #17 1.73e-02
TAC, chrX, #4 1.59e-03
-100 -80 -60 -40 -20 0
Position (nt)
Motif 1
Motif 2
Motif 3

B


Asp
Combined p-value
GTC, chr3, #3 6.51e-05
GTC, chr5, #22 6.00e-05
GTC, chr9, #6 7.96e-03
GTC, chr1, #69 3.27e-10
GTC, chr1, #72 3.27e-10
GTC, chr17, #38 2.87e-05
GTC, chr1, #75 3.27e-10
GTC, chr1, #78 3.27e-10
GTC, chr1, #81 3.27e-10
-100 -80 -60 -40 -20 0
Position (nt)
Motif 1
Motif 2
Motif 3


Cys
Combined p-value
GCA, chr17, #30 1.48e-09
GCA, chr7, #19 1.48e-03
GCA, chr7, #8 2.76e-04
GCA, chr7, #14 3.49e-05
GCA, chr7, #11 1.92e-04
GCA, chr7, #23 6.50e-04
GCA, chr7, #16 3.71e-03
GCA, chr17, #29 3.28e-02
GCA, chr7, #25 8.79e-04
GCA, chr7, #22 2.07e-04
GCA, chr3, #6 4.19e-05
GCA, chr7, #13 5.11e-02
GCA, chr7, #6 2.91e-04
GCA, chr14, #8 3.81e-09
GCA, chr1, #127 1.48e-09
GCA, chr3, #7 2.07e-02
GCA, chr15, #3 1.48e-09
-100 -80 -60 -40 -20 0
Position (nt)
Motif 1
Motif 2
Motif 3


His
Combined p-value
GTG, chr3, #4 2.44e-03
GTG, chr1, #106 2.61e-10
GTG, chr1, #111 4.76e-10
GTG, chr1, #118 3.34e-06
GTG, chr1, #16 4.76e-10
GTG, chr1, #21 4.76e-10
GTG, chr15, #1 4.37e-04
GTG, chr15, #8 2.30e-02
GTG, chr15, #9 2.80e-05
GTG, chr6, #33 4.30e-05
GTG, chr9, #7 5.13e-02
-100 -80 -60 -40 -20 0
Position (nt)
Motif 1
Motif 2
Motif 3


Met
Combined p-value
CAT, chr9, #1 1.24e-01
CAT, chr6, #92 6.04e-05
CAT, chr17, #20 3.99e-03
CAT, chr6, #129 3.14e-02
CAT, chr6, #142 1.82e-02
CAT, chr6, #169 5.33e-02
CAT, chr6, #171 2.64e-02
CAT, chr6, #21 5.04e-05
CAT, chr6, #162 5.09e-09
CAT, chr6, #164 4.49e-10
CAT, chr6, #27 4.49e-10
CAT, chr6, #75 5.15e-02
CAT, chr6, #97 1.97e-03
CAT, chr16, #20 3.75e-04
-100 -80 -60 -40 -20 0
Position (nt)
Motif 1
Motif 2
Motif 3

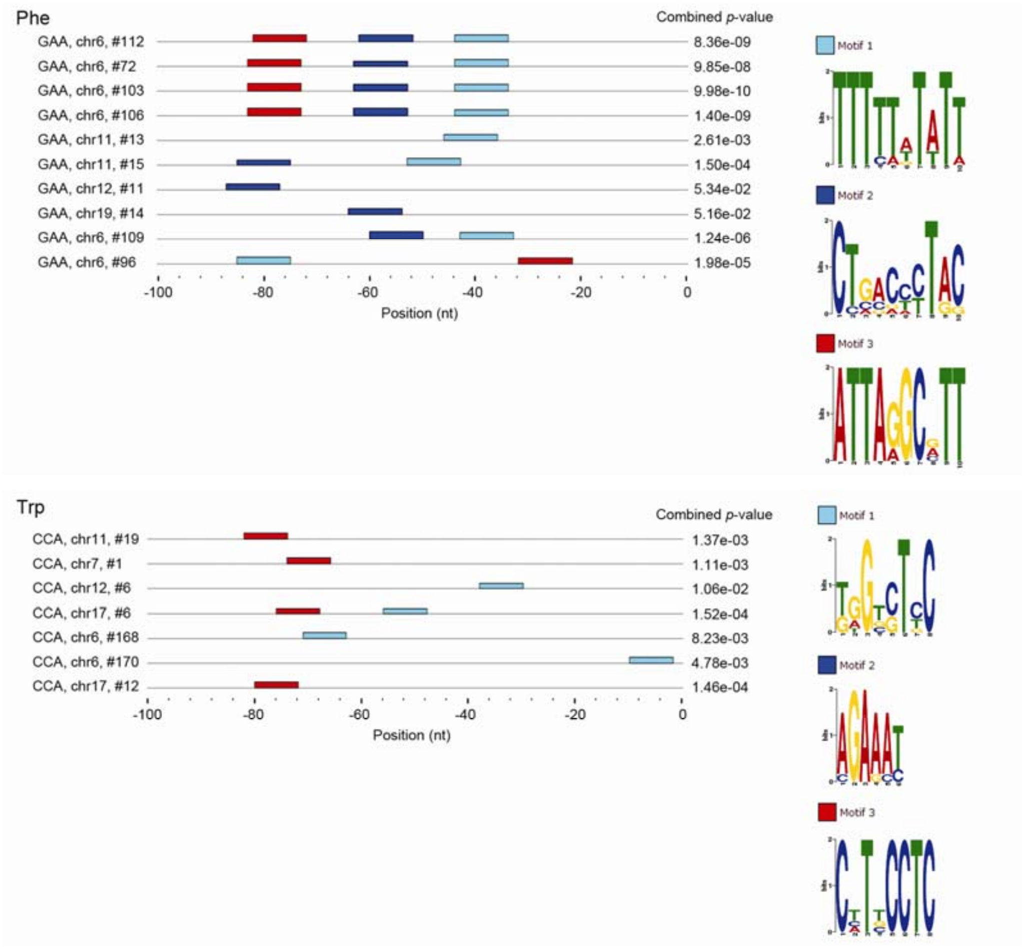


Figure S4. Majority of the human tRNA genes with the same anticodon bear conserved motifs in their upstream regions as identified by MEME 4.4. tRNA families are grouped based on the amino acid they bear and subdivided into amino acids encoded by more than one codon (A) and by one single codon (B). In panel A, the square brackets on the left group the tRNAs with the same anticodon. Each tRNA is specified by its anticodon, chromosome location and serial chromosome number (data were retrieved from the genomic tRNA database). The horizontal lines represent the 100-nucleotide-long region upstream of the 5'-end of the mature tRNA genes; position 0 determines the first nucleotide of the mature tRNA. The positions of the three most significant motifs are colored (cyan, blue, red) and the corresponding sequence logos are presented on the right side of each group. The combined *p*-value is the product of the *p*-values of all motifs detected within the upstream sequence; they are inversely proportional to the significance of the motifs. Motifs with *p*<0.05 were considered as significant.

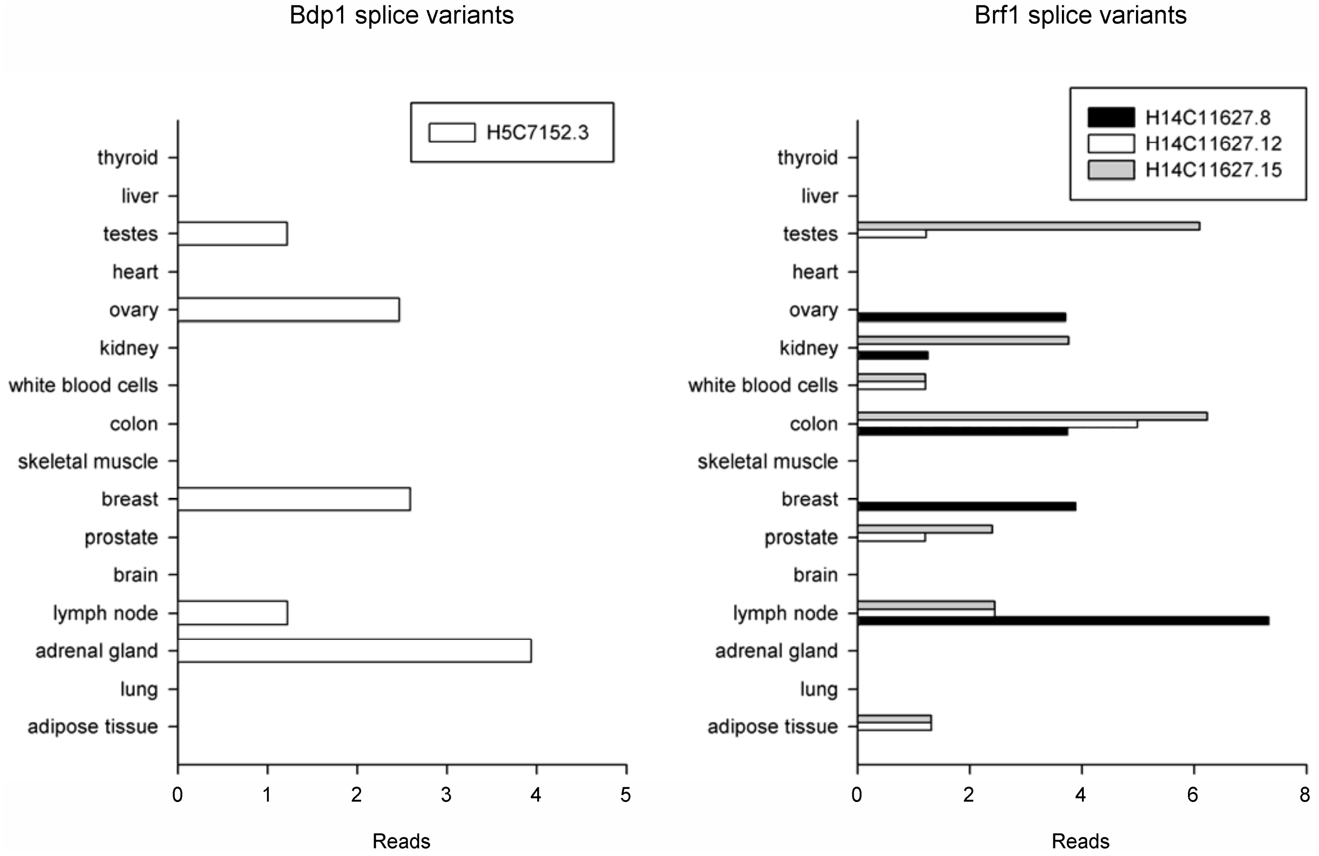
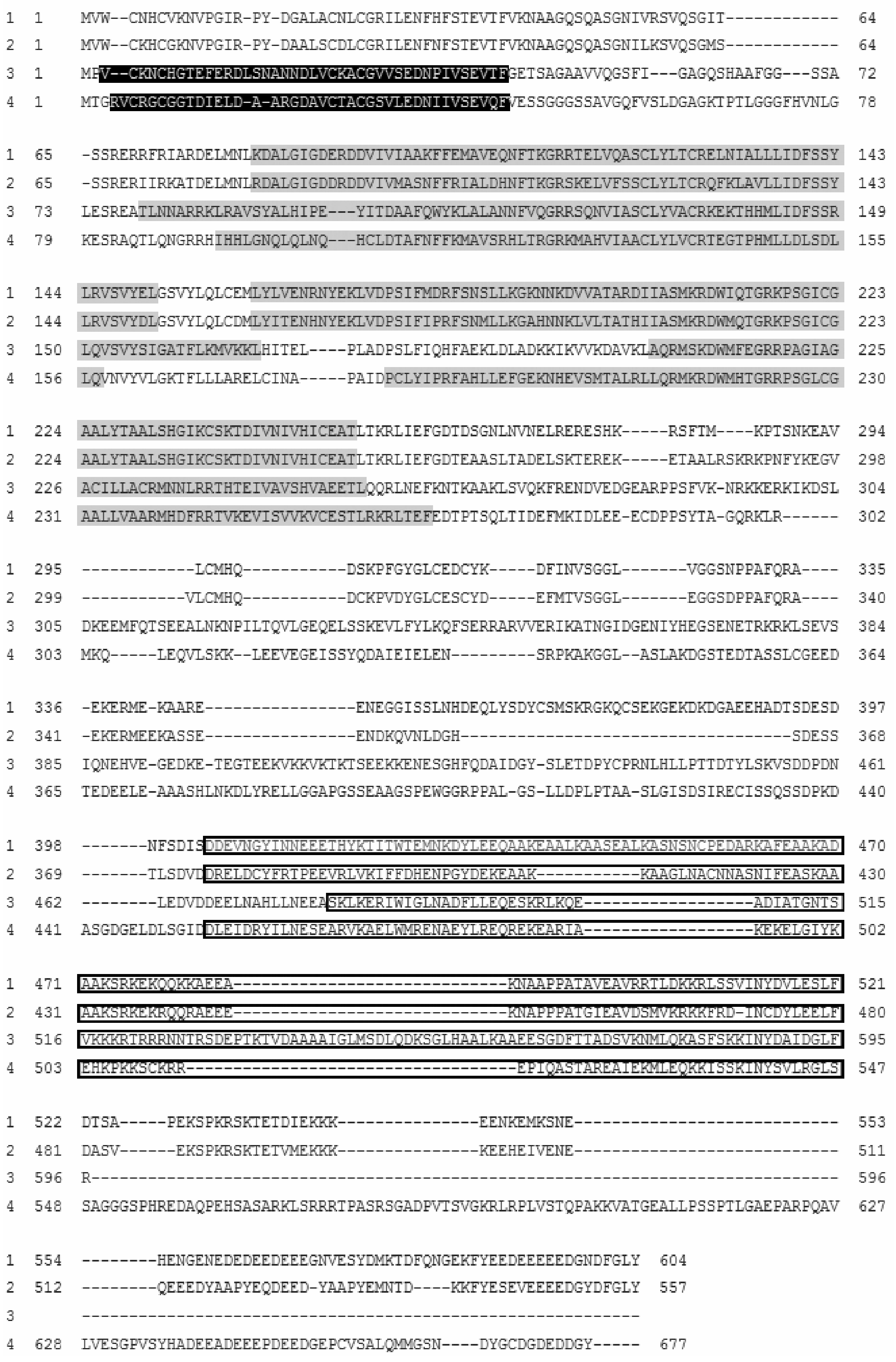


Figure S5. Distribution of the Bdp1 and Brf1 splice isoforms among different tissues. The read hits for specific splice variant were normalized to the total read count of the sequencing data sets (for details see the description in the Results section). Numbering of the splicing variants is according to the ECgene alternative splicing database (Table S2). Some other theoretically possible splicing variants of Bdp1 or Brf1 (Table S2) were also detected in the deep-sequencing datasets however at a very low level (less than 3 reads per 100 million reads) and were therefore not considered in the analysis.



```
1  1    MVW--CNHCVKNVPGIR-PY-DGALACNLCGRILENFHFSTEVTFVKNAAGQSQASGNIVRSVQSGIT------------  64
2  1    MVW--CKHCGKNVPGIR-PY-DAALSCDLCGRILENFNFSTEVTFVKNAAGQSQASGNILKSVQSGMS------------  64
3  1    MPV--CKNCHGTEFERDLSNANNDLVCKACGVVSEDNPIVSEVTFGETSAGAAVVQGSFI---GAGQSHAAFGG---SSA  72
4  1    MTGRVCRGCGGTDIELD-A-ARGDAVCTACGSVLEDNIIVSEVQFVESSGGGSSAVGQFVSLDGAGKTPTLGGGFHVNLG  78

1  65   -SSRERRFRIARDELMNLKDALGIGDERDDVIVIAAKFFEMAVEQNFTKGRRTELVQASCLYLTCRELNIALLLIDFSSY  143
2  65   -SSRERIIRKATDELMNLRDALGIGDDRDDVIVMASNFFRIALDHNFTKGRSKELVFSSCLYLTCRQFKLAVLLIDFSSY  143
3  73   LESREATLNNARRKLRAVSYALHIPE---YITDAAFQWYKLALANNFVQGRRSQNVIASCLYVACRKEKTHHMLIDFSSR  149
4  79   KESRAQTLQNGRRHIHHLGNQLQLNQ---HCLDTAFNFFKMAVSRHLTRGRKMAHVIAACLYLVCRTEGTPHMLLDLSDL  155

1  144  LRVSVYELGSVYLQLCEMLYLVENRNYEKLVDPSIFMDRFSNSLLKGKNNKDVVATARDIIASMKRDWIQTGRKPSGICG  223
2  144  LRVSVYDLGSVYLQLCDMLYITENHNYEKLVDPSIFIPRFSNMLLKGAHNNKLVLTATHIIASMKRDWMQTGRKPSGICG  223
3  150  LQVSVYSIGATFLKMVKKLHITEL----PLADPSLFIQHFAEKLDLADKKIKVVKDAVKLAQRMSKDWMFEGRRPAGIAG  225
4  156  LQVNVYVLGKTFLLLARELCINA-----PAIDPCLYIPRFAHLLEFGEKNHEVSMTALRLLQRMKRDWMHTGRRPSGLCG  230

1  224  AALYTAALSHGIKCSKTDIVNIVHICEATLTKRLIEFGDTDSGNLNVNELRERESHK-----RSFTM----KPTSNKEAV  294
2  224  AALYTAALSHGIKCSKTDIVNIVHICEATLTKRLIEFGDTEAASLTADELSKTEREK-----ETAALRSKRKPNFYKEGV  298
3  226  ACILLACRMNNLRRTHTEIVAVSHVAEETLQQRLNEFKNTKAAKLSVQKFRENDVEDGEARPPSFVK-NRKKERKIKDSL  304
4  231  AALLVAARMHDFRRTVKEVISVVKVCESTLRKRLTEFEDTPTSQLTIDEFMKIDLEE-ECDPPSYTA-GQRKLR------  302

1  295  ------------LCMHQ-----------DSKPFGYGLCEDCYK-----DFINVSGGL-------VGGSNPPAFQRA----  335
2  299  -----------VLCMHQ-----------DCKPVDYGLCESCYD-----EFMTVSGGL-------EGGSDPPAFQRA----  340
3  305  DKEEMFQTSEEALNKNPILTQVLGEQELSSKEVLFYLKQFSERRARVVERIKATNGIDGENIYHEGSENETRKRKLSEVS  384
4  303  MKQ-----LEQVLSKK--LEEVEGEISSYQDAIEIELEN---------SRPKAKGGL--ASLAKDGSTEDTASSLCGEED  364

1  336  -EKERME-KAARE----------------ENEGGISSLNHDEQLYSDYCSMSKRGKQCSEKGEKDKDGAEEHADTSDESD  397
2  341  -EKERMEEKASSE----------------ENDKQVNLDGH-----------------------------------SDESS  368
3  385  IQNEHVE-GEDKE-TEGTEEKVKKVKTKTSEEKKENESGHFQDAIDGY-SLETDPYCPRNLHLLPTTDTYLSKVSDDPDN  461
4  365  TEDEELE-AAASHLNKDLYRELLGGAPGSSEAAGSPEWGGRPPAL-GS-LLDPLPTAA-SLGISDSIRECISSQSSDPKD  440

1  398  -------NFSDISDDEVNGYINNEEETHYKTIIWTEMNKDYLEEQAAKEAALKAASEALKASNSNCPEDARKAFEAAKAD  470
2  369  -------TLSDVDDRELDCYFRTPEEVRLVKIFFDHENPGYDEKEAAK-----------KAAGLNACNNASNIFEASKAA  430
3  462  --------LEDVDDEELNAHLLNEEASKLKERIWIGLNADFLLEQESKRLKQE------------------ADIATGNTS  515
4  441  ASGDGELDLSGIDDLEIDRYILNESEARVKAELWMRENAEYLREQREKEARIA------------------KEKELGIYK  502

1  471  AAKSRKEKQQKKAEEA-----------------------------KNAAPPATAVEAVRRTLDKKRLSSVINYDVLESLF  521
2  431  AAKSRKEKRQQRAEEE-----------------------------KNAPPPATGIEAVDSMVKRKKFRD-INCDYLEELF  480
3  516  VKKKRTRRRNNTRSDEPTKTVDAAAAIGLMSDLQDKSGLHAALKAAEESGDFTTADSVKNMLQKASFSKKINYDAIDGLF  595
4  503  EHKPKKSCKRR-----------------------------------EPIQASTAREAIEKMLEQKKISSKINYSVLRGLS  547

1  522  DTSA-----PEKSPKRSKTETDIEKKK---------------EENKEMKSNE----------------------------  553
2  481  DASV------EKSPKRSKTETVMEKKK---------------KEEHEIVENE----------------------------  511
3  596  R-------------------------------------------------------------------------------  596
4  548  SAGGGSPHREDAQPEHSASARKLSRRRTPASRSGADPVTSVGKRLRPLVSTQPAKKVATGEALLPSSPTLGAEPARPQAV  627

1  554  --------HENGENEDEDEEDEEEGNVESYDMKTDFQNGEKFYEEDEEEEEDGNDFGLY  604
2  512  --------QEEEDYAAPYEQDEED-YAAPYEMNTD----KKFYESEVEEEEDGYDFGLY  557
3       -----------------------------------------------------------
4  628  LVESGPVSYHADEEADEEEPDEEDGEPCVSALQMMGSN----DYGCDGDEDDGY-----  677
```

Figure S6. Sequence alignment of the two putative Brf1-homologues from *Arabidopsis thaliana*. 1- AT3G09360 2 - AT2G45100, 3 – Brf1 from *Saccharomyces cerevisiae*, and 4 – Brf1 from *Homo sapiens*. Conserved regions are highlighted using the same color code as in Figure 4: black – represents the TFIIB zinc-binding domain, gray - cyclin fold boxes and white box - Brf1-like TBP-binding domain.