Supporting Materials for "Core promoter T-blocks correlate with gene expression levels in *C. elegans*"

Uladzislau Hryshkevich, Tamar Hashimshony and Itai Yanai

Supplemental Information

Figure captions

Figure S1. T-blocks and A-blocks have similar effect on nucleosome occupancy. To compare the effect of T-blocks and A-blocks on nucleosome positioning, we composed 4 groups of genes: 1. genes with many (≥ 4.5) T-blocks and many (≥ 3.3) A-blocks, 2. genes with many (≥ 4.5) T-blocks, but few (≤ 1.4) A-blocks, 3. genes with few (≤ 1.9) T-blocks and many (≥ 3.3) A-blocks, and 4. genes with few (≤ 1.9) T-blocks and few (≤ 1.4) A-blocks. The nucleosome occupancy of different groups was compared by Kolmagorov-Smirnov test. The distribution of mean $\log_2$ nucleosome occupancy of the core promoters in the four gene groups. *P*-values indicate the result of the Komagorov-Smirnov test between each pair of distributions.

Figure S2. Correlation between T-block occurrences and expression level in expression data from adult and L1 (first larvae) worms. In a similar fashion to Figure 4, for each set of genes with a given level of expression, the distribution of the number of T-blocks is shown as a heat map. Twenty equally populated groups of genes were defined based upon their gene expression levels. Most genes have a low level of expression, and thus the ten groups with the lowest expression levels were merged into one, '1' group. The remaining ten are shown in the figures as groups 11 through 20. The distribution of T-blocks of each category was compared with that of the '1' group and the P-value is indicated to the right. B. same as A for A-blocks. C. and D. are the same as A and B but for expression data from L1 worms. Expression data were obtained from GEO DataSets (adult expression: GSM543285, GSM543286 and GSM543287, L1 expression: GSM146422, GSM146423 and GSM147330) (Irazoqui et al. 2010; Kirienko and Fay 2007) .

Figure S3. The correlation between T-block occurrences and expression level is unique to the core promoter. The presentation is the same as in Figure 5. For each of 15 gene regions, including the proximal promoter, starts and ends of the first, second and third exons and introns, 3'UTR, and the 3' adjacent and 3' distal end regions the T-block and A-block frequency is shown for different expression groups. For exons and introns, Ensembl annotation was used to identify the 1^{st}, 2^{nd}, or 3^{rd}, ignoring alternatively spliced introns/exons.

Figure S4. 3'UTR length correlates with gene expression level. Genes were binned according to the length of the 3'UTR. The boxplots show the distribution of expression level for each 3'UTR length bin.

Figure S5. The T/A rich tail of the SL1 motif is not sufficient to explain the correlation between the number of T-blocks and the presence of the SL1 motif. We truncated the core promoter (Figure 1) by 40bp from the 3'end to remove the T/A rich region of the SL1 motif. We recomputed the SL1-fuzziness and again mapped it onto the heatmap of fuzziness vs. block numbers as in figure 7. Since the results are nearly identical we conclude that the T/A rich tail alone cannot account for the observed correlation.

Figure S6. The correlation between expression levels and the number of T-blocks is not explained by misannotated operons. We repeated the analysis shown in Figure 4 on strictly non-operon genes, defined as genes with an upstream and downstream intergenic region of at least 1kb that are not defined as operon genes in Wormbase. The correlation is preserved in this stringently filtered set of genes.

Figure S7. Thymines and T-blocks are enriched relative to adenines and A-blocks in the core promoters of many metazoans. A. The distribution of the thymine fraction within the forward strands of various metazoan core promoters. **B.** T-blocks are more frequent in the core promoters of most metazoans than A-blocks. Genomes were downloaded from Ensembl (http://www.ensembl.org/).

Figure S8. The distribution of the lengths of 5'UTRs. According to Wormbase release 210, 82% of the genes have shorter than or equal to 60bp.

Tables

Table S1. Motif frequencies across *Caenorhabditis* core promoters.

Frequency of different motifs across *Caenorhabditis* core promoters are shown as percentage of the core promoters containing each motif. Motif frequencies were calculated for motifs identified with $P < 0.005$, FDR-corrected.

Motif / Species	SL1	Kozak	TATA box	Sp1	T-blocks
C. remanei	56.88	29.21	5.74	1.84	43.91
C. briggsae	63.85	33.50	4.03	2.43	41.85
C. brenneri	61.76	32.69	6.74	1.5	43.36
C. elegans	58.95	37.19	6.34	1.23	41.87
C. japonica	50.53	31.46	4.85	1.2	22.98

Table S2. Distribution of T-blocks and A-blocks in the core promoters.

Numbers of genes with different quantities of T- and A-blocks are indicated. Difference between two distributions ($P<10^{-307}$) were computed using the Kolmagorov-Smirnov test.

# of blocks	T-blocks	A-blocks
0	670	1039
1	1517	2623
2	2324	3746
3	2820	3736
4	3129	3043
5	2712	1784
6	2137	892
7	1276	331
8	556	87
9	139	15
10	19	3

Table S3. Combinatorial analysis shows that SL1 and T-blocks are enriched for co-occurences, while both are depleted in co-occurences with TATA-box.

P-values are computed using the hypergeometric distribution. Occurrences were defined using the MEME-motifs and a MAST threshold of $P<0.005$, FDR-corrected. Green *P*-values indicate enrichment in motif co-occurrence, while red *P*-values indicate depletion in motif co-occurrences.

Motifs / Species	SL1 and T-blocks co-occurences	SL1 and TATA-box co-occurences	T-blocks and TATA-box co-occurences
C. remanei	10^{-16}	10^{-72}	10^{-150}
C. briggsae	10^{-16}	10^{-52}	10^{-89}
C. brenneri	10^{-16}	10^{-115}	10^{-101}
C. elegans	10^{-16}	10^{-35}	10^{-87}
C. japonica	10^{-16}	10^{-22}	10^{-32}

Table S4. GO analysis. The table indicates the top 15 GO terms enriched in genes containing both SL1 and T-blocks, and genes containing a TATA-box.

Blue GO terms indicate association with development, growth and reproduction; while red GO terms indicate association with stress.

#	SL1-T-blocks Gene Ontology Terms	*P* <	TATA-box Gene Ontology Terms	*P* <
1	reproduction	10^{-125}	nucleosome assembly	10^{-125}
2	embryonic development ending in birth or egg hatching	10^{-11}	nucleosome	10^{-12}
3	nematode larval development	10^{-11}	extracellular region	10^{-10}
4	growth	10^{-10}	lysozyme activity	10^{-5}
5	cell division	10^{-10}	aminoacylase activity	10^{-4}
6	structural constituent of ribosome	10^{-10}	transferase activity, transferring hexosyl groups	10^{-4}
7	cytokinesis	10^{-10}	monooxygenase activity	10^{-4}
8	ribosome	10^{-10}	binding	10^{-4}
9	translation	10^{-10}	iron ion binding	10^{-4}
10	intracellular	10^{-8}	body morphogenesis	10^{-4}
11	hermaphrodite genitalia development	10^{-7}	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase activity	10^{-4}
12	positive regulation of growth rate	10^{-6}	enterobactin biosynthetic process	10^{-4}
13	nucleic acid binding	10^{-6}	DNA binding	10^{-4}
14	molting cycle, collagen and cuticulin-based cuticule	10^{-5}	sugar binding	10^{-4}
15	negative regulation of vulval development	10^{-4}	heme binding	10^{-3}

Table S5. SNPs identified between the CB4856 (Hawaiian) and the reference N2 (Bristol) *C. elegans* strain. 78.119 SNPs were identified with a coverage of at least 5 and no heterogeneity. Chromosomal location is relative to WormBase release 195. Table is included as an Excel spreadsheet.

Reference:

Irazoqui, J.E., E.R. Troemel, R.L. Feinbaum, L.G. Luhachack, B.O. Cezairliyan, and F.M. Ausubel. 2010. Distinct pathogenesis and host responses during infection of *C. elegans* by *P. aeruginosa* and *S. aureus*. *PLoS Pathog* **6:** e1000982.

Kirienko, N.V. and D.S. Fay. 2007. Transcriptome profiling of the *C. elegans* Rb ortholog reveals diverse developmental roles. *Dev Biol* **305:** 674-684.

Figure S1

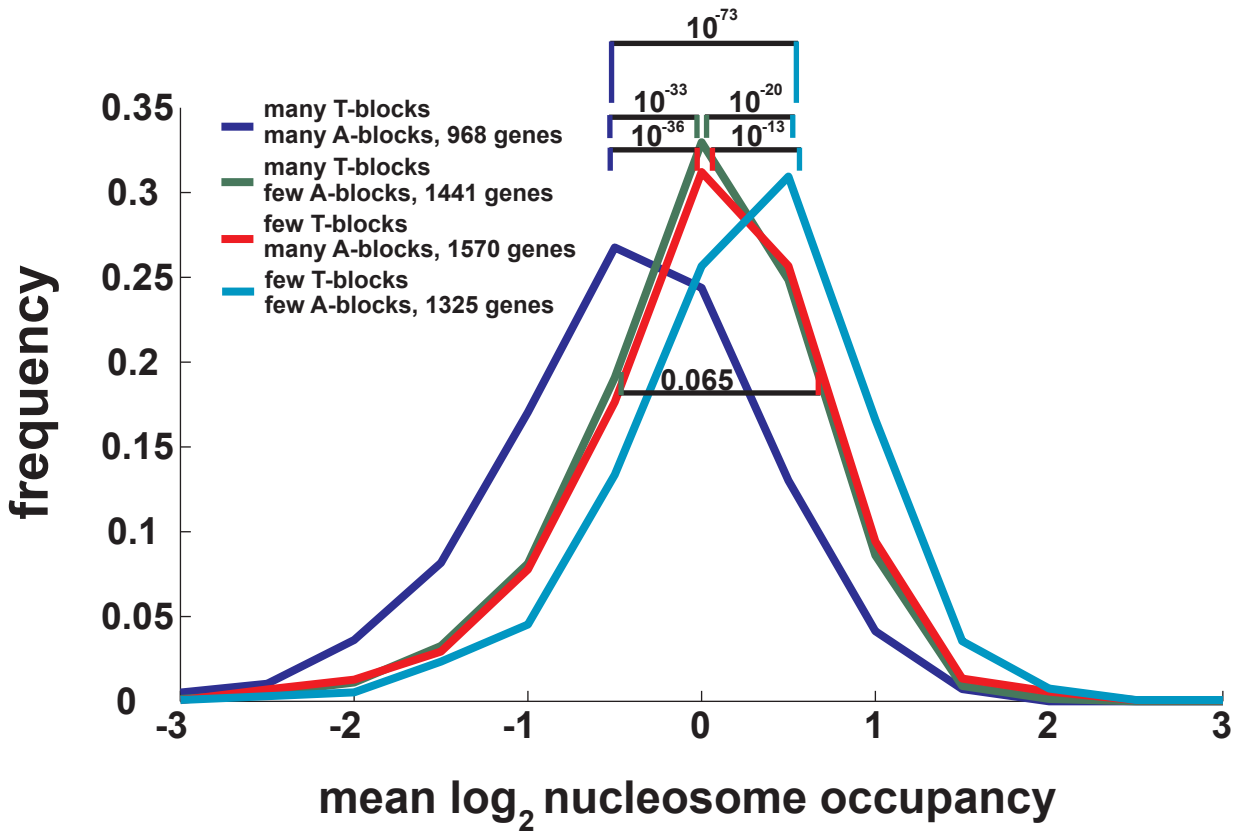

10^{-73}
10^{-33}
10^{-20}
10^{-36}
10^{-13}
0.065
many T-blocks
many A-blocks, 968 genes
many T-blocks
few A-blocks, 1441 genes
few T-blocks
many A-blocks, 1570 genes
few T-blocks
few A-blocks, 1325 genes
frequency
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-3
-2
-1
0
1
2
3
mean $\log_2$ nucleosome occupancy

Figure S2

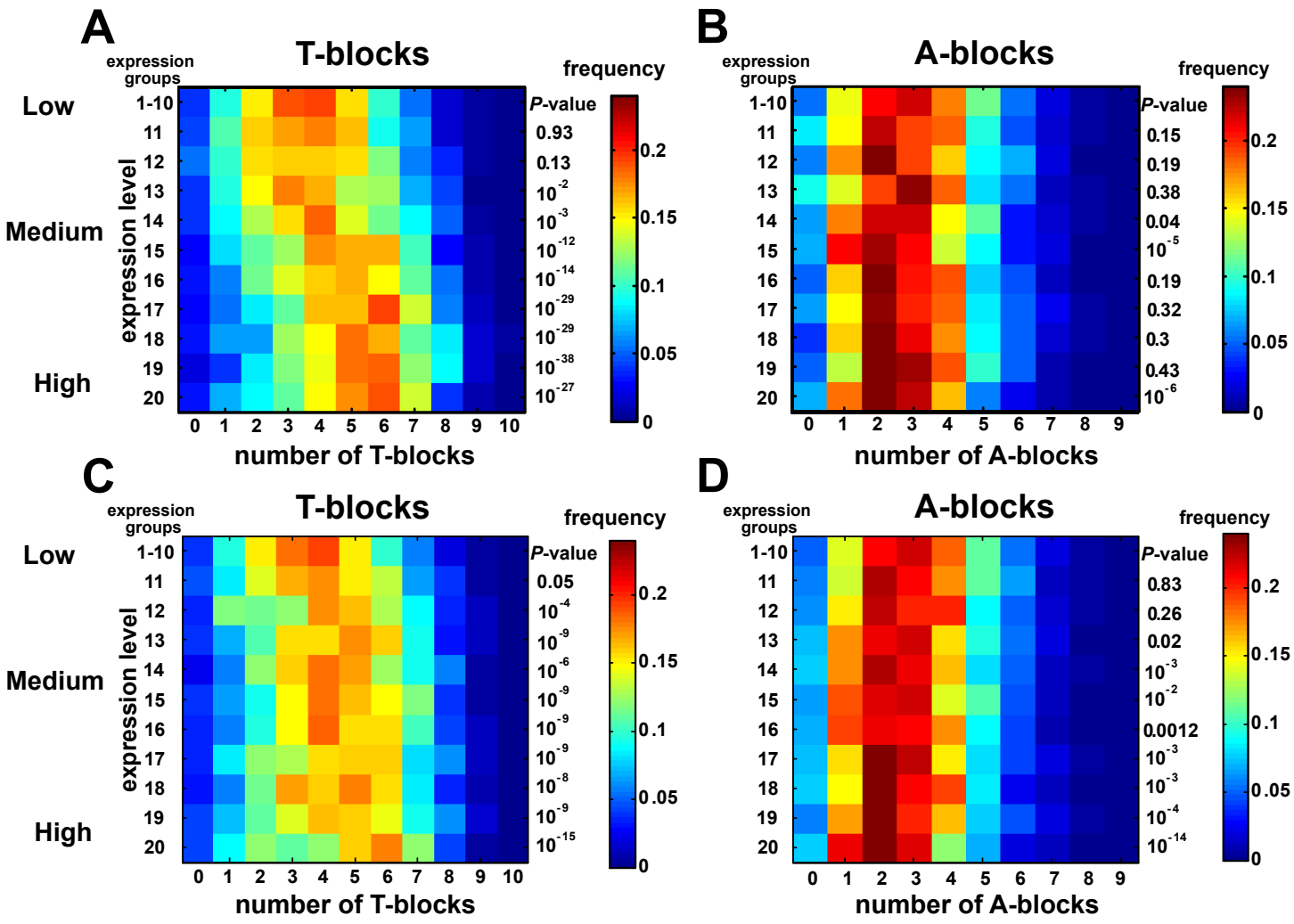

A
T-blocks
expression groups
frequency
Low
Medium
High
expression level
1-10
11
12
13
14
15
16
17
18
19
20
P-value
0.93
0.13
10-2
10-3
10-12
10-14
10-29
10-29
10-38
10-27
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
number of T-blocks
B
A-blocks
expression groups
frequency
1-10
11
12
13
14
15
16
17
18
19
20
P-value
0.15
0.19
0.38
0.04
10-5
0.19
0.32
0.3
0.43
10-6
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9
number of A-blocks
C
T-blocks
expression groups
frequency
Low
Medium
High
expression level
1-10
11
12
13
14
15
16
17
18
19
20
P-value
0.05
10-4
10-9
10-6
10-9
10-9
10-9
10-8
10-9
10-15
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9 10
number of T-blocks
D
A-blocks
expression groups
frequency
1-10
11
12
13
14
15
16
17
18
19
20
P-value
0.83
0.26
0.02
10-3
10-2
0.0012
10-3
10-3
10-4
10-14
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8 9
number of A-blocks

Figure S3

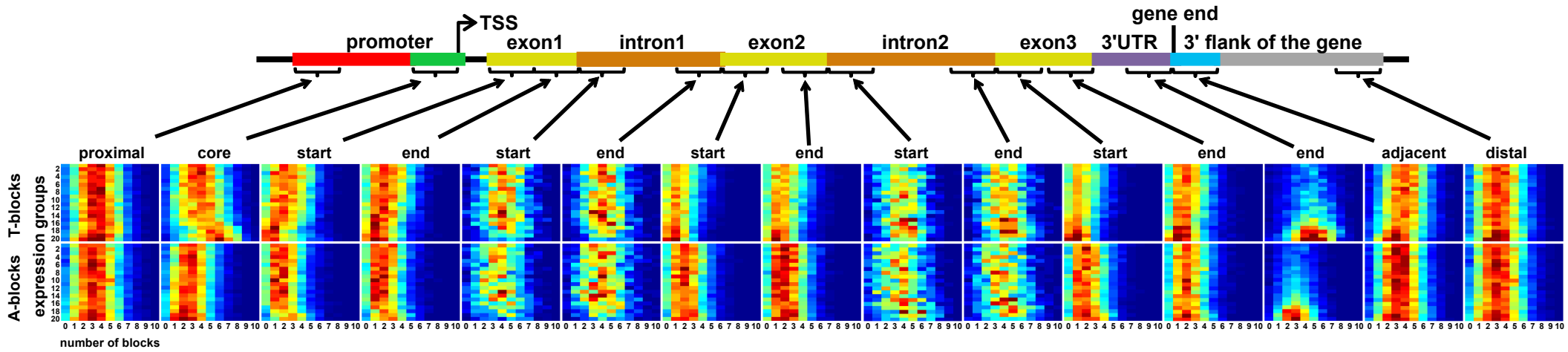

TSS
gene end
promoter
exon1
intron1
exon2
intron2
exon3
3'UTR
3' flank of the gene
proximal
core
start
end
start
end
start
end
start
end
start
end
end
adjacent
distal
T-blocks
A-blocks
expression groups
number of blocks

Figure S4

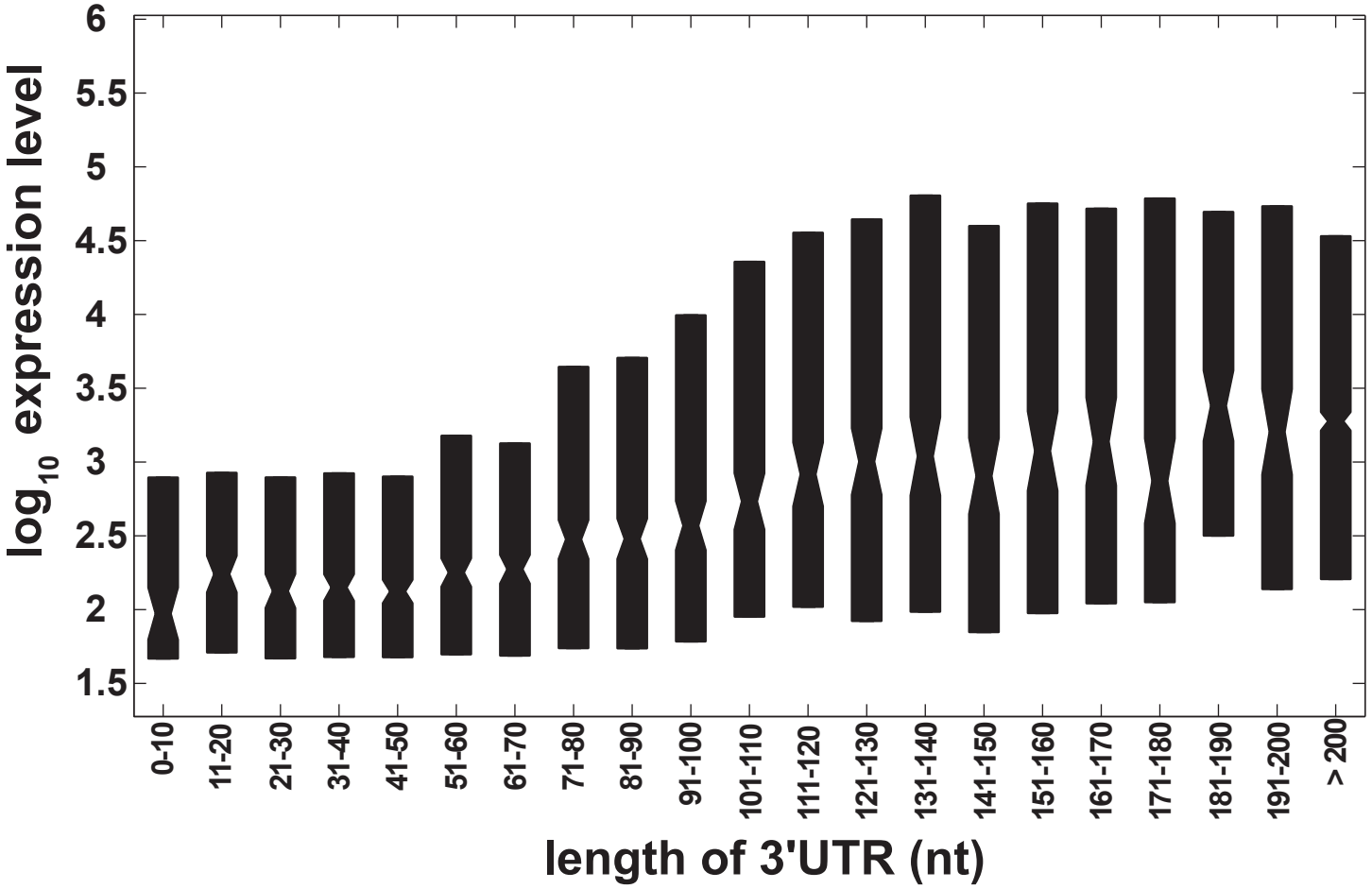

log10 expression level
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
0-10
11-20
21-30
31-40
41-50
51-60
61-70
71-80
81-90
91-100
101-110
111-120
121-130
131-140
141-150
151-160
161-170
171-180
181-190
191-200
> 200
length of 3'UTR (nt)

Figure S5

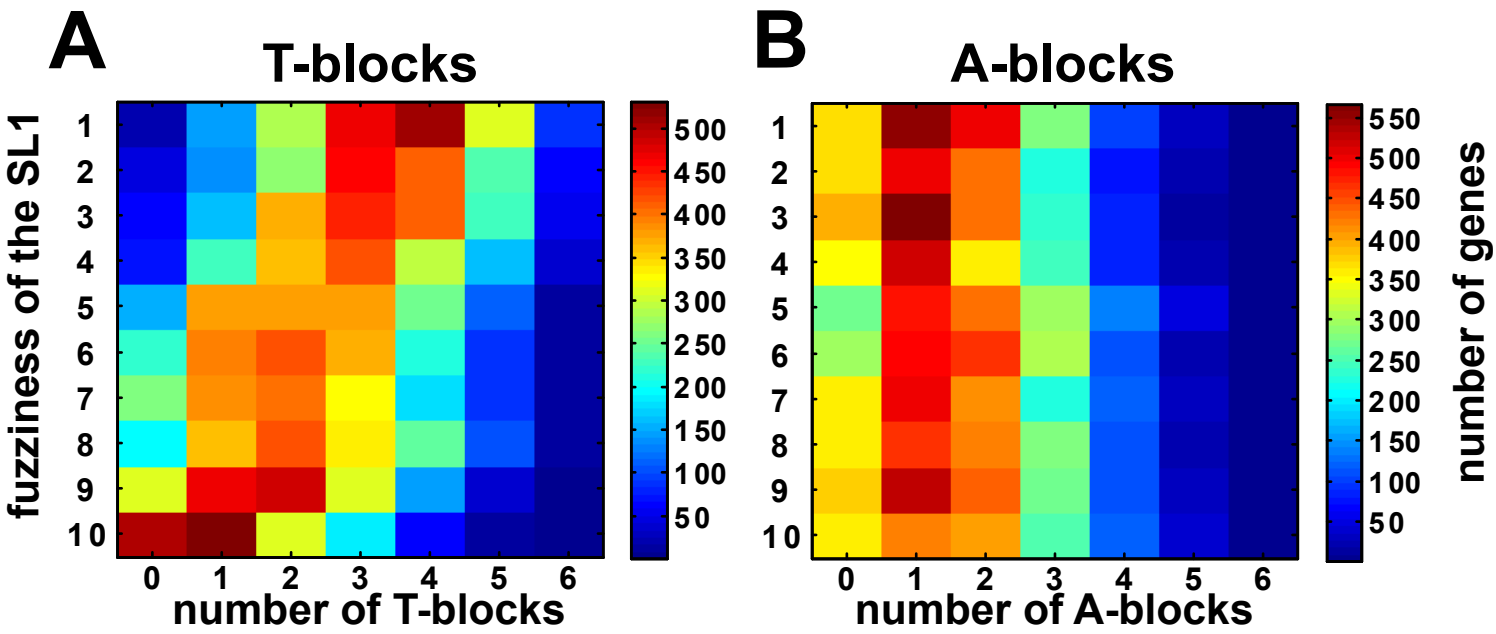

A
T-blocks
fuzziness of the SL1
1
2
3
4
5
6
7
8
9
10
0
1
2
3
4
5
6
number of T-blocks
500
450
400
350
300
250
200
150
100
50
B
A-blocks
number of A-blocks
550
500
450
400
350
300
250
200
150
100
50
number of genes

Figure S6

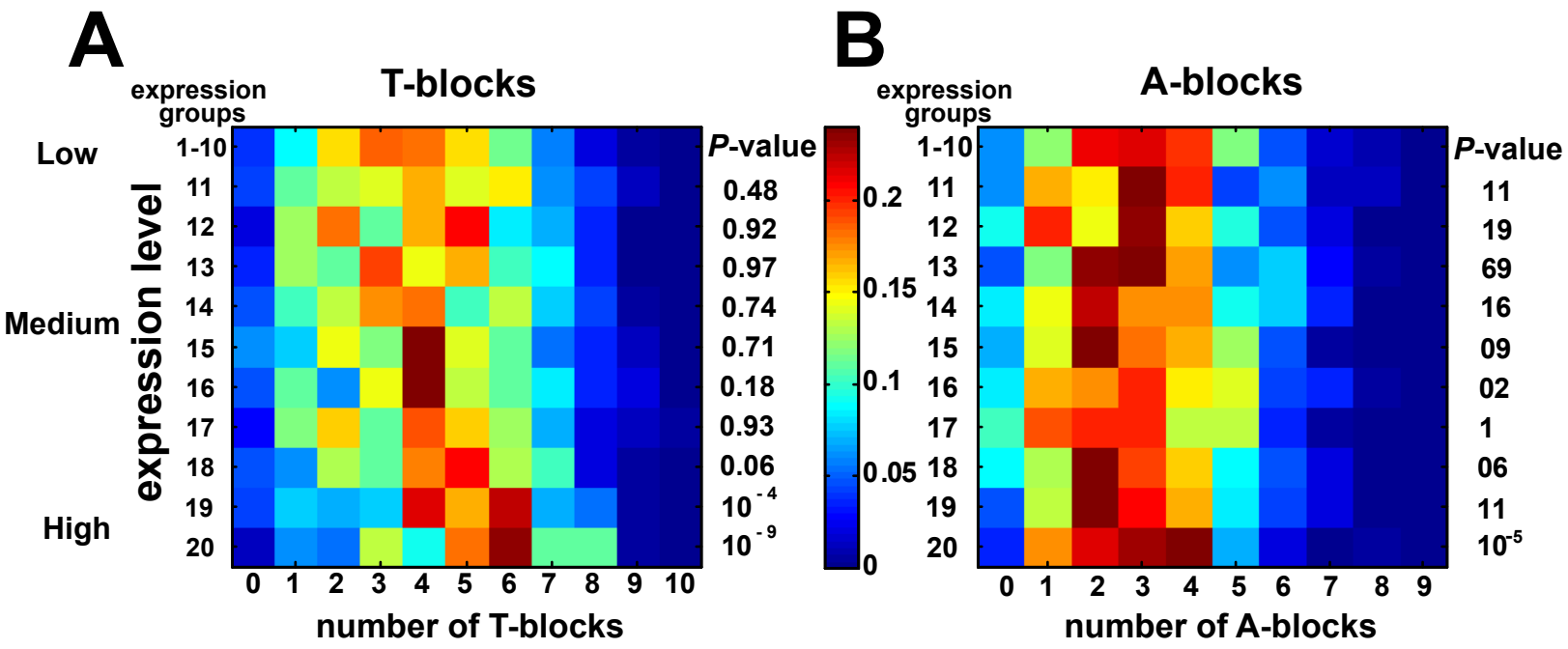

A
T-blocks
expression groups
expression level
Low
Medium
High
1-10
11
12
13
14
15
16
17
18
19
20
0 1 2 3 4 5 6 7 8 9 10
number of T-blocks
P-value
0.48
0.92
0.97
0.74
0.71
0.18
0.93
0.06
10 -4
10 -9
B
A-blocks
expression groups
1-10
11
12
13
14
15
16
17
18
19
20
0 1 2 3 4 5 6 7 8 9
number of A-blocks
P-value
11
19
69
16
09
02
1
06
11
10-5
0.2
0.15
0.1
0.05
0

Figure S7

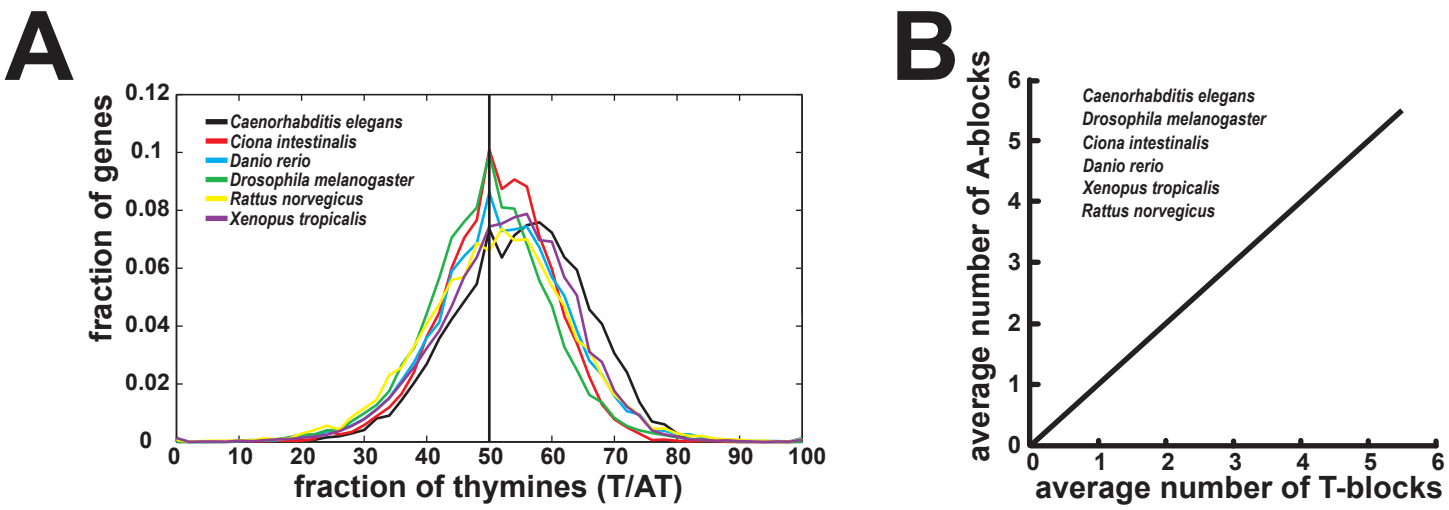
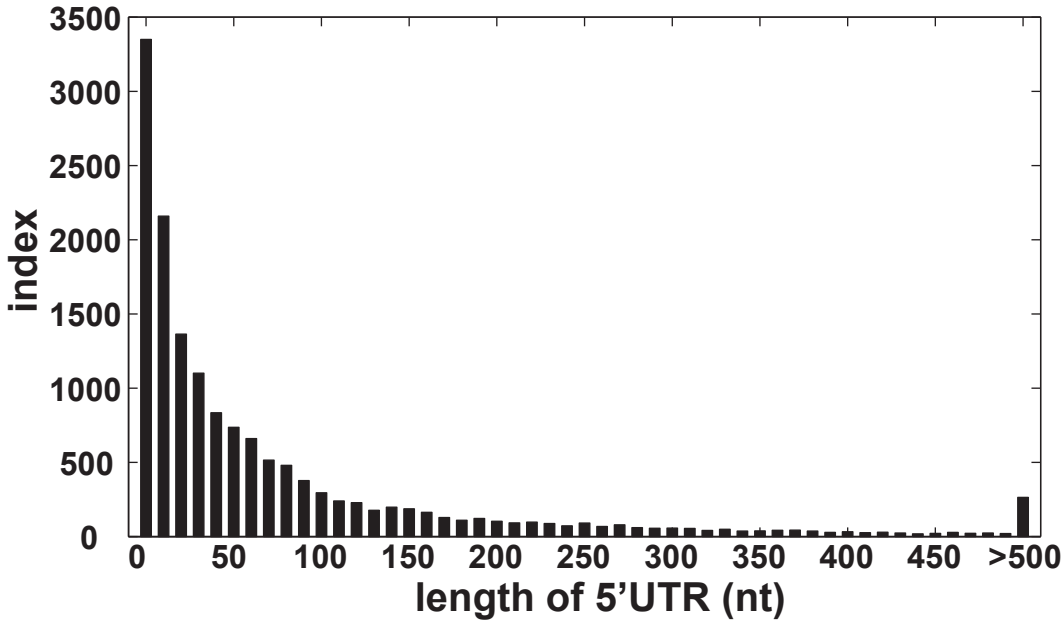

A
fraction of genes
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
100
fraction of thymines (T/AT)
Caenorhabditis elegans
Ciona intestinalis
Danio rerio
Drosophila melanogaster
Rattus norvegicus
Xenopus tropicalis
B
average number of A-blocks
6
5
4
3
2
1
0
0
1
2
3
4
5
6
average number of T-blocks
Caenorhabditis elegans
Drosophila melanogaster
Ciona intestinalis
Danio rerio
Xenopus tropicalis
Rattus norvegicus

Figure S8


3500
3000
2500
2000
1500
1000
500
0
index
0
50
100
150
200
250
300
350
400
450
>500
length of 5'UTR (nt)