Bits and pieces about transcription start sites

Hedi Peterson

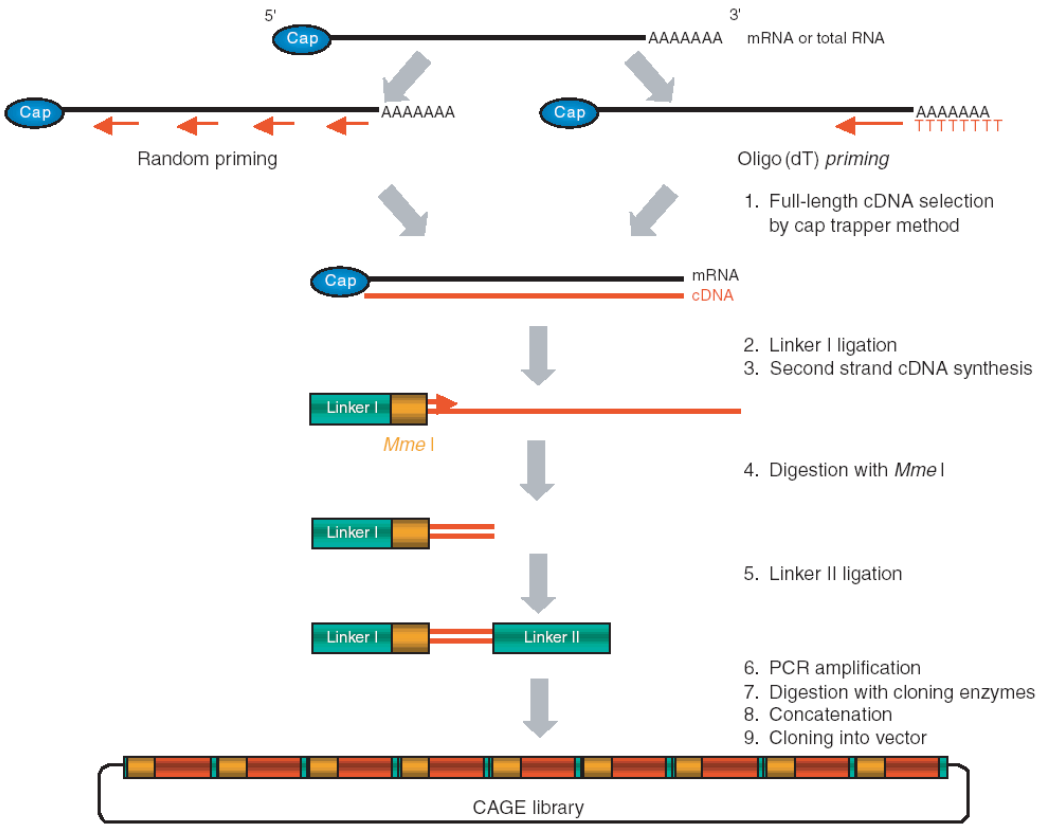
17.11.2006
JClub

Transcription start site

- Position on DNA where transcription starts - RNA polymerase (II) starts mRNA synthesis
- Transcription is triggered by transcription factors that bind to short DNA motifs upstream from the TSS
- Besides TFBS there are other common regulative elements - TATA-box, CCAAT-box, Gc-box and CpG islands

CAGE - what is it?

- **C**ap **A**nalysis **G**ene **E**xpression
- produces 20-21bp long tags



CAGE

- produces 20-21bp long tags form 5’ end of full-length cDNA
- tags will be mapped to corresponding genome
- overlapping tags from same strand form a tag cluster (TC)
- TC + surroundings form a core promoter

RIKEN - FANTOM3 - CAGE

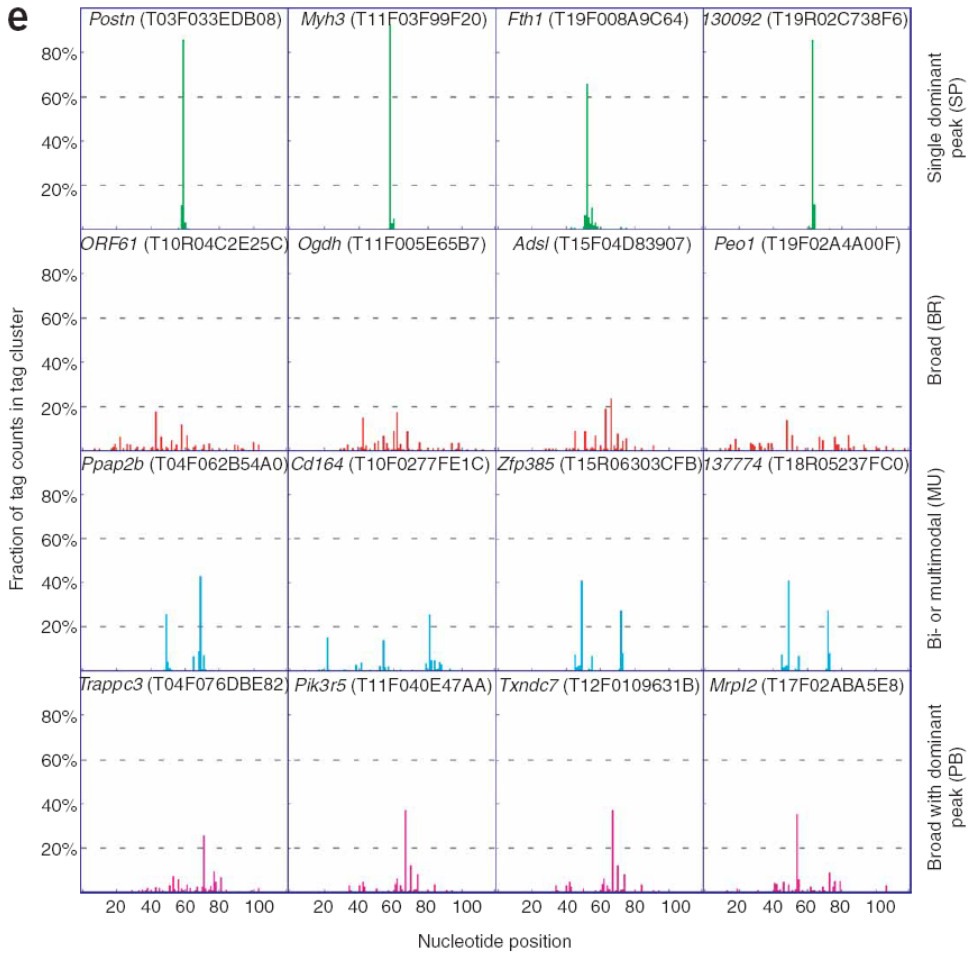
Human and mouse data

- 209(145) cDNA libraries for 23 tissues for mouse
- 43(41) libraries for human
- identified 729504 mouse and 665278 human potential TSSs (80-95% belong to TC)

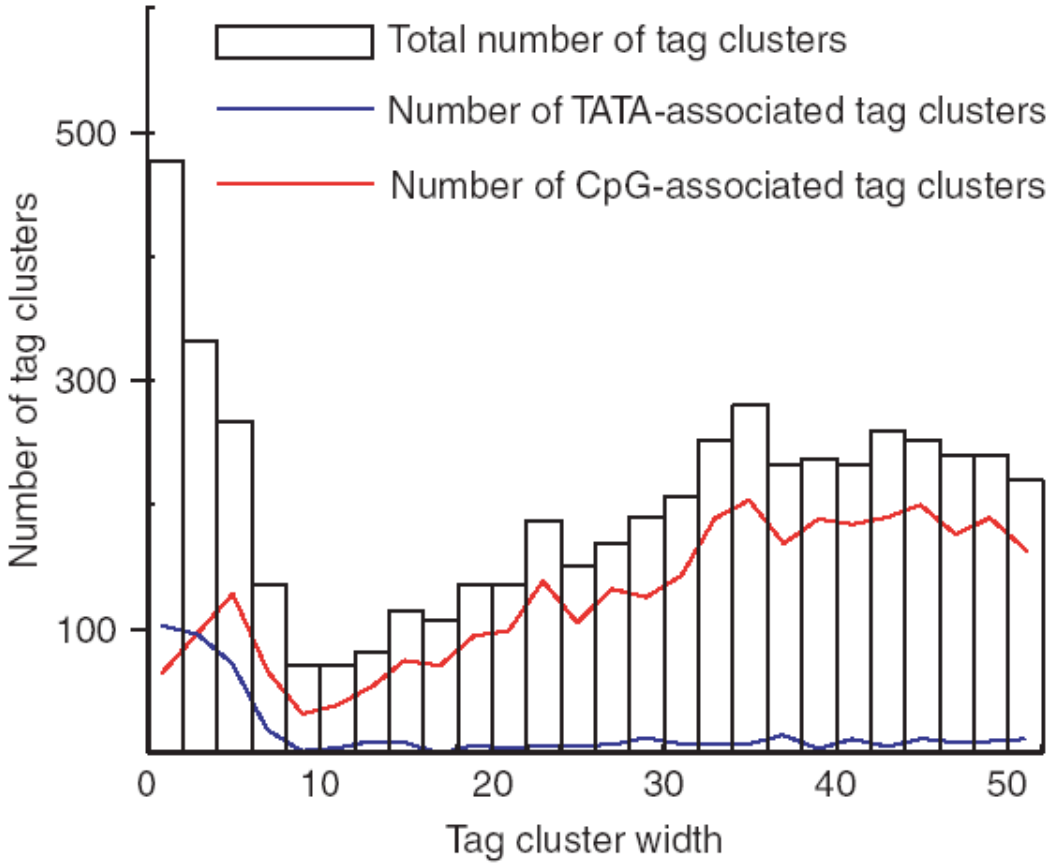
CAGE tag clusters

Four type of tag clusters (based on 8000 mouse and 6000 human TC, 100 tags per TC)

- single peak
- broad shape
- bi/multimodal
- broad with dominant peak

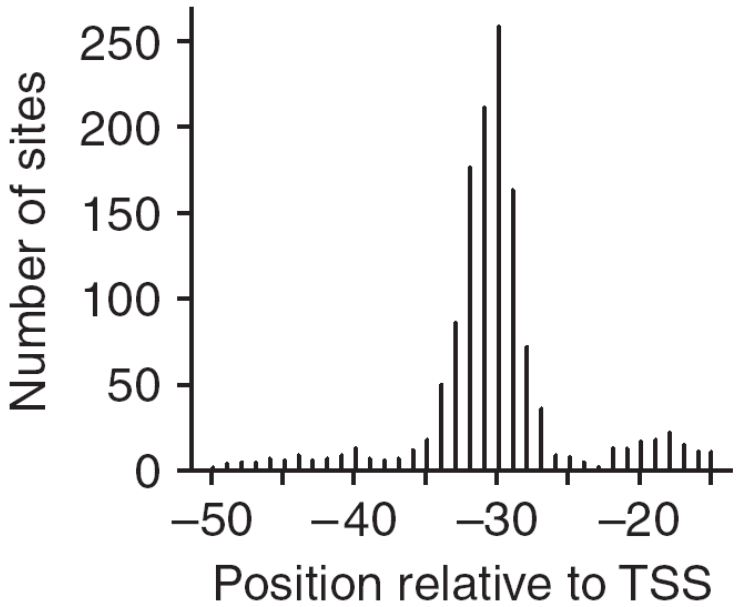


CAGE tag clusters



TATA-box

- strongly overrepresented in sharp TSS promoters
- highly conserved promoter region across species
- TATA-independent transc. init. occurs within a CpG island (90%)
- Sp1 recruits TBP when TATA-box is not present (in broad promoters)
- tissue specificity when PB promoters + strong TATA peak



CpG islands

- strongly associated with broad TSS regions
- more G nucleotides on "+" strand (could indicate promoter orientation)
- associated to house-keeping genes
- associated to bidirectional promoter activity
- central nervous system-specific promoters are especially CpG rich

Other elements

- CCAAT- and GC-box associated with sharp peak TSS class
- secondary peak in 3'UTR with Inr GGG

Statistics

Table 2 Overrepresentation and underrepresentation of transcriptional starting site sequence

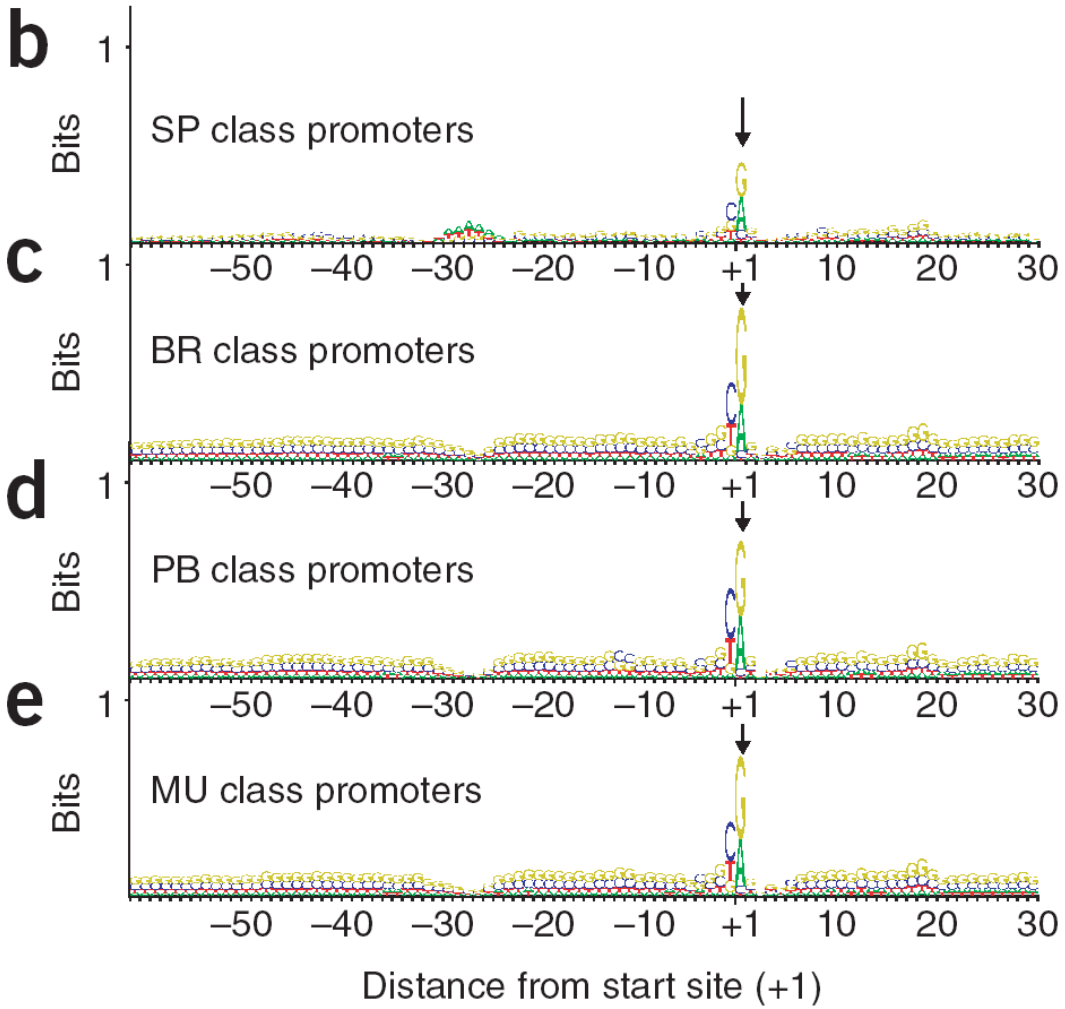
Overall analysis	SP	BR	PB	MU
TATA (all)	**3.1×10^{-73}**	***1.9×10^{-16}***	***1.8×10^{-10}***	***2.4×10^{-9}***
CCAAT (all)	0.04	0.42	0.37	0.49
GC (all)	***1×10^{-4}***	0.20	0.40	0.33
CpG (all)	***1.0×10^{-137}***	**1.4×10^{-65}**	**8.7×10^{-6}**	0.02
CpG promoters versus non-CpG promoters	**SP**	**BR**	**PB**	**MU**
TATA (no CpG)	**2.6×10^{-77}**	***1.6×10^{-16}***	***2.8×10^{-16}***	***1.0×10^{-9}***
CCAAT (no CpG)	**6.8×10^{-23}**	***9.2×10^{-16}***	0.11	0.42
GC (no CpG)	**7.8×10^{-25}**	***5.9×10^{-18}***	0.48	0.35
CpG (no TATA, CCAAT or GC)	***4.8×10^{-45}***	**4.7×10^{-17}**	**3.4×10^{-5}**	0.87

For each shape class, we determined whether a TATA box (within 50 bp) or a CCAAT, GC or CpG (within 200 bp) upstream of the start site of the clusters was present. *P* values were determined using the Fisher exact test (**Supplementary Note**). *P* values in boldface and italics indicate significant underrepresentation ($P < 0.01$); *P* values in boldface alone indicate significant overrepresentation ($P < 0.01$). In the lower part of the table, we separated pure CpG-island–overlapping promoters (without TATA, CCAAT and GC elements) and TATA, CCAAT and GC promoters (without CpG islands).

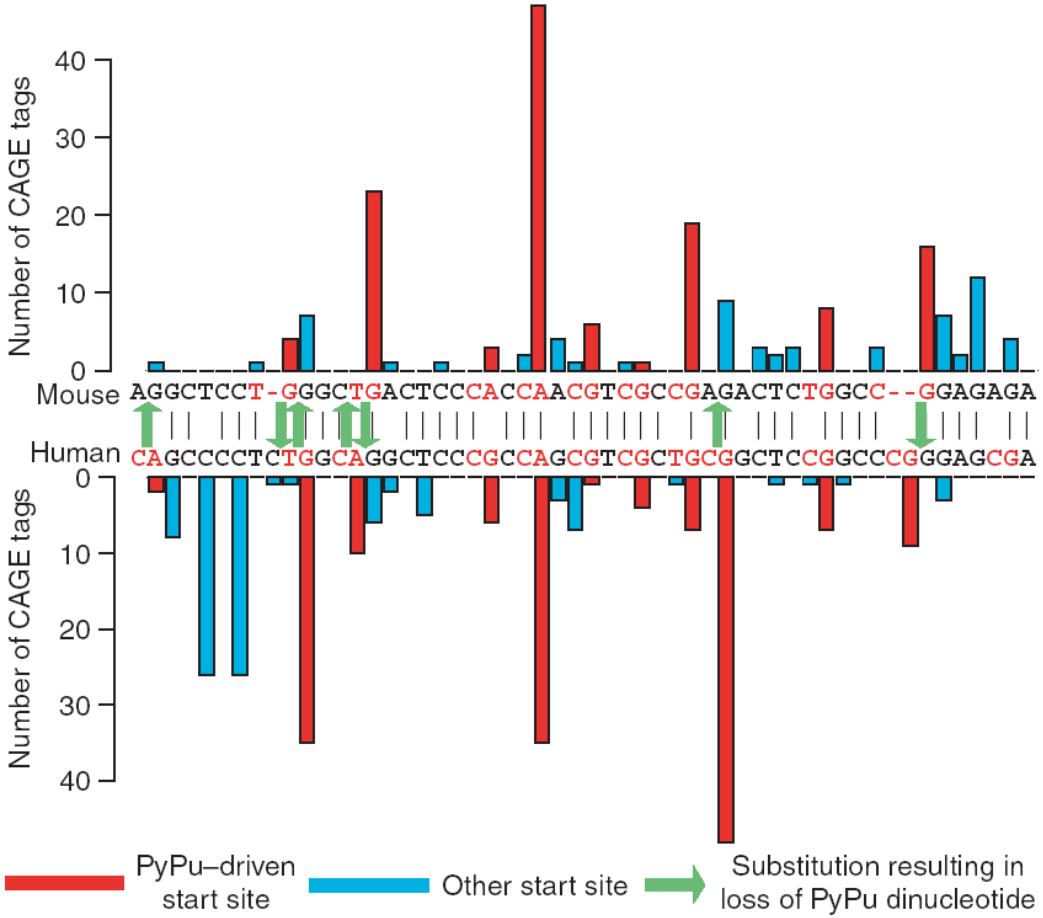
Initiator sequence (Inr element)

- previously known Inr element -> Py-Py-A-N-T/A-Py-Py
- Transcription preferentially starts (+1 pos) with a purine
- +1 pos preference for pyrimidine (A/G) (58.6%)
- CA for highly exp. transcripts
- AG, GG for rarely exp. transcripts

Inr elements for TC types



Pyrimidine/Purine substitutions



Promoter evolution

- Within 200bp upstream of TSS identity percentage declines
- TATA-containing promoters had a lower substitution rate than other 3 TC types
- human broad shape promoters have higher substitution rate than mouse promoter regions
- PyPu substitutions change the promoter expression
- Direct correlation between the global properties of pomoters and TFBS density
- 3’ UTR TSSs are at least some cases functional

TSS distances between human and mouse

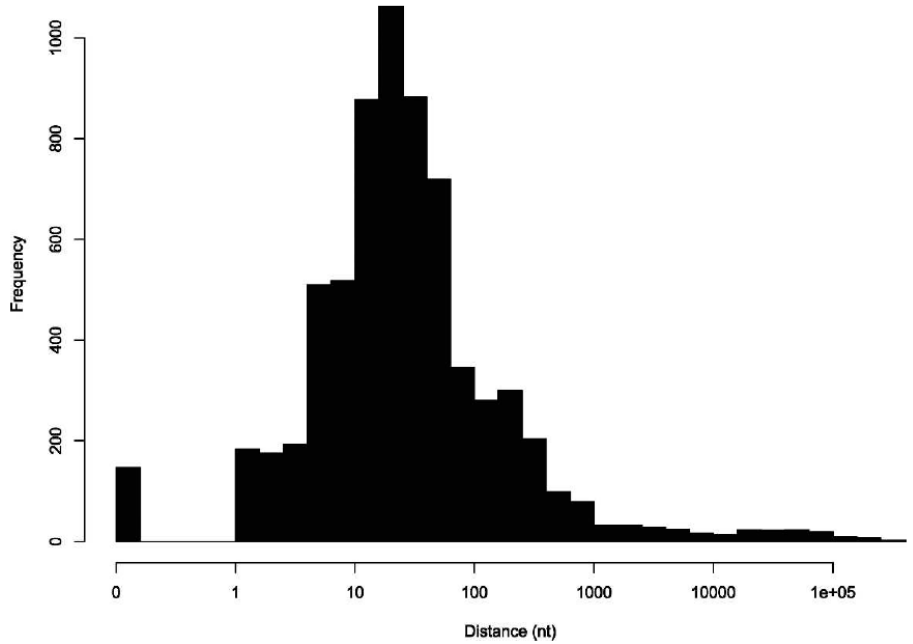
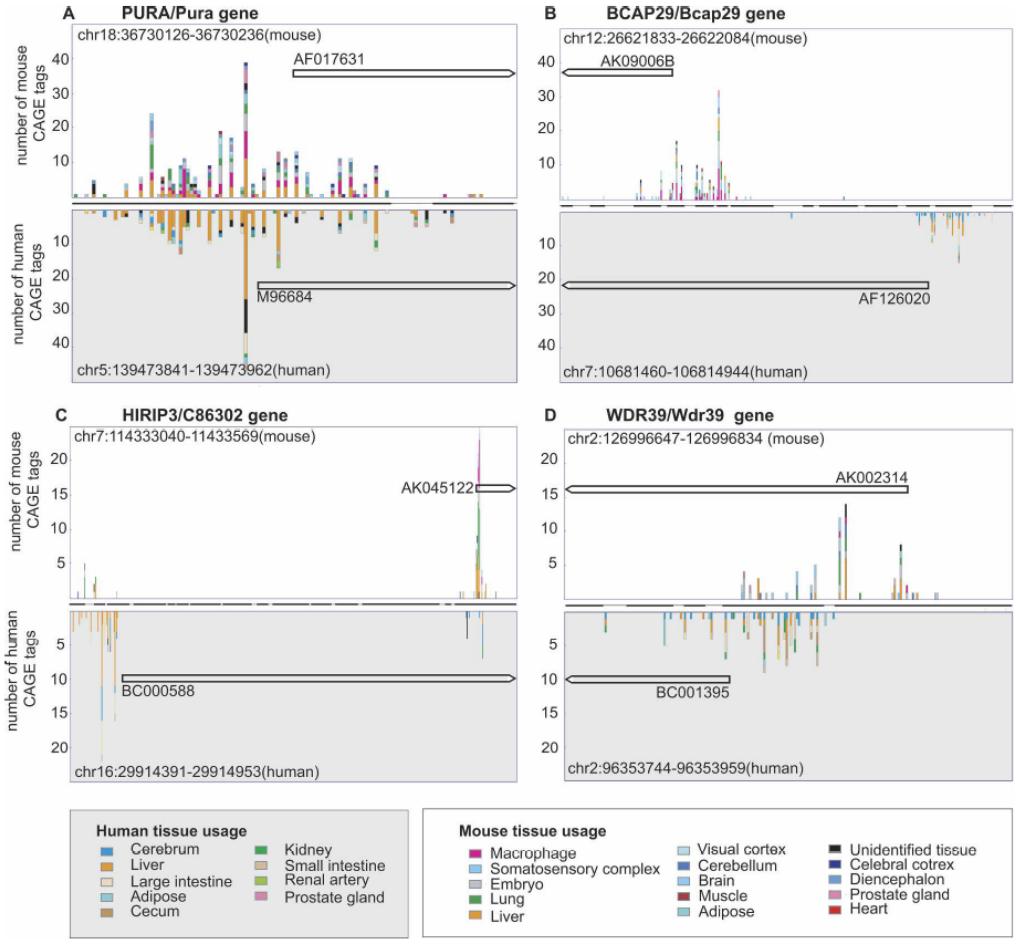


Figure 1. Histogram of distances between transcription start sites of homologous transcripts. The *x*-axis indicates the distance between the human TSS and the human position aligned to the mouse TSS.

TSS turnover



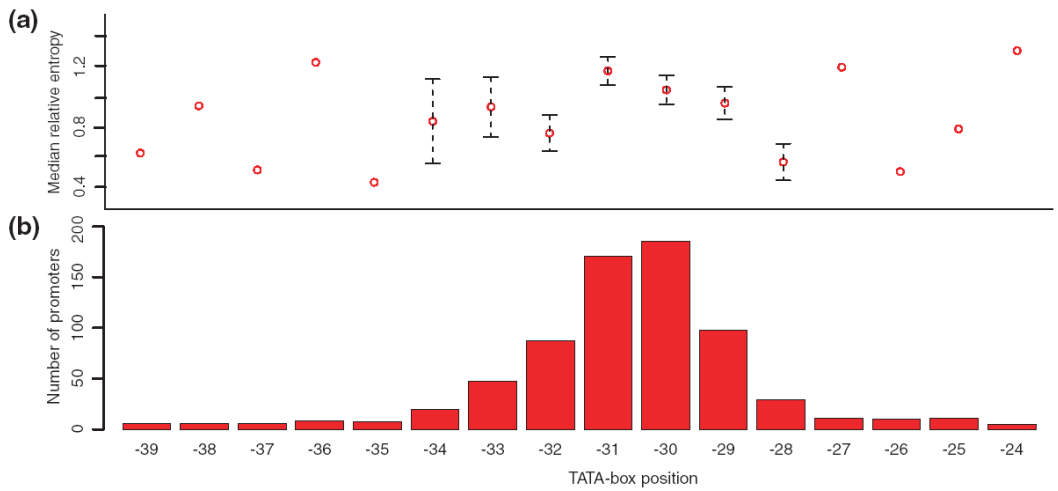
3’UTR promoters

- Specific motif - 3 guanines just before TSSs.
- Highy conserved region +40..+90bp
- Proposed sense-antisense regulatory mechanism for downstream genes
- 3’ TSS is rather tissue specific and prevalent in cerebellum and lung, reduced in embryo

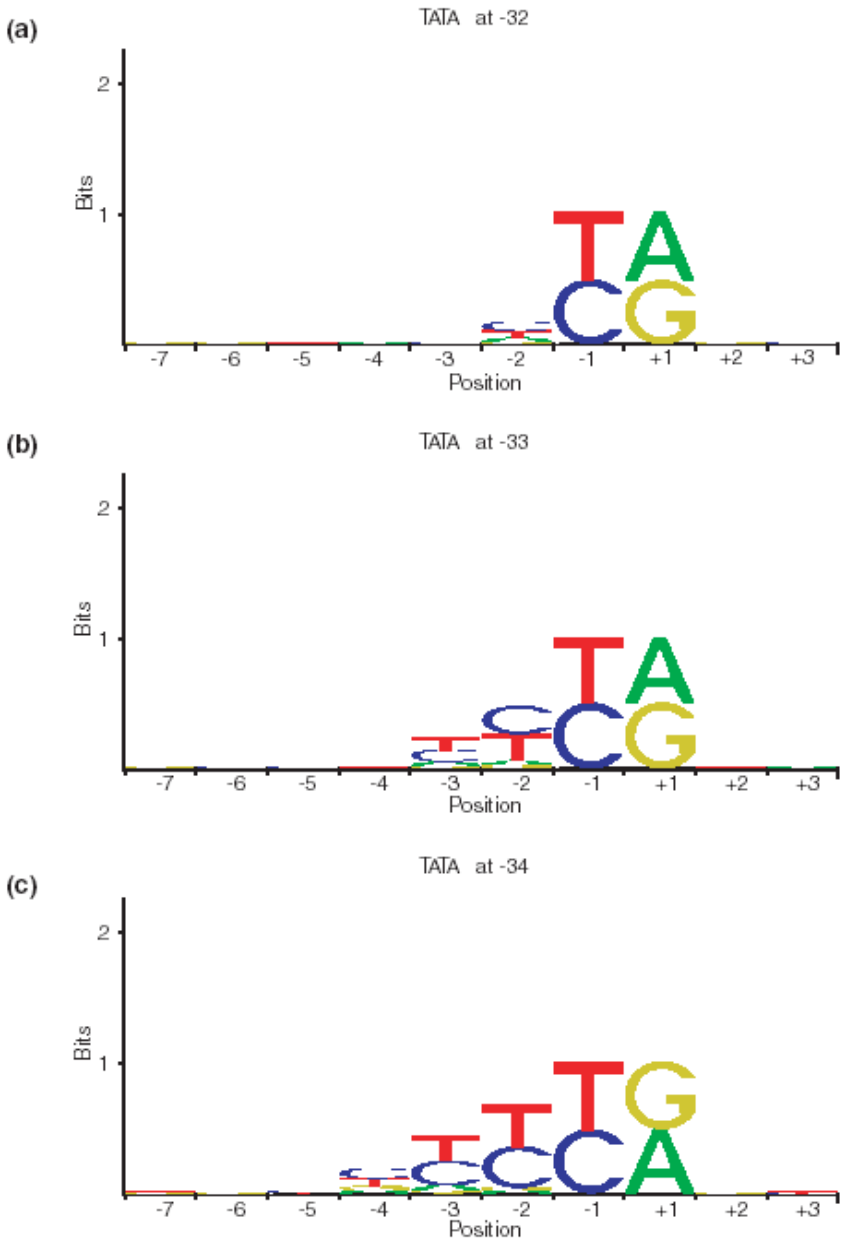
Back to TATA-box

- Tissue specificity is achived by having TATA-box -32..-29bp upstream of TSS
- positions -31 and -30 are optimal
- TATA-box motifs closer than -28bp are generally non-functional
- TATA-box at position -34 has consensus motif TATATAA (other positions have consensus TATAAA), so -34 can also be -32 TATA-box

TATA-box distribution



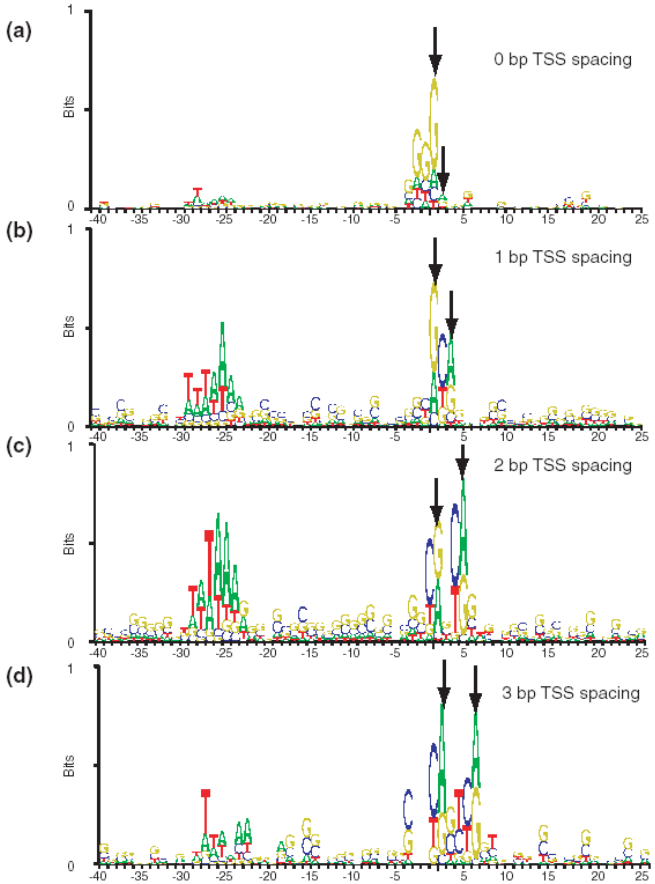
HMMs for upstream Inr



Twin-TSS

- Distance between two major TSS peaks are less than 5
- If the distance is 0 then specific Inr (GGG) is preferred

Twin TSSs


(a)
0 bp TSS spacing
(b)
1 bp TSS spacing
(c)
2 bp TSS spacing
(d)
3 bp TSS spacing
Bits
1
0
-40 -35 -30 -25 -20 -15 -10 -5 5 10 15 20 25

New knowledge

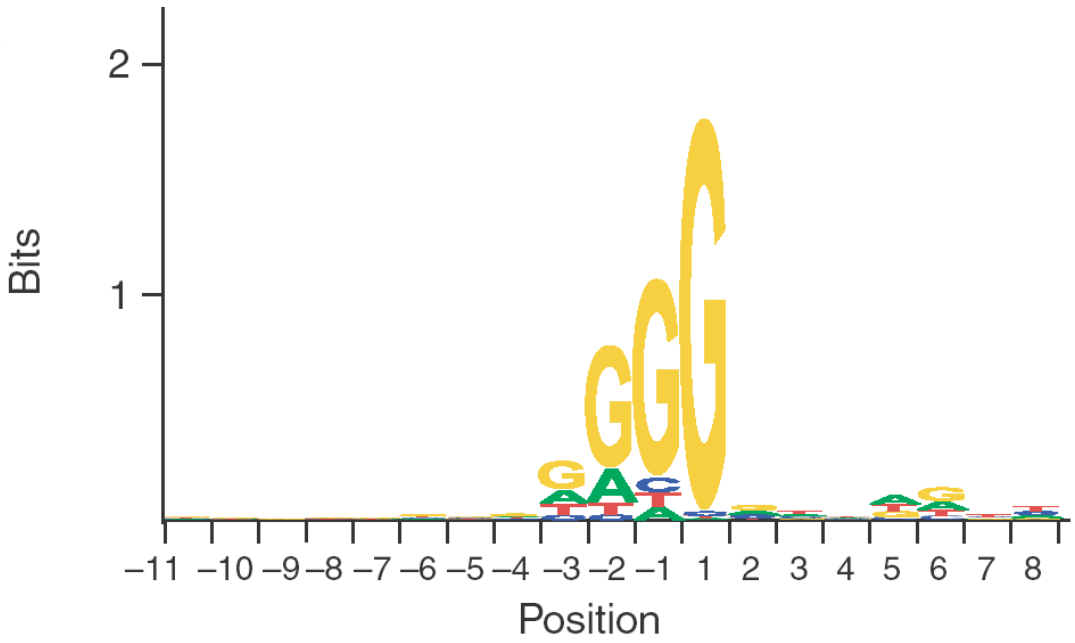
- 58% protein-coding transcriptional units use alternative promoters (previously known 20%)
- 93% transcriptional units, having at least 2 alternative promoters, have distrinct metionine start codons
- two positions (-31,-30) for TATA-box are very specific for highly expressed genes

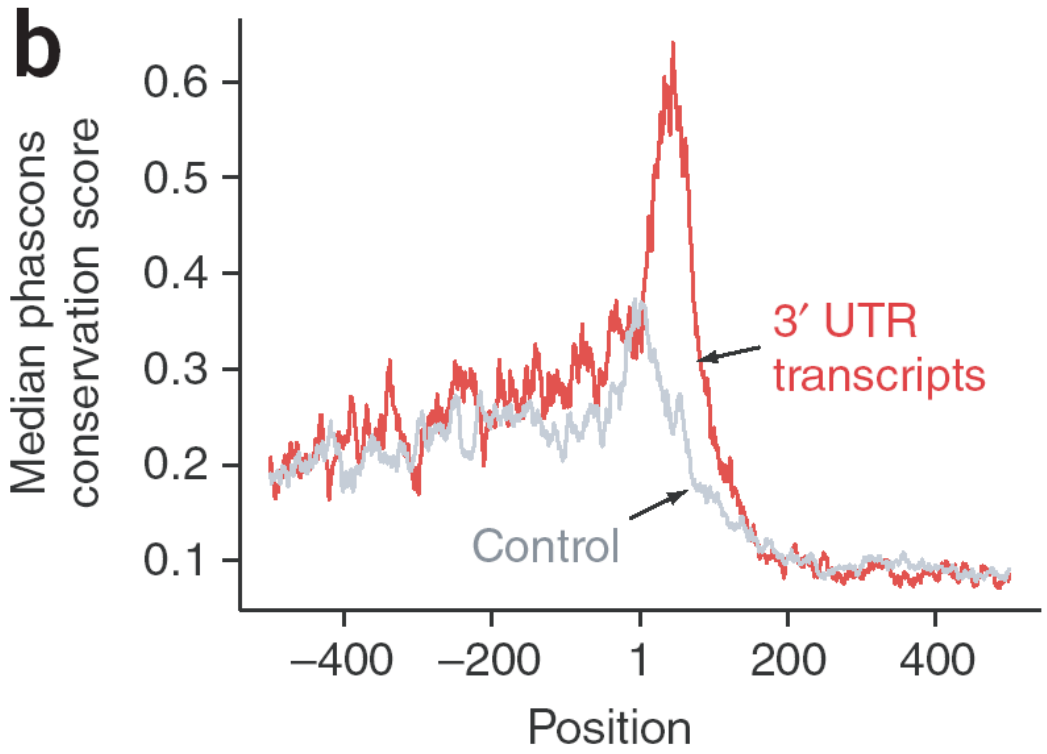
Used literature

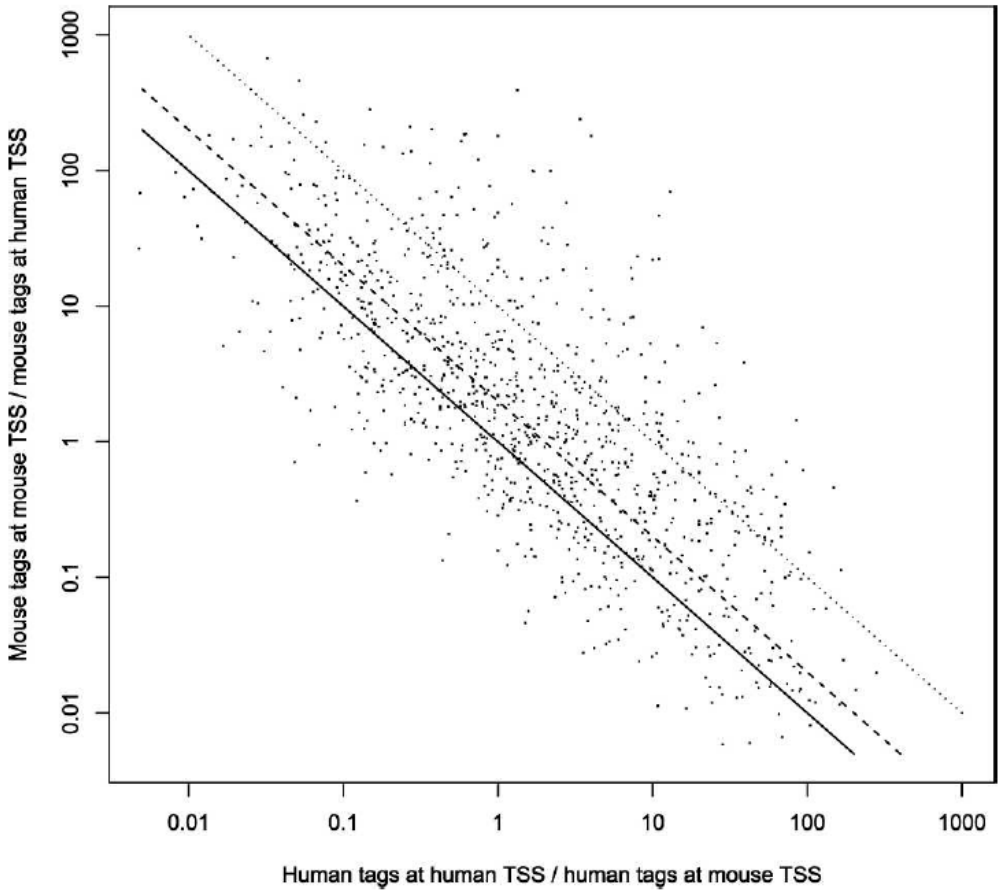
- Frith *et al* "**Evolutionary turnover of mammalian transcription start sites**", Genome Research, apr 2006
- Carninci *et al* "**Genome-wide analysis of mammalian promoter architecture and evolution**", Nature Genetics, jun 2006
- Ponjavic *et al* "**Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters**", Genome Biology, aug 2006

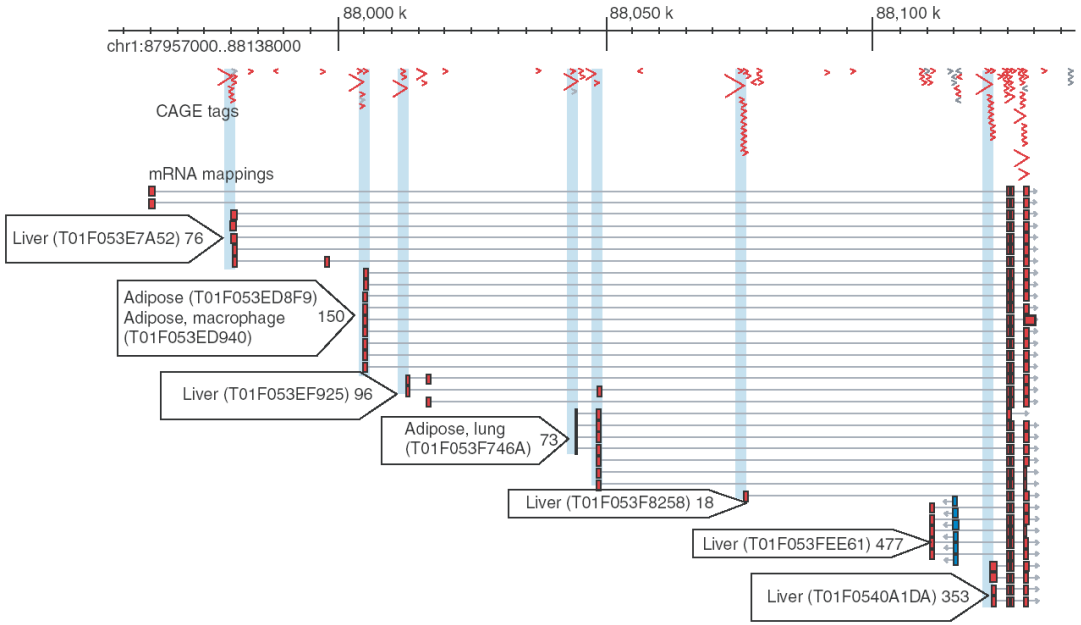
Questions?

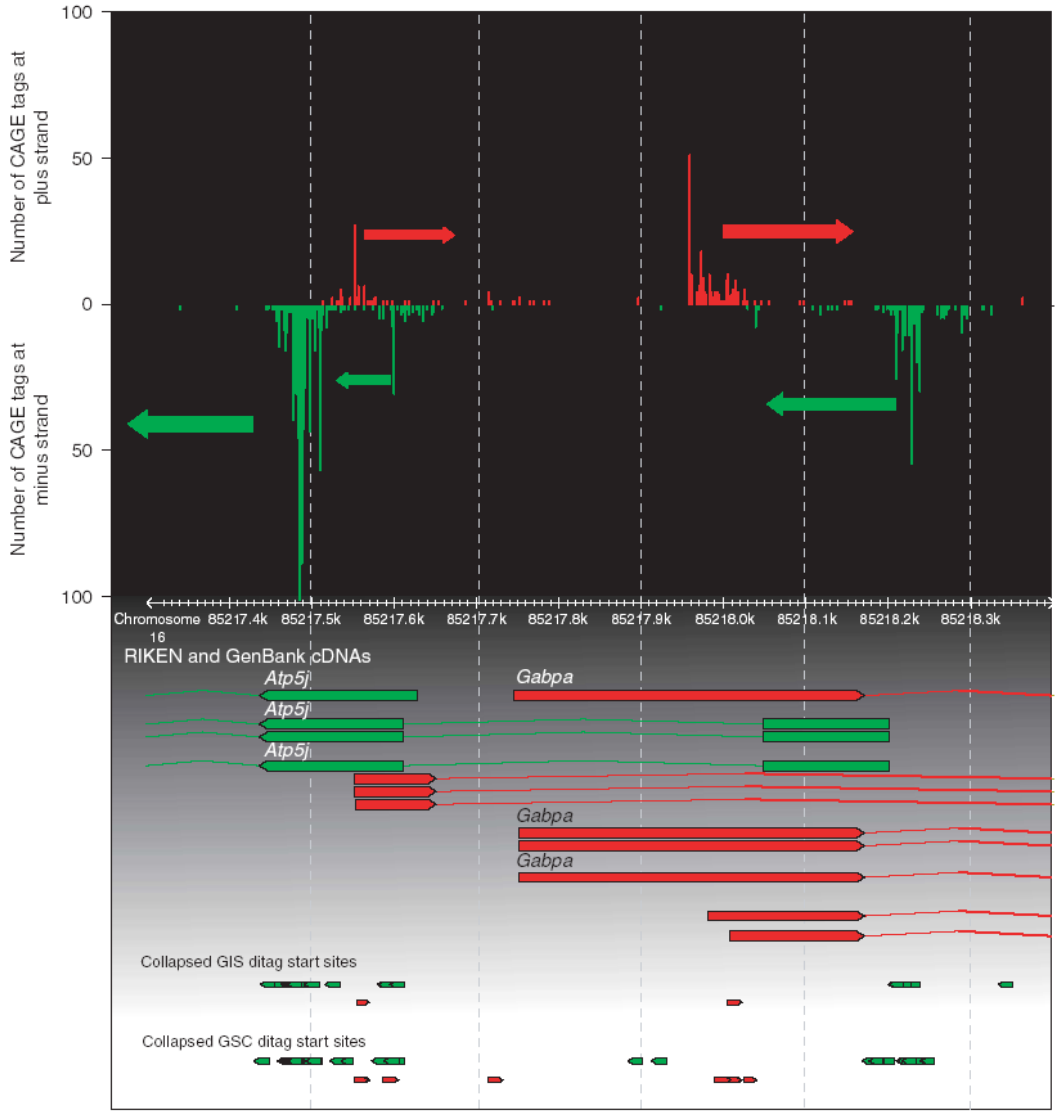
- Any questions?

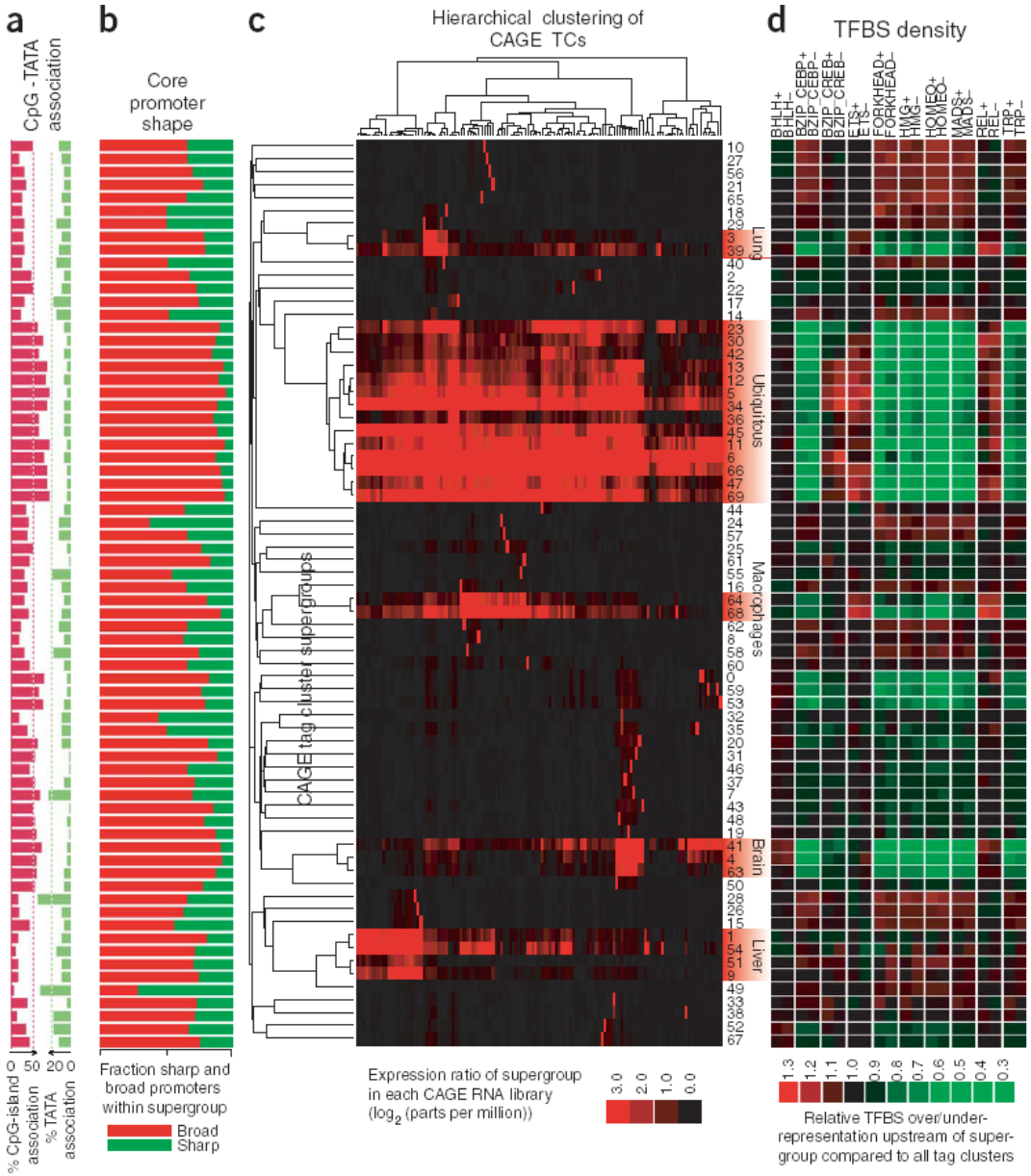

Bits
2
1
−11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1 1 2 3 4 5 6 7 8
Position

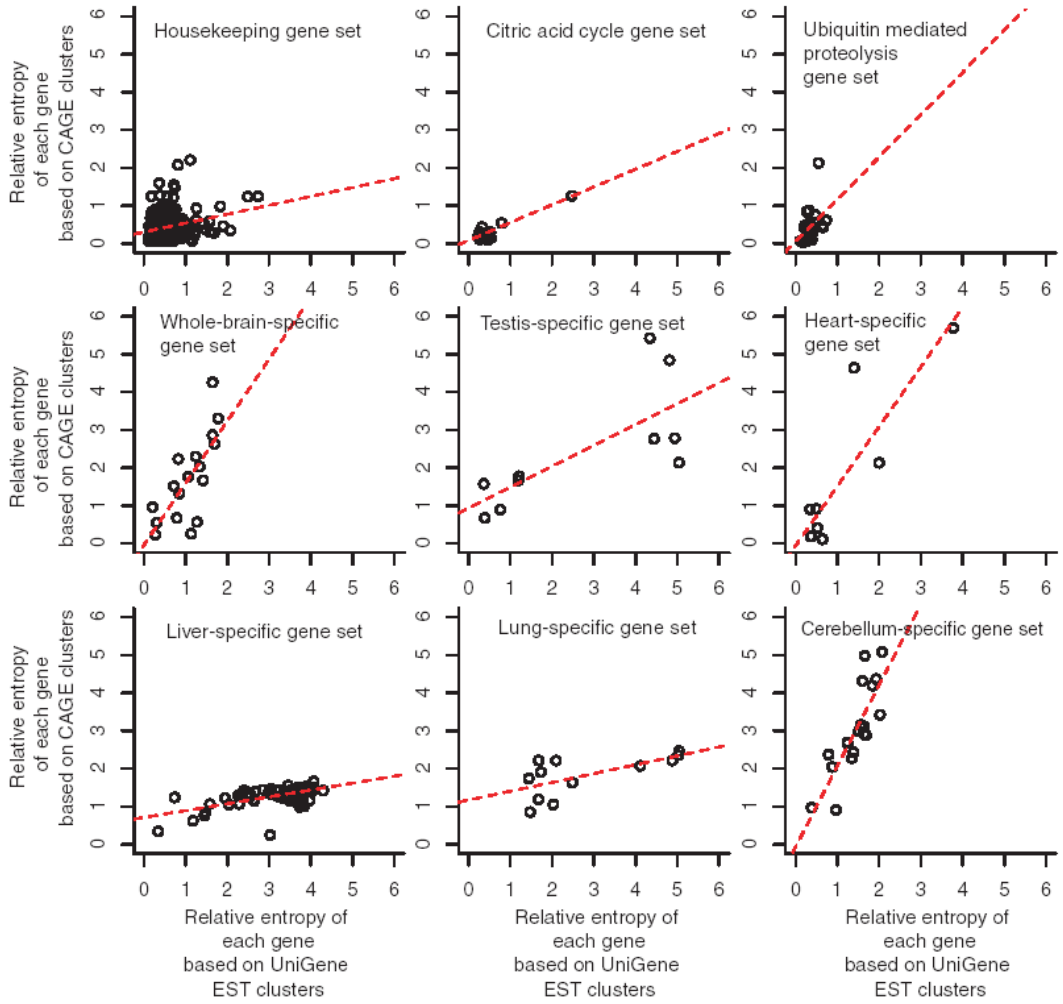

b
Median phascons conservation score
0.6
0.5
0.4
0.3
0.2
0.1
3′ UTR transcripts
Control
–400
–200
1
200
400
Position

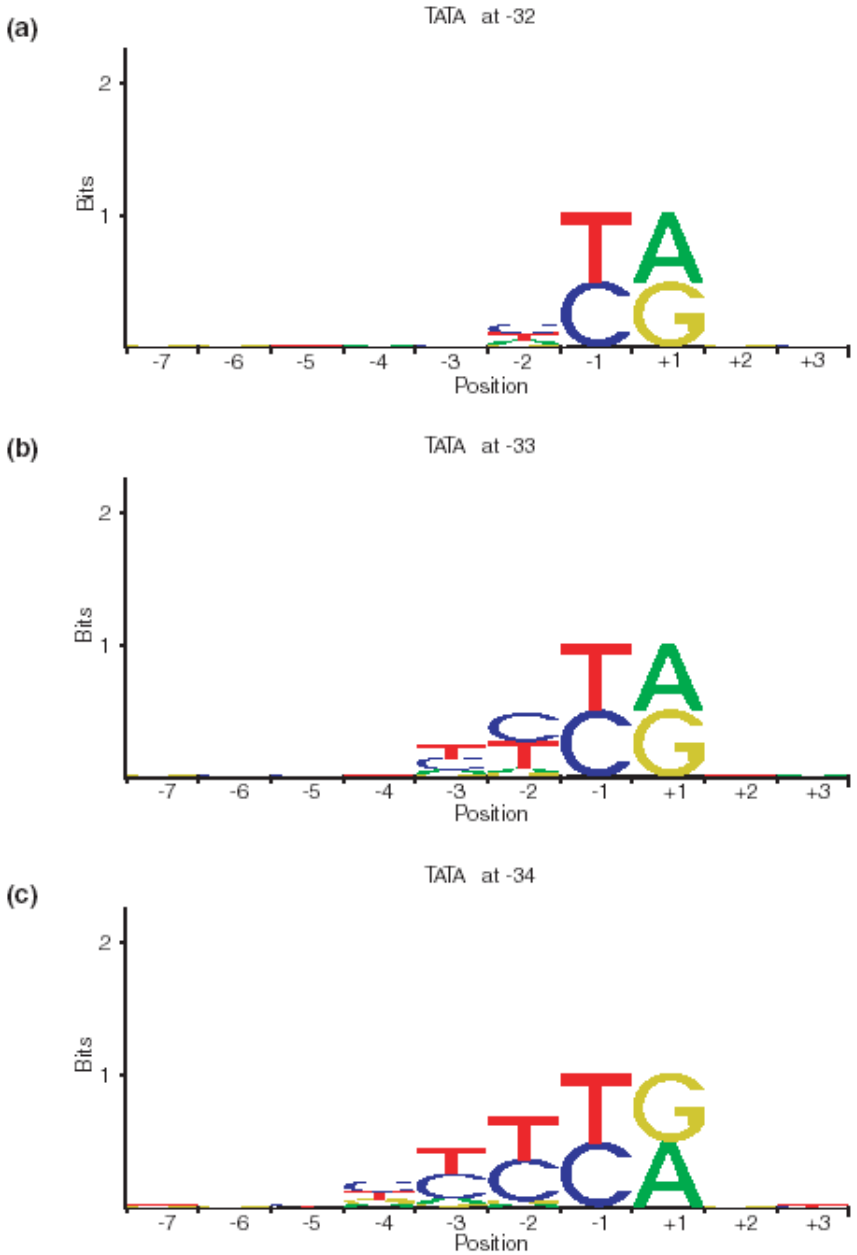

Mouse tags at mouse TSS / mouse tags at human TSS
1000
100
10
1
0.1
0.01
0.01
0.1
1
10
100
1000
Human tags at human TSS / human tags at mouse TSS

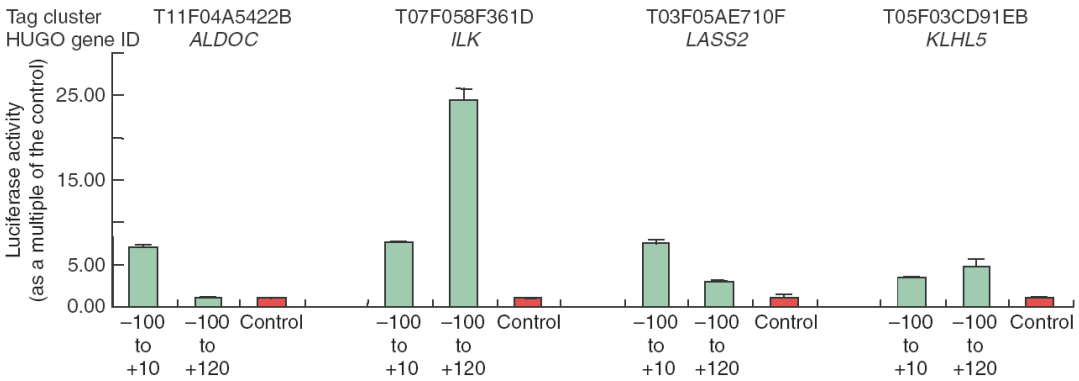

88,000 k
88,050 k
88,100 k
chr1:87957000..88138000
CAGE tags
mRNA mappings
Liver (T01F053E7A52) 76
Adipose (T01F053ED8F9)
Adipose, macrophage
(T01F053ED940)
150
Liver (T01F053EF925) 96
Adipose, lung
(T01F053F746A)
73
Liver (T01F053F8258) 18
Liver (T01F053FEE61) 477
Liver (T01F0540A1DA) 353

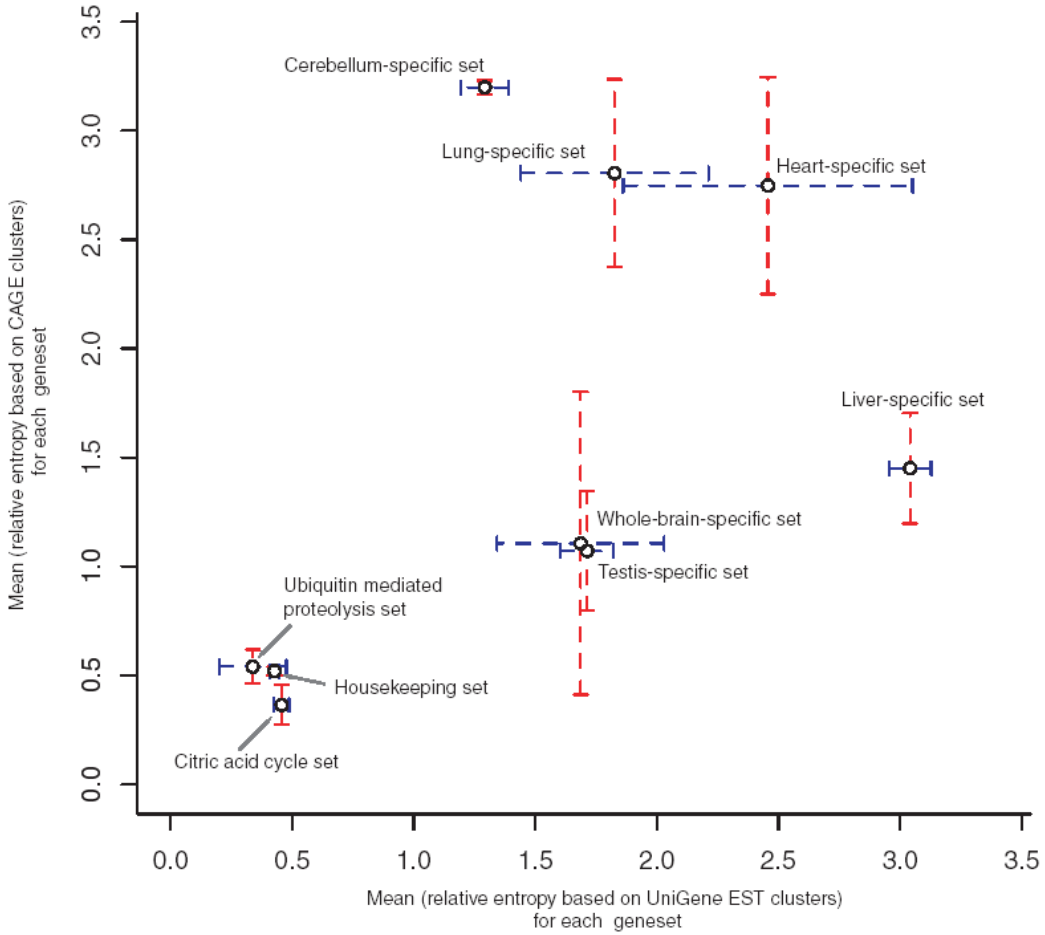

Number of CAGE tags at plus strand
Number of CAGE tags at minus strand
100
50
0
50
100
Chromosome 16
85217.4k
85217.5k
85217.6k
85217.7k
85217.8k
85217.9k
85218.0k
85218.1k
85218.2k
85218.3k
RIKEN and GenBank cDNAs
Atp5j
Atp5j
Atp5j
Gabpa
Gabpa
Gabpa
Collapsed GIS ditag start sites
Collapsed GSC ditag start sites


a
CpG - TATA association
% CpG-island association
% TATA association
b
Core promoter shape
Fraction sharp and broad promoters within supergroup
Broad
Sharp
c
Hierarchical clustering of CAGE TCs
CAGE tag cluster supergroups
Lung
Ubiquitous
Macrophages
Brain
Liver
Expression ratio of supergroup in each CAGE RNA library (log2 (parts per million))
d
TFBS density
BHLH+
BHLH−
BZIP_CEBP+
BZIP_CEBP−
BZIP_CREB+
BZIP_CREB−
ETS+
ETS−
FORKHEAD+
FORKHEAD−
HMG+
HMG−
HOMEO+
HOMEO−
MADS+
MADS−
REL+
REL−
TRP+
TRP−
Relative TFBS over/under-representation upstream of super-group compared to all tag clusters


Housekeeping gene set
Citric acid cycle gene set
Ubiquitin mediated proteolysis gene set
Whole-brain-specific gene set
Testis-specific gene set
Heart-specific gene set
Liver-specific gene set
Lung-specific gene set
Cerebellum-specific gene set
Relative entropy of each gene based on CAGE clusters
Relative entropy of each gene based on UniGene EST clusters


(a)
TATA at -32
Bits
2
1
-7 -6 -5 -4 -3 -2 -1 +1 +2 +3
Position
(b)
TATA at -33
Bits
2
1
-7 -6 -5 -4 -3 -2 -1 +1 +2 +3
Position
(c)
TATA at -34
Bits
2
1
-7 -6 -5 -4 -3 -2 -1 +1 +2 +3
Position


Tag cluster
HUGO gene ID
T11F04A5422B
ALDOC
T07F058F361D
ILK
T03F05AE710F
LASS2
T05F03CD91EB
KLHL5
Luciferase activity
(as a multiple of the control)
25.00
15.00
5.00
0.00
−100 to +10
−100 to +120
Control


Cerebellum-specific set
Lung-specific set
Heart-specific set
Liver-specific set
Whole-brain-specific set
Testis-specific set
Ubiquitin mediated
proteolysis set
Housekeeping set
Citric acid cycle set
Mean (relative entropy based on CAGE clusters)
for each geneset
Mean (relative entropy based on UniGene EST clusters)
for each geneset
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

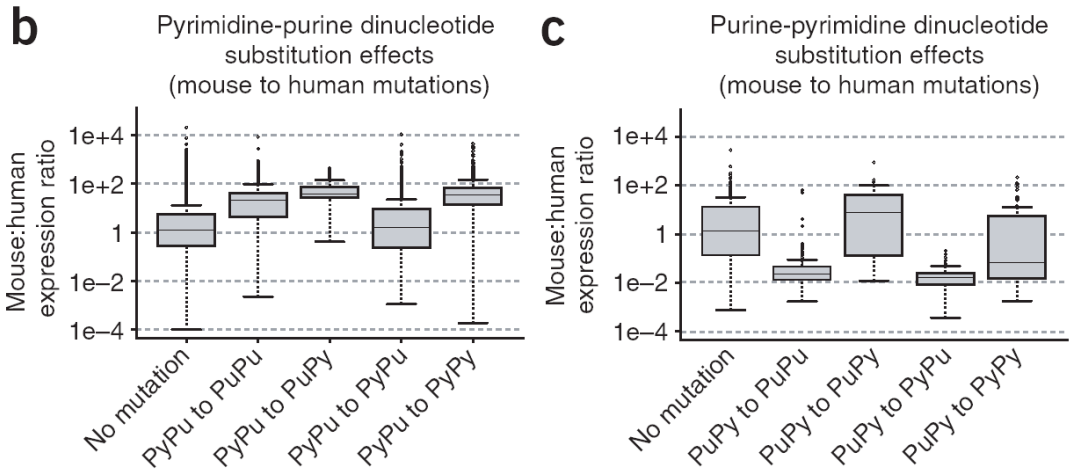
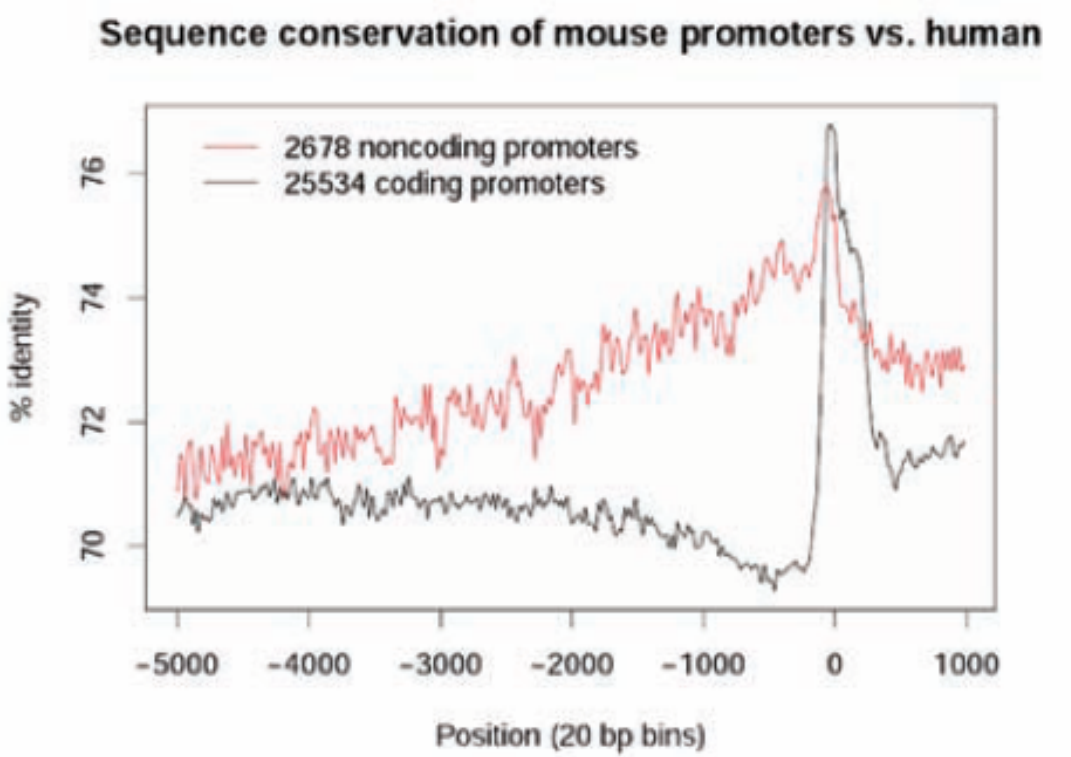
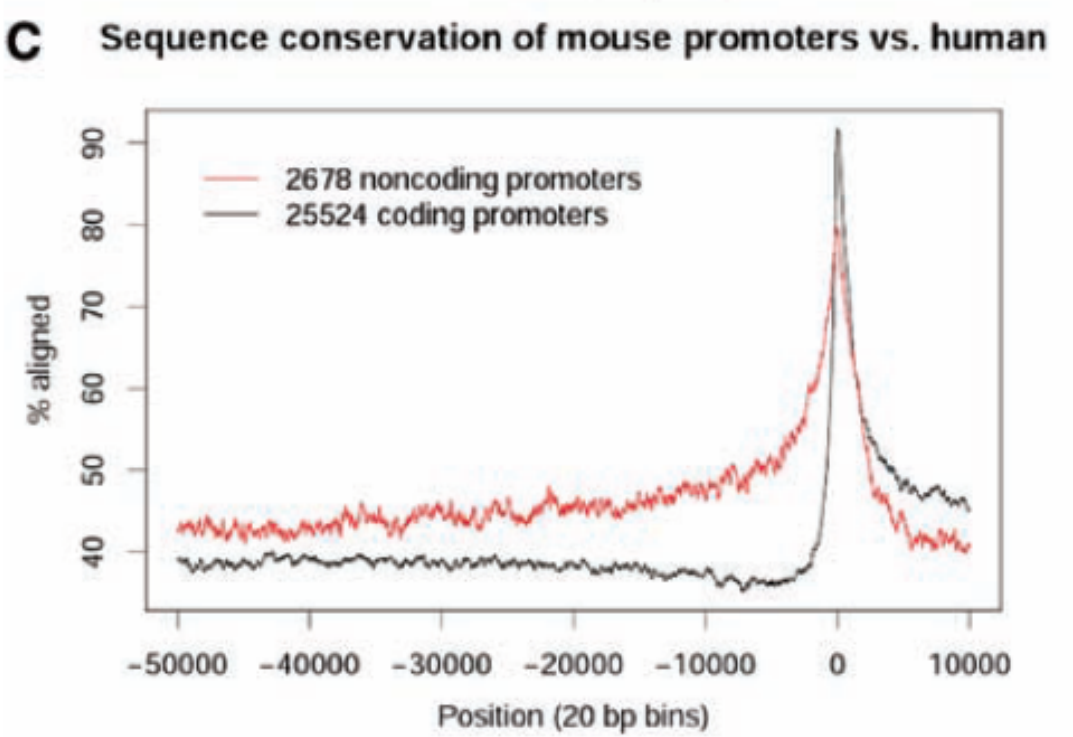

b
Pyrimidine-purine dinucleotide
substitution effects
(mouse to human mutations)
Mouse:human
expression ratio
1e+4
1e+2
1
1e−2
1e−4
No mutation
PyPu to PuPu
PyPu to PuPy
PyPu to PyPu
PyPu to PyPy
c
Purine-pyrimidine dinucleotide
substitution effects
(mouse to human mutations)
Mouse:human
expression ratio
1e+4
1e+2
1
1e−2
1e−4
No mutation
PuPy to PuPu
PuPy to PuPy
PuPy to PyPu
PuPy to PyPy

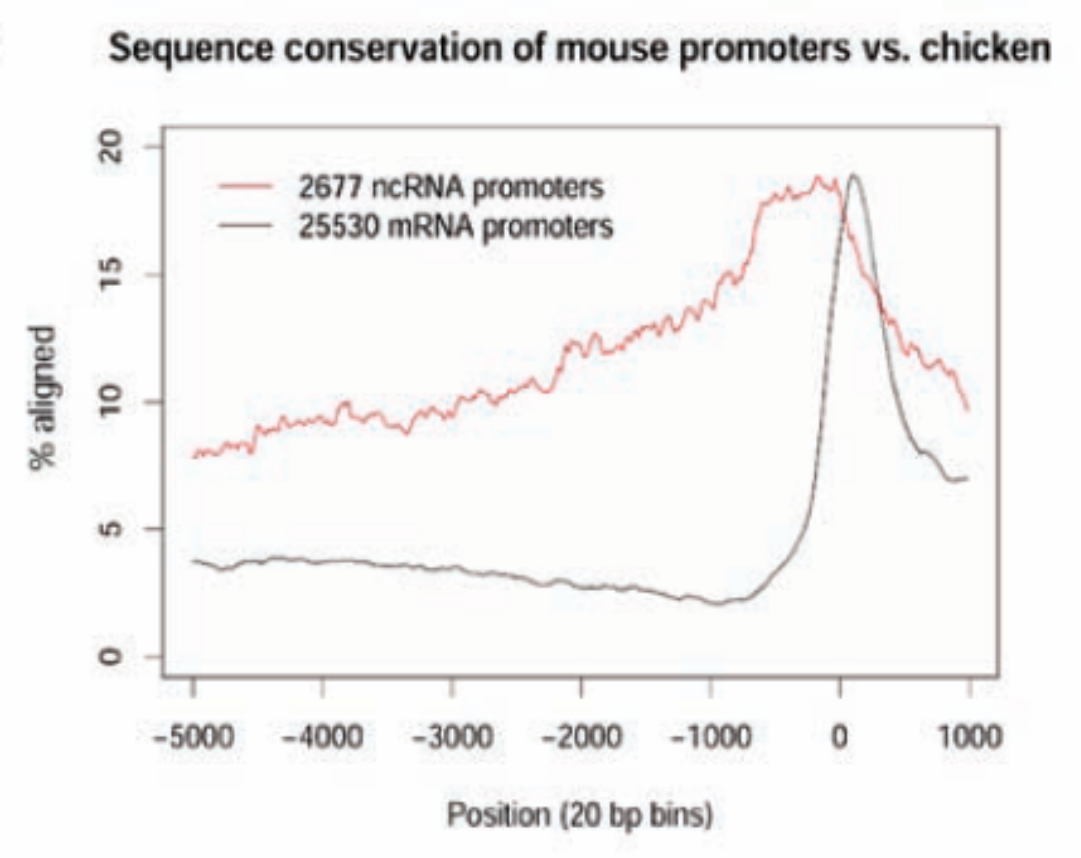
Table 1. Evolution of PyPu initiator dinucleotides at CAGE tag start sites in promoters with and without TSS turnover

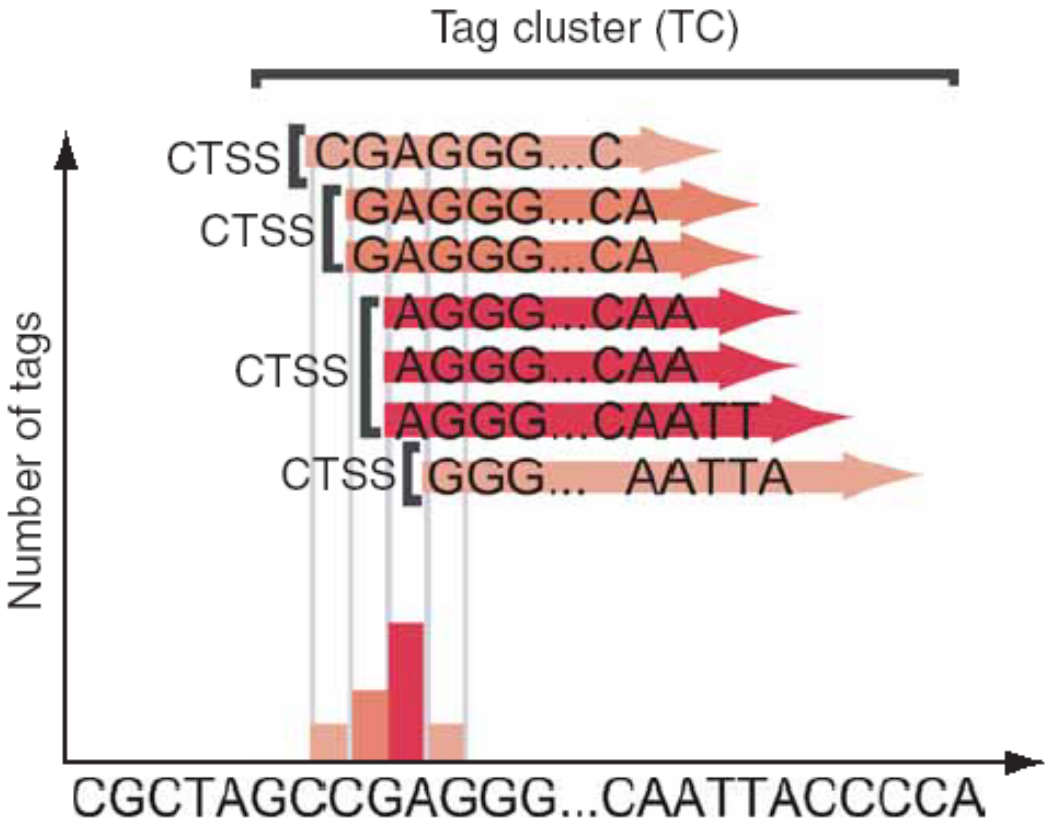
Type of change	PyPu→PuPy	PyPu→PuPu	PyPu→PyPy	PyPu→PyPu
Cases in turnover promoters	26	194	189	923
(%)	1.95	14.56	14.19	69.29
Cases in reference promoters	284	2544	2431	18,029
(%)	1.22	10.92	10.46	77.42
Turnover rate/reference rate	1.60	1.33	1.36	0.90
Significance of difference[a]	3.03×10^{-2}	7.76×10^{-5}	3.12×10^{-5}	3.04×10^{-11}

[a]Fisher's exact two-tail test.

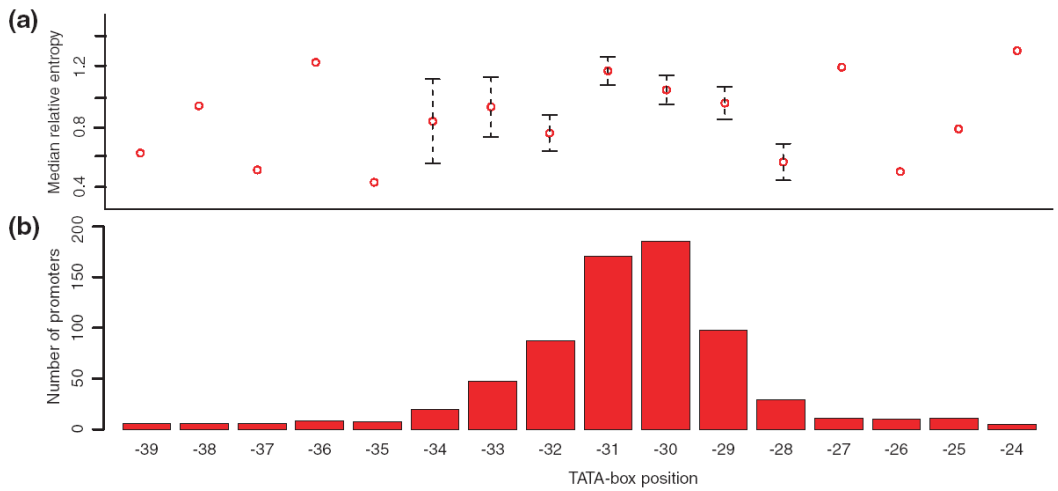

Sequence conservation of mouse promoters vs. human
2678 noncoding promoters
25534 coding promoters
% identity
70
72
74
76
-5000
-4000
-3000
-2000
-1000
0
1000
Position (20 bp bins)


C
Sequence conservation of mouse promoters vs. human
2678 noncoding promoters
25524 coding promoters
% aligned
40
50
60
70
80
90
−50000
−40000
−30000
−20000
−10000
0
10000
Position (20 bp bins)

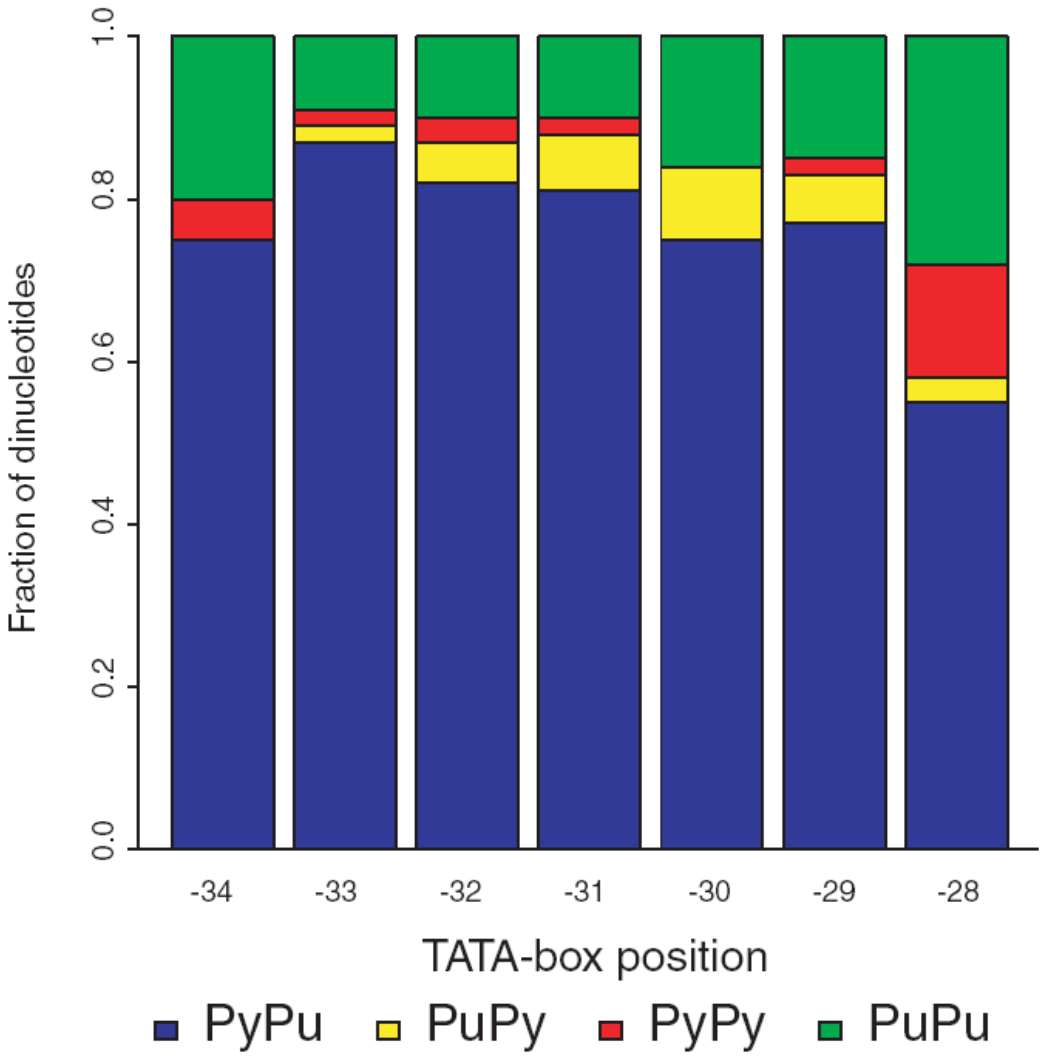

Sequence conservation of mouse promoters vs. chicken
2677 ncRNA promoters
25530 mRNA promoters
% aligned
0
5
10
15
20
-5000
-4000
-3000
-2000
-1000
0
1000
Position (20 bp bins)


Tag cluster (TC)
CTSS
CGAGGG...C
CTSS
GAGGG...CA
GAGGG...CA
AGGG...CAA
CTSS
AGGG...CAA
AGGG...CAATT
CTSS
GGG... AATTA
Number of tags
CGCTAGCCGAGGG...CAATTACCCCA

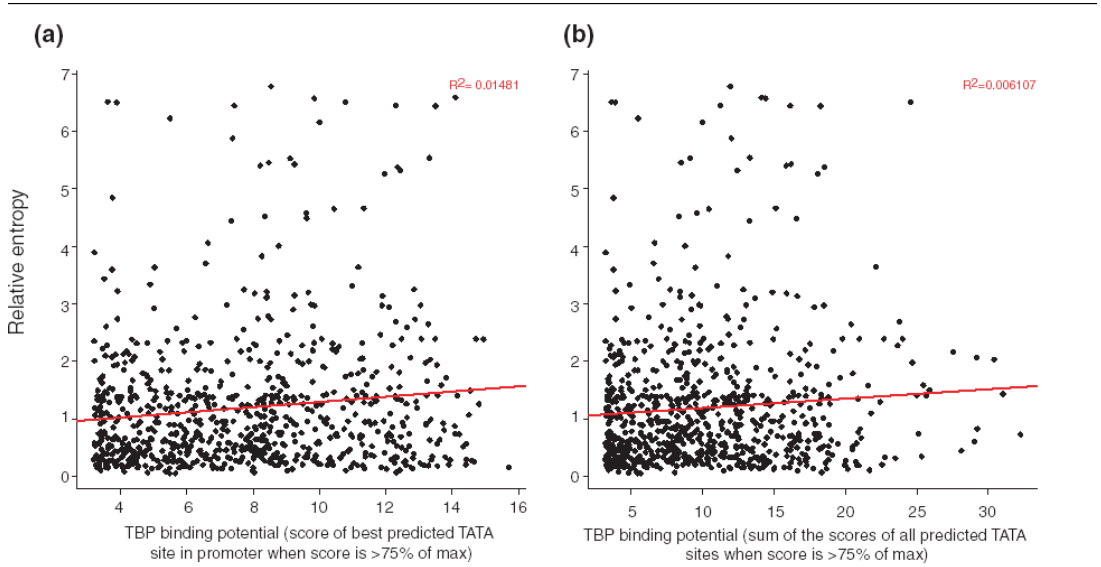
tataboxpos


(a)
Median relative entropy
0.4
0.8
1.2
(b)
Number of promoters
0
50
100
150
200
-39
-38
-37
-36
-35
-34
-33
-32
-31
-30
-29
-28
-27
-26
-25
-24
TATA-box position

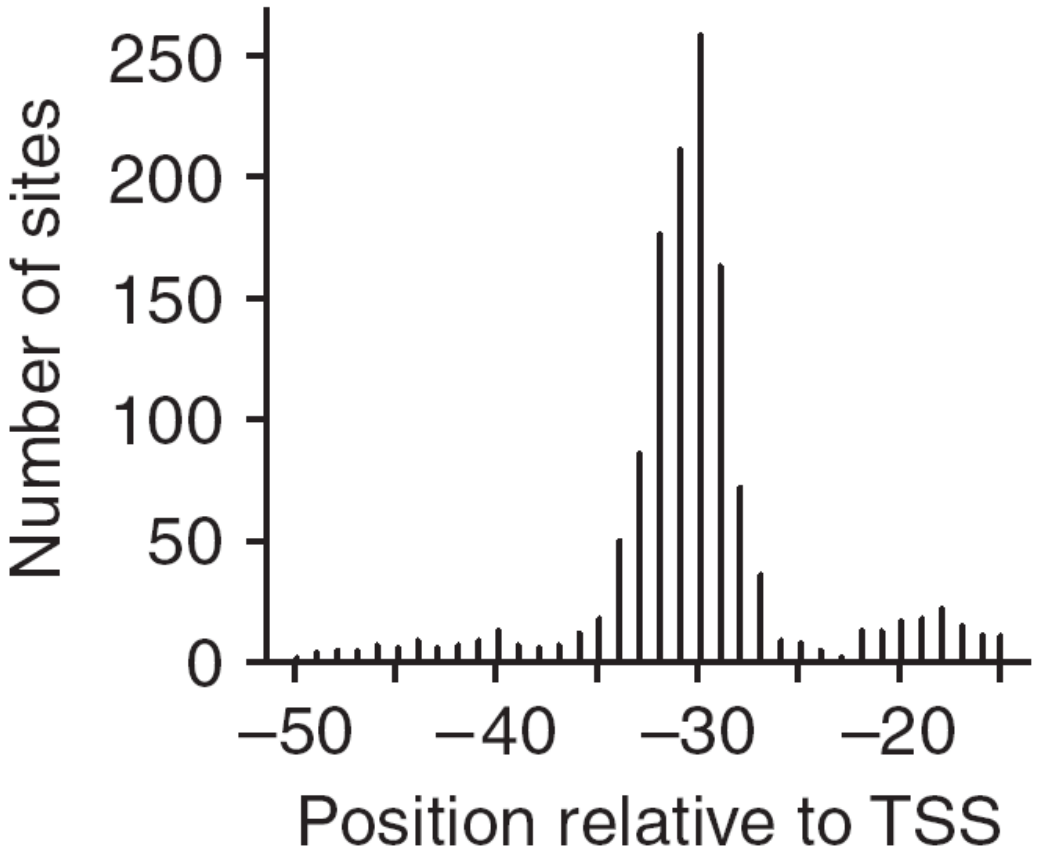
tataboxtss


Fraction of dinucleotides
0.0
0.2
0.4
0.6
0.8
1.0
-34
-33
-32
-31
-30
-29
-28
TATA-box position
PyPu
PuPy
PyPy
PuPu

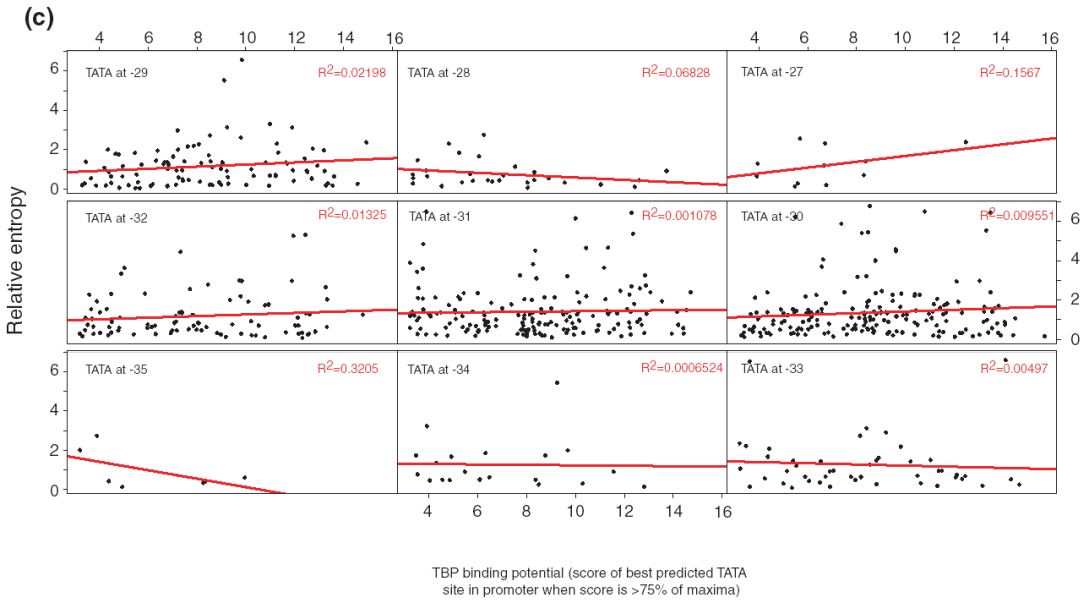
tataentropy



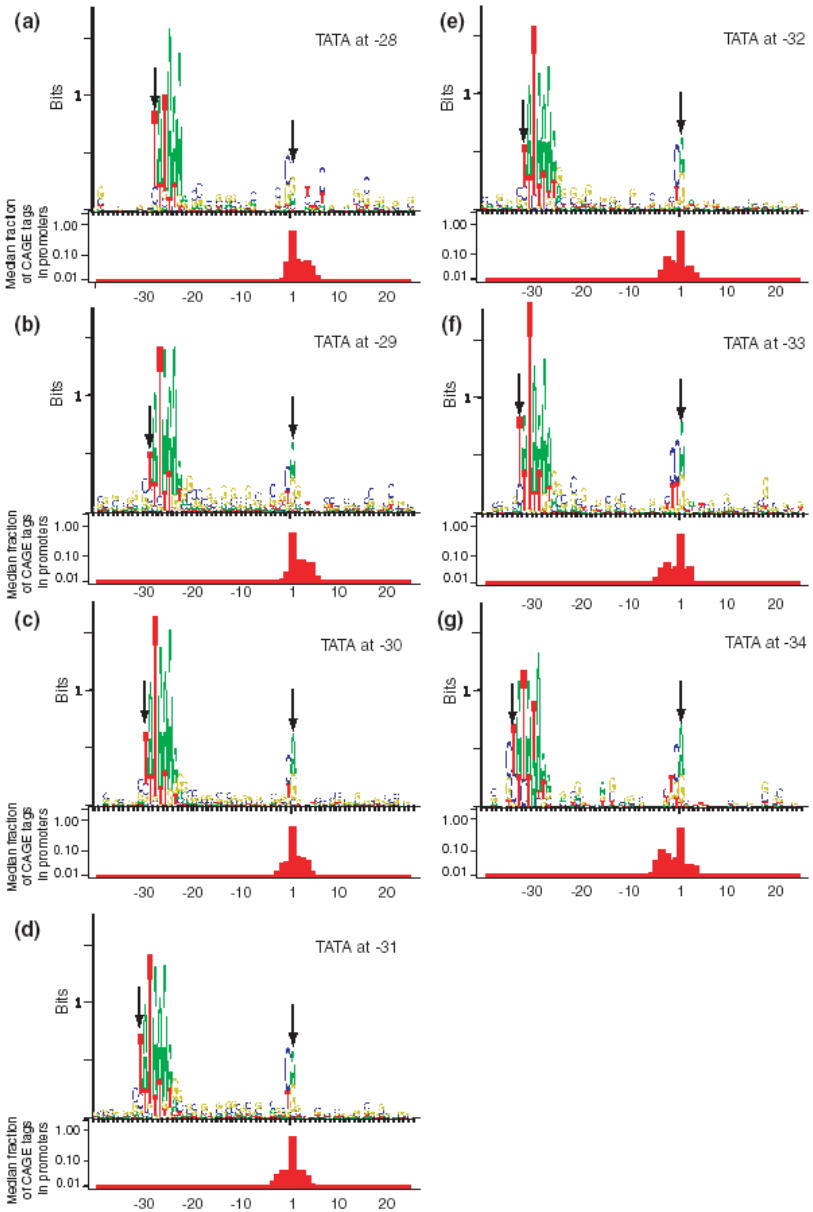
tata


Number of sites
250
200
150
100
50
0
−50
−40
−30
−20
Position relative to TSS

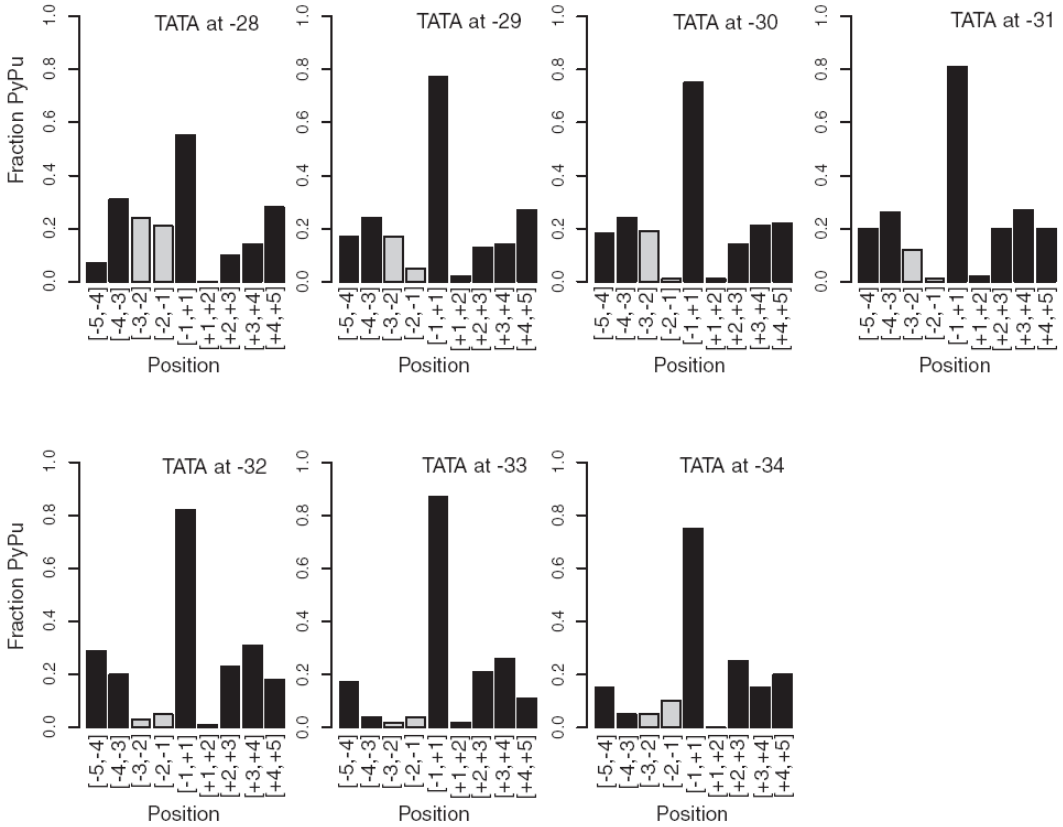
tatarelentropy



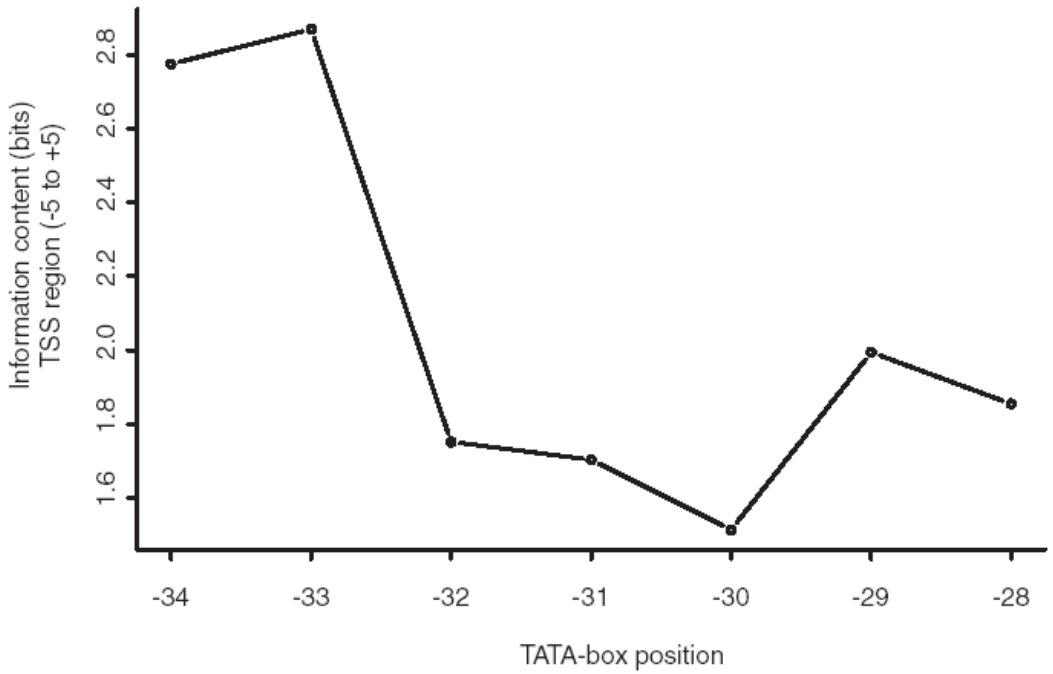
tataseqlogo



tatas


TATA at -28
TATA at -29
TATA at -30
TATA at -31
TATA at -32
TATA at -33
TATA at -34
Fraction PyPu
0.0
0.2
0.4
0.6
0.8
1.0
[-5,-4]
[-4,-3]
[-3,-2]
[-2,-1]
[-1,+1]
[+1,+2]
[+2,+3]
[+3,+4]
[+4,+5]
Position

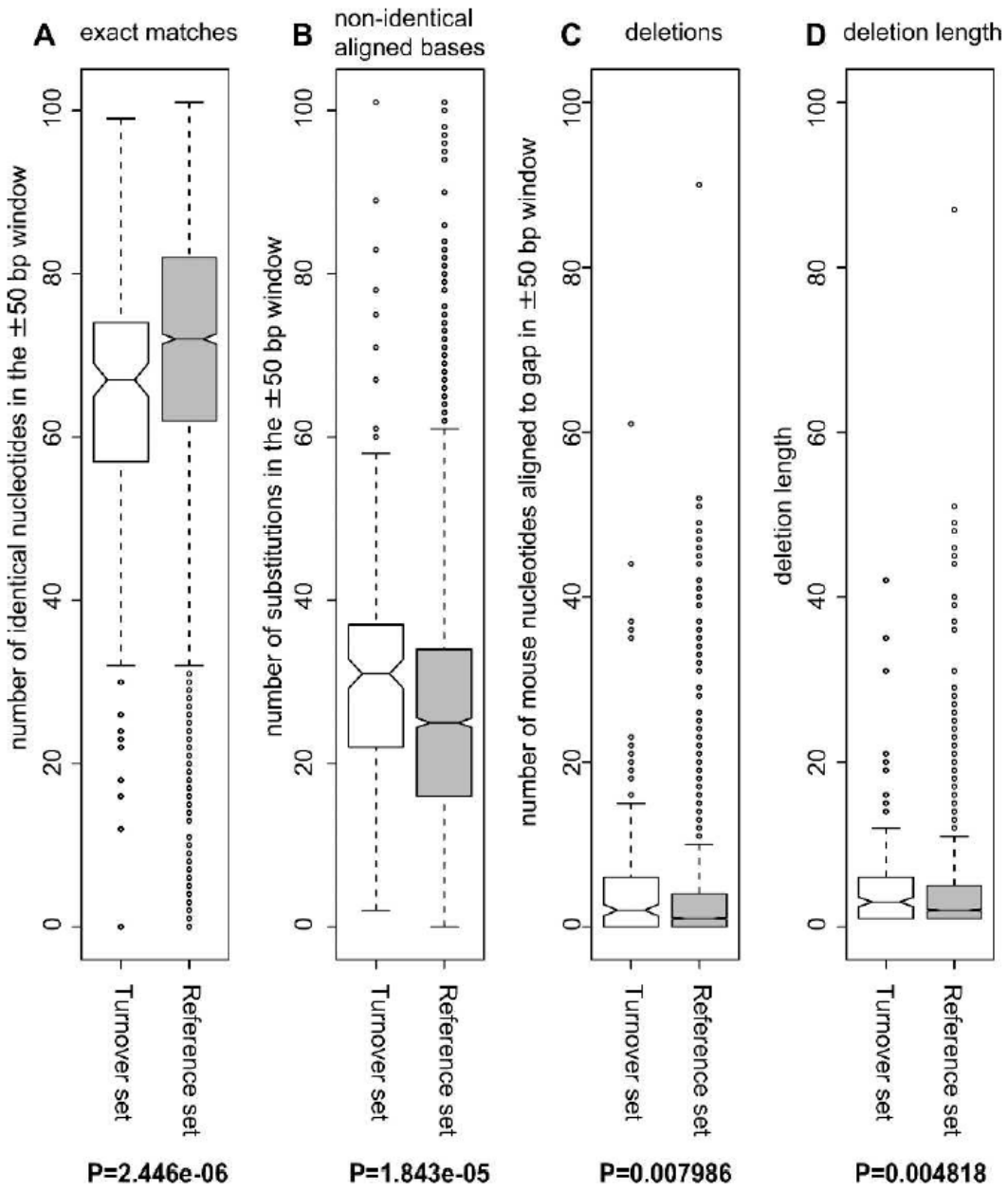
tatatssinf


Information content (bits)
TSS region (-5 to +5)
1.6
1.8
2.0
2.2
2.4
2.6
2.8
-34
-33
-32
-31
-30
-29
-28
TATA-box position

Correlation of tissue specificities measured by relative entropy in CAGE and UniGene EST clusters

Gene set	EST versus CAGE: Spearman rank correlation coefficient	Spearman rank correlation *P* value	Number of genes
Whole brain specific	216	1.10×10^{-3}	17
Testis specific	48	9.68×10^{-2}	9
Heart specific	40	1.48×10^{-2}	10
Liver specific	20,898	1.32×10^{-6}	66
Lung specific	92	1.81×10^{-2}	12
Cerebellum specific	186	$<2.20 \times 10^{-16}$	20
Citric acid cycle	318	2.90×10^{-1}	14
Ubiquitin-mediated proteolysis pathway	886	5.94×10^{-3}	23
Housekeeping genes	2,208,352	8.54×10^{-6}	263
All sets combined	5,269,164	$<2.20 \times 10^{-16}$	434

tsscons


A
exact matches
number of identical nucleotides in the ±50 bp window
0
20
40
60
80
100
Turnover set
Reference set
P=2.446e-06
B
non-identical aligned bases
number of substitutions in the ±50 bp window
0
20
40
60
80
100
Turnover set
Reference set
P=1.843e-05
C
deletions
number of mouse nucleotides aligned to gap in ±50 bp window
0
20
40
60
80
100
Turnover set
Reference set
P=0.007986
D
deletion length
deletion length
0
20
40
60
80
100
Turnover set
Reference set
P=0.004818

tssdist

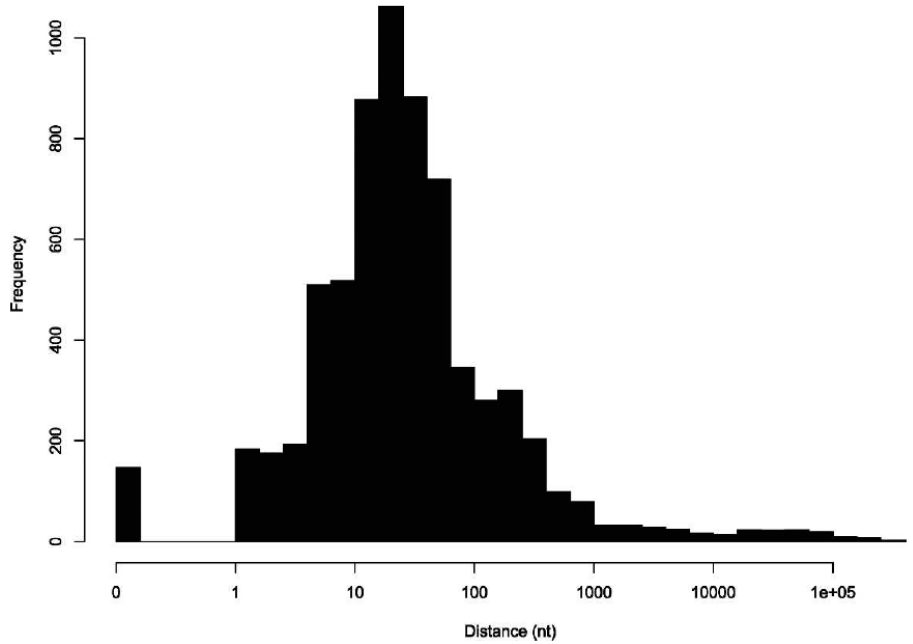
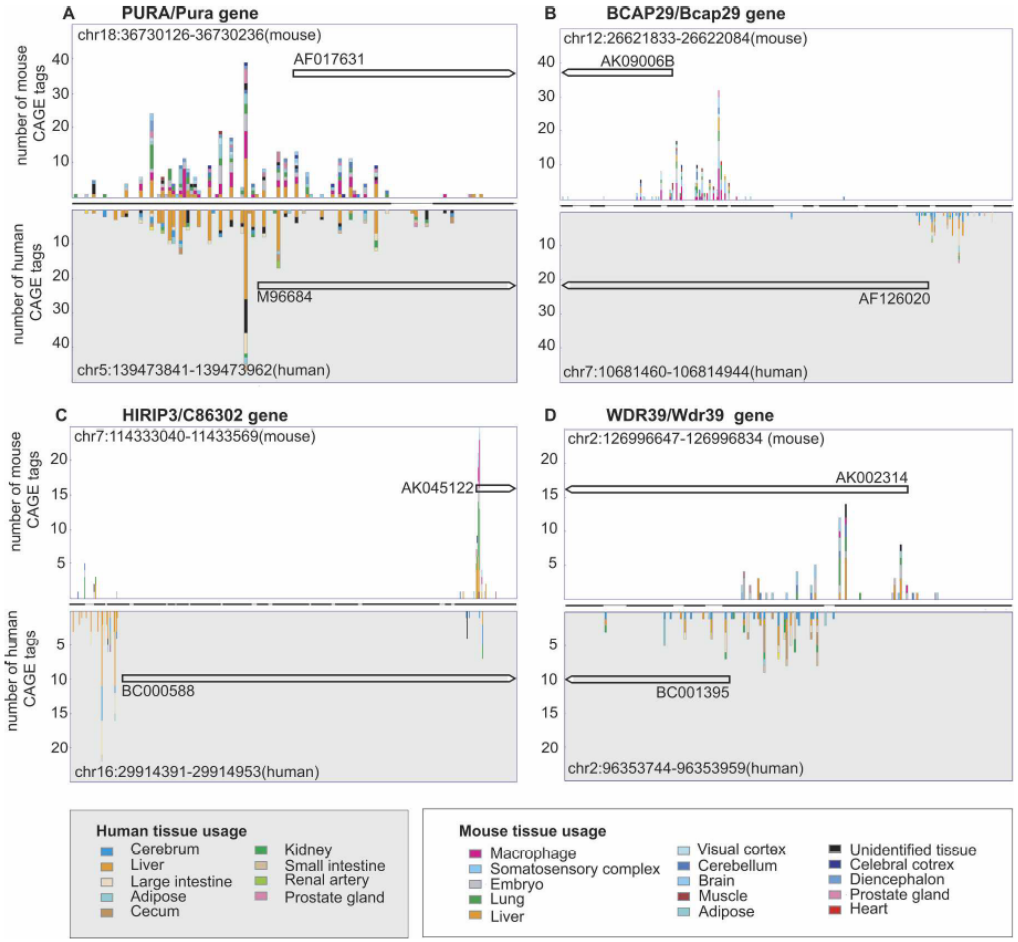
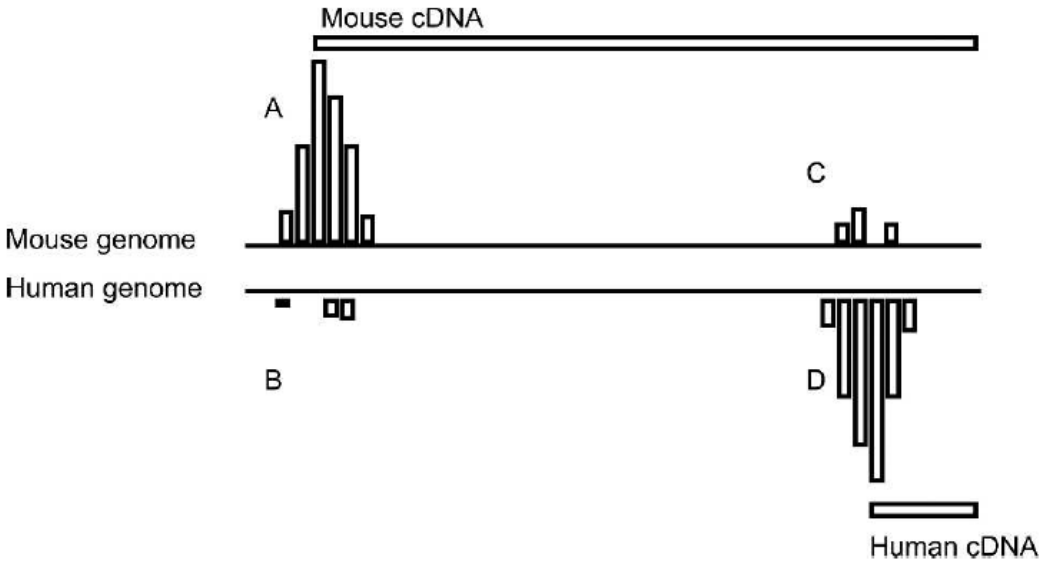


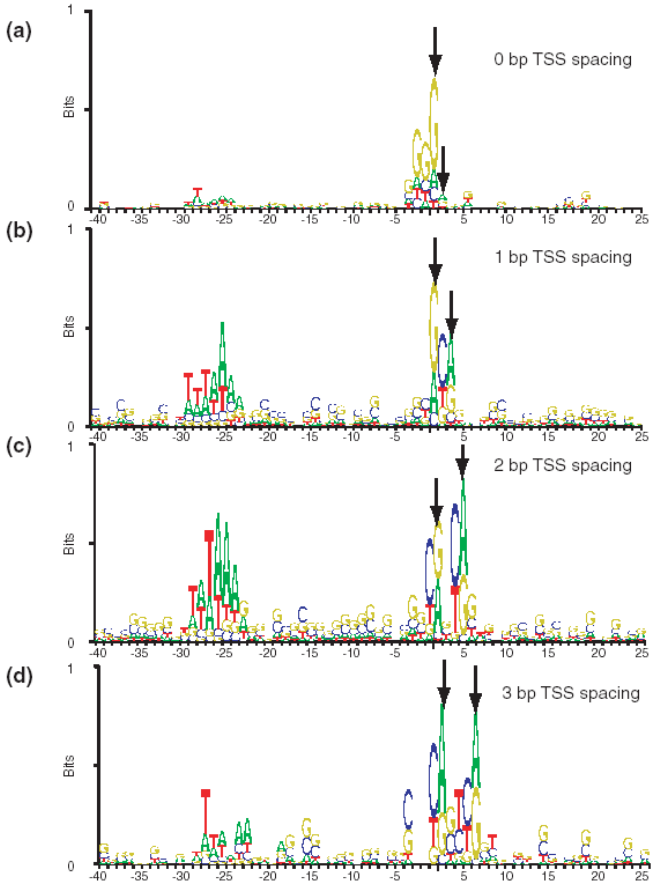
Figure 1. Histogram of distances between transcription start sites of homologous transcripts. The *x*-axis indicates the distance between the human TSS and the human position aligned to the mouse TSS.

tssusage


A
PURA/Pura gene
chr18:36730126-36730236(mouse)
AF017631
number of mouse CAGE tags
number of human CAGE tags
M96684
chr5:139473841-139473962(human)
B
BCAP29/Bcap29 gene
chr12:26621833-26622084(mouse)
AK09006B
AF126020
chr7:10681460-106814944(human)
C
HIRIP3/C86302 gene
chr7:114333040-11433569(mouse)
AK045122
BC000588
chr16:29914391-29914953(human)
D
WDR39/Wdr39 gene
chr2:126996647-126996834 (mouse)
AK002314
BC001395
chr2:96353744-96353959(human)
Human tissue usage
Cerebrum
Liver
Large intestine
Adipose
Cecum
Kidney
Small intestine
Renal artery
Prostate gland
Mouse tissue usage
Macrophage
Somatosensory complex
Embryo
Lung
Liver
Visual cortex
Cerebellum
Brain
Muscle
Adipose
Unidentified tissue
Celebral cotrex
Diencephalon
Prostate gland
Heart

turnoverprom




(a)
0 bp TSS spacing
(b)
1 bp TSS spacing
(c)
2 bp TSS spacing
(d)
3 bp TSS spacing
Bits

Numbers

TSS evolution/turnover

TATA-box