

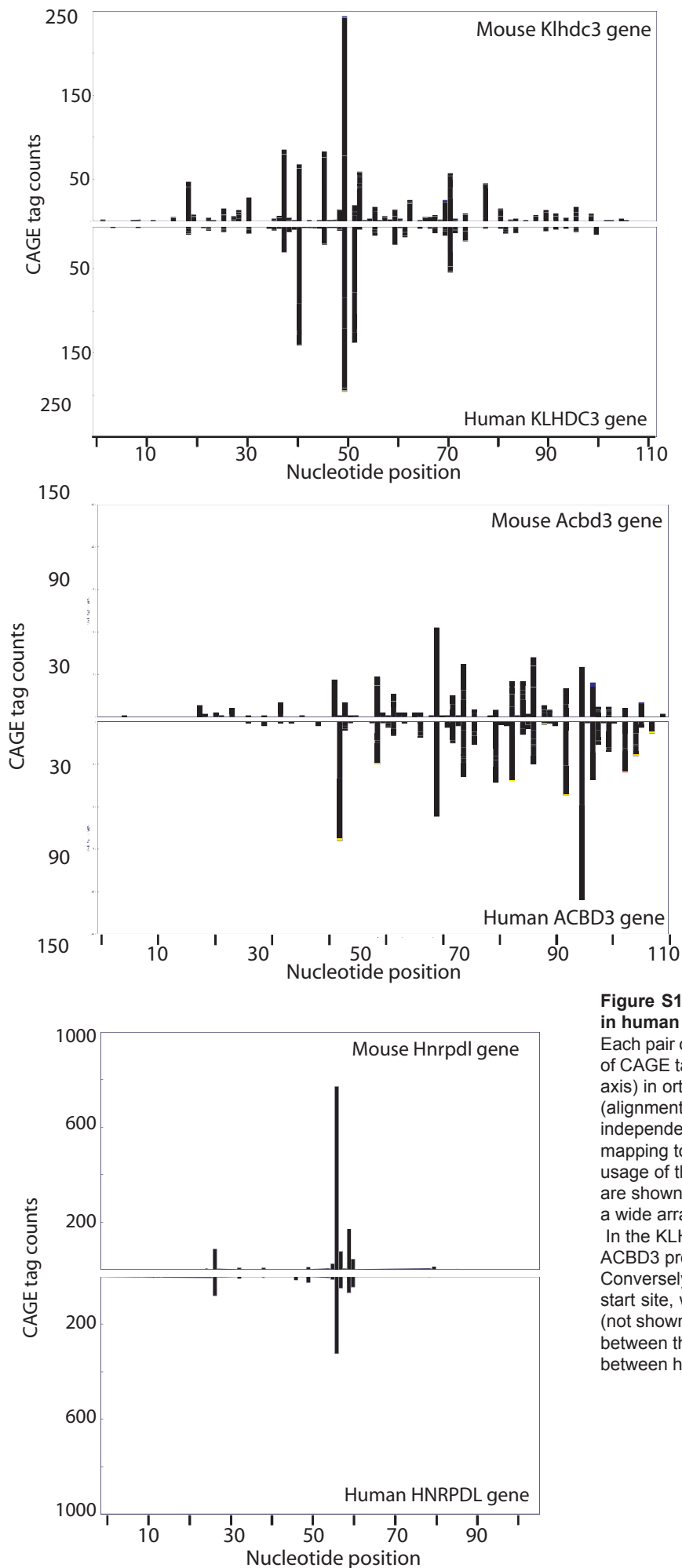


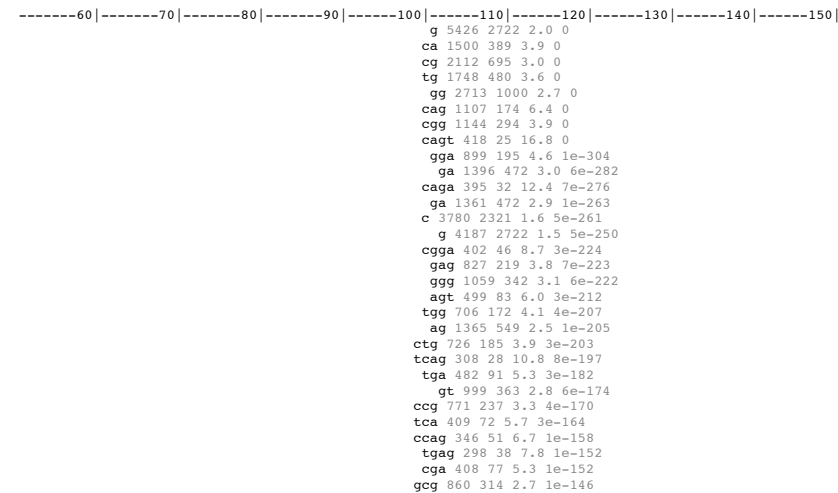
Figure S1: Examples of initiation site distributions in human and mouse

Each pair of panels show bar plots counting the number of CAGE tags (y axis) starting at a particular position (x axis) in orthologous mouse and human core promoters (alignments are from the UCSC BLASTZ net tracks independent of tag locations). The number of tags mapping to certain nucleotide is interpreted as the TSS usage of this position. Only tags on the relevant strand are shown. The KLHDC3 and ACBD3 genes both have a wide array of transcription start sites.

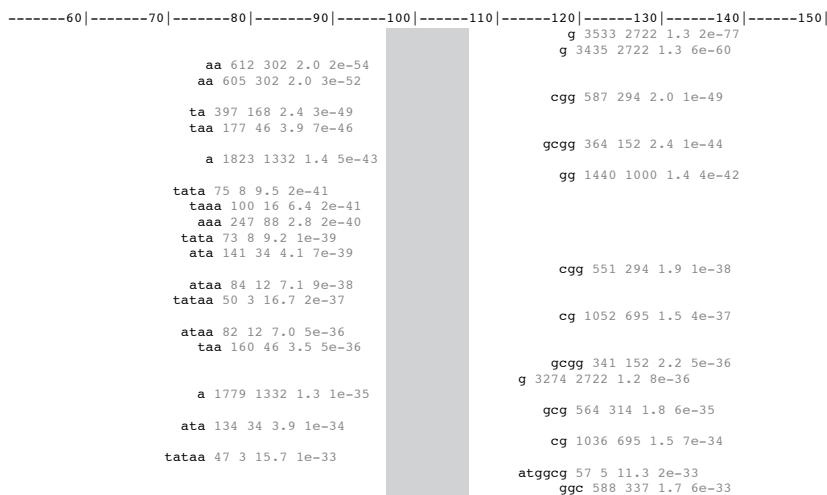
In the KLHDC3 promoter has a central peak, while the ACBD3 promoter has several similarly sized peaks.

Conversely, the HNRPD gene has a more dominant start site, which is directed by a TATA 30 nts upstream (not shown), but also has two smaller peaks conserved between the species. Note the striking correspondence between human and mouse TSS usage.

A



B



C



D

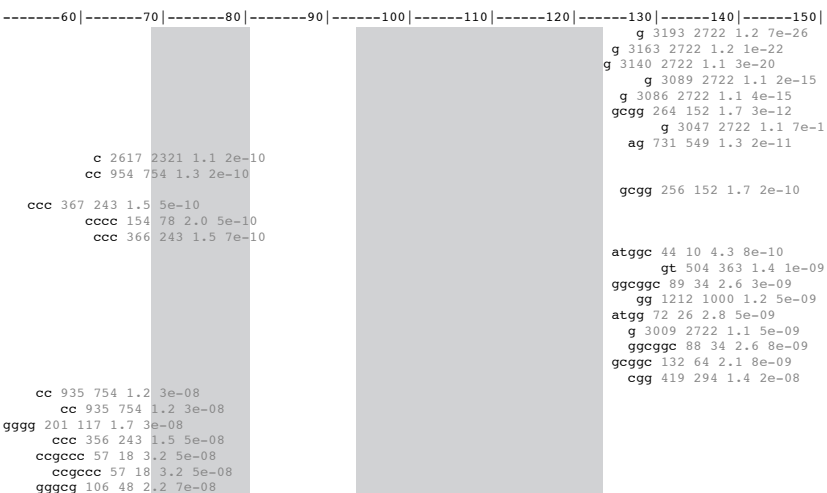


Figure S2
Overrepresented DNA k-mers at fixed distances from dominant transcription start sites in Hep G2 cells (HBY CAGE library) .

Overrepresented words, sorted by P-value, are displayed at the positions where they occur relative to the TSS. The TSS is at position 101. The four numbers after each word are: the observed count (out of 7734 sequences), the expected count (based on the word's overall abundance in the + -100 region), the fold overrepresentation, and the p-value. Shading indicates masked regions: words completely contained in these regions are excluded, allowing other words to rise to the top. A) Initiator motif/Pyrimidine-Purine . B) TATA box. C) gcg repeat. D) Sp1 motif (left) and gcg echo (right). Note that the words in each image are sorted after p-value in all shown columns – this means that in some cases k-mers situated at distinctly different positions will be shown in the same panel (for instance, in the TATA panel, part of the gcg repeat k-mers are shown since their significance range is overlapping with the TATA k-mers).

A

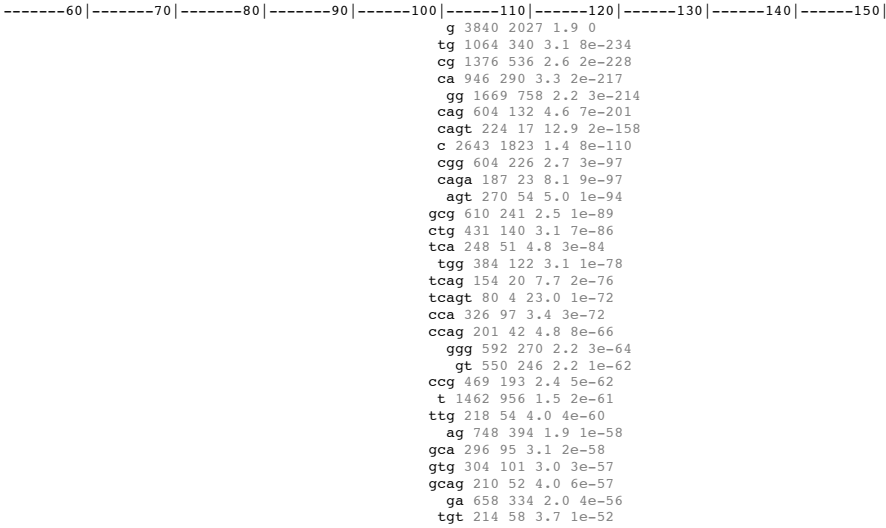
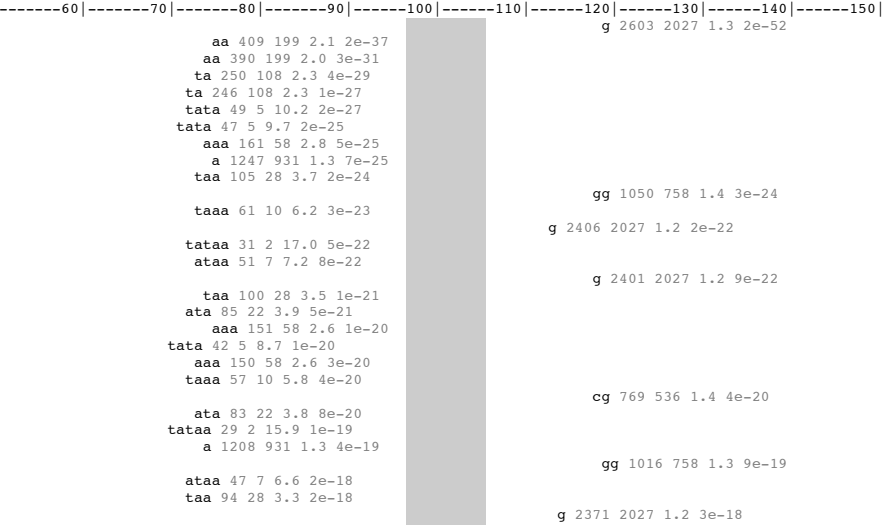
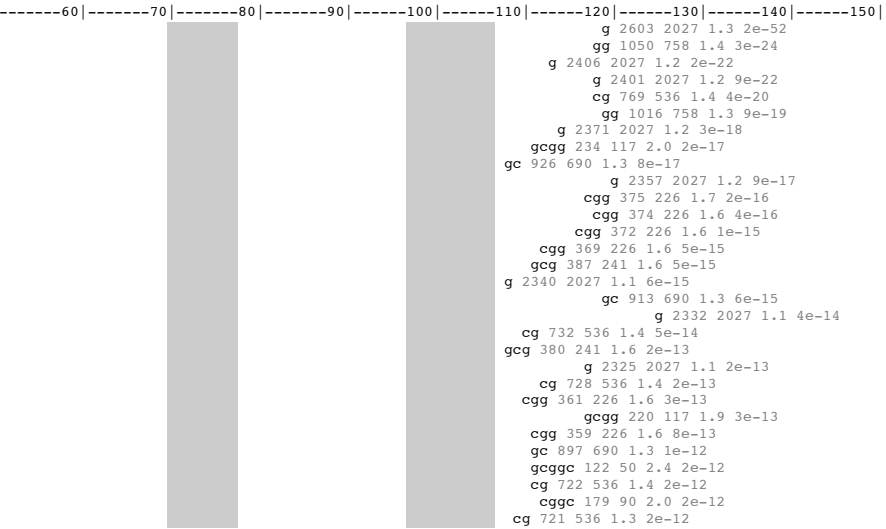


Figure S3:
Overrepresented DNA k-mers at
fixed distances from dominant
transcription start sites in RNA
library HBM (human fibroblasts):
5737 sequences. Please see the
legend for Fig. S2.

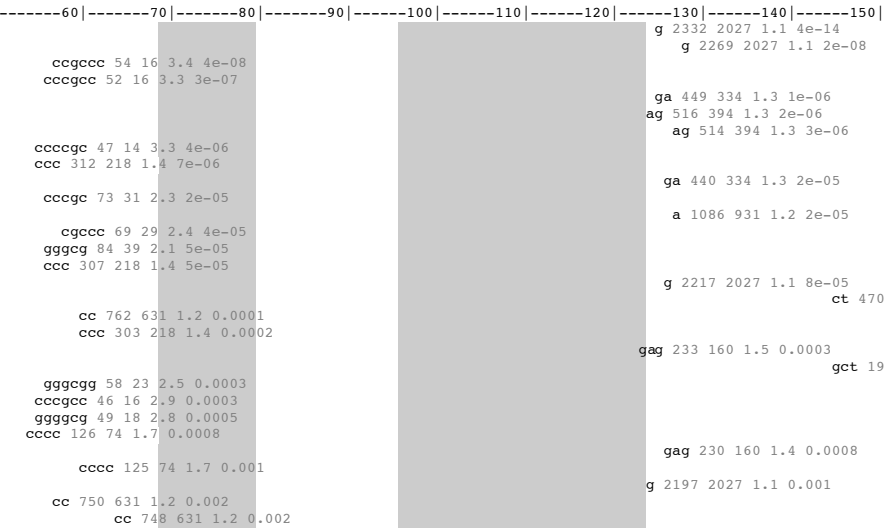
B



C



D



```

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|
g 2399 1265 1.9 1e-314
ca 802 184 4.4 1e-268
cag 546 83 6.6 6e-258
cagt 248 12 21.1 1e-226
cg 1017 340 3.0 2e-220
tg 777 220 3.5 4e-204
c 1949 1142 1.7 1e-164
agt 296 37 8.0 2e-156
tca 270 34 8.1 2e-143
tcag 180 13 13.6 9e-131
caga 179 15 11.9 3e-120
tcagt 100 3 39.2 2e-114
ag 644 254 2.5 7e-100
ctg 345 88 3.9 4e-95
ttg 219 38 5.8 1e-88
tgt 218 39 5.6 2e-85
ctca 127 12 10.8 2e-79
gt 457 166 2.8 8e-79
cagtc 84 4 23.8 9e-78
ccag 165 24 6.8 9e-75
ccg 365 119 3.1 9e-73
cca 245 58 4.2 8e-72
ctcag 90 5 17.2 2e-71
gca 251 62 4.0 6e-70
gt 431 166 2.6 1e-66
gcg 407 152 2.7 1e-65
gcag 181 34 5.3 1e-65
tc 480 200 2.4 7e-65
ccagt 69 3 23.2 2e-62
cagtt 65 2 26.6 2e-62

```

```

-60|-----60|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|
aa 355 136 2.6 2e-54
aa 339 136 2.5 2e-47
a 948 613 1.6 2e-41
tata 57 4 14.1 5e-40
ta 226 79 2.9 2e-39
ata 98 16 6.0 4e-39
aaa 153 40 3.8 1e-38
a 917 613 1.5 2e-34
ta 213 79 2.7 2e-33
ataa 58 6 10.2 3e-33
tata 51 4 12.7 3e-33
taaa 62 8 8.3 8e-31
ata 87 16 5.3 1e-30
taa 100 22 4.6 1e-30
taa 100 22 4.6 1e-30
ta 206 79 2.6 3e-30
ta 202 79 2.6 2e-28
ata 84 16 5.1 2e-28
ataaa 38 2 16.1 5e-27
aaa 132 40 3.3 8e-27
tata 45 4 11.2 1e-26
tat 80 16 5.0 5e-26
tataa 32 2 20.6 2e-25
ataa 50 6 8.8 2e-25
a 872 613 1.4 5e-25
tat 77 16 4.8 6e-24
tataa 30 2 19.3 8e-23
ataa 46 6 8.1 1e-21
ctata 26 1 21.7 2e-20
g 1695 1265 1.3 9e-46

```

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

g 1695 1265 1.3 9e-46
g 1552 1265 1.2 4e-20
cgg 261 139 1.9 3e-17
atggc 38 5 8.1 1e-16
cg 511 340 1.5 2e-16
gg 639 451 1.4 7e-16
g 1515 1265 1.2 5e-15
ggc 275 158 1.7 4e-14
atggcg 27 2 11.7 4e-14
cgg 247 139 1.8 2e-13
gg 623 451 1.4 4e-13
cggc 126 55 2.3 5e-12
g 1488 1265 1.2 9e-12
g 1487 1265 1.2 1e-11
gcgg 148 72 2.1 4e-11
cg 482 340 1.4 5e-11
atggcg 24 2 10.4 7e-11
atgg 45 11 4.1 6e-10
gc 581 434 1.3 1e-09
atggc 30 5 6.4 1e-09
aagatg 19 2 12.9 3e-09
gatggc 20 2 11.0 8e-09
g 1459 1265 1.1 1e-08
ggc 252 158 1.6 1e-08
agatgg 19 2 11.7 1e-08
g 1456 1265 1.1 2e-08
gcggc 80 31 2.6 2e-08
gc 569 434 1.3 6e-08
cgc 212 129 1.6 6e-08
atggcg 21 2 9.1 7e-08

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

g 1488 1265 1.2 9e-12
g 1459 1265 1.1 1e-08
g 1456 1265 1.1 2e-08

c 1319 1142 1.1 2e-07

cg 212 129 1.6 6e-08

g 1443 1265 1.1 4e-07
ag 356 254 1.4 6e-07
ga 307 214 1.4 1e-06
ag 351 254 1.4 3e-06

cc 494 379 1.3 4e-06
cc 492 379 1.3 7e-06

ccc 197 126 1.6 2e-05

g 741 613 1.2 2e-05

cccgcc 36 10 3.8 4e-05

ga 299 214 1.4 3e-05

gc 546 434 1.3 5e-05
gcg 226 152 1.5 7e-05

ag 344 254 1.4 4e-05

ggcg 42 13 3.3 7e-05

gga 225 152 1.5 0.0001

gaga 49 18 2.7 0.0001

ggg 225 152 1.5 0.0001

a 731 613 1.2 0.0002

ccgcc 35 10 3.6 0.0002
c 1285 1142 1.1 0.0002

ag 339 254 1.3 0.0002
ga 293 214 1.4 0.0002

ccattt 12 1 12.8 0.0003

gcagc 42 14 2.9 0.0004

ccccg 44 16 2.9 0.0004

Figure S4:
Overrepresented DNA k-mers at fixed distances from dominant transcription start sites in RNA library HAM (human occipital cortex): 3664 sequences. Please see the legend for Fig. S2.

A

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

cag 805 95 8.5 0
ca 1047 218 4.8 0
g 2610 1377 1.9 0
caga 340 19 17.8 1e-288
cagt 302 14 20.8 4e-275
tg 938 264 3.5 4e-250
ag 958 299 3.2 6e-224
cg 1004 341 3.0 3e-209
c 2126 1207 1.8 1e-194
agt 356 47 7.6 1e-182
tcag 237 17 14.2 1e-177
aga 401 64 6.3 2e-177
ga 786 251 3.1 9e-173
cacag 179 8 21.2 7e-162
gg 1114 474 2.4 2e-160
tca 307 42 7.3 1e-151
ga 723 251 2.9 8e-139
ccag 235 26 8.9 3e-132
cgg 499 138 3.6 5e-128
gcag 265 37 7.1 2e-127
g 2105 1377 1.5 4e-118
tcagt 109 3 33.6 7e-118
ctg 410 102 4.0 2e-117
gt 593 204 2.9 7e-114
gag 422 114 3.7 2e-109
agag 208 26 7.9 9e-107
gag 417 114 3.6 2e-106
gga 380 96 4.0 3e-106
cca 318 68 4.7 1e-105
gca 319 70 4.6 5e-104

Figure S5:
Overrepresented DNA k-mers at
fixed distances from dominant
transcription start sites in RNA
library IN (mouse 17.5 day
embryo): 4090 sequences.
Please see the legend for Fig.
S2.

B

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

aa 474 164 2.9 1e-87
aa 449 164 2.7 1e-75
aaa 224 47 4.8 5e-75
taaa 109 9 11.8 3e-71
aaa 205 47 4.4 5e-62
tata 83 6 14.4 3e-60
ata 135 21 6.6 2e-59
ata 135 21 6.6 2e-59
taa 149 26 5.8 8e-59
ataa 87 7 12.6 2e-58
tata 81 6 14.0 6e-58
tata 81 6 14.0 6e-58
ataaaa 68 3 20.2 1e-57
ta 295 98 3.0 4e-56
ataa 83 7 12.0 4e-54
ta 290 98 3.0 1e-53
ata 128 21 6.2 1e-53
taa 142 26 5.5 2e-53
ta 285 98 2.9 3e-51
ta 284 98 2.9 9e-51
a 1129 733 1.5 1e-49
tataaa 54 2 24.4 3e-49
ataa 77 7 11.1 1e-47

taa 130 26 5.0 1e-44
a 1105 733 1.5 5e-44
aa 375 164 2.3 6e-44
tataaa 50 2 22.6 1e-43
aa 373 164 2.3 4e-43
taaa 81 9 8.7 1e-42

g 1824 1377 1.3 8e-45

C

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

g 1824 1377 1.3 8e-45
gc 675 448 1.5 3e-23
gc 670 448 1.5 3e-22
g 1671 1377 1.2 4e-19
tttgc 34 3 10.4 7e-18
gcggc 94 28 3.4 9e-18
cg 517 341 1.5 2e-17
cg 515 341 1.5 5e-17
gc 639 448 1.4 2e-16
atggcg 31 3 11.2 3e-16
cggc 128 50 2.6 4e-16
gcgg 160 70 2.3 6e-16
cgg 254 138 1.9 9e-16
cggc 127 50 2.5 1e-15
ggc 281 158 1.8 2e-15
cg 506 341 1.5 3e-15
gcg 272 153 1.8 4e-15
cg 251 138 1.8 6e-15
atggc 39 6 6.8 1e-14
ggc 277 158 1.8 2e-14
cgg 249 138 1.8 2e-14
tctgac 20 1 21.8 2e-14
gcggc 87 28 3.1 4e-14
gg 656 474 1.4 5e-14
atggc 38 6 6.6 9e-14
gc 624 448 1.4 1e-13
gg 652 474 1.4 2e-13
g 1623 1377 1.2 4e-13
gcg 264 153 1.7 5e-13
gcg 263 153 1.7 9e-13

D

-----60|-----70|-----80|-----90|-----100|-----110|-----120|-----130|-----140|-----150|

gggcgg 51 13 3.9 1e-09
ggg 245 153 1.6 2e-08
ggcgg 65 22 2.9 2e-08

cccc 85 35 2.4 3e-08

cc 501 374 1.3 9e-08
ccc 192 115 1.7 2e-07

ccc 191 115 1.7 3e-07
ggg 237 153 1.6 9e-07
ggg 235 153 1.5 2e-06

ccc 186 115 1.6 4e-06

cc 487 374 1.3 8e-06
ggcgg 59 22 2.7 1e-05

ccc 183 115 1.6 2e-05
gggg 97 49 2.0 3e-05
gggcgg 43 13 3.3 4e-05

taaga 20 2 9.6 2e-08
ttaaga 13 1 23.3 4e-08

aaa 99 47 2.1 2e-07
ttaaaa 15 1 13.6 9e-07

g 1554 1377 1.1 3e-06
ggcggc 47 14 3.3 4e-06

aaaa 46 15 3.0 7e-06
ttaa 29 7 4.3 1e-05
cgc 193 122 1.6 2e-05
aaga 45 15 3.0 2e-05
atggcg 20 3 7.2 2e-05

gc 563 448 1.3 5e-05
g 1539 1377 1.1 5e-05

A

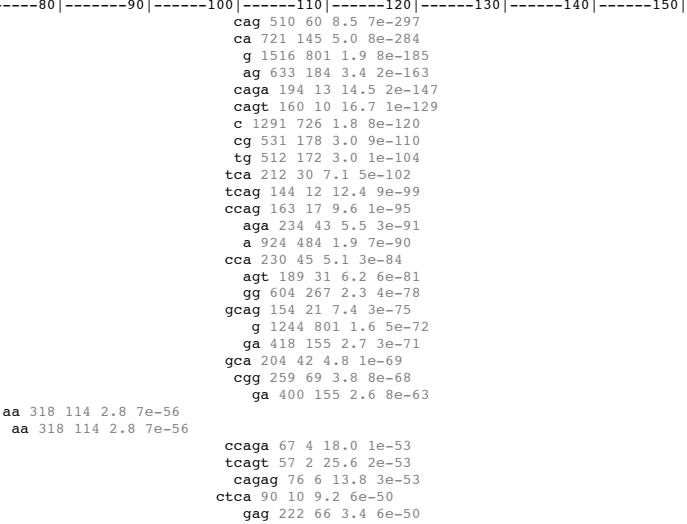
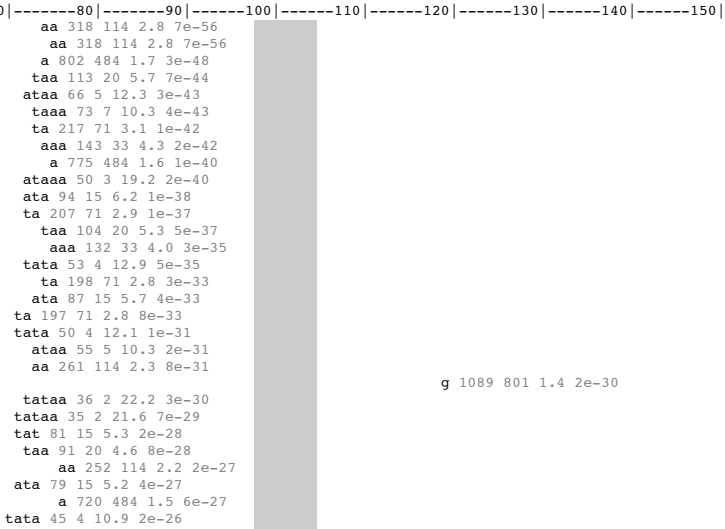
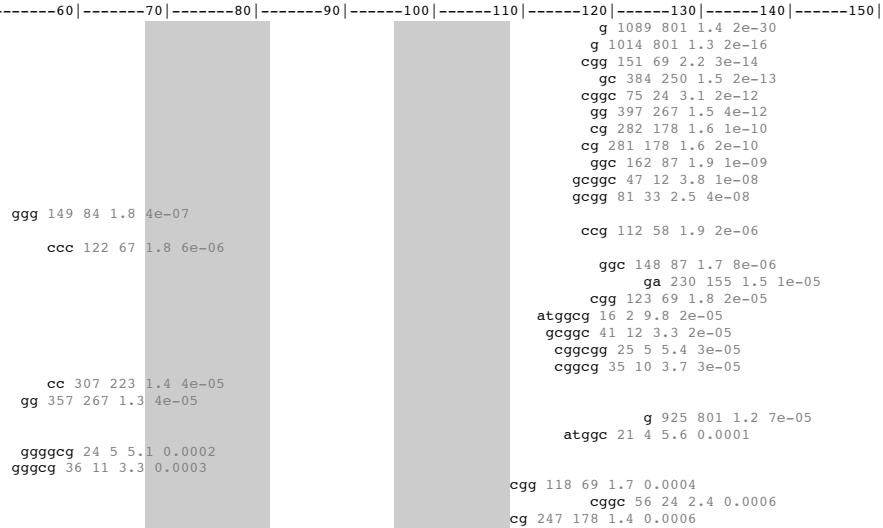


Figure S 6:
Overrepresented DNA k-mers at fixed distances from dominant transcription start sites in RNA library CBR (mouse liver): 2527 sequences. Please see the legend for Fig. S2.

B



C



D

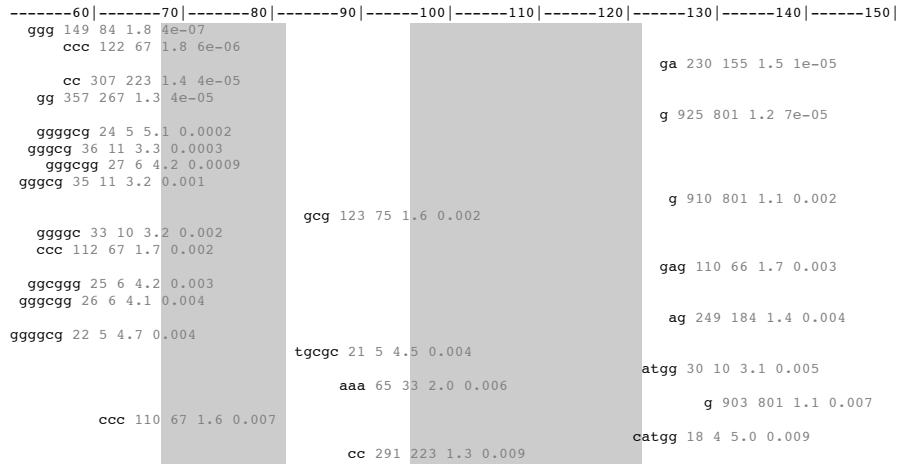




Figure S7:
Overrepresented DNA k-mers at fixed distances from dominant transcription start sites in RNA library BC (mouse cerebellum): 2601 sequences. Please see the legend for Fig. S2.