Methylated cytosines mutate to transcription factor binding sites that drive tetrapod evolution

Ximiao He[1], Desiree Tillo[1], Jeff Vierstra[2], Syed Khund-Sayeed[1], Callie Deng[1], G. Jordan Ray[1], John Stamatoyannopoulos[2], Peter C. Fitzgerald[3], Charles Vinson[1@]

[1]Laboratory of Metabolism, NCI, NIH,

[2]Department of Genome Sciences, University of Washington, WA, 98195,

[3]Genome Analysis Unit, Genetics Branch, NCI, NIH, Bethesda, MD 20892, USA.

[@]Corresponding author, Tel: (301) 496-8753, Fax: (301) 496-8419, E-mail: Vinsonc@mail.nih.gov

Running Title: 5mC mutate to TFBS that drives tetrapod evolution

Keywords: CG dinucleotide, tissue specific, C/EBP, TFBS, AP-1, CG methylation, coelacanth, TG dinucleotide.

Supplementary Information

Supplementary Figure Legend:

Figure S1. The bimodal distribution of abundance of 8-mers initially occurs in the coelacanth. Histogram showing the abundance of 32,896 continuous 8-mers in eukaryotic genomes: **(A)** Human, **(B)** Mouse, **(C)** Dog, **(D)** Elephant, **(E)** Opossum, **(F)** Wallaby, **(G)** Mink whale, **(H)** Dolphin**, (I)** Chicken**, (J)** Lizard, **(K)** *Xenopus*, **(L)** Coelacanth, **(M)** Elephant shark, **(N)** Zebrafish, **(O)** Nile Tilapia **(P)** Cod, **(Q)** Stickleback, **(R)** Fugu, **(S)** Tetraodon, **(T)** Fly, **(U)** Bee, **(V)** C. Elegans, **(W)** Arabidopsis and **(X)** Yeast. The x-axis displays the frequency of each 8-mer in the genome (8-mer occurrence per 100Kb), and the y-axis indicates the number of 8-mers with that frequency in the genome.

Figure S2. Abundance of 8-mers in human genome in conserved non-coding elements. (A) For each continuous 8-mer, the occurrence in human genome is plotted on the horizontal axis vs. the percentage in the CNEs representing 2.6% of the genome, plotted on the vertical axis. The histogram on the left is shows the distribution of percentage of 8-mers in CNEs. The 8-mers are divided into 3 groups based on the presence of the CG and TG dinucleotide. **(B)** Same as in **(A)**, but with log scale axes. **(C)** Same as in **Figure S2B**, but for CG-containing 8-mers only.

Figure S3. Abundance of 8-mers in human genome in different groups of CNEs. (A-G) For each continuous 8-mer, the occurrence in human genome is plotted on the horizontal axis vs. the enrichment in the each group of species-specific CNEs (form stickleback to dog) plotted on the vertical-axis. The 8-mers are divided into 3 groups based on the presence of the CG and TG dinucleotide. The species-specific CNEs are including: **(A)** Stickleback, **(B)** Coelacanth, **(C)** *Xenopus*, **(D)** Lizard/Chicken, **(E)** Opossum, **(F)** Elephant, and **(G)** Dog. Calculation of 8-mers in the mouse-specific CNEs is not shown due to the low coverage in the human genome (0.03% of human genome), resulting in non-significant values of 8-mer enrichment.

Figure S4. Abundance of three groups of 8-mers in the human genome versus their occurrence in DHSs. (A-C) For each continuous 8-mer, the occurrence in human genome is plotted on the horizontal axis vs. the percentage in the regulatory regions represented 9.0% of the genome (DNase I hypersensitive sites, DHSs) plotted on the vertical-axis. The histogram below the scatterplot shows the distribution of abundance of all 8-mers in human, and the histogram on the left shows the distribution of the percentage of 8-mers within DHSs. The 8-mers are divided into 3 groups: **(A)** no CG or TG; **(B)** with TG but not CG; and **(C)** with CG.

Figure S5. Abundance of 8-mers in the mouse genome versus their occurrence in DHSs. For each continuous 8-mer, the occurrence in mouse genome is plotted on the horizontal axis vs. the percentage in the regulatory regions (DNase I hypersensitive sites, DHSs) representing 10.2% of the genome plotted on the vertical-axis. The histogram below the scatterplot shows the distribution of abundance of all 8-mers in mouse, and the histogram on the left shows the distribution of the percentage of 8-mers within DHSs. 8-mers are divided into 3 groups based on the presence of the CG and TG dinucleotides as indicated in the legend.

Figure S6. Abundance of CG containing 8-mers in human genome versus their occurrence in DHSs. **(A)** For each continuous 8-mer, the occurrence in human genome is plotted on the horizontal axis vs. their percentage within DHSs covering 9.0% of the genome plotted on the vertical-axis. The 8-mers with CGs are divided into 4 groups based on the number of CGs in each 8-mer. **(B)** same as in **(A)**, but with log scale axes.

Figure S7. Abundance of 8-mers in human genome in DHSs +/- CNEs. **(A-D)** For each continuous 8-mer, the occurrence in human genome is plotted on the horizontal axis vs. the percentage within DHSs **(A)** overlapping conserved non-coding elements (DHSs+CNEs) representing 1.6% of the genome, and **(B)** those that do not (DHSs-CNEs) representing 7.4% of the genome is plotted on the vertical-axis. The 8-mers are divided into 3 groups based on the presence of the CG and TG dinucleotide. **(C-D)** Similar to **(A-B)**, but vertical-axes is enrichment of 8-mers and the DHSs are divided into two groups: **(C)** DHSs+CNEs and **(D)** DHSs-CNEs.

Figure S8. Enrichment of continuous 8-mers in human DHSs identified in 125 samples. **(A-C)** All the DHS are divided into 125 groups based on how many samples the DHS overlapped. The boxplots of enrichment of 8-mer in each sample number DHSs for the three groups of 8-mers: **(A)** no CG or TG; **(B)** with TG but not CG; and **(C)** with CG. **(D)** The percentage of DHSs representing in human genome for 125 DHS groups based on sample numbers. **(E)** The percentage of DHSs representing in human genome for six DHS groups based on sample numbers: Group 1 (only 1 sample); Group 2 (2 to 5 samples); Group 3 (6 to 20 samples); Group 4 (21 to 60 samples); Group 5 (61 to 124 samples) and Group 6 (all 125 samples, HK-DHSs).

Figure S9. Enrichment of continuous 8-mers in human DHSs identified in 125 samples. All the DHS are divided into 6 groups based on how many samples the DHS overlapped: Group 1 (only 1 sample); Group 2 (2 to 5 samples); Group 3 (6 to 20 samples); Group 4 (21 to 60 samples); Group 5 (61 to 124 samples) and Group 6 (all 125 samples, HK-DHSs). The 8-mer enrichment was plotted between Groups 1-5 (**A-E**) against Group 6.

Figure S10. TG containing 8-mers are enriched in tissue-specific DHSs in mouse. (A) For each continuous 8-mer, the percentage in the DHSs common to 55 mouse samples that represents 0.2% of the genome (HK-DHS) is plotted on the horizontal axis vs. the percentage in the tissue specific regulatory regions representing 10.0% of the genome (TS-DHS) plotted on the vertical-axis. The 8-mers are divided into 3 groups based on the presence of the CG and TG dinucleotide. **(B)** Same as in **(A)** but with the x-axis zoomed in.

Figure S11. NF-1 motifs are enriched in tissue-specific DHSs in human. For each continuous (NNNN-K_0-NNNN) and discontinuous 8-mer containing pairs of 4-mers separated by 1-bp to 14-bps (NNNNK$_{(1-14)}$NNNN), the enrichment in HK-DHSs is plotted on the horizontal axis vs. enrichment in TS-DHS. The 8-mers are color-coded based on the number of CGs. The AP-1 motifs (TGA-TCA) and NF-1 motif (TTGGCA---TGCCAA) are identified by different spacers between two 4-mers.

Figure S12. GR motifs are enriched in tissue-specific DHSs in human and mouse. For each discontinuous 8-mer with GR-like format (N-NNN---NNN-N), the enrichment in HK-DHSs is plotted on the horizontal axis vs. enrichment in TS-DHS for **(A)** mouse and **(B)** human genome. The 8-mers are color-coded based on the number of CGs. The GR motif (G-ACA---TGT-C) was identified as TS-DHS enriched.

Figure S13. Genome coverage for human DHSs identified in 125 samples and mouse DHSs identified in 55 samples. All the human DHS are divided into 6 groups based on how many samples the DHS overlapped (tissue-specificity): Group 1 (only 1 sample); Group 2 (2 to 5 samples); Group 3 (6 to 20 samples); Group 4 (21 to 60 samples); Group 5 (61 to 124) and Group 6 (all 125 samples, HK-DHSs). **(A)**

Percentage of the genome covered by group. Similar to mouse DHSs. (**A)** The DHS in each group identified by tissue specificity are divided into evolutionary types (Common, Human/Mouse-specific and Human/Mouse-unique DHSs) for human (**A**) and mouse (**B**).

Figure S14. AP-1 motifs are most enriched in the tissue-specific DHSs that are also unique and specific to human genome. Enrichment of 8-mers in the overlap of DHSs types by evolution (Common, Human-specific and Human-unique) and by specificity in cells in human genome. The enrichment in HK-DHSs (Group 6: all 125 cells) is plotted on the horizontal axis vs. enrichment in TS-DHS (Group 3: 6 to 20 cells) for all 8-mers **(A-C)** color coded by CG/TG occurrence in: **(A)** Human-unique DHSs, **(B)** Human-specific DHSs, and **(C)** Common DHSs. **(D-F)** The enrichment for 8-mers without CGs (-TG-CG, and +TG-CG) was plotted in **(D)** Human-unique, **(E)** Human-specific and **(F)** Common DHSs in human.

Figure S15. AP-1 motifs are most enriched in the tissue-specific DHSs that are also unique and specific to mouse genome. Enrichment of 8-mers in the overlap of DHSs types by evolution (Common, Mouse-specific and Mouse-unique) and by specificity in cells in mouse genome. The enrichment in HK-DHSs (Group 6 : all 55 samples) is plotted on the horizontal axis vs. enrichment in TS-DHS (Group 3: 5 to 12 samples) for all 8-mers **(A-C)** color coded by CG/TG occurrence in: **(A)** Mouse-unique DHSs, **(B)** Mouse-specific DHSs, and **(C)** Common DHSs. **(D-F)** The enrichment for 8-mers without CGs -TG-CG, and +TG-CG) was plotted in **(D)** Mouse-unique, **(E)** Mouse-specific and **(F)** Common DHSs in mouse.

Figure S16 Intrinsic nucleosome occupancy scores for AP-1 motifs in human and mouse. (A) Average intrinsic nucleosome occupancy score (INOS) for 177,730 AP-1 motifs (TGA$^{C}/_{G}$TCA) in DNase I hypersensitive sites (+DHS, blue), and 282,498 motifs not in DHSs (-DHS, black) within ±150-bps from the center of AP-1 motif (TGA$^{C}/_{G}$TCA) in human. **(B)** Similar to **(A)** but for AP-1 motifs in mouse. **(C)** Average INOS near (±150-bps) AP-1 motifs in common DHSs in human for the 10,501 motifs with exact AP-1 motif in mouse genome (purple), and the 17,938 motifs without exact AP-1 motif in mouse genome (red). **(D)** Similar to **(C)** but for AP-1 motifs in common DHSs in mouse. **(E)** Average INOS near (±150-bps) AP-1 motifs in human-specific DHSs for the 5,591 motifs with exact AP-1 motif in mouse genome (purple), and the 55,338 motifs without exact AP-1 motif in mouse genome (red). **(F)** Similar to **(E),** but for AP-1 motifs in mouse-specific DHSs.

Supplementary Tables:

Table S1. (A) Occurrence of 8-mers in human genome (hg19). **(B)** Occurrence of 8-mers in mouse genome (mm9). **(C)** Occurrence of 8-mers in simulated mouse genome with the same 2-mer frequency

Table S2. The genome percentage of CNEs covered by DHSs (Specific for each lineage).

Table S3. Description of 75 samples of human DHSs.

Table S4. The most enriched TG-containing 8-mers in TS-DHSs in Human (Top 20)

Table S5. The outlier 8-mers in **Figure 4** and their statistical significance.

Table S6. (A) The enrichment of E-Box motifs in human DHSs (TOP 20). **(B)** The enrichment of AP-1 motifs in human DHSs (TOP 20).

Table S7. The MRCA (most recent common ancestor) sequence to variations of 11mers with TGACTCA in human.

Table S8. (A-I) the MRCA (most recent common ancestor sequence) to variations of 11mers with TGACTCA in three groups 1) in CNEs and in DHSs; 2) in CNEs and no DHSs and 3) no CNEs for **(A)** Stickleback, **(B)** Coelacanth, **(C)** Xenopus, **(D)** Chicken, **(E)** Lizard, **(F)** Opossum, **(G)** Elephant, **(H)** Dog, **(I)** Mouse. The first two sections for each table are number and percentage of the base component of each position. The “Conserved” section shows the number and percentage of the AP-1 consensus motif in each group. For the “Variations” section, the first and last 2 “N”, the percentage indicates the base component at each position. For the middle 7 positions of AP-1 motif, the “0.0%” indicates the consensus base, while the percentage of each base indicates the variation component at each position. For example, in position 3, the variation component is A→T, C→T, and G→T.

Figure S1. Abundance of 8-mers in eukaryotic genomes

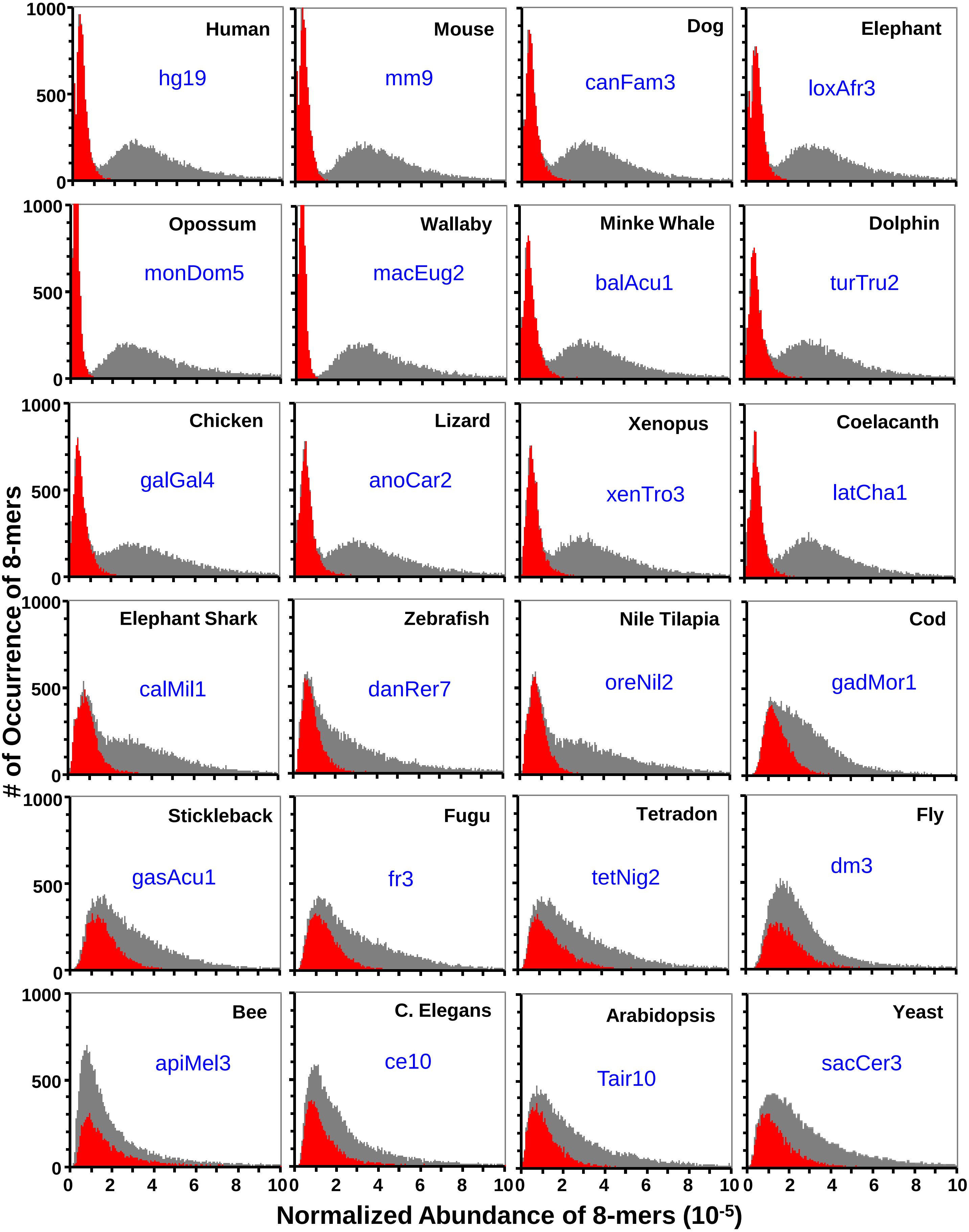


Figure S2

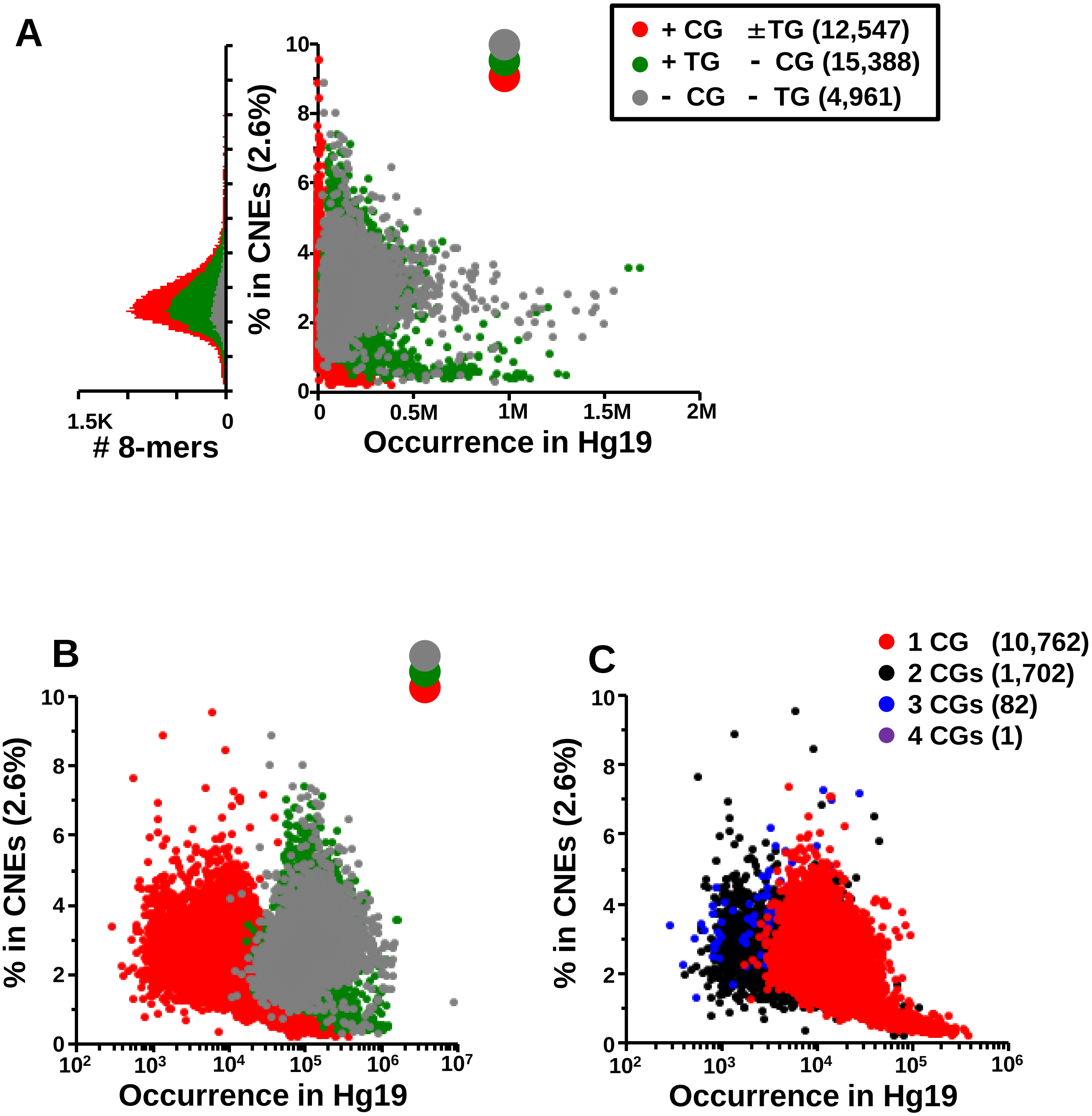

A
● + CG ±TG (12,547)
● + TG - CG (15,388)
● - CG - TG (4,961)
1.5K
0
8-mers
% in CNEs (2.6%)
10
8
6
4
2
0
0
0.5M
1M
1.5M
2M
Occurrence in Hg19
B
% in CNEs (2.6%)
10^2
10^3
10^4
10^5
10^6
10^7
Occurrence in Hg19
C
● 1 CG (10,762)
● 2 CGs (1,702)
● 3 CGs (82)
● 4 CGs (1)
% in CNEs (2.6%)
10^2
10^3
10^4
10^5
10^6
Occurrence in Hg19

Figure S3 Enrichment of 8-mers in CNEs

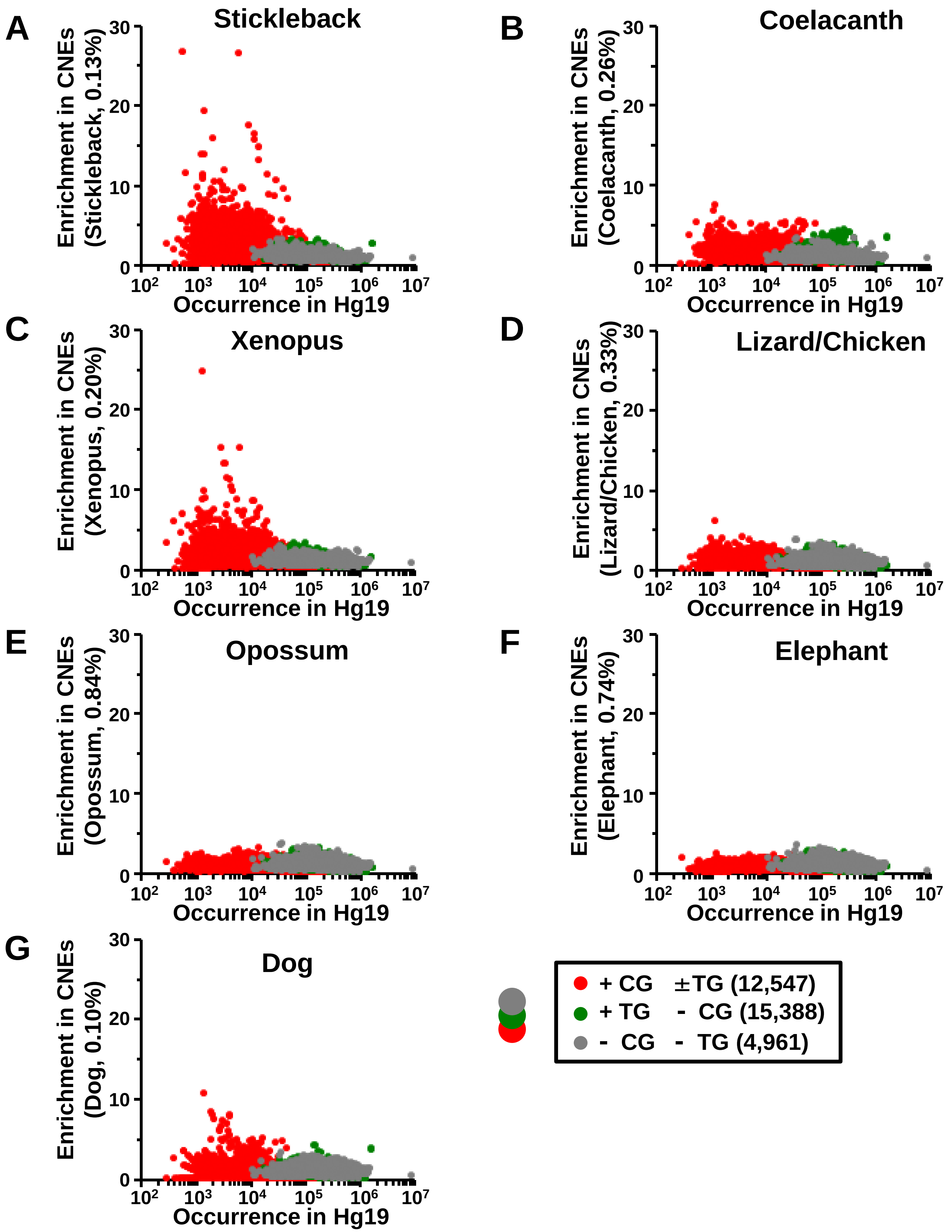


Figure S4 8-mer in different groups in human genome and DHSs

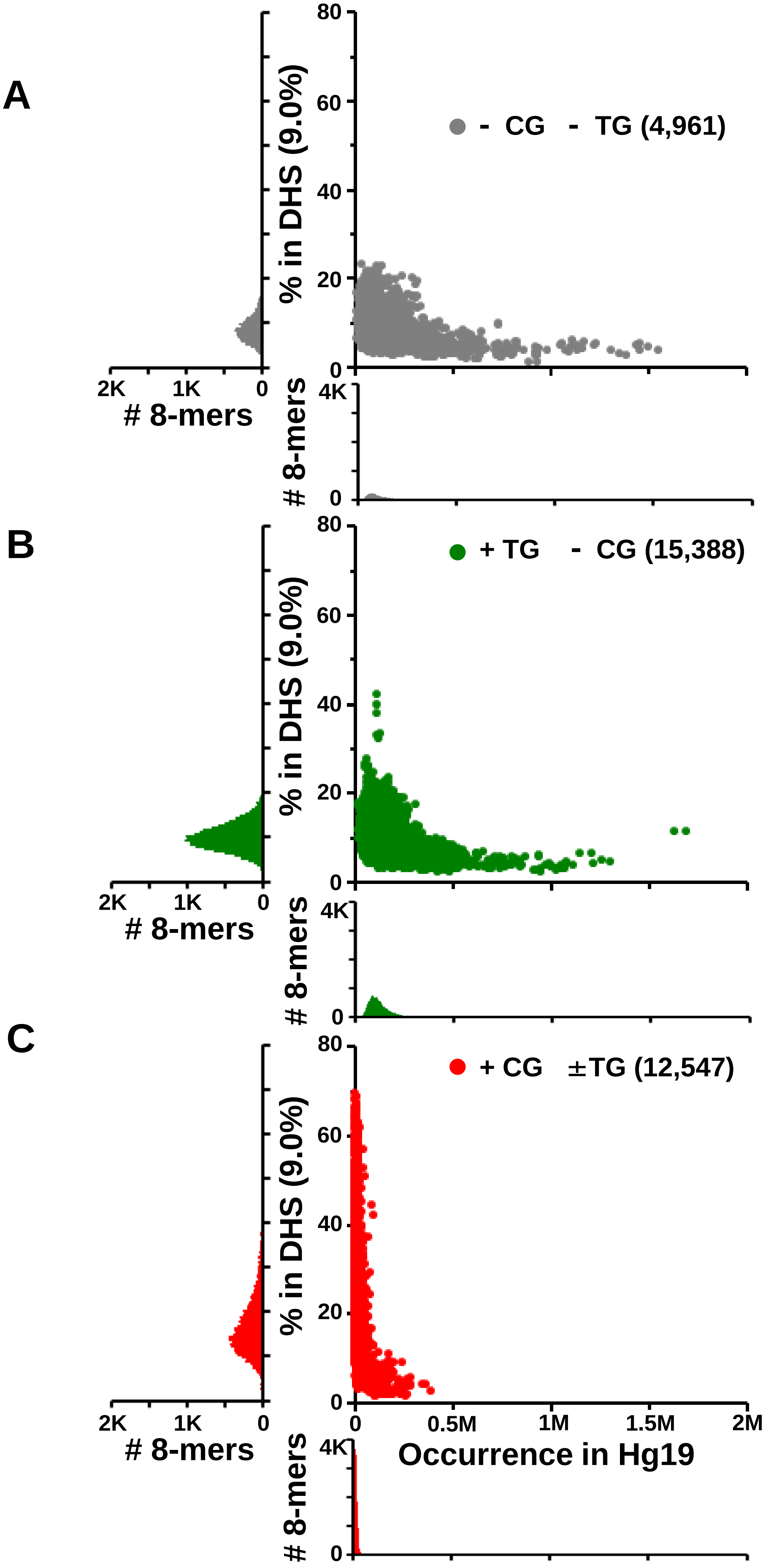


Figure S5

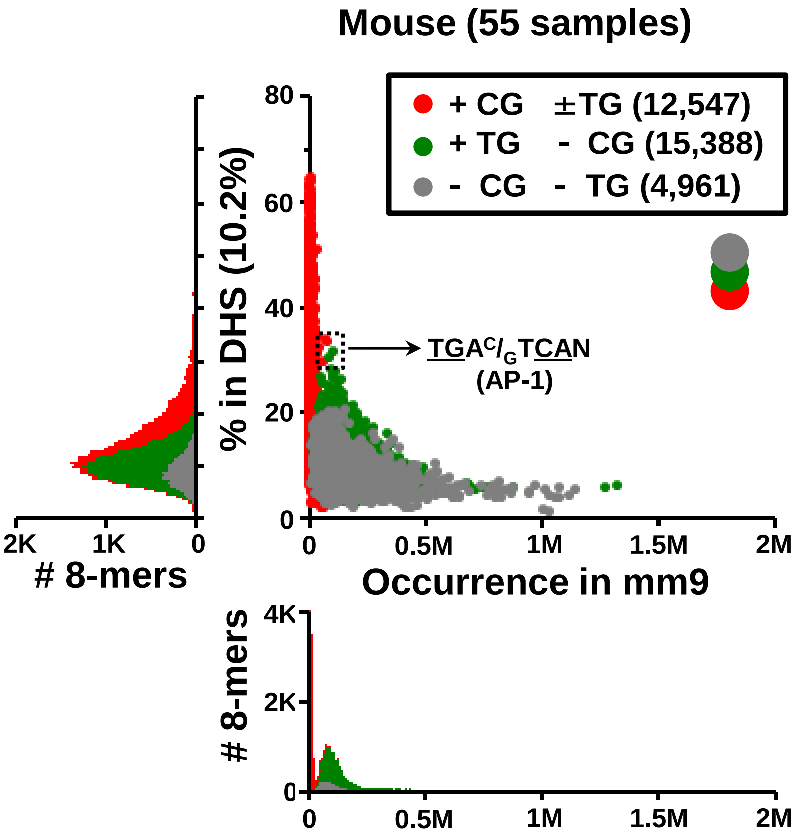

Mouse (55 samples)
● + CG ±TG (12,547)
● + TG - CG (15,388)
● - CG - TG (4,961)
TGA^C/_GTCAN
(AP-1)
% in DHS (10.2%)
80
60
40
20
0
0.5M
1M
1.5M
2M
Occurrence in mm9
2K
1K
8-mers
4K
2K

Figure S6. 8-mer with CGs in human genome and DHSs

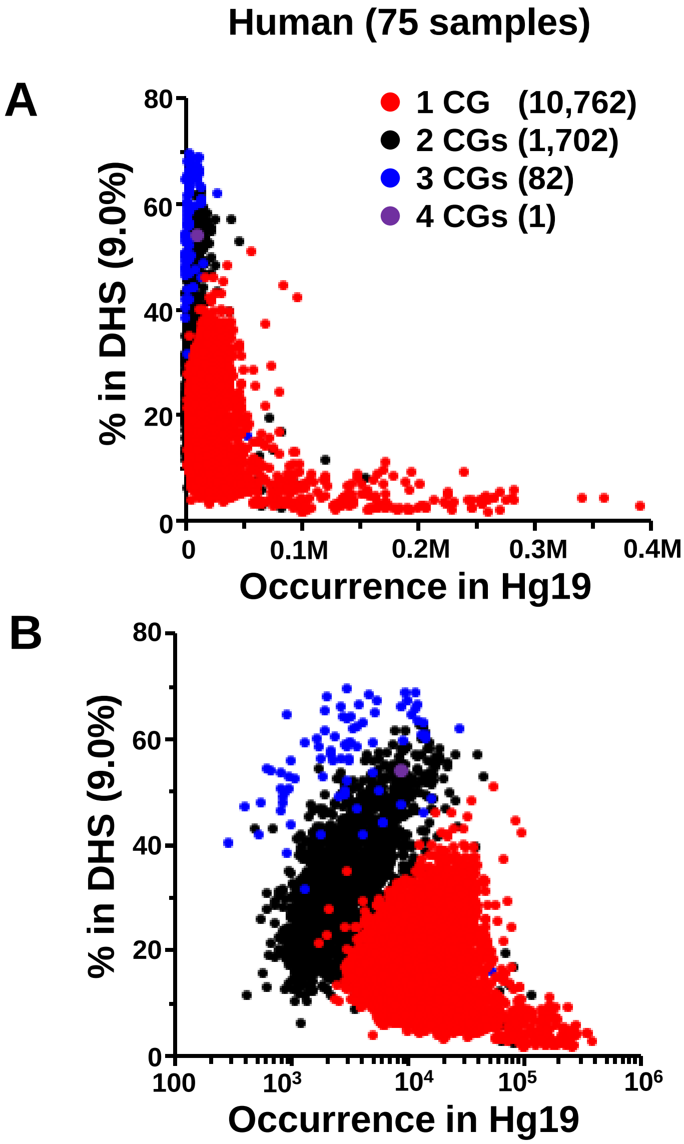
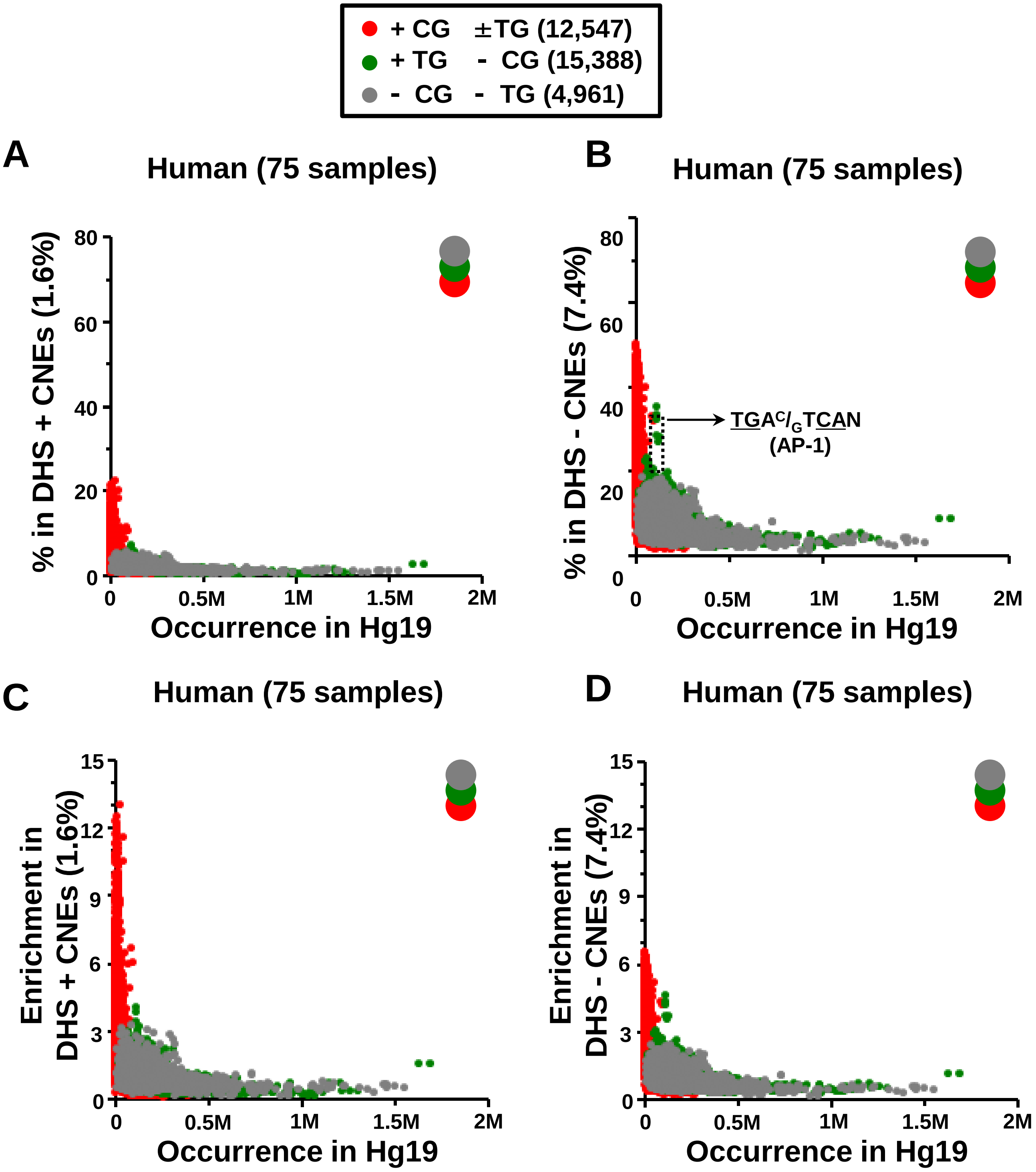


Figure S7



● + CG ±TG (12,547)
● + TG - CG (15,388)
● - CG - TG (4,961)
A
Human (75 samples)
% in DHS + CNEs (1.6%)
80
60
40
20
0
0
0.5M
1M
1.5M
2M
Occurrence in Hg19
B
Human (75 samples)
% in DHS - CNEs (7.4%)
80
60
40
20
0
TGAC/GTCAN
(AP-1)
0
0.5M
1M
1.5M
2M
Occurrence in Hg19
C
Human (75 samples)
Enrichment in DHS + CNEs (1.6%)
15
12
9
6
3
0
0
0.5M
1M
1.5M
2M
Occurrence in Hg19
D
Human (75 samples)
Enrichment in DHS - CNEs (7.4%)
15
12
9
6
3
0
0
0.5M
1M
1.5M
2M
Occurrence in Hg19

Figure S8: Enrichment of 8-mers in human DHSs by shared sample numbers

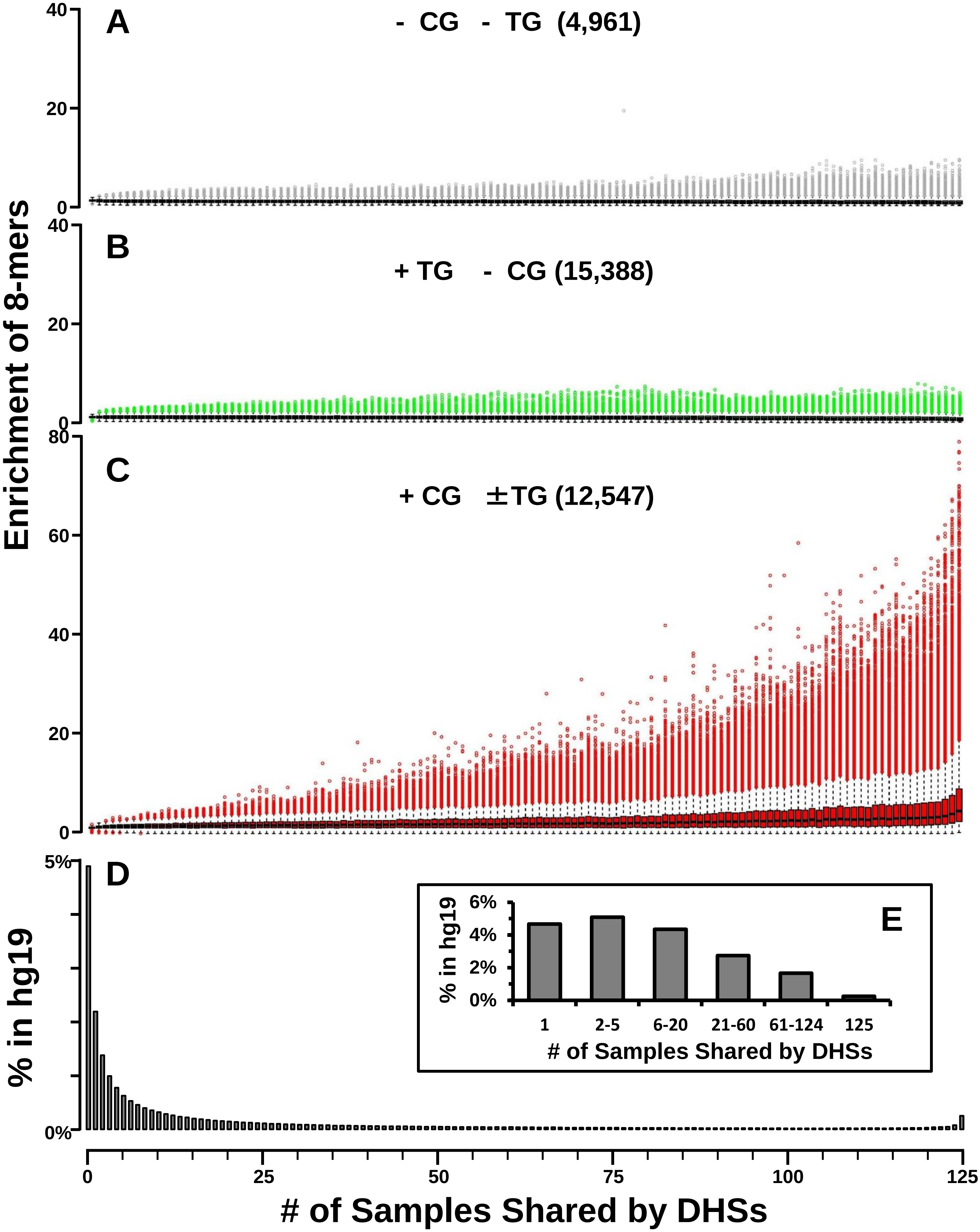


Figure S9: Enrichment of 8-mers in human DHSs

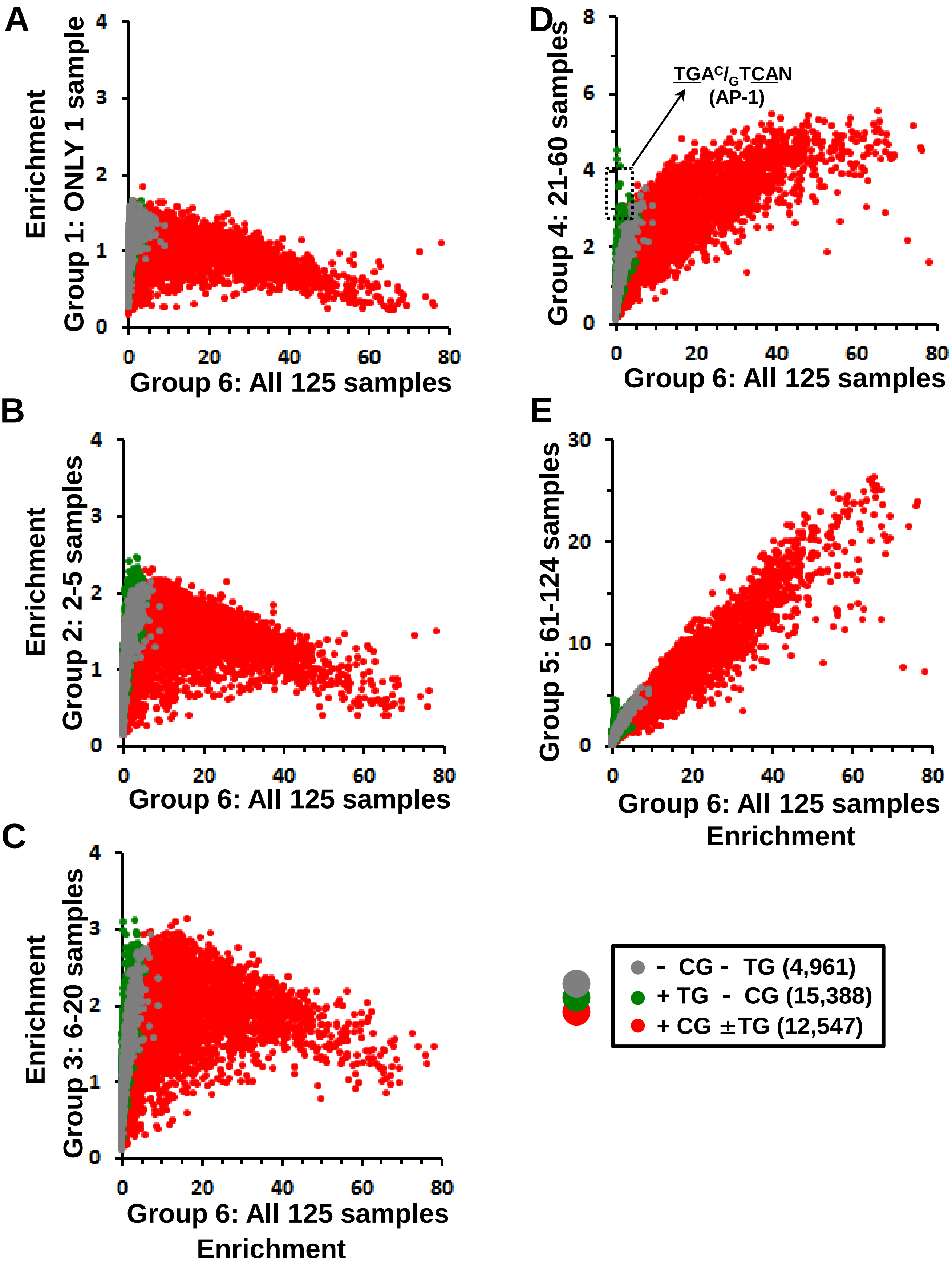


Figure S10

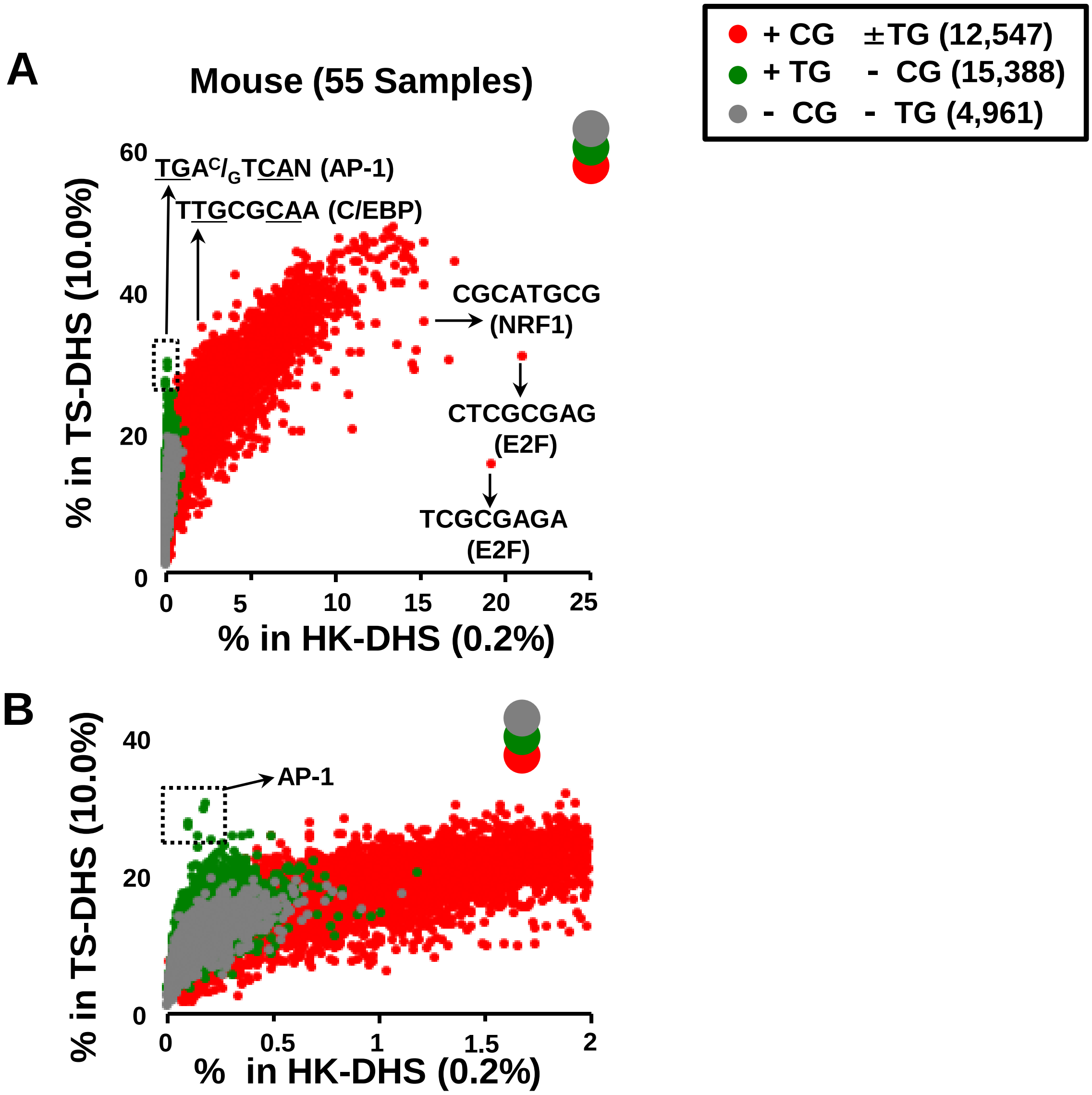

● + CG ±TG (12,547)
● + TG - CG (15,388)
● - CG - TG (4,961)
A
Mouse (55 Samples)
TGA^{C}/$_{G}$TCAN (AP-1)
TTGCGCAA (C/EBP)
CGCATGCG (NRF1)
CTCGCGAG (E2F)
TCGCGAGA (E2F)
60
40
20
0
0
5
10
15
20
25
% in TS-DHS (10.0%)
% in HK-DHS (0.2%)
B
AP-1
40
20
0
0
0.5
1
1.5
2
% in TS-DHS (10.0%)
% in HK-DHS (0.2%)

Figure S11. split 8-mers enriched in tissue specific DHSs in Human

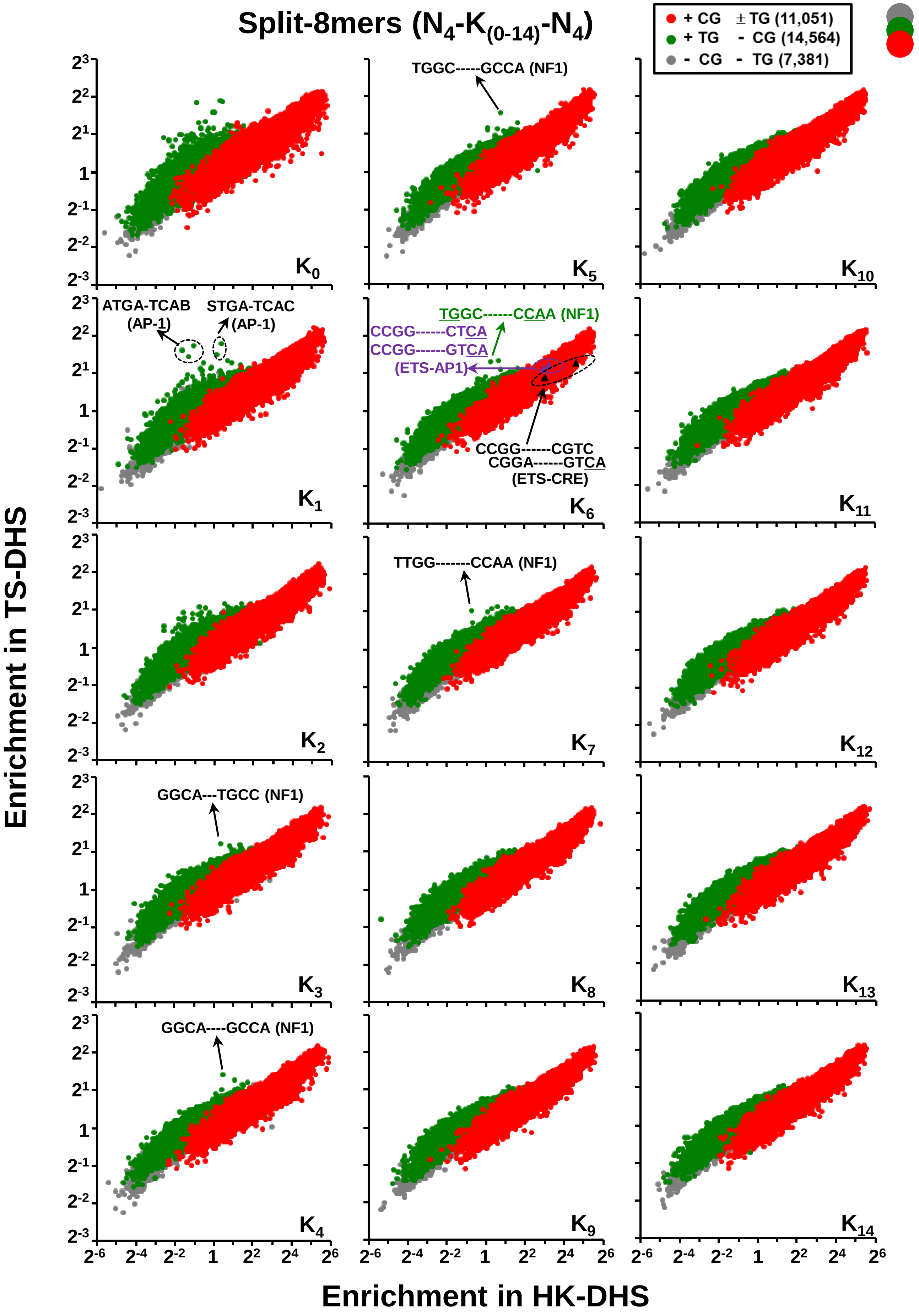


Figure S12. GR consensus motif is enriched in tissue specific DHSs in both Human and Mouse

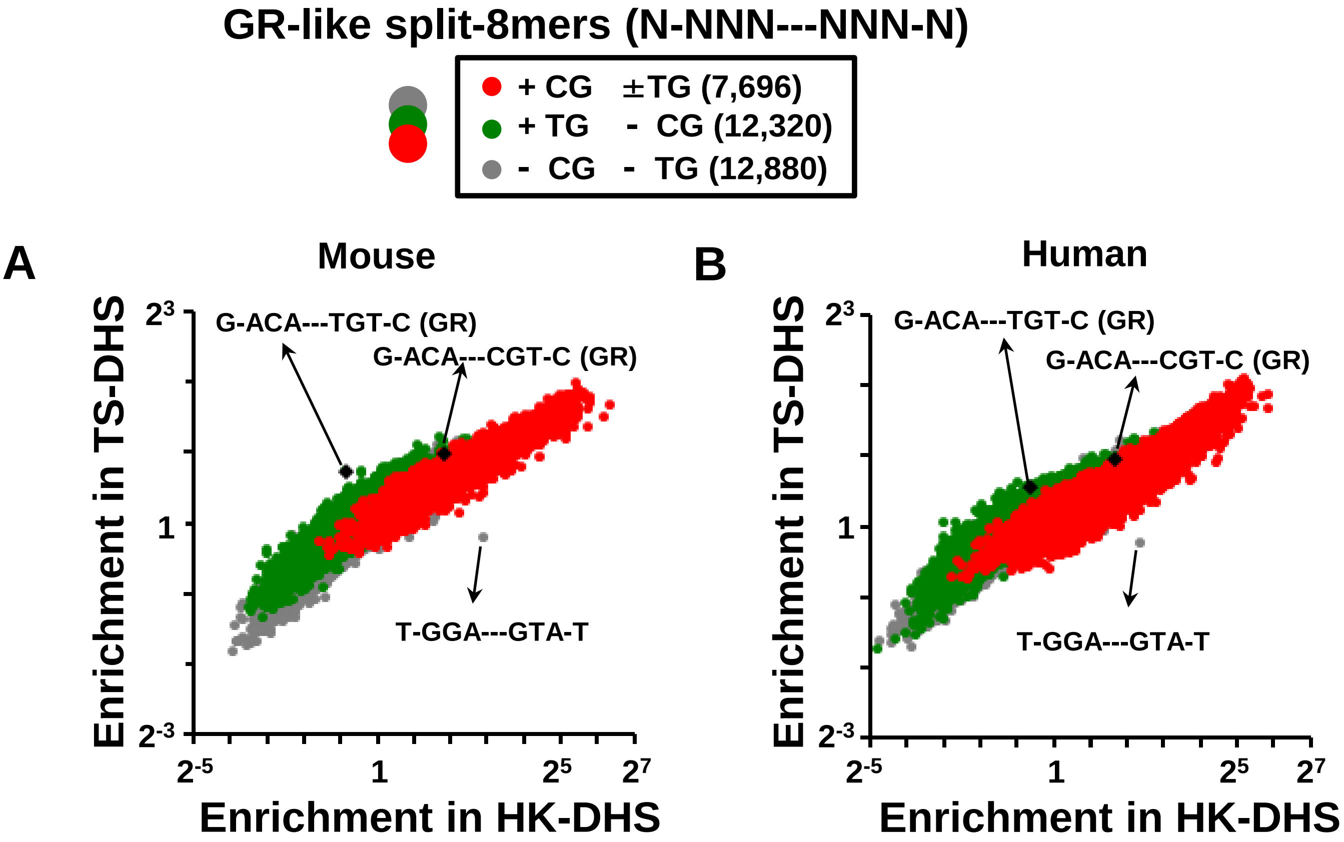
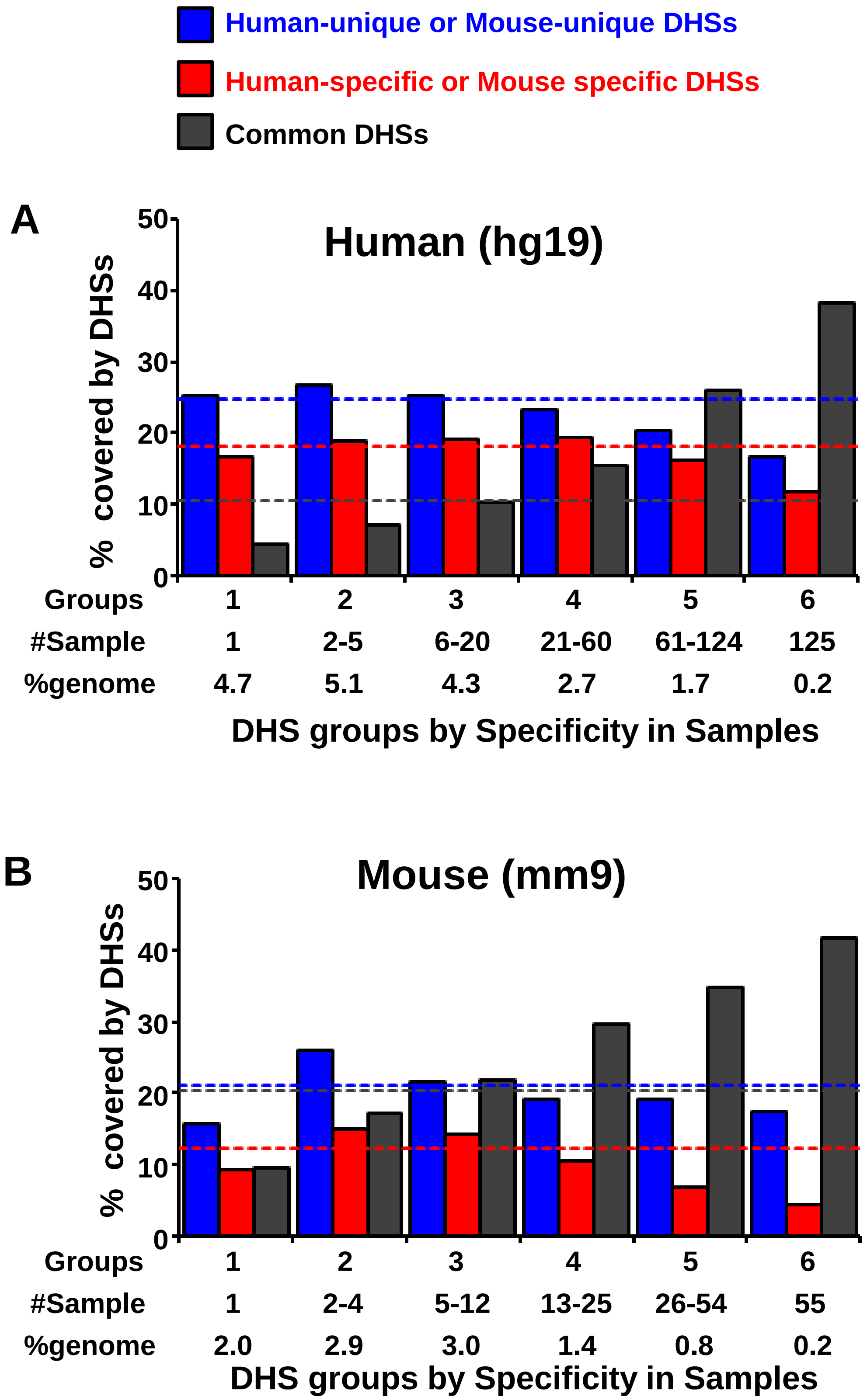


Figure S13.



Human-unique or Mouse-unique DHSs
Human-specific or Mouse specific DHSs
Common DHSs
A
Human (hg19)
% covered by DHSs
0
10
20
30
40
50
Groups
1
2
3
4
5
6
#Sample
1
2-5
6-20
21-60
61-124
125
%genome
4.7
5.1
4.3
2.7
1.7
0.2
DHS groups by Specificity in Samples
B
Mouse (mm9)
% covered by DHSs
0
10
20
30
40
50
Groups
1
2
3
4
5
6
#Sample
1
2-4
5-12
13-25
26-54
55
%genome
2.0
2.9
3.0
1.4
0.8
0.2
DHS groups by Specificity in Samples

Figure S14. Enrichment of 8-mers in the overlap of human DHSs types by evolution and by specificity in cells

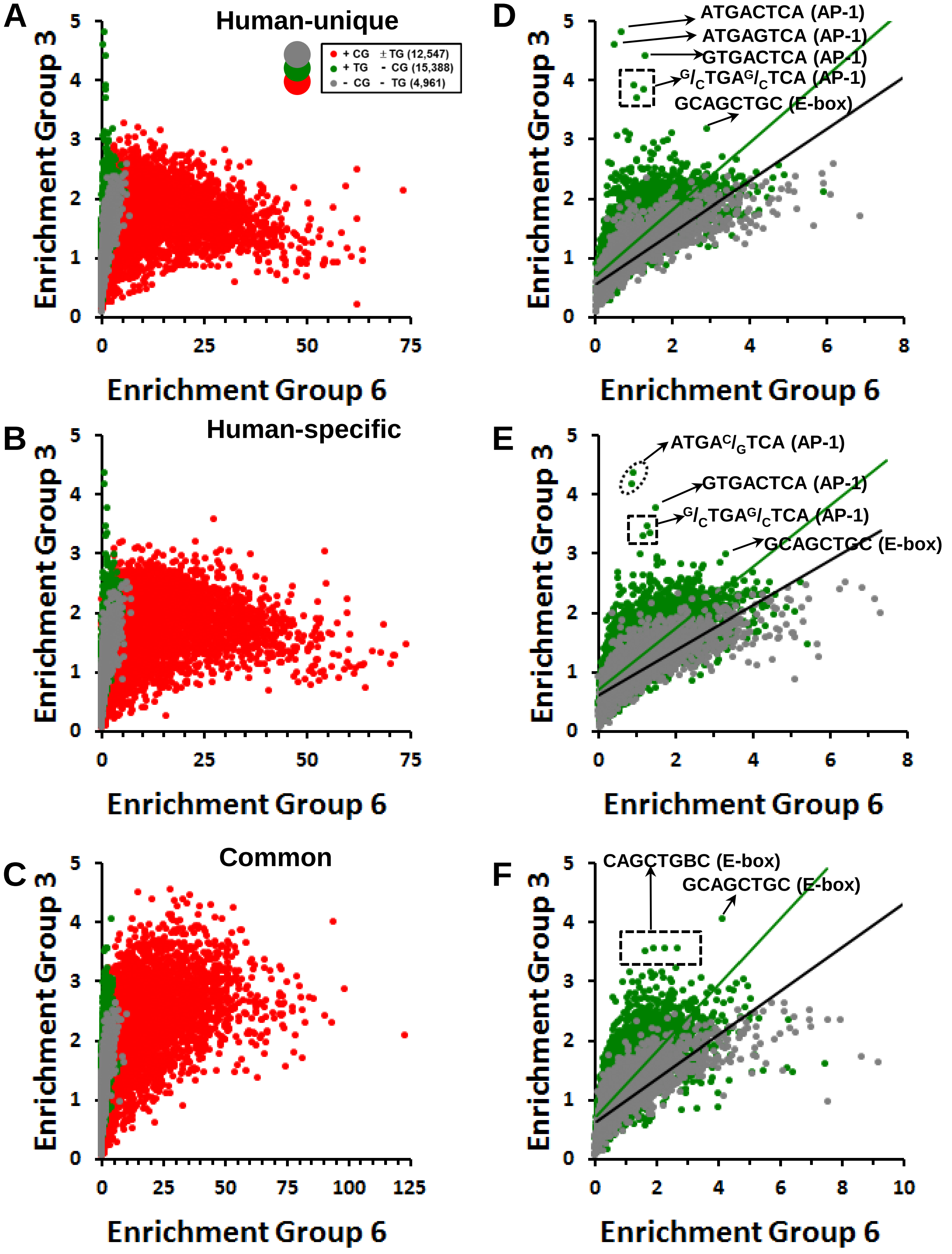


Figure S15. Enrichment of 8-mers in the overlap of mouse DHSs types by evolution and by cell specificity

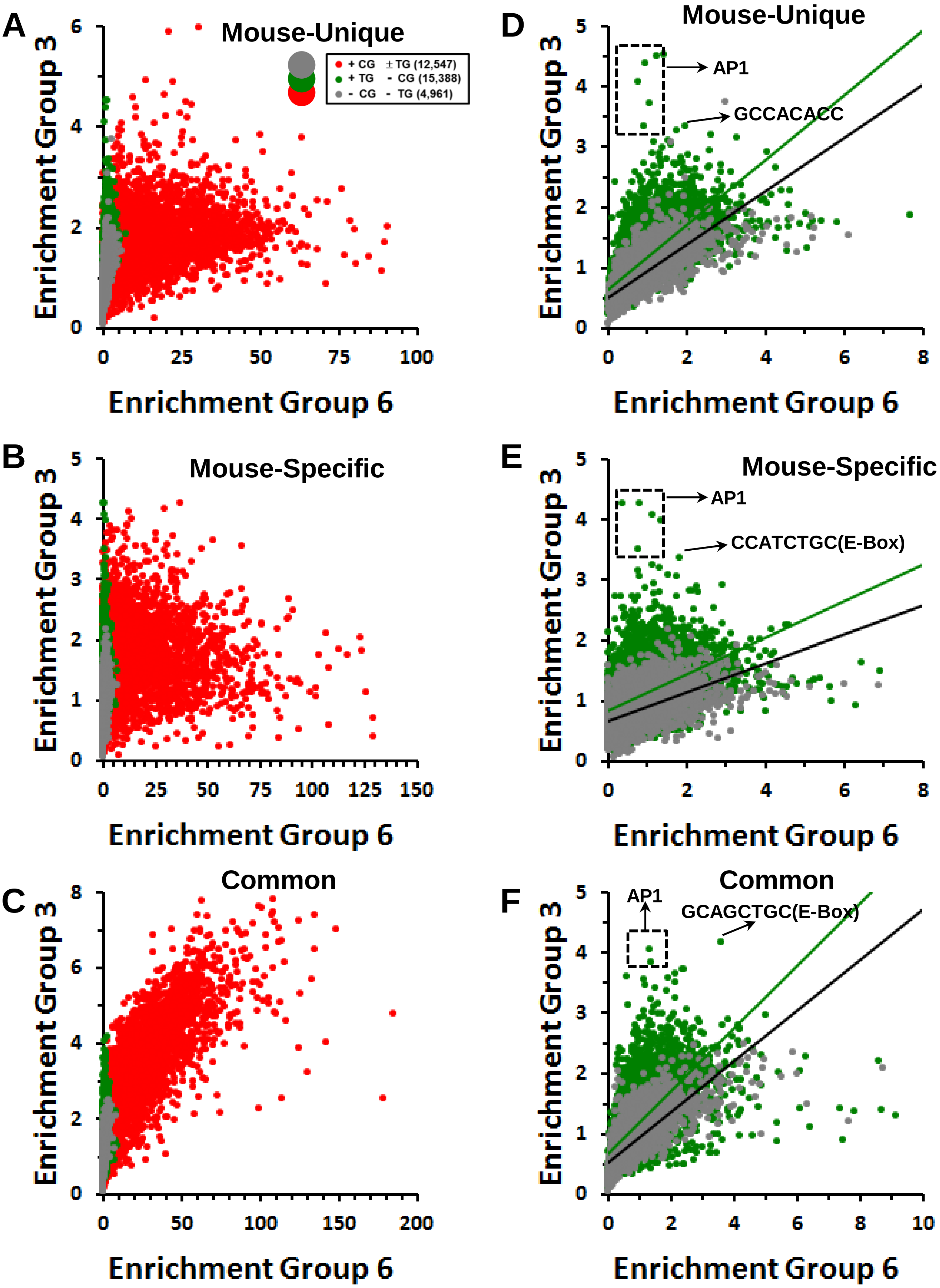


Figure S16. INOS around AP-1 motifs in human and mouse genomes

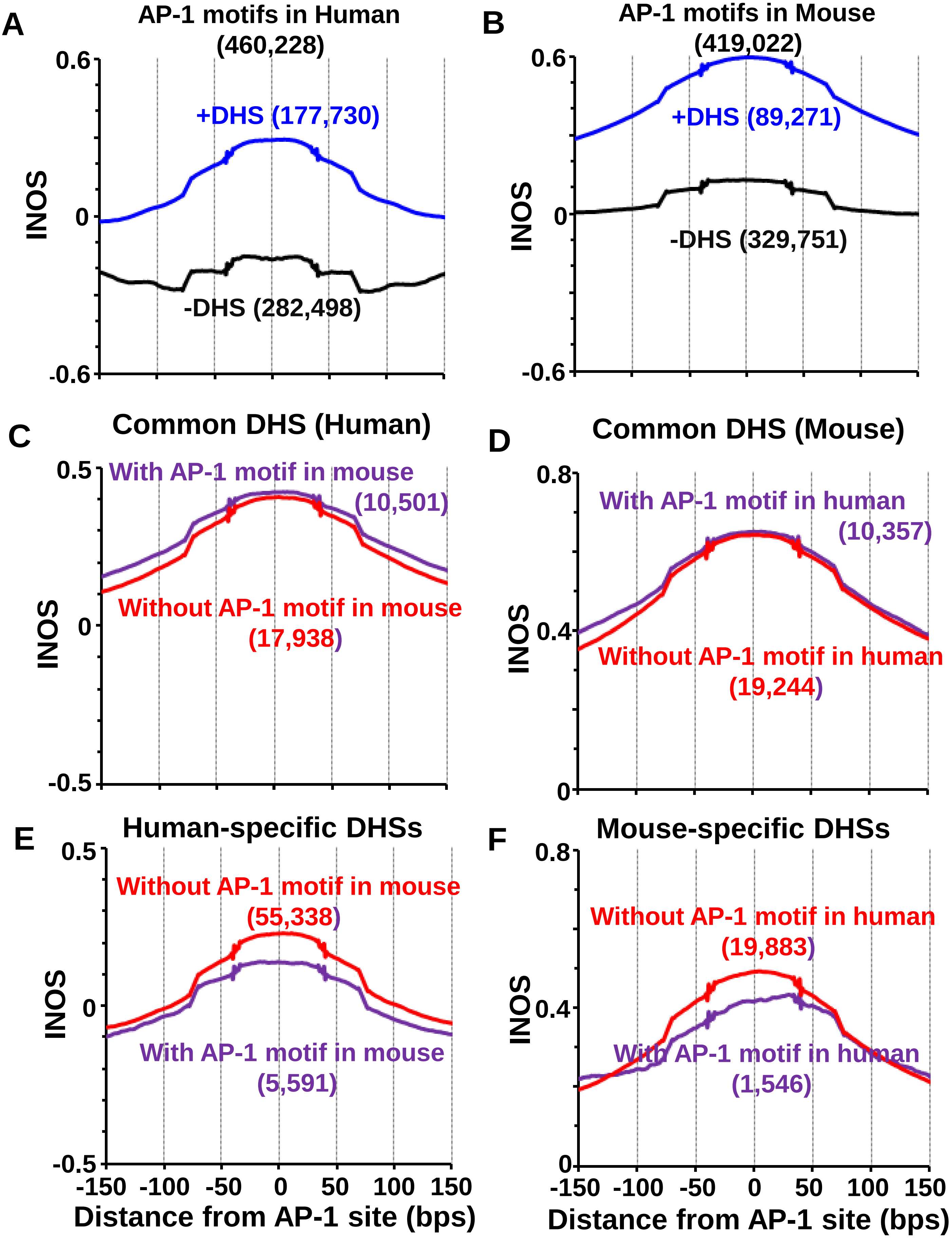


Table S1A. Occurrence of 8-mers in human genome (hg19)

Type	# of 8-mers	hg19		
		mean	SD	median
4-CpGs	1	8,749	-	8,749
3-CpGs	82	5,314	7,254	3,047
2-CpGs	1,702	5,274	10,003	2,961
1-CpGs	10,762	16,561	21,201	12,411
0-CpGs No CA|TG	4,961	139,309	187,376	**96,700**
0-CpGs With CA|TG	15,388	128,839	87,414	**108,974**
0-CpGs with 1 CA|TG	9,844	129,290	92,529	106,215
0-CpGs with 2 CA|TG	4,944	125,383	70,981	110,474
0-CpGs with 3 CA|TG	592	147,268	97,781	130,663
0-CpGs with 4 CA|TG	10	302,558	493,921	183,265
0-CpGs All	20,349	131,392	119,820	106,864

Table S1B. Occurrence of 8-mers in mouse genome (mm9)

Type	# of 8-mers	mm9		
		mean	SD	median
4-CpGs	1	25,805	-	25,805
3-CpGs	82	3,869	5,000	1,827
2-CpGs	1,702	3,134	3,007	2,307
1-CpGs	10,762	12,793	7,403	11,108
0-CpGs No CA|TG	4,961	122,936	134,830	92,315
0-CpGs With CA|TG	15,388	117,317	85,390	104,257
0-CpGs with 1 CA|TG	9,844	115,657	63,165	100,985
0-CpGs with 2 CA|TG	4,944	115,810	51,348	106,470
0-CpGs with 3 CA|TG	592	146,927	224,763	124,061
0-CpGs with 4 CA|TG	10	744,386	1,684,815	236,315
0-CpGs All	20,349	118,687	99,754	102,034

Table S1C. Occurrence of 8-mers in simulated mouse genome with the same 2-mer frequency

Type	# of 8-mers	mm9.2mer.simulated		
		mean	SD	median
4-CpGs	1	9	-	9
3-CpGs	82	168	38	177
2-CpGs	1,702	1,512	372	1,498
1-CpGs	10,762	13,347	3,892	12,901
0-CpGs No CA|TG	4,961	114,682	44,772	107,315
0-CpGs With CA|TG	15,388	119,792	38,227	114,410
0-CpGs with 1 CA|TG	9,844	120,040	40,878	113,302
0-CpGs with 2 CA|TG	4,944	119,674	33,999	116,694
0-CpGs with 3 CA|TG	592	117,166	22,711	119,834
0-CpGs with 4 CA|TG	10	88,877	28,176	105,910
0-CpGs All	20,349	118,546	39,981	112,486

Table S2 The genome percentage of CNEs covered by DHSs (Specific for each lineage)

	Total			CNEs covered by DHSs				CNEs no DHSs			
Referred Species	**#**	**(bps)**	**(%)**	**#**	**%**	**bps**	**%**	**#**	**%**	**bps**	**%**
Sea urchin	**19,384**	**1,498,590**	**0.05%**	**5,146**	**26.5%**	**488,000**	**0.02%**	**14,238**	**73.5%**	**1,010,590**	**0.03%**
Stickleback	29,268	3,998,819	0.13%	10,574	36.1%	2,110,947	0.07%	18,694	63.9%	1,887,872	0.06%
Coelacanth	53,985	8,148,181	0.26%	27,613	51.1%	5,475,222	0.18%	26,372	48.9%	2,672,959	0.09%
Xenopus	44,200	6,161,506	0.20%	17,638	39.9%	3,178,145	0.10%	26,562	60.1%	2,983,361	0.10%
Chicken/Lizard	92,263	10,253,282	0.33%	46,190	50.1%	6,009,197	0.20%	46,073	49.9%	4,244,085	0.14%
Opossum	254,019	25,820,535	0.84%	136,678	53.8%	15,826,549	0.51%	117,341	46.2%	9,993,986	0.32%
Elephant	225,958	22,755,844	0.74%	114,320	50.6%	13,283,245	0.43%	111,638	49.4%	9,472,599	0.31%
Dog	29,755	2,965,269	0.10%	12,986	43.6%	1,468,401	0.05%	16,769	56.4%	1,496,868	0.05%
Mouse	10,149	931,843	0.03%	5,062	49.9%	505,467	0.02%	5,087	50.1%	426,376	0.01%
Human	739,597	81,035,297	2.63%	371,061	50.2%	47,857,173	1.55%	368,536	49.8%	33,178,106	1.08%

	Total			CNEs in DHSs and G+C% <= 42%				CNEs in DHSs and G+C% > 42%			
Referred Species	**#**	**(bps)**	**(%)**	**#**	**%**	**bps**	**%**	**#**	**%**	**bps**	**%**
Sea urchin	**19,384**	**1,498,590**	**0.05%**	**4,035**	**20.8%**	**372,306**	**0.01%**	**1,111**	**5.7%**	**115,694**	**0.00%**
Stickleback	29,268	3,998,819	0.13%	4,310	14.7%	1,070,080	0.03%	6,264	21.4%	1,040,867	0.03%
Coelacanth	53,985	8,148,181	0.26%	17,374	32.2%	3,793,336	0.12%	10,239	19.0%	1,681,886	0.05%
Xenopus	44,200	6,161,506	0.20%	10,731	24.3%	2,031,331	0.07%	6,907	15.6%	1,146,814	0.04%
Chicken/Lizard	92,263	10,253,282	0.33%	33,081	35.9%	4,629,093	0.15%	13,109	14.2%	1,380,104	0.04%
Opossum	254,019	25,820,535	0.84%	93,643	36.9%	11,626,903	0.38%	43,035	16.9%	4,199,646	0.14%
Elephant	225,958	22,755,844	0.74%	75,361	33.4%	9,208,637	0.30%	38,959	17.2%	4,074,608	0.13%
Dog	29,755	2,965,269	0.10%	8,768	29.5%	1,041,380	0.03%	4,218	14.2%	427,021	0.01%
Mouse	10,149	931,843	0.03%	3,011	29.7%	308,071	0.01%	2,051	20.2%	197,396	0.01%
Human	739,597	81,035,297	2.63%	246,279	33.3%	33,708,831	1.09%	124,782	16.9%	14,148,342	0.46%

	Total			CNEs no DHSs and G+C% < 42%				CNEs no DHSs and G+C% > 42%			
Referred Species	**#**	**(bps)**	**(%)**	**#**	**%**	**bps**	**%**	**#**	**%**	**bps**	**%**
Sea urchin	**19,384**	**1,498,590**	**0.05%**	**5,146**	**26.5%**	**488,000**	**0.02%**	**14,238**	**73.5%**	**1,010,590**	**0.03%**
Stickleback	29,268	3,998,819	0.13%	10,889	37.2%	1,038,322	0.03%	7,805	26.7%	849,550	0.03%
Coelacanth	53,985	8,148,181	0.26%	21,224	39.3%	2,162,601	0.07%	5,148	9.5%	510,358	0.02%
Xenopus	44,200	6,161,506	0.20%	18,772	42.5%	1,928,453	0.06%	7,790	17.6%	1,054,908	0.03%
Chicken/Lizard	92,263	10,253,282	0.33%	39,986	43.3%	3,763,800	0.12%	6,087	6.6%	480,285	0.02%
Opossum	254,019	25,820,535	0.84%	102,515	40.4%	8,935,679	0.29%	14,826	5.8%	1,058,307	0.03%
Elephant	225,958	22,755,844	0.74%	97,499	43.1%	8,461,032	0.27%	14,139	6.3%	1,011,567	0.03%
Dog	29,755	2,965,269	0.10%	14,841	49.9%	1,355,497	0.04%	1,928	6.5%	141,371	0.00%
Mouse	10,149	931,843	0.03%	4,089	40.3%	356,276	0.01%	998	9.8%	70,100	0.00%
Human	739,597	81,035,297	2.63%	309,815	41.9%	28,001,660	0.91%	58,721	7.9%	5,176,446	0.17%

Table S3. Description of 75 samples of human DHSs

Cells	#Samples	Types	Notes
AG04449	2	fibroblasts	fetal buttock/thigh fibroblast
AG04450	2	fibroblasts	fetal lung fibroblast
AG09309	2	fibroblasts	adult toe fibroblast from apparently healthy 21 year old
AG09319	2	fibroblasts	gum tissue fibroblasts from apparently heathly 24 year old
AG10803	2	fibroblasts	abdominal skin fibroblasts from apparently heathly 22 year old
BJ	2	fibroblasts	skin fibroblast, (PMID: 9916803)
Caco2	2	cancer cells	colorectal adenocarcinoma. (PMID: 1939345)
CMK	1	cancer cells	acute megakaryocytic leukemia cells (PMID: 3016165)
GM06990	2	B cells	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed
GM12865	2	B cells	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed
GM12878	2	B cells	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed
HAEpiC	2	epithelial cells	amniotic epithelial cells
HCF	2	fibroblasts	cardiac fibroblasts
HCM	2	myocytes	cardiac myocytes
HCPEpiC	2	epithelial cells	choroid plexus epithelial cells
HEEpiC	2	epithelial cells	esophageal epithelial cells
HeLaS3	2	cancer cells	cervical carcinoma
HepG2	2	cancer cells	hepatocellular carcinoma
HGF	2	fibroblasts	gingival fibroblasts
HIPEpiC	2	epithelial cells	iris pigment epithelial cells
HL60	2	cancer cells	promyelocytic leukemia cells, (PMID: 276884)
HMEC	1	epithelial cells	mammary epithelial cells
HNPCEpiC	2	epithelial cells	non-pigment ciliary epithelial cells
HRCEpiC	2	epithelial cells	renal cortical epithelial cells
HRE	2	epithelial cells	renal epithelial cells
HRPEpiC	2	epithelial cells	retinal pigment epithelial cells
HUVEC	1	endothelial cells	umbilical vein endothelial cells
Jurkat	2	cancer cells	T lymphoblastoid derived from an acute T cell leukemia. (PMID: 68013)
Jurkat	1	cancer cells	T lymphoblastoid derived from an acute T cell leukemia. (PMID: 68013)
K562	2	cancer cells	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC
MCF7	2	cancer cells	mammary gland, adenocarcinoma. (PMID: 4357757), newly promoted to tier 2: not in 2011 analysis
NB4	2	cancer cells	acute promyelocytic leukemia cell line. (PMID: 1995093)
NHDFneo	2	fibroblasts	neonatal dermal fibroblasts
NHEK	1	keratinocytes	epidermal keratinocytes
NHLF	2	fibroblasts	lung fibroblasts
PANC1	2	cancer cells	pancreatic carcinoma, (PMID: 1140870)
SAEC	2	epithelial cells	small airway epithelial cells
SKMC	2	skeletal muscle cells	skeletal muscle cells
SKNSHRA	2	cancer cells	neuroblastoma cell line, treatment: differentiated with retinoic acid
Th1	1	T cells	primary Th1 T cells
Th2	1	T cells	primary Th2 T cells

Table S4. The most enriched TG-containing 8-mers in TS-DHSs in Human (Top 20)

8mer	Rvs_cmp_8mer	TFBSs	#in_Genome	%in_DHS_HK	%in_DHS_TS
ATGACTCA	TGAGTCAT	AP-1	109,976	0.1%	41.70%
ATGAGTCA	TGACTCAT	AP-1	112,487	0.1%	39.60%
GTGACTCA	TGAGTCAC	AP-1	108,976	0.3%	37.50%
CTGAGTCA	TGACTCAG	AP-1	127,041	0.2%	33.00%
GTGAGTCA	TGACTCAC	AP-1	115,946	0.2%	32.80%
CTGACTCA	TGAGTCAG	AP-1	124,674	0.2%	31.90%
GAGTCATC	GATGACTC	AP-1	61,429	0.2%	27.30%
GACTCATC	GATGAGTC	AP-1	52,483	0.1%	26.40%
GAGTCAGC	GCTGACTC	AP-1	67,729	0.4%	25.80%
CATGACTC	GAGTCATG	AP-1	73,203	0.2%	25.40%
GACTCACC	GGTGAGTC	AP-1	53,567	0.5%	25.30%
GCAGCTGC	GCAGCTGC	E-Box	69,467	0.8%	25.10%
CATGAGTC	GACTCATG	AP-1	70,593	0.1%	24.70%
GAGTCACC	GGTGACTC	AP-1	70,528	0.4%	24.70%
AGTGACTC	GAGTCACT	AP-1	91,863	0.2%	24.10%
GACTCAGC	GCTGAGTC	AP-1	86,224	0.4%	23.70%
AGGAAGTG	CACTTCCT	ETS1	177,209	0.3%	23.10%
GAGTCATA	TATGACTC	AP-1	62,505	0.1%	23.10%
AATGAGTC	GACTCATT	AP-1	84,909	0.1%	22.90%
AATGACTC	GAGTCATT	AP-1	92,558	0.1%	22.50%

Table S5. The outlier 8-mers in Figure 4B and their statistical significance

8-mers	rMers	TFBS	#in Genome	p.in.HK_DHS	p.in.TS_DHS
TGAGTCAA	TTGACTCA	AP-1	116,602	1.70E-18	<1.0E-324
TGACTCAA	TTGAGTCA	AP-1	104,754	1.50E-12	<1.0E-324
GTGACTCA	TGAGTCAC	AP-1	108,976	4.52E-08	<1.0E-324
ATGAGTCA	TGACTCAT	AP-1	112,487	1.69E-07	<1.0E-324
ATGACTCA	TGAGTCAT	AP-1	109,976	7.30E-07	<1.0E-324
GCAGCTGC	GCAGCTGC	E-Box	69,467	6.16E-244	<1.0E-324
TCGCGAGA	TCTCGCGA	E2F	2,252	<1.0E-324	3.22E-90
CGCATGCG	CGCATGCG	NRF1	3,172	<1.0E-324	9.59E-228
CTCGCGAG	CTCGCGAG	E2F	2,195	<1.0E-324	3.99E-109
GCCCCCCC	GGGGGGGC	unknown/SP-1/KLF	31,007	<1.0E-324	<1.0E-324
GATTGGCC	GGCCAATC	CCAAT	38,430	<1.0E-324	3.30E-216
GAGGGGGC	GCCCCCTC	unknown/SP-1/KLF	84,959	<1.0E-324	<1.0E-324

Table S6A The enrichment of E-Box motifs in human DHSs (TOP 20)

cells	Enrichment	types	notes
HSMMtube	8.38	myotubes	skeletal muscle myotubes differentiated from the HSMM cell line
HSMMemb	7.52	myocytes	embryonic myoblast
HSMM	6.59	myocytes	skeletal muscle myoblasts
Caco2	6.59	cancer cells	colorectal adenocarcinoma. (PMID: 1939345)
SKNSHRA	5.98	cancer cells	neuroblastoma cell line, treatment: differentiated with retinoic acid
iPS	5.91	ESC	induced pluripotent stem cell derived from skin fibroblast
SKMC	5.45	skeletal muscle cells	skeletal muscle cells
BE2C	5.28	cancer cells	neuroblastoma, BE-C is a clone of the SK-N-BE neuroblastoma cell line (see ATCC CRL-2271)
GM12891	4.99	B cells	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed
GM12892	4.92	B cells	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed
H1hESC	4.84	ESC	embryonic stem cells
H9ES	4.62	ESC	embryonic stem cell (hESC) H9
8988T	4.61	cancer cells	pancreas adenocarcinoma (PA-TU-8988T)
CLL	4.58	cancer cells	chronic lymphocytic leukemia cell, T-cell lymphocyte
Urothelia	4.52	ureter cells	primary ureter cell culture of urothelial cells derived from a 12 year-old girl and immortalized by transfection with a temperature-sensitive SV-40 large T antigen gene
GM19239	4.51	B cells	lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed
GM19238	4.47	B cells	lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed
HepG2	4.40	cancer cells	hepatocellular carcinoma
Jurkat	4.24	cancer cells	T lymphoblastoid derived from an acute T cell leukemia. (PMID: 68013)
WERIRb1	4.20	cancer cells	retinoblastoma (PMID: 844036)

Table S6B The enrichment of AP-1 motifs in human DHSs (TOP 20)

Cells	Enrichment	Types	Notes
SAEC	15.53	epithelial cells	small airway epithelial cells
PrEC	14.30	epithelial cells	prostate epithelial cell line
HEEpiC	14.25	epithelial cells	esophageal epithelial cells
HBMEC	14.12	endothelial cells	brain microvascular endothelial cells
HMF	13.99	fibroblasts	mammary fibroblasts
HRE	13.92	epithelial cells	renal epithelial cells
RWPE1	13.86	epithelial cells	prostate epithelial
HNPCEpiC	13.59	epithelial cells	non-pigment ciliary epithelial cells
NHLF	13.57	fibroblasts	lung fibroblasts
WI38Tamoxifen	13.42	fibroblasts	embryonic lung fibroblast cells, hTERT immortalized, includes Raf1 construct
NHA	13.35	astrocytes	astrocytes (also called Astrocy) [from brain]
HCFaa	13.16	fibroblasts	cardiac fibroblasts- adult atrial
HIPEpiC	13.07	epithelial cells	iris pigment epithelial cells
HRCEpiC	12.75	epithelial cells	renal cortical epithelial cells
WI38	12.56	fibroblasts	embryonic lung fibroblast cells, hTERT immortalized, includes Raf1 construct
HCT116	12.47	cancer cells	colorectal carcinoma (PMID: 7214343)
HPDE6E6E7	12.13	pancreatic duct cells	pancreatic duct cells immortalized with E6E7 gene of HPV
RPTEC	11.67	epithelial cells	renal proximal tubule epithelial cells
HMVECLBl	11.41	endothelial cells	blood microvascular endothelial cells, lung-derived
BJ	11.37	fibroblasts	skin fibroblast

Table S7. Statistics of the MRCA sequence to variations of 11mers with TGACTCA in CNEs with DHSs in human

The Most Recent Common Ancestor (MRCA)	Referred Species	Genome assembly	in CNEs and DHSs				in CNEs not in DHSs			
			#	#≤1SNPs	#TGACTCA	%TGACTCA	#	#≤1SNPs	#TGACTCA	%TGACTCA
01.Osteichthyes	Stickleback	gasAcu1	313	95	39	12.5%	187	31	13	7.0%
02.Sarcopterygii	Coelacanth	latCha1	798	435	267	33.5%	186	55	20	10.8%
03.Tetrapods	Xenopus	xenTro3	1,090	415	252	23.1%	360	67	23	6.4%
04.Amniotes	Chicken	galGal3	2,384	1,374	896	37.6%	645	280	116	18.0%
04.Amniotes	Lizard	anoCar2	1,907	993	637	33.4%	557	214	89	16.0%
05.Theria	Opossum	monDom5	8,176	5,748	4,015	49.1%	2,128	1,138	618	29.0%
06.Eutheria	Elephant	loxAfr3	13,130	11,741	9,083	69.2%	3,713	2,966	1,908	51.4%
07.Boreoeutheria	Dog	canFam3	13,339	12,125	9,775	73.3%	3,764	3,090	2,032	54.0%
08.Euarchontoglires	Mouse	mm9	11,657	9,566	7,109	61.0%	3,034	2,161	1,247	41.1%
ALL	Human	hg19	14,691				4,626			

The Most Recent Common Ancestor (MRCA)	Referred Species	Genome assembly	not in CNEs				All			
			#	#≤1SNPs	#TGACTCA	%TGACTCA	#	#≤1SNPs	#TGACTCA	%TGACTCA
01.Osteichthyes	Stickleback	gasAcu1	4,310	830	192	4.5%	4,810	956	244	5.1%
02.Sarcopterygii	Coelacanth	latCha1	4,694	1,224	335	7.1%	5,678	1,714	622	11.0%
03.Tetrapods	Xenopus	xenTro3	6,645	1,277	291	4.4%	8,095	1,759	566	7.0%
04.Amniotes	Chicken	galGal3	8,268	1,981	697	8.4%	11,297	3,635	1,709	15.1%
04.Amniotes	Lizard	anoCar2	7,946	1,811	556	7.0%	10,410	3,018	1,282	12.3%
05.Theria	Opossum	monDom5	42,560	10,053	4,287	10.1%	52,864	16,939	8,920	16.9%
06.Eutheria	Elephant	loxAfr3	191,040	95,610	41,821	21.9%	207,883	**110,317**	**52,812**	25.4%
07.Boreoeutheria	Dog	canFam3	207,455	102,966	45,674	22.0%	224,558	118,181	57,481	25.6%
08.Euarchontoglires	Mouse	mm9	125,928	45,561	19,025	15.1%	140,619	57,288	27,381	19.5%
ALL	Human	hg19	440,911				460,228			

Table S8A1. Statistics of the MRCA sequence (01. Ost. Stackleback) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
95	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	26	32	4	2	88	0	1	1	79	26	19
	C	24	23	7	1	1	95	2	87	6	15	28
	G	23	21	3	90	5	0	3	3	6	21	15
	T	22	19	81	2	1	0	89	4	4	33	33
%	A	27.4%	33.7%	4.2%	2.1%	92.6%	0.0%	1.1%	1.1%	83.2%	27.4%	20.0%
	C	25.3%	24.2%	7.4%	1.1%	1.1%	100.0%	2.1%	91.6%	6.3%	15.8%	29.5%
	G	24.2%	22.1%	3.2%	94.7%	5.3%	0.0%	3.2%	3.2%	6.3%	22.1%	15.8%
	T	23.2%	20.0%	85.3%	2.1%	1.1%	0.0%	93.7%	4.2%	4.2%	34.7%	34.7%
		Conserved										
	#	0	0	81	90	88	95	89	87	79	0	0
	%	0.0%	0.0%	85.3%	94.7%	92.6%	100.0%	93.7%	91.6%	83.2%	0.0%	0.0%
		Variations										
%	A	27.4%	33.7%	28.6%	**40.0%**	0.0%	0.0%	16.7%	12.5%	0.0%	27.4%	20.0%
	C	25.3%	24.2%	**50.0%**	20.0%	14.3%	0.0%	33.3%	0.0%	37.5%	15.8%	29.5%
	G	24.2%	22.1%	21.4%	0.0%	71.4%	0.0%	50.0%	37.5%	**37.5%**	22.1%	15.8%
	T	23.2%	20.0%	0.0%	40.0%	14.3%	0.0%	0.0%	**50.0%**	25.0%	34.7%	34.7%

Table S8A2. Statistics of the MRCA sequence (01. Ost. Stackleback) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
31	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	2	6	1	0	29	0	1	1	28	5	3
	C	13	11	0	0	1	31	4	29	3	4	4
	G	13	4	1	31	1	0	4	1	0	19	14
	T	3	10	29	0	0	0	22	0	0	3	10
%	A	6.5%	19.4%	3.2%	0.0%	93.5%	0.0%	3.2%	3.2%	90.3%	16.1%	9.7%
	C	41.9%	35.5%	0.0%	0.0%	3.2%	100.0%	12.9%	93.5%	9.7%	12.9%	12.9%
	G	41.9%	12.9%	3.2%	100.0%	3.2%	0.0%	12.9%	3.2%	0.0%	61.3%	45.2%
	T	9.7%	32.3%	93.5%	0.0%	0.0%	0.0%	71.0%	0.0%	0.0%	9.7%	32.3%
		Conserved										
	#	0	0	29	31	29	31	22	29	28	0	0
	%	0.0%	0.0%	93.5%	100.0%	93.5%	100.0%	71.0%	93.5%	90.3%	0.0%	0.0%
		Variations										
%	A	6.5%	19.4%	50.0%	**0.0%**	0.0%	0.0%	11.1%	50.0%	0.0%	16.1%	9.7%
	C	41.9%	35.5%	**0.0%**	0.0%	50.0%	0.0%	44.4%	0.0%	100.0%	12.9%	12.9%
	G	41.9%	12.9%	50.0%	0.0%	50.0%	0.0%	44.4%	50.0%	**0.0%**	61.3%	45.2%
	T	9.7%	32.3%	0.0%	0.0%	0.0%	0.0%	0.0%	**0.0%**	0.0%	9.7%	32.3%

Table S8A3. Statistics of the MRCA sequence (01. Ost. Stackleback) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
830	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	161	195	31	13	729	12	35	12	720	115	217
	C	209	278	34	35	7	802	122	782	48	225	233
	G	263	218	30	767	78	0	36	24	37	319	205
	T	197	139	735	15	16	16	637	12	25	171	175
%	A	19.4%	23.5%	3.7%	1.6%	87.8%	1.4%	4.2%	1.4%	86.7%	13.9%	26.1%
	C	25.2%	33.5%	4.1%	4.2%	0.8%	96.6%	14.7%	94.2%	5.8%	27.1%	28.1%
	G	31.7%	26.3%	3.6%	92.4%	9.4%	0.0%	4.3%	2.9%	4.5%	38.4%	24.7%
	T	23.7%	16.7%	88.6%	1.8%	1.9%	1.9%	76.7%	1.4%	3.0%	20.6%	21.1%
		Conserved										
	#	0	0	735	767	729	802	637	782	720	0	0
	%	0.0%	0.0%	88.6%	92.4%	87.8%	96.6%	76.7%	94.2%	86.7%	0.0%	0.0%
		Variations										
%	A	19.4%	23.5%	32.6%	**20.6%**	0.0%	42.9%	18.1%	25.0%	0.0%	13.9%	26.1%
	C	25.2%	33.5%	**35.8%**	55.6%	6.9%	0.0%	63.2%	0.0%	43.6%	27.1%	28.1%
	G	31.7%	26.3%	31.6%	0.0%	77.2%	0.0%	18.7%	50.0%	**33.6%**	38.4%	24.7%
	T	23.7%	16.7%	0.0%	23.8%	15.8%	57.1%	0.0%	**25.0%**	22.7%	20.6%	21.1%

Table S8B1. Statistics of the MRCA sequence (02. Sar. Coelacanth) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
435	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	138	151	5	5	407	20	4	13	407	97	118
	C	69	91	9	2	3	395	11	407	7	77	81
	G	95	110	1	423	21	0	2	6	7	104	61
	T	133	83	420	5	4	20	418	9	14	157	175
%	A	31.7%	34.7%	1.1%	1.1%	93.6%	4.6%	0.9%	3.0%	93.6%	22.3%	27.1%
	C	15.9%	20.9%	2.1%	0.5%	0.7%	90.8%	2.5%	93.6%	1.6%	17.7%	18.6%
	G	21.8%	25.3%	0.2%	97.2%	4.8%	0.0%	0.5%	1.4%	1.6%	23.9%	14.0%
	T	30.6%	19.1%	96.6%	1.1%	0.9%	4.6%	96.1%	2.1%	3.2%	36.1%	40.2%
							Conserved					
	#	0	0	420	423	407	395	418	407	407	0	0
	%	0.0%	0.0%	96.6%	97.2%	93.6%	90.8%	96.1%	93.6%	93.6%	0.0%	0.0%
							Variations					
%	A	31.7%	34.7%	33.3%	**41.7%**	0.0%	50.0%	23.5%	46.4%	0.0%	22.3%	27.1%
	C	15.9%	20.9%	**60.0%**	16.7%	10.7%	0.0%	64.7%	0.0%	25.0%	17.7%	18.6%
	G	21.8%	25.3%	6.7%	0.0%	75.0%	0.0%	11.8%	21.4%	**25.0%**	23.9%	14.0%
	T	30.6%	19.1%	0.0%	41.7%	14.3%	50.0%	0.0%	**32.1%**	50.0%	36.1%	40.2%

Table S8B2. Statistics of the MRCA sequence (02. Sar. Coelacanth) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
55	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	22	19	0	5	50	2	0	5	51	14	12
	C	9	9	2	0	2	53	3	45	0	17	9
	G	4	11	0	50	2	0	4	1	2	10	9
	T	20	16	53	0	1	0	48	4	2	14	25
%	A	40.0%	34.5%	0.0%	9.1%	90.9%	3.6%	0.0%	9.1%	92.7%	25.5%	21.8%
	C	16.4%	16.4%	3.6%	0.0%	3.6%	96.4%	5.5%	81.8%	0.0%	30.9%	16.4%
	G	7.3%	20.0%	0.0%	90.9%	3.6%	0.0%	7.3%	1.8%	3.6%	18.2%	16.4%
	T	36.4%	29.1%	96.4%	0.0%	1.8%	0.0%	87.3%	7.3%	3.6%	25.5%	45.5%
							Conserved					
	#	0	0	53	50	50	53	48	45	51	0	0
	%	0.0%	0.0%	96.4%	90.9%	90.9%	96.4%	87.3%	81.8%	92.7%	0.0%	0.0%
							Variations					
%	A	40.0%	34.5%	0.0%	**100.0%**	0.0%	100.0%	0.0%	50.0%	0.0%	25.5%	21.8%
	C	16.4%	16.4%	**100.0%**	0.0%	40.0%	0.0%	42.9%	0.0%	0.0%	30.9%	16.4%
	G	7.3%	20.0%	0.0%	0.0%	40.0%	0.0%	57.1%	10.0%	**50.0%**	18.2%	16.4%
	T	36.4%	29.1%	0.0%	0.0%	20.0%	0.0%	0.0%	**40.0%**	50.0%	25.5%	45.5%

Table S8B3. Statistics of the MRCA sequence (02. Sar. Coelacanth) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1224	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	304	356	56	23	1100	59	72	28	1064	251	366
	C	265	384	23	23	17	1099	108	1133	58	241	293
	G	343	255	31	1152	84	0	27	33	35	451	281
	T	312	229	1114	26	23	66	1017	30	67	281	284
%	A	24.8%	29.1%	4.6%	1.9%	89.9%	4.8%	5.9%	2.3%	86.9%	20.5%	29.9%
	C	21.7%	31.4%	1.9%	1.9%	1.4%	89.8%	8.8%	92.6%	4.7%	19.7%	23.9%
	G	28.0%	20.8%	2.5%	94.1%	6.9%	0.0%	2.2%	2.7%	2.9%	36.8%	23.0%
	T	25.5%	18.7%	91.0%	2.1%	1.9%	5.4%	83.1%	2.5%	5.5%	23.0%	23.2%
							Conserved					
	#	0	0	1114	1152	1100	1099	1017	1133	1064	0	0
	%	0.0%	0.0%	91.0%	94.1%	89.9%	89.8%	83.1%	92.6%	86.9%	0.0%	0.0%
							Variations					
%	A	24.8%	29.1%	50.9%	**31.9%**	0.0%	47.2%	34.8%	30.8%	0.0%	20.5%	29.9%
	C	21.7%	31.4%	**20.9%**	31.9%	13.7%	0.0%	52.2%	0.0%	36.3%	19.7%	23.9%
	G	28.0%	20.8%	28.2%	0.0%	67.7%	0.0%	13.0%	36.3%	**21.9%**	36.8%	23.0%
	T	25.5%	18.7%	0.0%	36.1%	18.5%	52.8%	0.0%	**33.0%**	41.9%	23.0%	23.2%

Table S8C1. Statistics of the MRCA sequence (03. Tet. Xenopus) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
415	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	136	152	8	13	393	10	11	6	380	80	120
	C	81	122	8	2	4	390	8	392	16	93	67
	G	91	72	1	393	16	0	0	2	8	81	65
	T	107	69	398	7	2	15	396	15	11	161	163
%	A	32.8%	36.6%	1.9%	3.1%	94.7%	2.4%	2.7%	1.4%	91.6%	19.3%	28.9%
	C	19.5%	29.4%	1.9%	0.5%	1.0%	94.0%	1.9%	94.5%	3.9%	22.4%	16.1%
	G	21.9%	17.3%	0.2%	94.7%	3.9%	0.0%	0.0%	0.5%	1.9%	19.5%	15.7%
	T	25.8%	16.6%	95.9%	1.7%	0.5%	3.6%	95.4%	3.6%	2.7%	38.8%	39.3%
		Conserved										
	#	0	0	398	393	393	390	396	392	380	0	0
	%	0.0%	0.0%	95.9%	94.7%	94.7%	94.0%	95.4%	94.5%	91.6%	0.0%	0.0%
		Variations										
%	A	32.8%	36.6%	47.1%	**59.1%**	0.0%	40.0%	57.9%	26.1%	0.0%	19.3%	28.9%
	C	19.5%	29.4%	**47.1%**	9.1%	18.2%	0.0%	42.1%	0.0%	45.7%	22.4%	16.1%
	G	21.9%	17.3%	5.9%	0.0%	72.7%	0.0%	0.0%	8.7%	**22.9%**	19.5%	15.7%
	T	25.8%	16.6%	0.0%	31.8%	9.1%	60.0%	0.0%	**65.2%**	31.4%	38.8%	39.3%

Table S8C2. Statistics of the MRCA sequence (03. Tet. Xenopus) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
67	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	29	12	0	1	55	3	2	1	62	10	23
	C	13	20	2	2	2	63	1	60	2	16	13
	G	8	14	6	62	8	0	0	1	2	22	11
	T	17	21	59	2	2	1	64	5	1	19	20
%	A	43.3%	17.9%	0.0%	1.5%	82.1%	4.5%	3.0%	1.5%	92.5%	14.9%	34.3%
	C	19.4%	29.9%	3.0%	3.0%	3.0%	94.0%	1.5%	89.6%	3.0%	23.9%	19.4%
	G	11.9%	20.9%	9.0%	92.5%	11.9%	0.0%	0.0%	1.5%	3.0%	32.8%	16.4%
	T	25.4%	31.3%	88.1%	3.0%	3.0%	1.5%	95.5%	7.5%	1.5%	28.4%	29.9%
		Conserved										
	#	0	0	59	62	55	63	64	60	62	0	0
	%	0.0%	0.0%	88.1%	92.5%	82.1%	94.0%	95.5%	89.6%	92.5%	0.0%	0.0%
		Variations										
%	A	43.3%	17.9%	0.0%	**20.0%**	0.0%	75.0%	66.7%	14.3%	0.0%	14.9%	34.3%
	C	19.4%	29.9%	**25.0%**	40.0%	16.7%	0.0%	33.3%	0.0%	40.0%	23.9%	19.4%
	G	11.9%	20.9%	75.0%	0.0%	66.7%	0.0%	0.0%	14.3%	**40.0%**	32.8%	16.4%
	T	25.4%	31.3%	0.0%	40.0%	16.7%	25.0%	0.0%	**71.4%**	20.0%	28.4%	29.9%

Table S8C3. Statistics of the MRCA sequence (03. Tet. Xenopus) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1277	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	368	413	62	45	1099	59	87	19	1115	239	383
	C	254	383	13	34	9	1161	81	1166	56	263	294
	G	295	262	38	1150	145	0	11	41	36	488	237
	T	360	219	1164	48	24	57	1098	51	70	287	363
%	A	28.8%	32.3%	4.9%	3.5%	86.1%	4.6%	6.8%	1.5%	87.3%	18.7%	30.0%
	C	19.9%	30.0%	1.0%	2.7%	0.7%	90.9%	6.3%	91.3%	4.4%	20.6%	23.0%
	G	23.1%	20.5%	3.0%	90.1%	11.4%	0.0%	0.9%	3.2%	2.8%	38.2%	18.6%
	T	28.2%	17.1%	91.2%	3.8%	1.9%	4.5%	86.0%	4.0%	5.5%	22.5%	28.4%
		Conserved										
	#	0	0	1164	1150	1099	1161	1098	1166	1115	0	0
	%	0.0%	0.0%	91.2%	90.1%	86.1%	90.9%	86.0%	91.3%	87.3%	0.0%	0.0%
		Variations										
%	A	28.8%	32.3%	54.9%	**35.4%**	0.0%	50.9%	48.6%	17.1%	0.0%	18.7%	30.0%
	C	19.9%	30.0%	**11.5%**	26.8%	5.1%	0.0%	45.3%	0.0%	34.6%	20.6%	23.0%
	G	23.1%	20.5%	33.6%	0.0%	81.5%	0.0%	6.1%	36.9%	**22.2%**	38.2%	18.6%
	T	28.2%	17.1%	0.0%	37.8%	13.5%	49.1%	0.0%	**45.9%**	43.2%	22.5%	28.4%

Table S8D1. Statistics of the MRCA sequence (04. Amn. Chicken) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1374	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	472	509	3	21	1312	66	26	28	1302	264	421
	C	239	307	18	6	14	1242	15	1301	15	284	251
	G	290	322	10	1310	39	0	3	20	26	322	207
	T	373	236	1343	37	9	66	1330	25	31	504	495
%	A	34.4%	37.0%	0.2%	1.5%	95.5%	4.8%	1.9%	2.0%	94.8%	19.2%	30.6%
	C	17.4%	22.3%	1.3%	0.4%	1.0%	90.4%	1.1%	94.7%	1.1%	20.7%	18.3%
	G	21.1%	23.4%	0.7%	95.3%	2.8%	0.0%	0.2%	1.5%	1.9%	23.4%	15.1%
	T	27.1%	17.2%	97.7%	2.7%	0.7%	4.8%	96.8%	1.8%	2.3%	36.7%	36.0%
		Conserved										
	#	0	0	1343	1310	1312	1242	1330	1301	1302	0	0
	%	0.0%	0.0%	97.7%	95.3%	95.5%	90.4%	96.8%	94.7%	94.8%	0.0%	0.0%
		Variations										
%	A	34.4%	37.0%	9.7%	**32.8%**	0.0%	50.0%	59.1%	38.4%	0.0%	19.2%	30.6%
	C	17.4%	22.3%	**58.1%**	9.4%	22.6%	0.0%	34.1%	0.0%	20.8%	20.7%	18.3%
	G	21.1%	23.4%	32.3%	0.0%	62.9%	0.0%	6.8%	27.4%	**36.1%**	23.4%	15.1%
	T	27.1%	17.2%	0.0%	57.8%	14.5%	50.0%	0.0%	**34.2%**	43.1%	36.7%	36.0%

Table S8D2. Statistics of the MRCA sequence (04. Amn. Chicken) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
280	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	112	81	2	15	257	18	12	11	263	77	89
	C	54	60	8	3	4	250	7	245	6	45	35
	G	44	53	6	257	11	0	1	5	8	71	57
	T	70	86	264	5	8	12	260	19	3	87	99
%	A	40.0%	28.9%	0.7%	5.4%	91.8%	6.4%	4.3%	3.9%	93.9%	27.5%	31.8%
	C	19.3%	21.4%	2.9%	1.1%	1.4%	89.3%	2.5%	87.5%	2.1%	16.1%	12.5%
	G	15.7%	18.9%	2.1%	91.8%	3.9%	0.0%	0.4%	1.8%	2.9%	25.4%	20.4%
	T	25.0%	30.7%	94.3%	1.8%	2.9%	4.3%	92.9%	6.8%	1.1%	31.1%	35.4%
		Conserved										
	#	0	0	264	257	257	250	260	245	263	0	0
	%	0.0%	0.0%	94.3%	91.8%	91.8%	89.3%	92.9%	87.5%	93.9%	0.0%	0.0%
		Variations										
%	A	40.0%	28.9%	12.5%	**65.2%**	0.0%	60.0%	60.0%	31.4%	0.0%	27.5%	31.8%
	C	19.3%	21.4%	**50.0%**	13.0%	17.4%	0.0%	35.0%	0.0%	35.3%	16.1%	12.5%
	G	15.7%	18.9%	37.5%	0.0%	47.8%	0.0%	5.0%	14.3%	**47.1%**	25.4%	20.4%
	T	25.0%	30.7%	0.0%	21.7%	34.8%	40.0%	0.0%	**54.3%**	17.6%	31.1%	35.4%

Table S8D3. Statistics of the MRCA sequence (04. Amn. Chicken) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1981	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	483	549	74	65	1772	109	74	29	1800	404	606
	C	421	601	58	45	38	1757	109	1853	56	408	466
	G	506	431	39	1819	130	0	26	40	47	704	421
	T	571	400	1810	52	41	115	1772	59	78	465	488
%	A	24.4%	27.7%	3.7%	3.3%	89.4%	5.5%	3.7%	1.5%	90.9%	20.4%	30.6%
	C	21.3%	30.3%	2.9%	2.3%	1.9%	88.7%	5.5%	93.5%	2.8%	20.6%	23.5%
	G	25.5%	21.8%	2.0%	91.8%	6.6%	0.0%	1.3%	2.0%	2.4%	35.5%	21.3%
	T	28.8%	20.2%	91.4%	2.6%	2.1%	5.8%	89.4%	3.0%	3.9%	23.5%	24.6%
		Conserved										
	#	0	0	1810	1819	1772	1757	1772	1853	1800	0	0
	%	0.0%	0.0%	91.4%	91.8%	89.4%	88.7%	89.4%	93.5%	90.9%	0.0%	0.0%
		Variations										
%	A	24.4%	27.7%	43.3%	**40.1%**	0.0%	48.7%	35.4%	22.7%	0.0%	20.4%	30.6%
	C	21.3%	30.3%	**33.9%**	27.8%	18.2%	0.0%	52.2%	0.0%	30.9%	20.6%	23.5%
	G	25.5%	21.8%	22.8%	0.0%	62.2%	0.0%	12.4%	31.3%	**26.0%**	35.5%	21.3%
	T	28.8%	20.2%	0.0%	32.1%	19.6%	51.3%	0.0%	**46.1%**	43.1%	23.5%	24.6%

Table S8E1. Statistics of the MRCA sequence (04. Amn. Lizard) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
993	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	335	367	8	17	937	49	12	18	939	190	291
	C	166	213	7	5	12	895	14	940	20	176	193
	G	225	219	6	948	31	0	3	11	16	249	161
	T	267	194	972	23	13	49	964	24	18	378	348
%	A	33.7%	37.0%	0.8%	1.7%	94.4%	4.9%	1.2%	1.8%	94.6%	19.1%	29.3%
	C	16.7%	21.5%	0.7%	0.5%	1.2%	90.1%	1.4%	94.7%	2.0%	17.7%	19.4%
	G	22.7%	22.1%	0.6%	95.5%	3.1%	0.0%	0.3%	1.1%	1.6%	25.1%	16.2%
	T	26.9%	19.5%	97.9%	2.3%	1.3%	4.9%	97.1%	2.4%	1.8%	38.1%	35.0%
		Conserved										
	#	0	0	972	948	937	895	964	940	939	0	0
	%	0.0%	0.0%	97.9%	95.5%	94.4%	90.1%	97.1%	94.7%	94.6%	0.0%	0.0%
		Variations										
%	A	33.7%	37.0%	38.1%	**37.8%**	0.0%	50.0%	41.4%	34.0%	0.0%	19.1%	29.3%
	C	16.7%	21.5%	**33.3%**	11.1%	21.4%	0.0%	48.3%	0.0%	37.0%	17.7%	19.4%
	G	22.7%	22.1%	28.6%	0.0%	55.4%	0.0%	10.3%	20.8%	**29.6%**	25.1%	16.2%
	T	26.9%	19.5%	0.0%	51.1%	23.2%	50.0%	0.0%	**45.3%**	33.3%	38.1%	35.0%

Table S8E2. Statistics of the MRCA sequence (04. Amn. Lizard) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
214	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	68	62	1	4	188	15	9	8	199	59	55
	C	43	51	2	4	4	196	8	188	5	26	31
	G	36	36	4	198	14	0	0	8	6	51	48
	T	67	65	207	8	8	3	197	10	4	78	80
%	A	31.8%	29.0%	0.5%	1.9%	87.9%	7.0%	4.2%	3.7%	93.0%	27.6%	25.7%
	C	20.1%	23.8%	0.9%	1.9%	1.9%	91.6%	3.7%	87.9%	2.3%	12.1%	14.5%
	G	16.8%	16.8%	1.9%	92.5%	6.5%	0.0%	0.0%	3.7%	2.8%	23.8%	22.4%
	T	31.3%	30.4%	96.7%	3.7%	3.7%	1.4%	92.1%	4.7%	1.9%	36.4%	37.4%
		Conserved										
	#	0	0	207	198	188	196	197	188	199	0	0
	%	0.0%	0.0%	96.7%	92.5%	87.9%	91.6%	92.1%	87.9%	93.0%	0.0%	0.0%
		Variations										
%	A	31.8%	29.0%	14.3%	**25.0%**	0.0%	83.3%	52.9%	30.8%	0.0%	27.6%	25.7%
	C	20.1%	23.8%	**28.6%**	25.0%	15.4%	0.0%	47.1%	0.0%	33.3%	12.1%	14.5%
	G	16.8%	16.8%	57.1%	0.0%	53.8%	0.0%	0.0%	30.8%	**40.0%**	23.8%	22.4%
	T	31.3%	30.4%	0.0%	50.0%	30.8%	16.7%	0.0%	**38.5%**	26.7%	36.4%	37.4%

Table S8E3. Statistics of the MRCA sequence (04. Amn. Lizard) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1811	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	458	524	67	45	1598	116	82	22	1625	409	539
	C	401	491	34	53	25	1584	118	1701	57	353	435
	G	435	363	39	1663	144	0	31	42	38	653	383
	T	517	433	1671	50	44	111	1580	46	91	396	454
%	A	25.3%	28.9%	3.7%	2.5%	88.2%	6.4%	4.5%	1.2%	89.7%	22.6%	29.8%
	C	22.1%	27.1%	1.9%	2.9%	1.4%	87.5%	6.5%	93.9%	3.1%	19.5%	24.0%
	G	24.0%	20.0%	2.2%	91.8%	8.0%	0.0%	1.7%	2.3%	2.1%	36.1%	21.1%
	T	28.5%	23.9%	92.3%	2.8%	2.4%	6.1%	87.2%	2.5%	5.0%	21.9%	25.1%
		Conserved										
	#	0	0	1671	1663	1598	1584	1580	1701	1625	0	0
	%	0.0%	0.0%	92.3%	91.8%	88.2%	87.5%	87.2%	93.9%	89.7%	0.0%	0.0%
		Variations										
%	A	25.3%	28.9%	47.9%	**30.4%**	0.0%	51.1%	35.5%	20.0%	0.0%	22.6%	29.8%
	C	22.1%	27.1%	**24.3%**	35.8%	11.7%	0.0%	51.1%	0.0%	30.6%	19.5%	24.0%
	G	24.0%	20.0%	27.9%	0.0%	67.6%	0.0%	13.4%	38.2%	**20.4%**	36.1%	21.1%
	T	28.5%	23.9%	0.0%	33.8%	20.7%	48.9%	0.0%	**41.8%**	48.9%	21.9%	25.1%

Table S8F1. Statistics of the MRCA sequence (05. The. Opossum) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
5748	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	1825	2015	48	80	5533	281	47	109	5536	1009	1629
	C	948	1453	44	35	52	5192	78	5458	65	1219	1281
	G	1321	1354	42	5561	133	0	15	52	63	1479	972
	T	1654	926	5614	72	30	275	5608	129	84	2041	1866
%	A	31.8%	35.1%	0.8%	1.4%	96.3%	4.9%	0.8%	1.9%	96.3%	17.6%	28.3%
	C	16.5%	25.3%	0.8%	0.6%	0.9%	90.3%	1.4%	95.0%	1.1%	21.2%	22.3%
	G	23.0%	23.6%	0.7%	96.7%	2.3%	0.0%	0.3%	0.9%	1.1%	25.7%	16.9%
	T	28.8%	16.1%	97.7%	1.3%	0.5%	4.8%	97.6%	2.2%	1.5%	35.5%	32.5%
		Conserved										
	#	0	0	5614	5561	5533	5192	5608	5458	5536	0	0
	%	0.0%	0.0%	97.7%	96.7%	96.3%	90.3%	97.6%	95.0%	96.3%	0.0%	0.0%
		Variations										
%	A	31.8%	35.1%	35.8%	**42.8%**	0.0%	50.5%	33.6%	37.6%	0.0%	17.6%	28.3%
	C	16.5%	25.3%	**32.8%**	18.7%	24.2%	0.0%	55.7%	0.0%	30.7%	21.2%	22.3%
	G	23.0%	23.6%	31.3%	0.0%	61.9%	0.0%	10.7%	17.9%	**29.7%**	25.7%	16.9%
	T	28.8%	16.1%	0.0%	38.5%	14.0%	49.5%	0.0%	**44.5%**	39.6%	35.5%	32.5%

Table S8F2. Statistics of the MRCA sequence (05. The. Opossum) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
1138	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	385	378	11	25	1054	64	7	31	1066	306	309
	C	190	252	13	4	23	1005	34	1037	27	162	206
	G	200	183	9	1088	48	0	6	21	16	301	206
	T	363	325	1105	21	13	69	1091	49	29	369	417
%	A	33.8%	33.2%	1.0%	2.2%	92.6%	5.6%	0.6%	2.7%	93.7%	26.9%	27.2%
	C	16.7%	22.1%	1.1%	0.4%	2.0%	88.3%	3.0%	91.1%	2.4%	14.2%	18.1%
	G	17.6%	16.1%	0.8%	95.6%	4.2%	0.0%	0.5%	1.8%	1.4%	26.4%	18.1%
	T	31.9%	28.6%	97.1%	1.8%	1.1%	6.1%	95.9%	4.3%	2.5%	32.4%	36.6%
		Conserved										
	#	0	0	1105	1088	1054	1005	1091	1037	1066	0	0
	%	0.0%	0.0%	97.1%	95.6%	92.6%	88.3%	95.9%	91.1%	93.7%	0.0%	0.0%
		Variations										
%	A	33.8%	33.2%	33.3%	**50.0%**	0.0%	48.1%	14.9%	30.7%	0.0%	26.9%	27.2%
	C	16.7%	22.1%	**39.4%**	8.0%	27.4%	0.0%	72.3%	0.0%	37.5%	14.2%	18.1%
	G	17.6%	16.1%	27.3%	0.0%	57.1%	0.0%	12.8%	20.8%	**22.2%**	26.4%	18.1%
	T	31.9%	28.6%	0.0%	42.0%	15.5%	51.9%	0.0%	**48.5%**	40.3%	32.4%	36.6%

Table S8F3. Statistics of the MRCA sequence (05. The. Opossum) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
10053	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	2864	2917	226	333	9185	682	195	272	9308	2064	2912
	C	1808	2755	108	171	204	8684	420	9172	290	2106	2374
	G	2419	2270	168	9297	476	0	34	201	190	3081	1870
	T	2962	2111	9551	252	188	687	9404	408	265	2802	2897
%	A	28.5%	29.0%	2.2%	3.3%	91.4%	6.8%	1.9%	2.7%	92.6%	20.5%	29.0%
	C	18.0%	27.4%	1.1%	1.7%	2.0%	86.4%	4.2%	91.2%	2.9%	20.9%	23.6%
	G	24.1%	22.6%	1.7%	92.5%	4.7%	0.0%	0.3%	2.0%	1.9%	30.6%	18.6%
	T	29.5%	21.0%	95.0%	2.5%	1.9%	6.8%	93.5%	4.1%	2.6%	27.9%	28.8%
		Conserved										
	#	0	0	9551	9297	9185	8684	9404	9172	9308	0	0
	%	0.0%	0.0%	95.0%	92.5%	91.4%	86.4%	93.5%	91.2%	92.6%	0.0%	0.0%
		Variations										
%	A	28.5%	29.0%	45.0%	**44.0%**	0.0%	49.8%	30.0%	30.9%	0.0%	20.5%	29.0%
	C	18.0%	27.4%	**21.5%**	22.6%	23.5%	0.0%	64.7%	0.0%	38.9%	20.9%	23.6%
	G	24.1%	22.6%	33.5%	0.0%	54.8%	0.0%	5.2%	22.8%	**25.5%**	30.6%	18.6%
	T	29.5%	21.0%	0.0%	33.3%	21.7%	50.2%	0.0%	**46.3%**	35.6%	27.9%	28.8%

Table S8G1. Statistics of the MRCA sequence (06. Eut. Elephant) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
11,741	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	3582	3797	71	128	11350	307	51	102	11326	2127	3431
	C	2028	3084	130	50	80	11117	152	11326	119	2593	2607
	G	2672	2921	78	11456	266	0	47	89	192	3380	2181
	T	3459	1939	11462	107	45	317	11491	224	104	3641	3522
%	A	30.5%	32.3%	0.6%	1.1%	96.7%	2.6%	0.4%	0.9%	96.5%	18.1%	29.2%
	C	17.3%	26.3%	1.1%	0.4%	0.7%	94.7%	1.3%	96.5%	1.0%	22.1%	22.2%
	G	22.8%	24.9%	0.7%	97.6%	2.3%	0.0%	0.4%	0.8%	1.6%	28.8%	18.6%
	T	29.5%	16.5%	97.6%	0.9%	0.4%	2.7%	97.9%	1.9%	0.9%	31.0%	30.0%
		Conserved										
	#	0	0	11462	11456	11350	11117	11491	11326	11326	0	0
	%	0.0%	0.0%	97.6%	97.6%	96.7%	94.7%	97.9%	96.5%	96.5%	0.0%	0.0%
		Variations										
%	A	30.5%	32.3%	25.4%	**44.9%**	0.0%	49.2%	20.4%	24.6%	0.0%	18.1%	29.2%
	C	17.3%	26.3%	**46.6%**	17.5%	20.5%	0.0%	60.8%	0.0%	28.7%	22.1%	22.2%
	G	22.8%	24.9%	28.0%	0.0%	68.0%	0.0%	18.8%	21.4%	**46.3%**	28.8%	18.6%
	T	29.5%	16.5%	0.0%	37.5%	11.5%	50.8%	0.0%	**54.0%**	25.1%	31.0%	30.0%

Table S8G2. Statistics of the MRCA sequence (06. Eut. Elephant) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
2,966	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	999	870	35	67	2796	118	38	35	2849	840	962
	C	512	705	42	32	33	2724	68	2794	35	458	487
	G	478	490	32	2835	113	0	11	46	39	789	513
	T	977	901	2857	32	24	124	2849	91	43	879	1004
%	A	33.7%	29.3%	1.2%	2.3%	94.3%	4.0%	1.3%	1.2%	96.1%	28.3%	32.4%
	C	17.3%	23.8%	1.4%	1.1%	1.1%	91.8%	2.3%	94.2%	1.2%	15.4%	16.4%
	G	16.1%	16.5%	1.1%	95.6%	3.8%	0.0%	0.4%	1.6%	1.3%	26.6%	17.3%
	T	32.9%	30.4%	96.3%	1.1%	0.8%	4.2%	96.1%	3.1%	1.4%	29.6%	33.9%
		Conserved										
	#	0	0	2857	2835	2796	2724	2849	2794	2849	0	0
	%	0.0%	0.0%	96.3%	95.6%	94.3%	91.8%	96.1%	94.2%	96.1%	0.0%	0.0%
		Variations										
%	A	33.7%	29.3%	32.1%	**51.1%**	0.0%	48.8%	32.5%	20.3%	0.0%	28.3%	32.4%
	C	17.3%	23.8%	**38.5%**	24.4%	19.4%	0.0%	58.1%	0.0%	29.9%	15.4%	16.4%
	G	16.1%	16.5%	29.4%	0.0%	66.5%	0.0%	9.4%	26.7%	**33.3%**	26.6%	17.3%
	T	32.9%	30.4%	0.0%	24.4%	14.1%	51.2%	0.0%	**52.9%**	36.8%	29.6%	33.9%

Table S8G3. Statistics of the MRCA sequence (06. Eut. Elephant) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
95,610	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	26653	24786	2013	3913	86299	4786	1631	1723	88605	23024	29896
	C	19620	27962	1548	1529	2034	85977	4922	87544	2372	19297	20023
	G	19260	21052	1838	88202	5886	0	428	1944	1942	29611	18920
	T	30077	21810	90211	1966	1391	4847	88629	4399	2691	23678	26771
%	A	27.9%	25.9%	2.1%	4.1%	90.3%	5.0%	1.7%	1.8%	92.7%	24.1%	31.3%
	C	20.5%	29.2%	1.6%	1.6%	2.1%	89.9%	5.1%	91.6%	2.5%	20.2%	20.9%
	G	20.1%	22.0%	1.9%	92.3%	6.2%	0.0%	0.4%	2.0%	2.0%	31.0%	19.8%
	T	31.5%	22.8%	94.4%	2.1%	1.5%	5.1%	92.7%	4.6%	2.8%	24.8%	28.0%
		Conserved										
	#	0	0	90211	88202	86299	85977	88629	87544	88605	0	0
	%	0.0%	0.0%	94.4%	92.3%	90.3%	89.9%	92.7%	91.6%	92.7%	0.0%	0.0%
		Variations										
%	A	27.9%	25.9%	37.3%	**52.8%**	0.0%	49.7%	23.4%	21.4%	0.0%	24.1%	31.3%
	C	20.5%	29.2%	**28.7%**	20.6%	21.8%	0.0%	70.5%	0.0%	33.9%	20.2%	20.9%
	G	20.1%	22.0%	34.0%	0.0%	63.2%	0.0%	6.1%	24.1%	**27.7%**	31.0%	19.8%
	T	31.5%	22.8%	0.0%	26.5%	14.9%	50.3%	0.0%	**54.5%**	38.4%	24.8%	28.0%

Table S8H1. Statistics of the MRCA sequence (07. Bor. Dog) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
12,125	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	3753	4009	58	111	11791	327	43	91	11749	2229	3536
	C	2054	3152	129	37	61	11485	132	11817	129	2659	2709
	G	2821	2977	73	11889	219	0	23	49	153	3424	2220
	T	3497	1987	11865	88	54	313	11927	168	94	3813	3660
%	A	31.0%	33.1%	0.5%	0.9%	97.2%	2.7%	0.4%	0.8%	96.9%	18.4%	29.2%
	C	16.9%	26.0%	1.1%	0.3%	0.5%	94.7%	1.1%	97.5%	1.1%	21.9%	22.3%
	G	23.3%	24.6%	0.6%	98.1%	1.8%	0.0%	0.2%	0.4%	1.3%	28.2%	18.3%
	T	28.8%	16.4%	97.9%	0.7%	0.4%	2.6%	98.4%	1.4%	0.8%	31.4%	30.2%
		Conserved										
	#	0	0	11865	11889	11791	11485	11927	11817	11749	0	0
	%	0.0%	0.0%	97.9%	98.1%	97.2%	94.7%	98.4%	97.5%	96.9%	0.0%	0.0%
		Variations										
%	A	31.0%	33.1%	22.3%	**47.0%**	0.0%	51.1%	21.7%	29.5%	0.0%	18.4%	29.2%
	C	16.9%	26.0%	**49.6%**	15.7%	18.3%	0.0%	66.7%	0.0%	34.3%	21.9%	22.3%
	G	23.3%	24.6%	28.1%	0.0%	65.6%	0.0%	11.6%	15.9%	**40.7%**	28.2%	18.3%
	T	28.8%	16.4%	0.0%	37.3%	16.2%	48.9%	0.0%	**54.5%**	25.0%	31.4%	30.2%

Table S8H2. Statistics of the MRCA sequence (07. Bor. Dog) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
3,090	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	1014	939	31	72	2941	142	16	32	2937	917	974
	C	550	712	48	31	35	2828	69	2949	61	463	496
	G	513	530	38	2947	90	0	8	34	56	800	573
	T	1013	909	2973	40	24	120	2997	75	36	910	1047
%	A	32.8%	30.4%	1.0%	2.3%	95.2%	4.6%	0.5%	1.0%	95.0%	29.7%	31.5%
	C	17.8%	23.0%	1.6%	1.0%	1.1%	91.5%	2.2%	95.4%	2.0%	15.0%	16.1%
	G	16.6%	17.2%	1.2%	95.4%	2.9%	0.0%	0.3%	1.1%	1.8%	25.9%	18.5%
	T	32.8%	29.4%	96.2%	1.3%	0.8%	3.9%	97.0%	2.4%	1.2%	29.4%	33.9%
		Conserved										
	#	0	0	2973	2947	2941	2828	2997	2949	2937	0	0
	%	0.0%	0.0%	96.2%	95.4%	95.2%	91.5%	97.0%	95.4%	95.0%	0.0%	0.0%
		Variations										
%	A	32.8%	30.4%	26.5%	**50.3%**	0.0%	54.2%	17.2%	22.7%	0.0%	29.7%	31.5%
	C	17.8%	23.0%	**41.0%**	21.7%	23.5%	0.0%	74.2%	0.0%	39.9%	15.0%	16.1%
	G	16.6%	17.2%	32.5%	0.0%	60.4%	0.0%	8.6%	24.1%	**36.6%**	25.9%	18.5%
	T	32.8%	29.4%	0.0%	28.0%	16.1%	45.8%	0.0%	**53.2%**	23.5%	29.4%	33.9%

Table S8H3. Statistics of the MRCA sequence (07. Bor. Dog) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
102,996	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	28962	27599	2090	4005	93401	5453	1570	1519	95063	24477	31664
	C	21108	28669	1421	1527	2027	91936	5655	95296	3323	20721	21436
	G	20799	23587	2633	95683	6273	0	373	1646	1878	31780	20791
	T	32127	23141	96852	1781	1295	5607	95398	4535	2732	26018	29105
%	A	28.1%	26.8%	2.0%	3.9%	90.7%	5.3%	1.5%	1.5%	92.3%	23.8%	30.7%
	C	20.5%	27.8%	1.4%	1.5%	2.0%	89.3%	5.5%	92.5%	3.2%	20.1%	20.8%
	G	20.2%	22.9%	2.6%	92.9%	6.1%	0.0%	0.4%	1.6%	1.8%	30.9%	20.2%
	T	31.2%	22.5%	94.0%	1.7%	1.3%	5.4%	92.6%	4.4%	2.7%	25.3%	28.3%
		Conserved										
	#	0	0	96852	95683	93401	91936	95398	95296	95063	0	0
	%	0.0%	0.0%	94.0%	92.9%	90.7%	89.3%	92.6%	92.5%	92.3%	0.0%	0.0%
		Variations										
%	A	28.1%	26.8%	34.0%	**54.8%**	0.0%	49.3%	20.7%	19.7%	0.0%	23.8%	30.7%
	C	20.5%	27.8%	**23.1%**	20.9%	21.1%	0.0%	74.4%	0.0%	41.9%	20.1%	20.8%
	G	20.2%	22.9%	42.9%	0.0%	65.4%	0.0%	4.9%	21.4%	**23.7%**	30.9%	20.2%
	T	31.2%	22.5%	0.0%	24.4%	13.5%	50.7%	0.0%	**58.9%**	34.4%	25.3%	28.3%

Table S8I1. Statistics of the MRCA sequence (08. Eua. Mouse) to variations of 11mers with TGACTCA (in CNEs and in DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
9,566	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	2825	2797	84	110	9164	239	63	99	9058	1640	2696
	C	1751	2527	117	51	78	9073	118	9259	154	2202	2258
	G	2259	2760	100	9336	279	0	38	58	228	2877	1876
	T	2731	1482	9265	69	45	254	9347	150	126	2847	2736
%	A	29.5%	29.2%	0.9%	1.1%	95.8%	2.5%	0.7%	1.0%	94.7%	17.1%	28.2%
	C	18.3%	26.4%	1.2%	0.5%	0.8%	94.8%	1.2%	96.8%	1.6%	23.0%	23.6%
	G	23.6%	28.9%	1.0%	97.6%	2.9%	0.0%	0.4%	0.6%	2.4%	30.1%	19.6%
	T	28.5%	15.5%	96.9%	0.7%	0.5%	2.7%	97.7%	1.6%	1.3%	29.8%	28.6%
		Conserved										
	#	0	0	9265	9336	9164	9073	9347	9259	9058	0	0
	%	0.0%	0.0%	96.9%	97.6%	95.8%	94.8%	97.7%	96.8%	94.7%	0.0%	0.0%
		Variations										
%	A	29.5%	29.2%	27.9%	**47.8%**	0.0%	48.5%	28.8%	32.2%	0.0%	17.1%	28.2%
	C	18.3%	26.4%	**38.9%**	22.2%	19.4%	0.0%	53.9%	0.0%	30.3%	23.0%	23.6%
	G	23.6%	28.9%	33.2%	0.0%	69.4%	0.0%	17.4%	18.9%	**44.9%**	30.1%	19.6%
	T	28.5%	15.5%	0.0%	30.0%	11.2%	51.5%	0.0%	**48.9%**	24.8%	29.8%	28.6%

Table S8I2. Statistics of the MRCA sequence (08. Eua. Mouse) to variations of 11mers with TGACTCA (in CNEs and no DHSs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
2,161	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	676	637	25	66	2000	79	37	21	2017	600	683
	C	383	538	40	28	36	1986	57	2045	51	348	353
	G	392	388	32	2039	104	0	5	29	50	596	405
	T	710	598	2064	28	21	96	2062	66	43	617	720
%	A	31.3%	29.5%	1.2%	3.1%	92.5%	3.7%	1.7%	1.0%	93.3%	27.8%	31.6%
	C	17.7%	24.9%	1.9%	1.3%	1.7%	91.9%	2.6%	94.6%	2.4%	16.1%	16.3%
	G	18.1%	18.0%	1.5%	94.4%	4.8%	0.0%	0.2%	1.3%	2.3%	27.6%	18.7%
	T	32.9%	27.7%	95.5%	1.3%	1.0%	4.4%	95.4%	3.1%	2.0%	28.6%	33.3%
		Conserved										
	#	0	0	2064	2039	2000	1986	2062	2045	2017	0	0
	%	0.0%	0.0%	95.5%	94.4%	92.5%	91.9%	95.4%	94.6%	93.3%	0.0%	0.0%
		Variations										
%	A	31.3%	29.5%	25.8%	**54.1%**	0.0%	45.1%	37.4%	18.1%	0.0%	27.8%	31.6%
	C	17.7%	24.9%	**41.2%**	23.0%	22.4%	0.0%	57.6%	0.0%	35.4%	16.1%	16.3%
	G	18.1%	18.0%	33.0%	0.0%	64.6%	0.0%	5.1%	25.0%	**34.7%**	27.6%	18.7%
	T	32.9%	27.7%	0.0%	23.0%	13.0%	54.9%	0.0%	**56.9%**	29.9%	28.6%	33.3%

Table S8I3. Statistics of the MRCA sequence (08. Eua. Mouse) to variations of 11mers with TGACTCA (no CNEs)

Total	Position	1	2	3	4	5	6	7	8	9	10	11
45,561	Consensus	N	N	T	G	A	C	T	C	A	N	N
#	A	12030	11347	1177	1856	40791	2159	969	864	41639	9777	14072
	C	9580	13322	697	905	910	41355	2010	41678	1527	9952	9957
	G	9864	11854	1163	42059	2971	0	246	970	957	14850	9377
	T	14087	9038	42524	741	889	2047	42336	2049	1438	10982	12155
%	A	26.4%	24.9%	2.6%	4.1%	89.5%	4.7%	2.1%	1.9%	91.4%	21.5%	30.9%
	C	21.0%	29.2%	1.5%	2.0%	2.0%	90.8%	4.4%	91.5%	3.4%	21.8%	21.9%
	G	21.7%	26.0%	2.6%	92.3%	6.5%	0.0%	0.5%	2.1%	2.1%	32.6%	20.6%
	T	30.9%	19.8%	93.3%	1.6%	2.0%	4.5%	92.9%	4.5%	3.2%	24.1%	26.7%
		Conserved										
	#	0	0	42524	42059	40791	41355	42336	41678	41639	0	0
	%	0.0%	0.0%	93.3%	92.3%	89.5%	90.8%	92.9%	91.5%	91.4%	0.0%	0.0%
		Variations										
%	A	26.4%	24.9%	38.8%	**53.0%**	0.0%	51.3%	30.0%	22.3%	0.0%	21.5%	30.9%
	C	21.0%	29.2%	**23.0%**	25.8%	19.1%	0.0%	62.3%	0.0%	38.9%	21.8%	21.9%
	G	21.7%	26.0%	38.3%	0.0%	62.3%	0.0%	7.6%	25.0%	**24.4%**	32.6%	20.6%
	T	30.9%	19.8%	0.0%	21.2%	18.6%	48.7%	0.0%	**52.8%**	36.7%	24.1%	26.7%