Supporting Information

Branciamore et al. 10.1073/pnas.1010506107

SI Text

CpG Island Cluster Analysis. We define a CGI cluster as a set of CGIs with distance between consecutive CGIs less than a given threshold (T) (Fig. S5). The distance between two CGIs is expressed as the number of bases between the last nucleotide in the preceding and the first nucleotide of the following element. A CGI cluster of size one corresponds to an isolated CGI. The total extension of a CGI cluster is defined by the positions of nucleotides at 5′ of the first CGI element and 3′ of the last CGI element.

Simulation of CGI Cluster Distribution. The distribution of CGI cluster for each human chromosome was simulated under the hypothesis that CGI elements were evenly spread with an average distance D being equal to what was observed in the corresponding chromosome.

For each chromosome we simulate 1,000 replicas with the same number of CGIs present as in the real chromosome. The distance between consecutive CGI elements was obtained by sampling from an exponential distribution with mean value D. Then we evaluated the CGI cluster distribution and (for each chromosome) the distribution of CGIs of different size as an average of the distributions found in the single simulation.

Analytical Approximation of CGI Cluster Distribution. Given a threshold (T), the analytical approximation of the CGI cluster distribution can be obtained under the assumption that the distance between consecutive elements follows an exponential distribution with average value D.

The probability that the distance of two consecutive CGIs is less than a given threshold T is determined by the cumulative distribution function:

$$p = F(T, D) = 1 - e^{-T/D}.$$

Let us define $P_1, P_2, P_3, \ldots, P_k$ as the probability that a CGI belongs to a cluster of size $1, 2, 3, \ldots, k$. Then the probability P_1 that the CGI is isolated would be

$$P_1 = (1-p)^2$$

This is actually the probability that the CGI is positioned at the distance larger than D simultaneously from the previous and next elements (Fig. S5).

If the number of CpG islands present in the chromosome is N, then (disregarding errors associated with the chromosome boundaries), the expected number of CGIs in clusters of size 1 (i.e., single CGIs) would be

$$C_1 \approx P_1 N.$$

Similarly, the probability P_2 that the CGI element belongs to clusters of size 2 is given by

$$P_2 = p(1-p)^2,$$

and the expected number of CpG islands in clusters of size 2 would be

$$C_2 \approx P_2 (N-1).$$

By analogous iterating considerations, one can evaluate the probability and the expected CGI number in clusters of larger sizes.

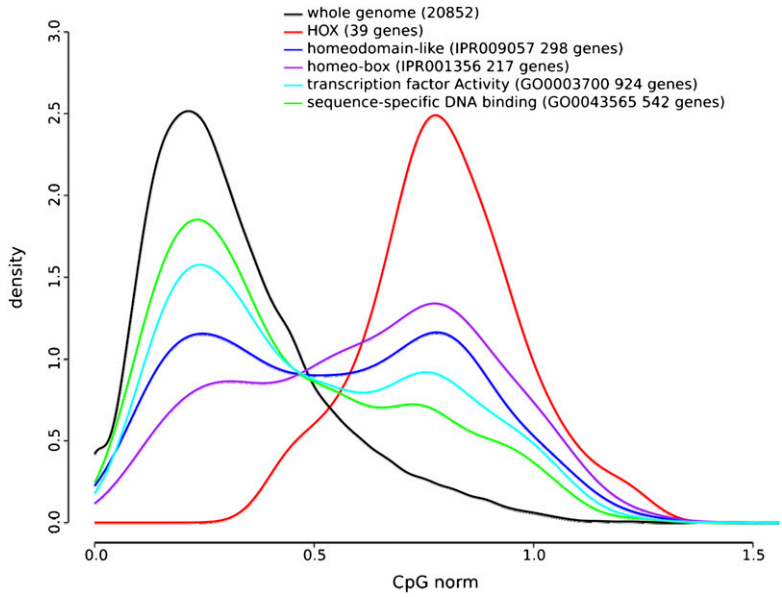

Fig. S1. The frequency distributions of genes built in accordance with their CpG normalized value (*Materials and Methods*). Shown are different subsets of genes obtained according to their Gene Ontology or interpro database association (*Materials and Methods*). All curves are normalized to have area = 1.

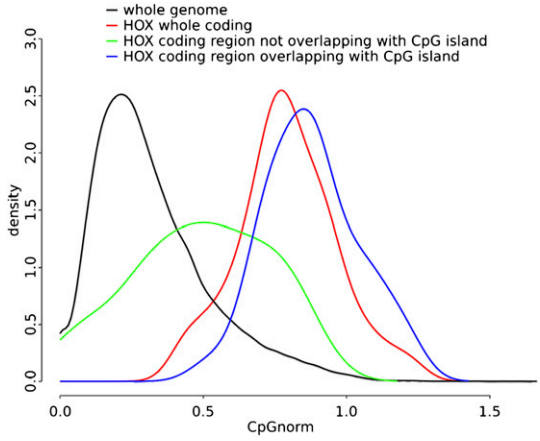

Fig. S2. The frequency distributions of genes built in accordance with their CpG normalized value for different subversions of human Hox genes. The frequency distributions for coding portions overlapping or not overlapping with CpG islands are shown in blue and green, respectively. Also shown are the CpG_{norm} distributions for the entire coding portion of the Hox genes (red) and the whole human genome (black). All curves are normalized to have area = 1.

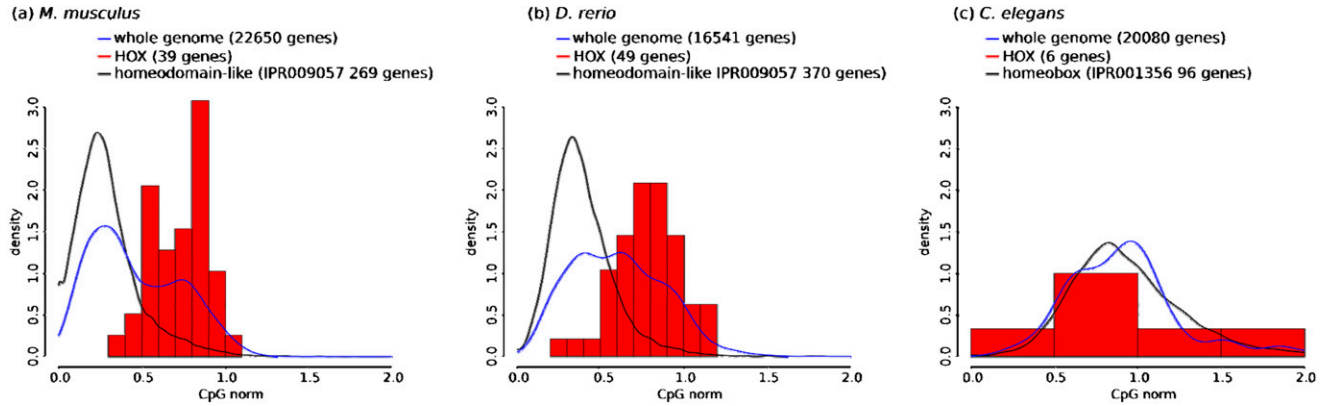

Fig. S3. The frequency distributions of genes built in accordance with their CpG normalized value (*Materials and Methods*) for *M. musculus* (*a*), *D. rerio* (*b*), and *C. elegans* (*c*). Interpro name IPR009057 for homeodomain-like genes was not avaible for *C. elegans*; therefore we used (as the closest subset) the homeobox genes IPR001356. All curves are normalized to have area = 1.

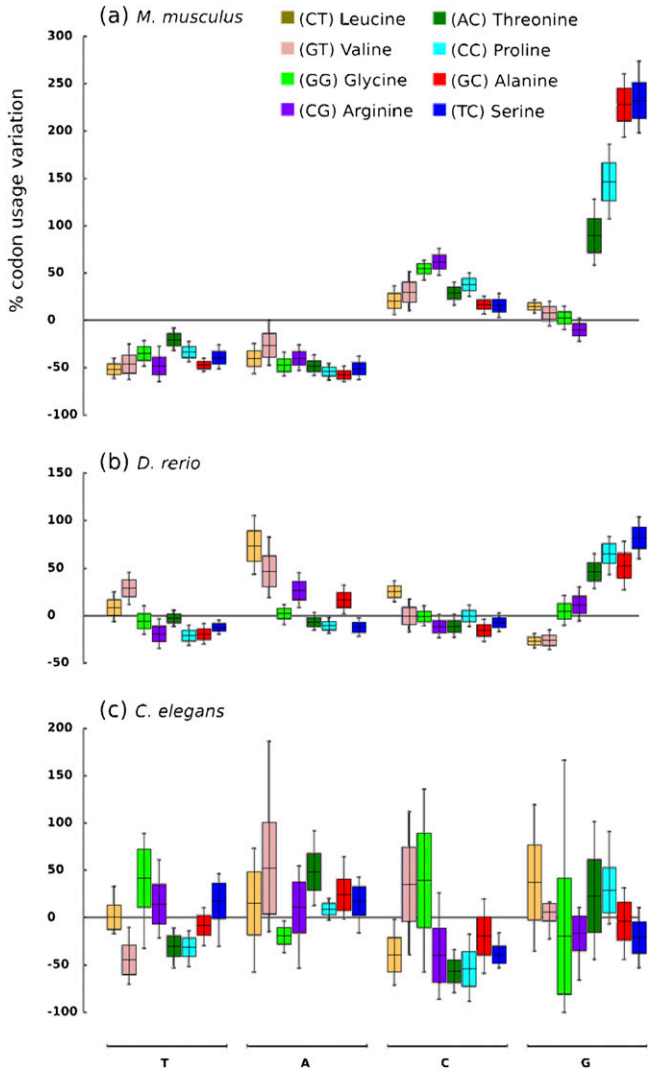

Fig. S4. Variation in usage of 4d codons in *M. musculus* (*a*), *D. rerio* (*b*), and *C. elegans* (*c*) *Hox* genes compared with the average genome values. Primary data were retrieved from the Codon Usage Database at http://www.kazusa.or.jp/codon/. All abbreviations are the same as in Fig. 3.

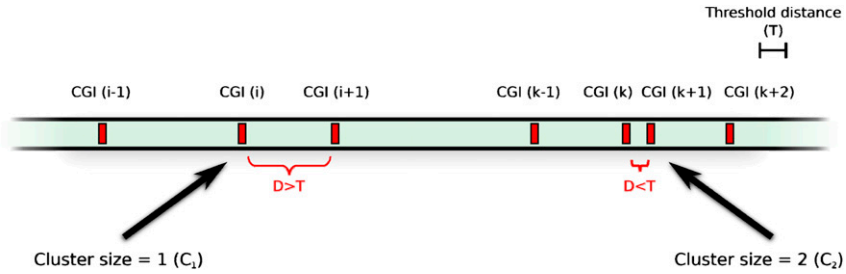

Fig. S5. Schematic representation of CGI cluster.

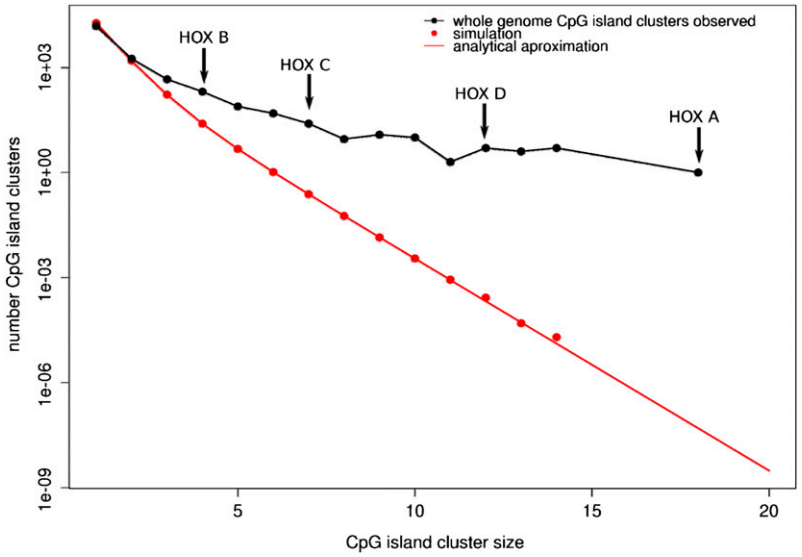

Fig. S6. Observed (black), expected by simulation (red circles), and analytically approximated (red line) numbers of CpG islands in clusters of different sizes (*Materials and Methods*). Numbers of CGI clusters are represented in the logarithmic scale. Arrows indicate the largest CGI clusters overlapped with the corresponding Hox gene loci. Obviously, the observed CGI clusterization significantly exceeds the simulated and analytical ones obtained under the assumption of their randomness.

Table S1. List of the genes randomly chosen from the human genome for comparison with *Hox* genes

Table S1 (DOC)

Table S2. Nucleotides frequencies in silent sites of 4d codons calculated for human Hox gene clusters as opposed to the average genome-wide

Table S2 (DOC)

Table S3. Enriched GO terms in genes not overlapping with CpG islands (noCGI), in genes overlapping with CpG islands (CGI), in genes overlapping with CpG island clusters (CGIcl), and in genes overlapping with CGI but not belonging to CGIcl (CGI minus CGIcl)

Table S3 (DOC)