Supporting online material for:

Human gene organization driven by the coordination of replication and transcription

The wavelet transform as a multi-scale shape detector

The continuous wavelet transform (Mallat 1998) (WT) is a space-scale analysis which consists in expanding a signal or *S* profile in terms of wavelets $\psi_{b,a}(x) = \frac{1}{\sqrt{a}}\psi(\frac{x-b}{a})$ that are constructed from a single function, the analyzing wavelet ψ, by means of translations (*b*) and dilations (*a*>0):

$$T_\psi[S](b,a) = \int S(x)\psi_{b,a}(x)\mathrm{d}x\,.$$

The wavelet coefficient $T_\psi[S](b,a)$ quantifies to which extent, around position *b* over a distance *a, S* has a similar shape as the analyzing wavelet ψ. In other words, by looking for the maxima of $T_\psi[S](b,a)$ over the space-scale half-plane, the WT can be used as multi-scale shape detector. The great freedom of choice for the shape of the analyzing wavelet ψ is a fundamental property of the WT allowing this signal-processing tool to be adapted to a large number of problems. Here, we take advantage of this flexibility to detect the factory-roof profile predicted for N-domains. Indeed, the linearly decreasing shape between two upward jumps characteristic of this profile can be used as an adapted analyzing wavelet Ξ (Fig. S1).

References

Mallat, S. 1998. *A Wavelet Tour in Signal Processing*. Academic Press, New York.

Semon, M., D. Mouchiroud, and L. Duret. 2005. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* **14:** 421-427.

	length (masked Mbp)	δ (%)	c_g (%)
domain 1	0.61	7.2	4.3, -15.4, -9.2, -10.8
domain 2	0.21	13.2	2.1, -6.7
domain 3	0.48	6.0	4.7, -16.3, 1.5, 8.2, 3.8, -3.5

Table S1. Replication and transcription bias model parameters obtained for the 3 N-domains presented in Fig. 3 (see main text, Eq. (2)). δ is the slope estimated for the replication bias; c_g is the transcription bias estimated for each gene.

a

n(R+) *n*(R-)	0	1	2	3	4	5	6	7	8
0	222	373	148	49	14	1	5	0	1
1	71	122	68	41	11	8	2	0	1
2	14	39	34	16	14	0	4	3	0
3	6	10	11	12	4	7	0	0	1
4	0	2	6	2	2	4	0	0	1
5	1	2	2	0	0	0	1	0	1
6	1	0	0	1	0	0	1	0	0
7	0	2	1	0	0	1	0	0	0
8	0	1	0	0	0	0	0	0	0

b

n(R+) *n*(R-)	0	1	2	3	4	5	6	7	8
0	218	355	139	41	10	1	3	0	0
1	67	112	62	35	9	3	1	0	1
2	9	35	29	12	11	0	1	1	0
3	3	6	9	9	1	3	0	0	0
4	0	1	4	1	0	2	0	0	0
5	0	1	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0

Table S2. (a) Repartition of the R+ and R- genes in the half N-domains containing defined gene numbers. In each cell, the number of half-domains containing *n*(R+) genes (column) and *n*(R-) genes (line). (b) Same as in (a) when the domains containing duplicated genes are not considered.

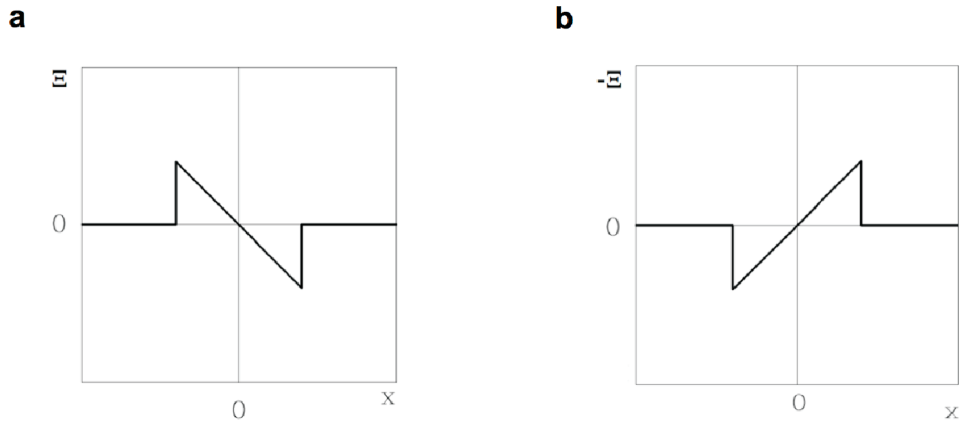

Figure S1. Analyzing wavelet adapted for the detection of N-domains. (a) The shape of the wavelet corresponds to the expected shape of the *S* profile between two well-positioned replication origins assuming a random terminus position (see main text Fig. 1). (b) Analyzing wavelet used to detect the inverted profiles.

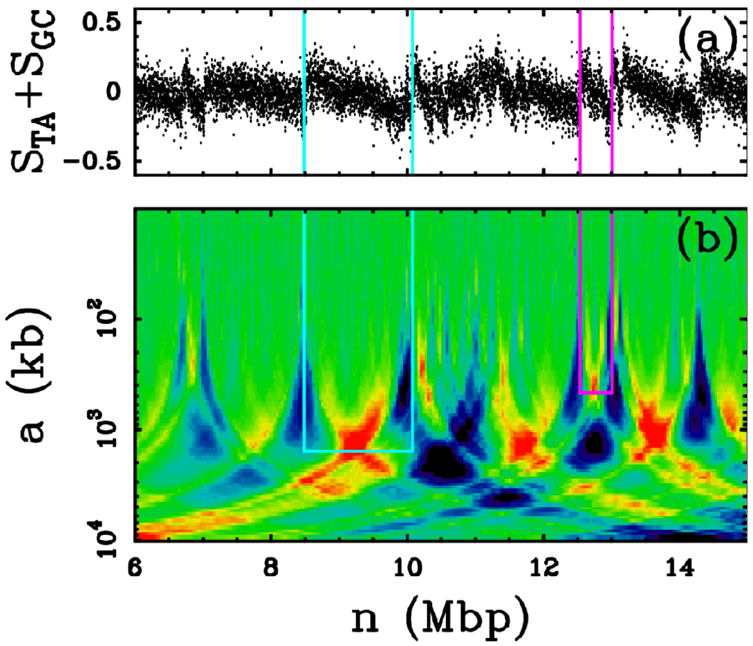

Figure S2. Wavelet-based analysis of genomic sequences. (a) Skew profile *S* of a 9Mbp repeat-masked fragment of human chromosome 21. (b) WT of *S* using Ξ; $T_{\Xi}[S](n,a)$ is color-coded from dark-blue (min; negative values) to red (max; positive values) through green (null values). Light-blue and purple lines illustrate the detection of two N-domains of significantly different sizes. Note that in (b), blue cone-shape areas signing upward jumps point at small scale (top) towards the putative replication origins and that the vertical positions of the WT maxima (red areas) corresponding to the two indicated N-domains match the distance between the putative replication origins (~1.6 Mbp and ~470 kbp respectively).

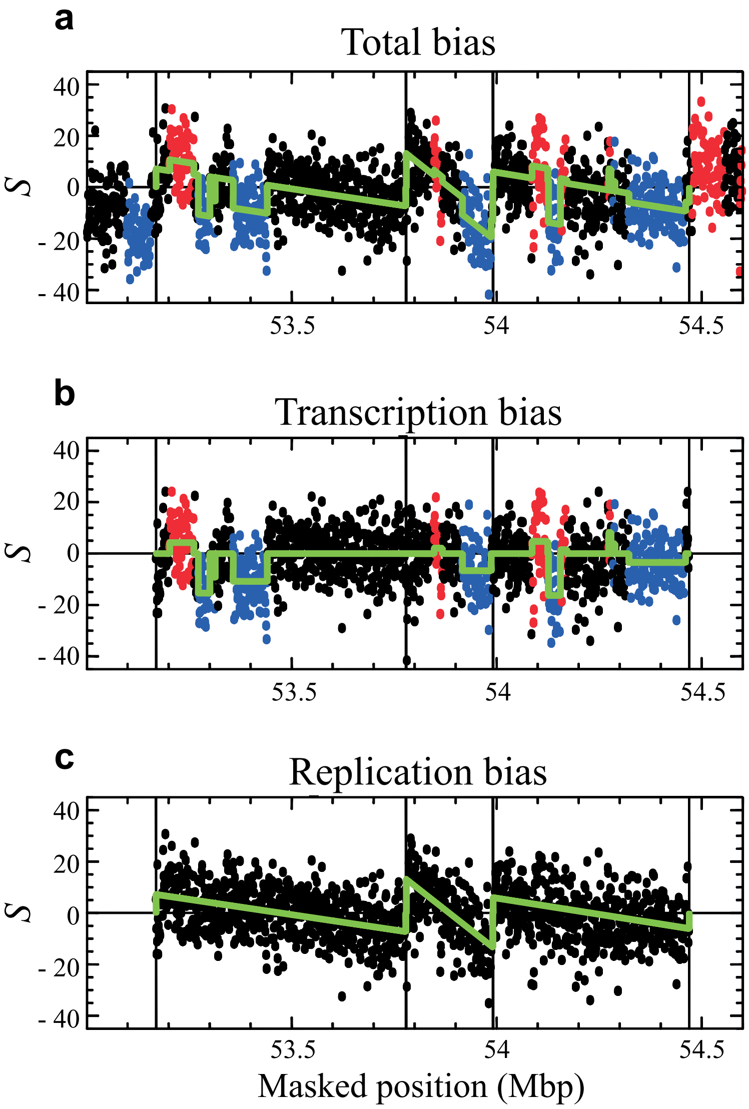

Figure S3. Disentangling transcription- and replication-associated skews in N-domains. (a) Skew profile, *S*, of a 1.6-Mbp, repeat-masked fragment of human chromosome 6 (each point corresponds to a 1-kbp window); red, + genes; blue, - genes; black, intergenic regions; the factory-roof profile estimated with Eq. (2) is shown in green; vertical lines correspond to the location of 4 putative replication origins that delimit 3 adjacent domains identified by the wavelet-based methodology. (b) Transcription-associated compositional asymmetry obtained by subtracting the estimated replication-associated profile (green lines in (c)) from the original *S* profile in (a); the estimated transcription step-like profile (second term in the right hand side of Eq. (2)) is shown in green. (c) Replication-associated compositional asymmetry obtained by subtracting the estimated transcription step-like profile (green lines in (b)) from the original *S* profile in (a); the estimated replication serrated profile (first term in the right hand side of Eq. (2)) is shown in green.

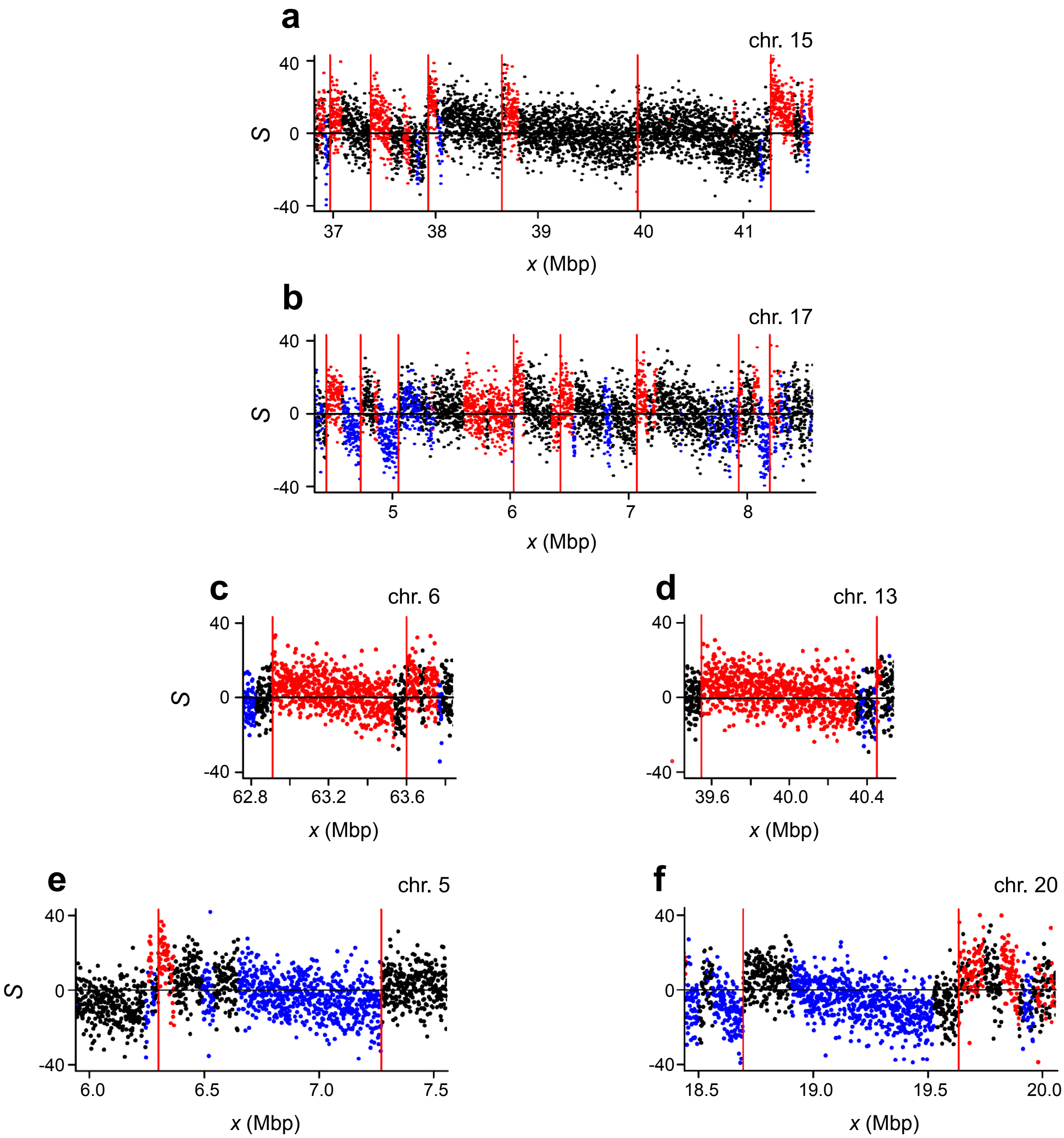

Figure S4. Examples of N-domains detected in the human genome. (a-f) Fragments of the indicated chromosomes. *S* values are computed in 1-kbp windows (without repeats); red, + genes; blue, - genes; black, intergenic regions; the domain borders are indicated by red vertical bars. On the abscissa, the position of a sequence window in Mbp. On the ordinate, the skew, *S*, as a percentage.

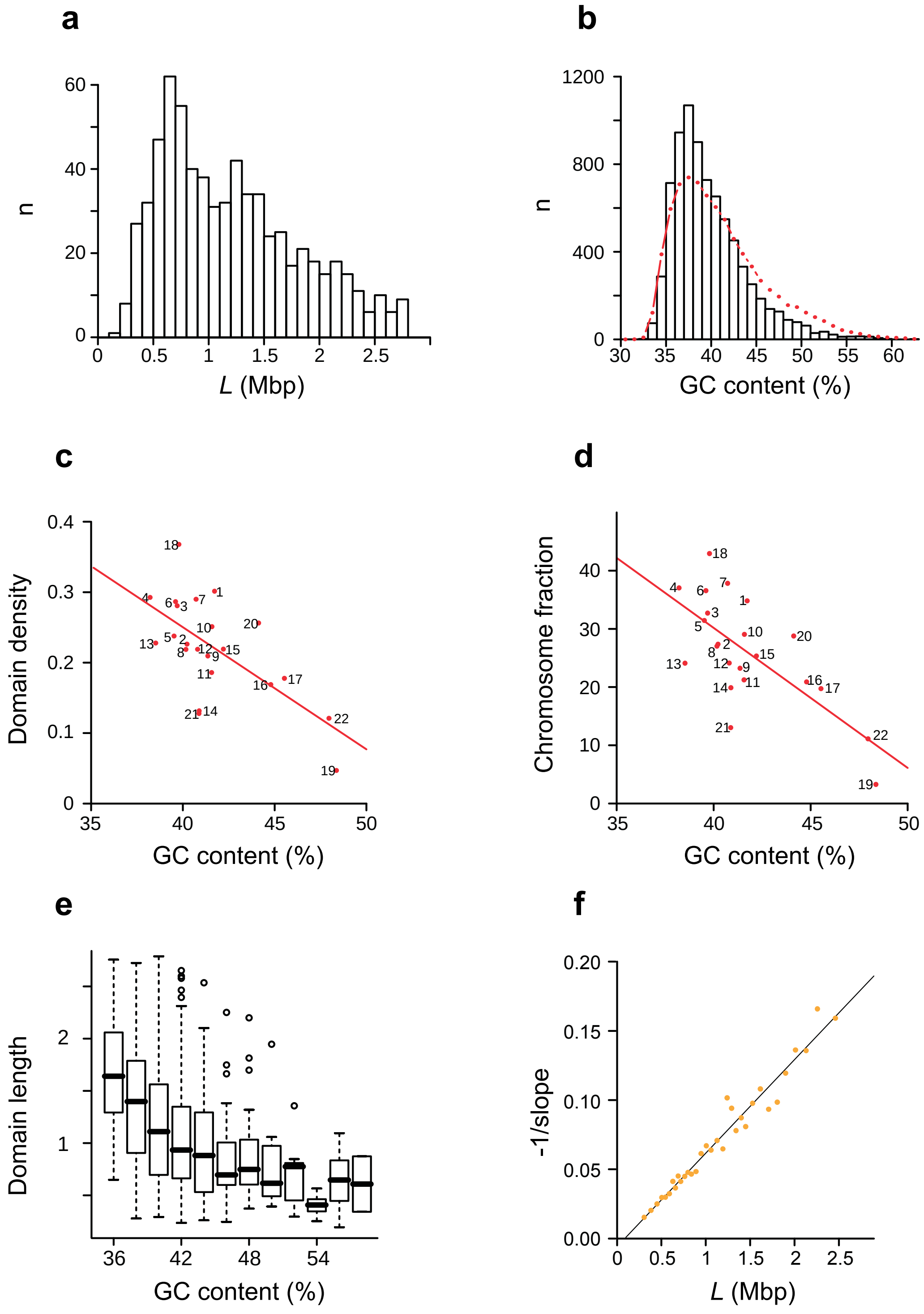

Figure S5. N-domain characteristics. (a) Histogram of the domain length *L* (Mbp, unmasked sequences). (b) Histogram of the GC content of the domains (black). Histogram of the GC content computed in 2-Mbp non-overlapping windows in the 22 autosomes (red). (c) Density of the domains (number of domains per Mbp) versus chromosome GC content; the correlation is $r = 0.62$, $P = 0.002$. Each number refers to the corresponding chromosome. (d) Fraction of chromosome length covered by the domains versus chromosome GC content; $r = 0.70$, $P = 3.10^{-4}$. (e) Domain length (*L*) versus domain GC content; $r = 0.4$, $P < 10^{-15}$. (f) Inverse of the slope of the skew profile versus domain length, *L*. Domains are ranked by *L* values, grouped by bins of 20 domains; the slope (in %/Mbp) is the mean slope of the linear fits of the skew profiles, computed for each set; the equation of the linear fit is y = 0.0675x-0.0058 (x correspond to *L* values and y to -1/slope values); $r = 0.9$, $P<10^{-15}$.

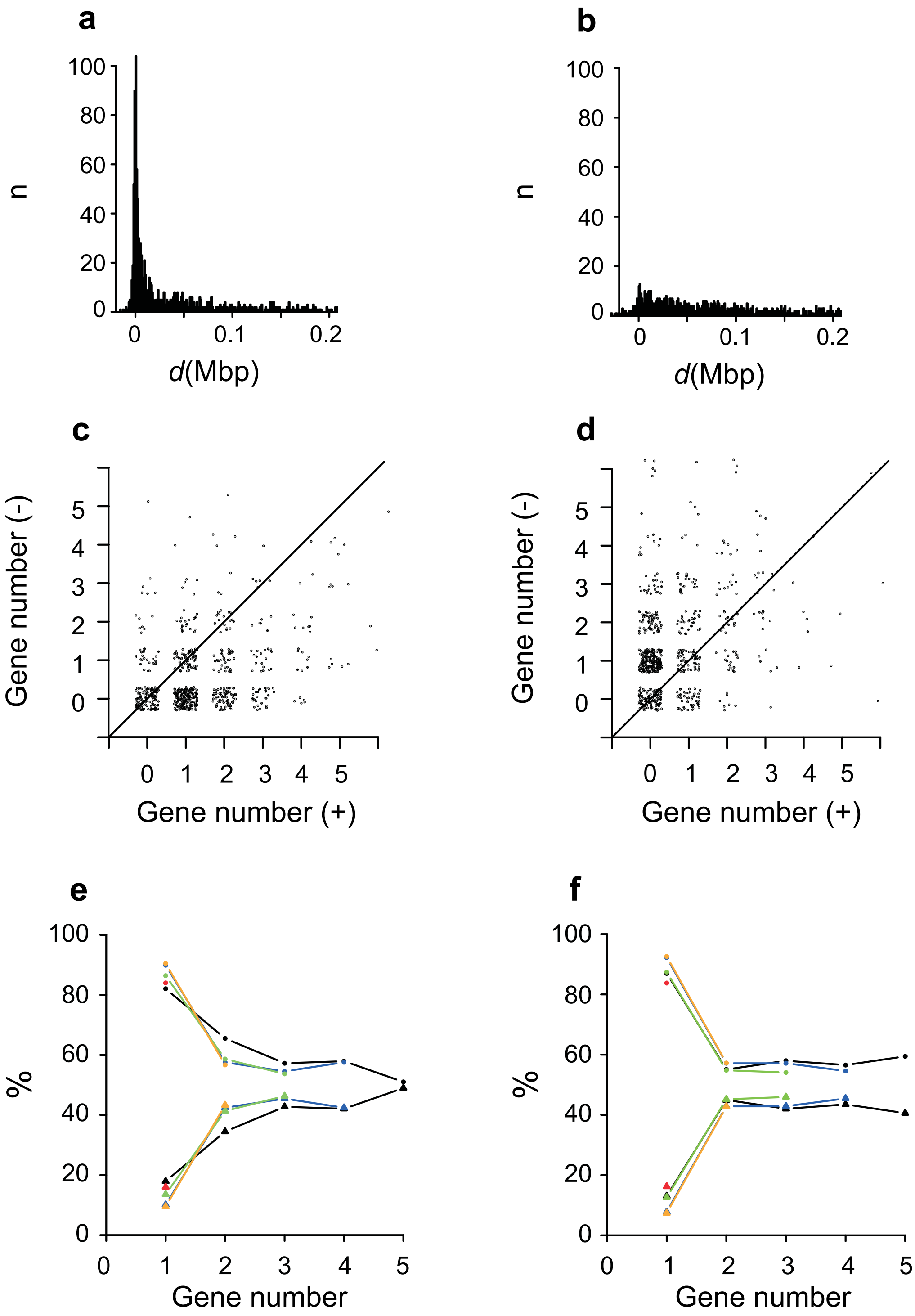

Figure S6. Gene organization in the N-domains. (a) Distribution of the distance *d* of the first gene to the corresponding domain extremity; for 49% of genes, the distance of the 5’ end is smaller than 10kbp. (b) As in (a) but with domains randomly positioned along the chromosomes. (c) Number and distribution of the + and - genes. Each point represents the number of + genes (on the abscissa) and of - genes (on the ordinate) located in the 5’ half of a domain (for clarity, the points corresponding to given abscissa and ordinate are dispersed around the corresponding position).(d) Same as in (a) but for the 3’ half of the domains. (e) Orientation of the successive genes in the half-domains. In each half-domain, the successive genes are numbered from the extremity toward the center. At each value of n (on the abscissa), the percentage of R+ genes (circle) and of R- genes (triangle) is reported on the ordinate. The color corresponds to half-domains containing 1 gene (red), 2 genes (orange), 3 genes (green) 4 genes (blue), more than 4 genes (black). (f) As in (e) after eliminating the domains containing duplicated genes (see main text).

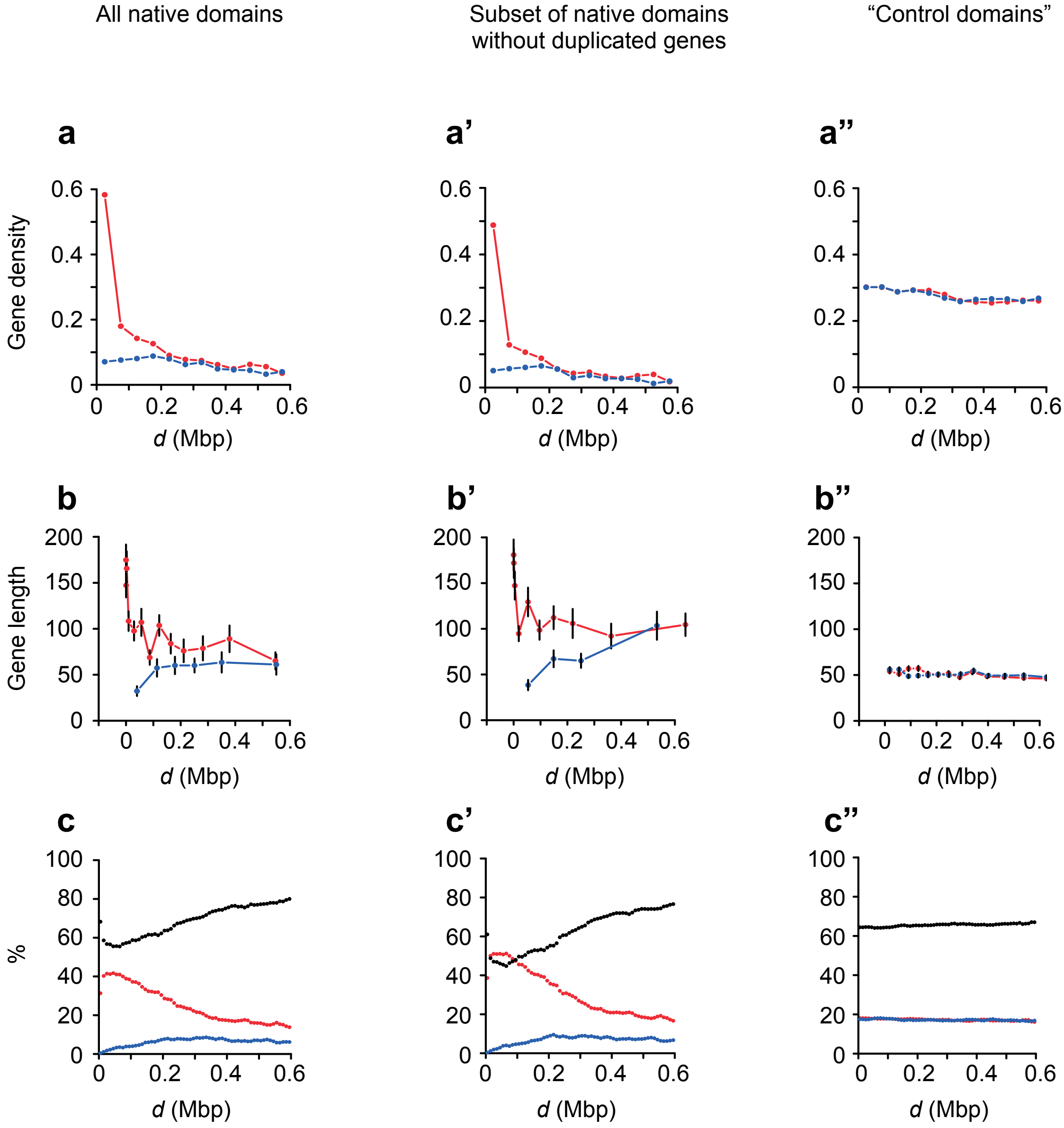

Figure S7. Comparison of the N-domains versus the control domains. We compared the domains detected in the human chromosomes (left column) to the subset of these domains that did not contain duplicated genes (central column), and to the control domains obtained after randomization of domain positions (see Methods in the main text) (right). To facilitate the comparison, the sequences in the 3' halves of the domains have been inverse-complemented, and are analyzed together with the 5' halves. (a-a") Gene density. The density is defined as the number of 5' ends for R+ genes (red) or of 3' ends for R- genes (blue) in adjacent 50-kbp windows, divided by the number of corresponding half-domains. (b-b") Mean gene length. Genes are ranked by their distance, *d*, from the closest half-domain extremity, grouped in sets of 150 genes, and the mean length (kbp) is computed for each of these sets; colors are as in (a-a"). (c-c") Relative numbers of base pairs that are transcribed in the R+ direction (red), R- direction (blue) and that are non-transcribed (black), determined in adjacent 10-kbp sequence windows.

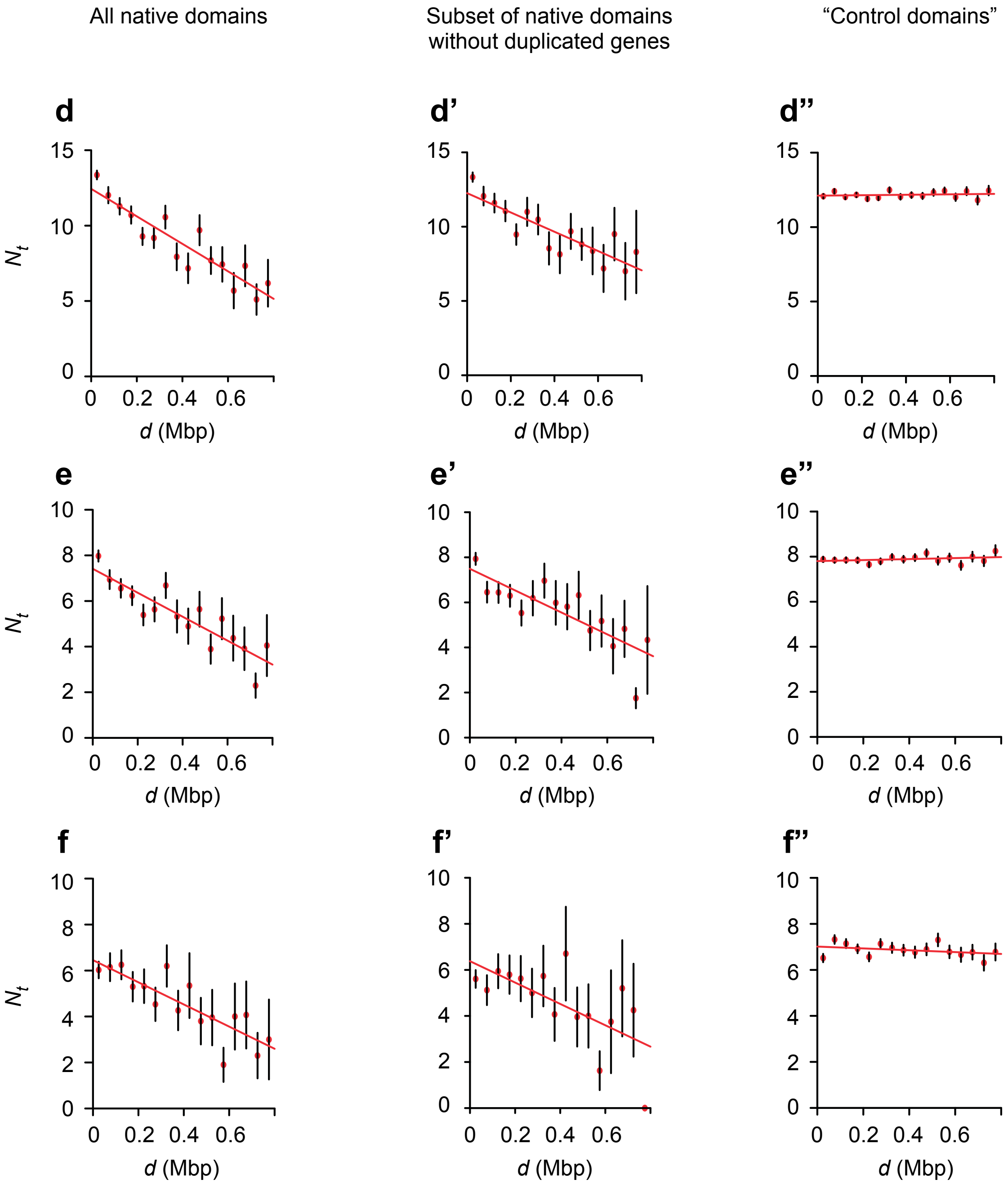

Figure S7 (continued). (d-d”) Mean expression breadth of the genes, calculated using EST data. Mean values are computed with R+ (resp. R-) genes whose 5’ ends (resp. 3’ ends) are located in adjacent 50-kbp windows. On the abscissa, the distance, *d*, in Mbp from the corresponding half-domain extremity (putative replication origin). (e-e”) same as in (d-d”) using SAGE data. (f-f”) same as in (d-d”) using microarray data. Expression data are described in Semon et al. 2005.

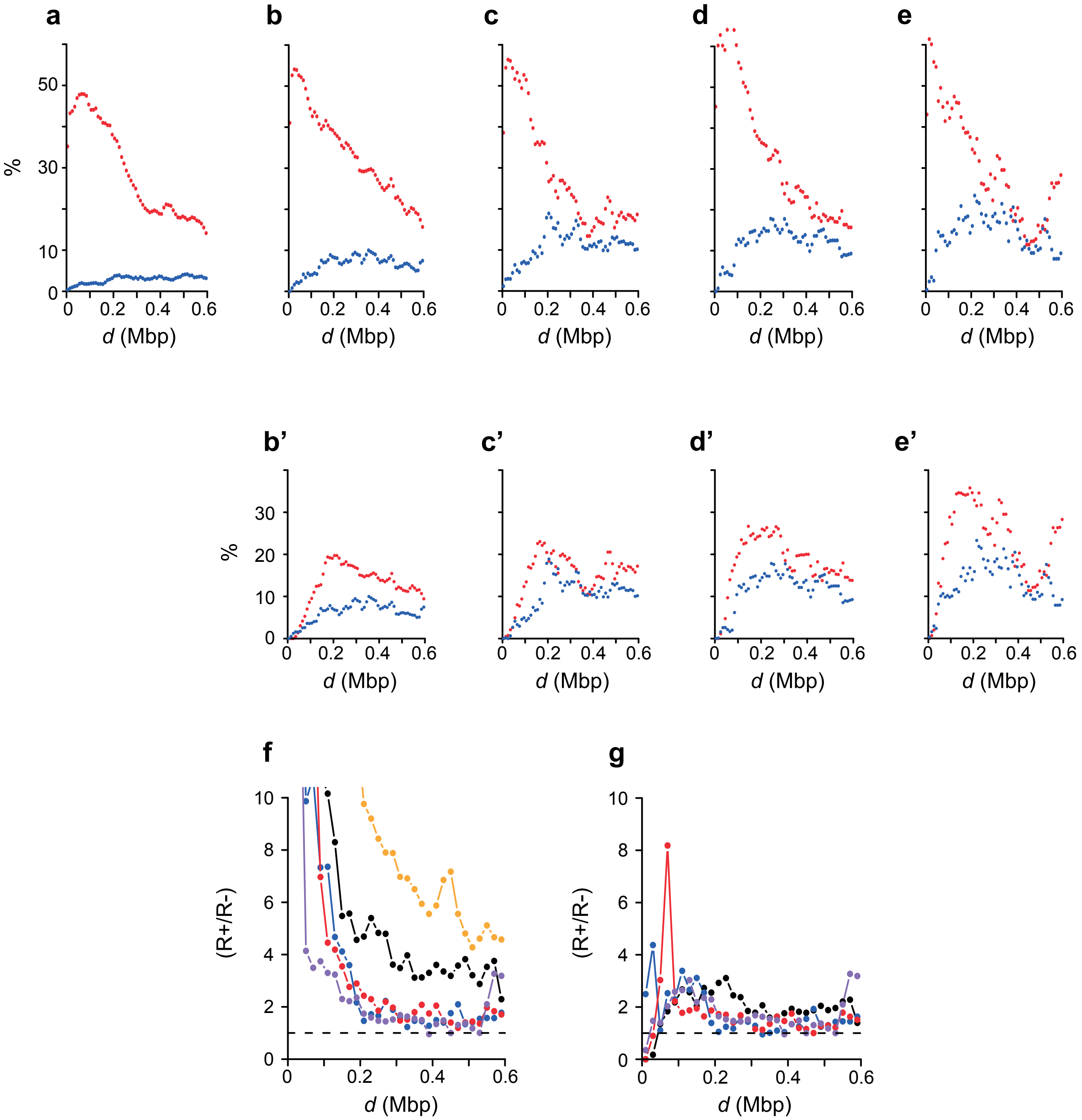

Figure S8. Polarity of transcription orientation in the N-domains. Relative numbers of base pairs transcribed in the R+ (red) and R- (blue) directions are computed in non-overlapping 10-kbp sequence windows in the half-domains (the 3' halves of the domains have been inverse complemented and added to the 5' halves). In this analysis, domains containing duplicated genes were not considered. The relative numbers have been computed in half-domains harboring 1 gene (a), 2 genes (b), 3 genes (c), 4 genes (d) and more than 4 genes (e). Similar measurements are performed after eliminating the first gene from the analysis, for half-domains harboring 2 genes (b'), 3 genes (c'), 4 genes (d') and more than 4 genes (e'). On the abscissa, the distance of the fragment from the corresponding half-domain extremity (putative replication origin) in Mbp. (f) Ratio of the number of R+ over R- base pairs (computed in non-overlapping 20-kbp sequence windows) in half-domains harboring 1 gene (yellow), 2 genes (black), 3 genes (blue), 4 genes (red) and more than 4 genes (purple); the dashes correspond to a ratio of 1. (g) Same as in (f) after eliminating the first gene (as in b'-e').

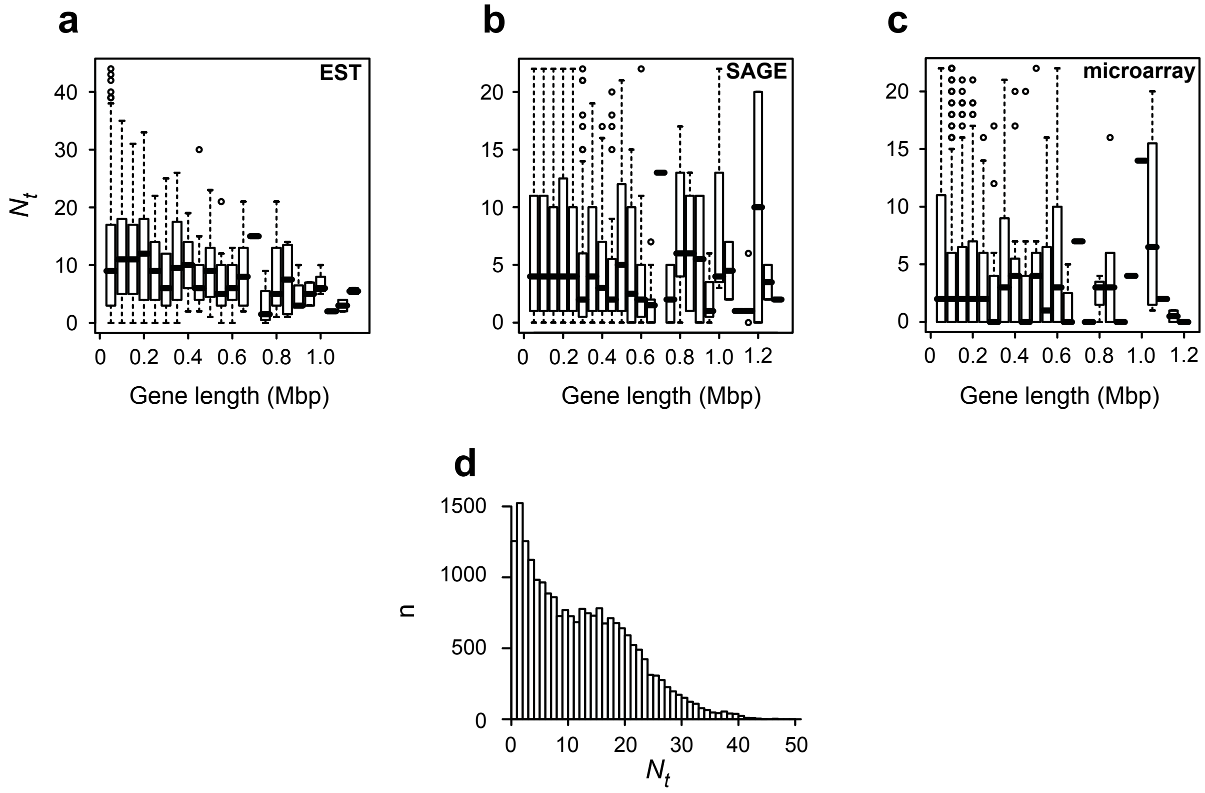

Figure S9. Correlation between gene length and breadth of expression for the genes located in the N-domains. (a) EST data, the correlation is r = 0.07, P = 4.10^{-4}. (b) SAGE data; the correlation is r = 0.06, P = 9.10^{-4}. (c) Microarray data, the correlation is r = 0.07, P = 7.10^{-3}. (d) Histogram of the expression breadth (determined using the EST data) of the whole genome gene set. Expression data are described in Semon et al. 2005.