Supporting Material for
Open chromatin encoded in DNA sequence is the signature of “master” replication origins in human cells

Benjamin Audit\textsuperscript{1,2}, Lamia Zaghloul\textsuperscript{1,2}, Cédric Vaillant\textsuperscript{1,2}, Guillaume Chevereau\textsuperscript{1,2}, Yves d’Aubenton-Carafa\textsuperscript{3}, Claude Thermes\textsuperscript{3} and Alain Arneodo\textsuperscript{1,2,*}

\textsuperscript{1} Université de Lyon, F-69000 Lyon, France
\textsuperscript{2} Laboratoire Joliot-Curie and Laboratoire de Physique, CNRS, Ecole Normale Supérieure de Lyon, F-69007 Lyon, France
\textsuperscript{3} Centre de Génétique Moléculaire, CNRS, F-91198 Gif-sur-Yvette, France

CONTENTS

Physical Modeling \hspace{1cm} 2
Cytosine methylation, CpG observed/expected ratio and CpG islands \hspace{1cm} 4
Supplementary Figures \hspace{1cm} 8

*To whom correspondence should be addressed. Tel: +33 4 7272 8757; Fax: +33 4 7272 8080; Email: alain.arneodo@ens-lyon.fr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
As proposed in (1), when focusing on the dynamical assembly of histone octamers along the DNA chain, chromatin can be reasonably modeled by a fluid of rods of finite extension $l$ (the DNA wrapping length around the octamer), binding and moving in an external potential $E(s)$ (the effective nucleosome formation potential) and interacting (potential $v(s, s')$) on a 1D substrate (the DNA chain). Within the grand canonical formalism, considering that the fluid is in contact with a thermal bath (at reciprocal temperature $\beta$) and a histone octamer reservoir (at chemical potential $\mu$), the thermodynamical equilibrium properties of the system are described by the grand partition function:

$$\Xi = \frac{1}{N!} \int D[s^{(N)}] \exp \left( -\beta \left( V(s^{(N)}) - \mu N \right) \right),$$  \hspace{1cm} (S1)

where $V(s^{(N)})$ is the total potential energy of the $N$ rods system:

$$V(s^{(N)}) = \sum_{k=1}^{N} E(s_k) + \frac{1}{2} \sum_{j \neq k} v(s_k, s_j).$$  \hspace{1cm} (S2)

From Eqs. (S1) and (S2), we get the nucleosome density profile as:

$$\rho(s) = \frac{1}{\Xi} \frac{\partial \ln \Xi}{\partial F(s)} = \frac{1}{\Xi} \sum_{N=1}^{N_{\text{max}}} e^{\beta \mu N} (N-1)! \int D[s^{(N-1)}] e^{-\beta V(s, s^{(N-1)})}. \hspace{1cm} (S3)$$

The thermodynamics of such system has been widely investigated in the literature. In the case of monodisperse hard rods on a uniform external potential, this is the well known Tonk gas (2). In the case of a non uniform external potential of interest here, the problem was partly solved in (3) where an exact functional relationship between the residual chemical potential $\mu - E(s)$ and the hard rods density $\rho(s)$ was derived.

**Nucleosome formation energy**

To compute the energy landscape associated to the formation of one nucleosome at a given position along DNA, we assumed that (i) DNA was an unshearable elastic rod whose conformation are described by the set of three local angles $\Omega_1(s)$ (tilt), $\Omega_2(s)$ (roll), $\Omega_3(s)$ (twist), and (ii) the DNA chain along the nucleosome at position $s$ was constrained to form an ideal superhelix of radius $R = 4.19$ nm and pitch $P = 2.59$ nm (4) over a total length $l$ which fixed the distribution of angular deformations $(\Omega_{i}^{\text{Nuc}}(u))_{i=1,2,3}, \ u = s, \cdots, s + l$. Within linear elasticity approximation,
the energy cost for nucleosome formation is given by:

$$\beta E(s, l) = \int_{s}^{s+l} \sum_{i=1}^{3} \frac{A_i}{2} (\Omega_{i}^{\text{nucl}}(u) - \Omega_{i}^{\text{o}}(u))^2 du,$$

where $A_1, A_2$ and $A_3$ are the stiffnesses associated to the tilt, roll and twist deformations around their intrinsic values $\Omega_{1}^{\text{o}}, \Omega_{2}^{\text{o}}$ and $\Omega_{3}^{\text{o}}$, respectively. Consistently with our previous works (1, 5–8), we used the “Pnuc” structural bending table (9) which is mainly a trinucleotide roll coding table ($\Omega_{2}^{\text{o}}$), with zero tilt ($\Omega_{1}^{\text{o}} = 0$) and constant twist ($\Omega_{3}^{\text{o}} = 2\pi/10.5$). As the values of this bending table were arbitrarily assigned between 0 and $\pi/18$ rad, we performed the following affine rescaling $\Omega_{2}^{\ast} = \gamma \Omega_{2}^{\text{o}} - \eta$ with $\gamma = 0.4, \eta = 0.06$, in order to get a comparable range of energy landscape fluctuations as in the experiments (1).

**Nucleosome density profile**

Since there are several nucleosomes, interactions between neighbouring nucleosomes have to be considered. We assumed that these interactions are dominated by steric hindrance, modeled by a hard core potential of size $l$ between 1D rods. According to (3), the equilibrium density $\rho(s)$ of hard rods in an external field $E(s, l)$ obeys the nonlinear integral equation:

$$\beta \mu = \beta E(s, l) + \ln \rho(s) - \ln \left(1 - \int_{s}^{s+l} \rho(s') ds'\right) + \int_{s-l}^{s} \frac{\rho(s')}{\rho(s') - \int_{s'}^{s+l} \rho(s'') ds''} ds'. \quad (S5)$$

This equation has an explicit solution (10) that required numerical integration. In order to assign values to the relative chemical potential $\mu - E$, the variance $\sigma^2(E)$ and $l$, we performed simulations on yeast chromosome 3 and compared them to *in vivo* nucleosome positioning data (11).

We imposed that 70% of the sequence was covered by nucleosomes and that the simulated and experimental nucleosome occupancy profiles had the same variance and autocorrelation profiles, we ended with the following parameter values:

$$\mu = -1.3 \, kT, \quad \sigma(E) = 1.95 \, kT, \quad l = 125 \, \text{bp}. \quad (S6)$$

Note that the hard rod size $l = 125$ bp is smaller than the well known length $L_c = 147$ bp of complexed DNA in the nucleosome core particle (4).
Nucleosome free regions

In order to predict the nucleosome free regions (NFR) observed in the experiment (11), we used both the energy profile $E(s)$ and the nucleosome occupancy probability profile $P(s)$ obtained by convolving the nucleosome density $\rho(s)$ with a square function $\cap$ of width $L_c = 147$ bp:

$$P(s) = \rho \ast \cap(s).$$  \hspace{1cm} (S7)

Two major features were required for a region to be qualified as free of nucleosome: (i) $P(s) < 0.35$; (ii) the energy $E(s)$ of this region is higher than the energy of the surroundings as the signature of the presence of an excluding energy barrier. This second requirement was added to prevent linkers from being labeled as NFR whereas the nucleosome depletion only stems from parking effects resulting in strong nucleosome positioning rather than true sequence impairment.

To specify the position of these excluding genomic energy barriers, we thresholded $P(s) < 0.35$ and only kept the minima of the resulting signal. Then we checked for the presence of a barrier in $E(s)$: after removing the low-frequency trends in $E(s)$ with a high-pass filter, if the energy profile was still higher than $3 \, kT$ at the position of the minima of $P(s)$, then we accepted it as a NFR. We assigned a score to the so-defined energy barriers as the mean number of nucleosomes in a $125$ bp wide window centered at the barrier position, the smaller the better. The width of a NFR was determined through the detection of the nearest inflexion points (IP) in $P(s)$. To detect these points of steep variation of $P(s)$, we investigated the correlation between $P(s)$ and a sliding first-order wavelet defined as the first derivative of a Gaussian function of width $30$ bp. Wherever the correlation coefficient was higher than $0.4$, the corresponding point was defined as an IP. The width of the NFR was then defined by the distance between the nearest left IP (steep rising of $E(s)$) and the nearest right IP (steep decrease of $E(s)$) bordering an energy barrier.

CYTOSINE METHYLATION, CpG OBSERVED/EXPECTED RATIO AND CpG ISLANDS

We took advantage of cytosine methylation data obtained by the Human Epigenome Project using bisulphite DNA sequencing (12–14) to check that the CpG observed/expected ratio (CpG o/e; See Material and Methods in the main text) can be used as an indicator of the methylation level in the germline (15). Methylation profiling of human chromosome 6, 20 and 22 was performed for 2524 amplicons ($\sim 1$ Mbp) across 12 different tissues (14). For each amplicon, we calculated the
mean and standard error of the mean (SEM) of the cytosine methylation level for each tissue (there are on average 16 CpGs per amplicon). For sperm, sufficient data were available for 1760 amplicons. As previously observed (13), the distribution of the methylation level presents a very clear bimodal shape, all the more pronounced when restricting the analysis to the 1042 amplicons with an homogeneous methylation level over all CpGs (SEM ≤ 5%): most amplicons are either hypomethylated (methylation level < 25%) or hypermethylated (methylation level > 75%) (Figure S5A). We analyzed the proportion of hypomethylated amplicons as a function of their overlap with CpG islands (CGIs) or their distance to the closest CGI (defined as smallest distance between the amplicon extremities and a CGI extremity) when they do not overlap a CGI. We recovered (16) that most CGIs are hypomethylated regions while, away from CGIs, the genome is mainly hypermethylated (Figure S5C). However, this analysis also shows that CGIs' majoritary hypomethylated state spreads out 1 kbp around the annotated CGIs (Figure S5C), so that the sequence coverage by CGIs enlarged 1 kbp at both extremities provides a marker for hypomethylated regions. Finally, we observed that hypomethylated amplicons correspond to higher values of the CpG o/e ratio, calculated over the amplicon sequence enlarged 1 kbp at both extremities and masked for CGIs, than hypermethylated amplicons (Figure S5B). This result was consistently recovered when considering only amplicons overlapping CGIs or only amplicons more than 1 kbp from the closest CGI. This comparison between CpG o/e and experimental methylation level in sperm provides further justification that CpG o/e is an adequate marker to study DNA methylation directly from the sequence.
REFERENCES

1. Vaillant, C., Audit, B., and Arneodo, A. (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.*, **99**, 218103.

2. Tonks, L. (1936) The complete equation of state of one, two and three-dimensional gases of hard elastic spheres. *Phys. Rev.*, **50**, 955–963.

3. Percus, J. K. (1976) Equilibrium state of a classical fluid of hard rods in an external-field. *J. Stat. Phys.*, **15**, 505–511.

4. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, **389**, 251–260.

5. Audit, B., Vaillant, C., Arneodo, A., d’AubentonCarafa, Y., and Thermes, C. (2002) Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J. Mol. Biol.*, **316**, 903–918.

6. Vaillant, C., Audit, B., and Arneodo, A. (2005) Thermodynamics of DNA loops with long-range correlated structural disorder. *Phys. Rev. Lett.*, **95**, 068101.

7. Vaillant, C., Audit, B., Thermes, C., and Arneodo, A. (2006) Formation and positioning of nucleosomes: effect of sequence-dependent long-range correlated structural disorder. *Eur. Phys. J. E*, **19**, 263–277.

8. Miele, V., Vaillant, C., d’Aubenton-Carafa, Y., Thermes, C., and Grange, T. (2008) DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.*, **36**, 3746–3756.

9. Goodsell, D. S. and Dickerson, R. E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.

10. Vanderlick, T. K., Scriven, L. E., and Davis, H. T. (1986) Solution of Percus equation for the density of hard-rods in an external-field. *Phys. Rev. A*, **34**, 5130–5131.

11. Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R., and Nislow, C. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, **39**, 1235–1244.

12. Lewin, J., Schmitt, A. O., Adorjan, P., Hildmann, T., and Piepenbrock, C. (2004) Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics*, **20**, 3005–3012.

13. Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Dan Andrews, T., Howe, K. L., Otto, T., Olek, A., Fischer, J., Gut, I. G., Berlin, K., and Beck, S. (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.*, **2**, e405.

14. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.

15. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.

16. Suzuki, M. M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

17. Audit, B., Nicolay, S., Huvet, M., Touchon, M., d’Aubenton-Carafa, Y., Thermes, C., and Arneodo, A. (2007) DNA replication timing data corroborate in silico human replication origin predictions. *Phys. Rev. Lett.*, **99**, 248102.

18. Huvet, M., Nicolay, S., Touchon, M., Audit, B., d’Aubenton-Carafa, Y., Arneodo, A., and Thermes, C. (2007) Human gene organization driven by the coordination of replication and transcription. *Genome Res.*, **17**, 1278–1285.
19. Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132, 887–898.

20. Lemaitre, C., Tannier, E., Gautier, C., and Sagot, M.-F. (2008) Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC Bioinformatics*, 9, 286.

21. Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E., and Audit, B. (2009) Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relations to genome organisation. Submitted to *BMC Genomics*.

22. Mitelman, F., Johansson, B., and Mertens, F., (eds.) (2008) Mitelman Database of Chromosome Aberrations in Cancer, , http://cgap.nci.nih.gov/Chromosomes/Mitelman.
Figure S1. Open chromatin markers and N-domains along human chromosomes (Continue next page).
Figure S1. Open chromatin markers and N-domains along human chromosomes (Continued). Six 19 Mbp long fragments of the human genome that contain replication N-domains (horizontal green lines) bordered by putative replication origins (vertical green lines) (17, 18). The different colored profiles correspond to the DNase I HS score (black, resolution 1 kbp), the NFR density (blue, resolution 100 kbp), the CpG o/e (red, resolution 100 kbp), the GC content (magenta, resolution 100 kbp) and the replication timing ratio $t_r$ (light blue, inhomogeneous spatial resolution $\sim 300$ kbp).
Figure S2. Nucleosome occupancy profiles and in silico Nucleosome Free Regions. Comparison between the theoretical nucleosome occupancy probability (green), theoretical NFR prediction (blue bars) and the experimental nucleosome score (19) (purple). (A) (resp. (B)) 2 kbp region of human chromosome 6 around position 77.427 Mbp close to the center of the N-domain bordered by $O_1$ and $O_2$ (resp. 76.372 Mbp close to N-domain border $O_1$). (C) (resp. (D)) Mean profiles around the 1017747 predicted NFRs in low GC content regions (≤ 41%) when aligned on their 5’ (resp. 3’) borders.
Figure S3. N-domain borders are enriched in CpG islands. Mean profiles of 1 kbp enlarged CpG island coverage over the 678 replication N-domains identified in the human genome (17, 18) as a function of the distance to the closest N-domain border. In (A), black line corresponds to the overall average; blue (resp. green) line corresponds to the average over NFR (resp. non-NFR) loci in low GC content regions (≤ 41%). In (B), color lines corresponds to the average over loci belonging to different isochores – blue line: GC< 37% (L1), green line: 37 <GC< 41% (L2) and red line GC> 41% (H1-3).
Figure S4. Genome organization around N-domain borders present a characteristic size. Mean profiles of 1-kbp-enlarged CpG island coverage (A), transcription start site (TSS) density (B), proportion of small intergenes (C, $l < 100$ kbp) and small genes (D, $l < 50$ kbp) as a function of the distance to the closest N-domain border for three N-domain size categories: $L < 0.8$ Mbp (red), $0.8 < L < 1.5$ Mbp (green) and $L > 1.5$ Mbp (blue). The proportion of small genes (resp. intergenes) corresponds to the ratio between the sequence length covered by small genes (resp. intergenes) and the total length covered by genes (resp. intergenes).
**Figure S5.** Cytosine methylation in sperm and CpG o/e ratio. (A) Histograms of the mean cytosine methylation level for the complete set of 1760 amplicons (filled black histogram) and for the 1042 amplicons having a standard error of the mean methylation (SEM) lower than 5% (grey). (B) CpG o/e of amplicons enlarged by 1 kbp on both sides versus the mean methylation level for amplicons having a SEM < 5%: (○) all amplicons; (▽) amplicons overlapping a CGI; (△) amplicons with a distance to the closest CGI greater than 1 kbp when masking nucleotides with a distance to the closest CGI smaller than 1 kbp. (C) Proportion of amplicons having a mean methylation level < 25% versus their coverage by CGIs when they overlap a CGI (left) or their distance to the closest CGI (right). In (B,C), data points were obtained by (i) sorting amplicons according to their mean methylation level, CGI coverage or distance to closest CGI, (ii) grouping them into classes following that order and (iii) computing CpG o/e ratio or the proportion of low methylated amplicons and average methylation level, CGI coverage or distance to closest CGI over each class; vertical bars represent the standard error; horizontal bars represent the ranges of methylation level, CGI coverage or distance to closest CGI over each class.
Figure S6. N-domain borders are potential regions of chromosome instability. Localisation of breakpoint regions (BPRs) (20, 21) and cancer gene RUNX1T1 (22) along the human genome in relation to replication N-domains. BPRs are shown in light blue: vertical arrows mark BPR midpoints, boxes mark BPR extents. RUNX1T1 transcription start site is marked by thick dark blue arrows. For the remaining color code signification, see Figure S1 and Figure 1A in main text.