Gene 419 (2008) 62–65

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Mapping DNase-I hypersensitive sites on human isochores

Miriam Di Filippo, Giorgio Bernardi *

Laboratory of Molecular Evolution, Stazione Zoologica Anton Dohrn, 80121 Naples, Italy

ARTICLE INFO

Article history:
Received 20 December 2007
Received in revised form 4 February 2008
Accepted 5 February 2008
Available online 21 February 2008

Received by T. Gojobori

Keywords:
DNase-I
Genomes
Hypersensitive sites
Isochores

ABSTRACT

Mapping DNase-I hypersensitive sites (HS) was used in the past to identify regulatory elements of specific genes. More recently, thousands of HS were identified in the human genome by using high-throughput methods. These approaches showed a general enrichment of HS near or within known genes, within CpG islands, within human–mouse conserved regions and in GC-rich regions of the genome. Here we show that HS: (i) are characterized by a much higher GC level (~56%) than the average GC level of the human genome (~41%); (ii) are overwhelmingly located in the GC-richest compartment of the genome, which is predominantly associated with an open chromatin structure; (iii) and are slightly more and slightly less frequent than genes, respectively, in the gene-rich and in the gene-poor isochore families.

1. Introduction

Eukaryotic DNA is packaged into a nucleoprotein complex known as chromatin, and this packaging has major functional consequences for most processes that involve DNA. A large number of investigations have analyzed the chromatin structure of individual genes (see, for example, Wu et al., 1979; Wu, 1980; McGhee et al., 1981; Gross and Garrard, 1988; Elgin, 1988; Boyes and Felsenfeld, 1996). Recently genome-wide analysis has been used employing different approaches, including mapping of DNase-I hypersensitive sites, HS (Crawford et al., 2004, 2006a,b; Sabo et al., 2004, 2006). The sites are so called because of their extreme sensitivity to enzymatic digestion by DNase-I, compared to the surrounding sequences. The sites are sensitive to enzymatic attack because they are poorly, if at all, protected by histones and are thus accessible to DNA-binding regulatory proteins (Gross and Garrard, 1988; Wolfe, 1993).

Over the past twenty years, many different regulatory elements, such as promotors, enhancers, suppressors, insulators and locus control regions have been shown to be associated with HS when active (Wu et al., 1979; Wu, 1980; McGhee et al., 1981; Szabo et al., 1987). Mapping of HS has been used to identify the precise location of these elements in specific genes. Until a few years ago, HS mapping was performed by Southern blotting and was generally limited to small regions of the genome. In this way, hundreds of HS associated with specific loci were described. Very recently, with the availability of complete genome sequences, high-throughput experimental methods were developed, and it became possible to map thousands of HS in human cells in a single study. Indeed, very recently, a genome-wide mapping of HS was performed by massive parallel signature sequencing (MPSS; Crawford et al., 2006a) which identified an estimated 20% of all HS in human CD4+T cells. HS mapping by MPSS was immediately followed (Crawford et al., 2006b) by a DNase-chip mapping, a higher-resolution method, which was used to identify HS by hybridizing captured DNase-digested ends to tiled microarrays. The DNase-chip was used to accurately identify HS within 1% human genome as selected by ENCODE (ENCyclopedia Of DNA Elements) Consortium (The Encode Consortium, 2004) from CD4+T cells and cycling B lymphoblastoid cell line, a primary and an immortalized cell type, respectively. Thirty of the ENCODE regions consist of randomly selected 500-kb segments stratified by different levels of gene density and sequence conservation, containing, therefore, an overall representation of the entire genome (The Encode Consortium, 2004; see, however, the preceding paper by Costantini et al., 2008).

These recent analyses showed a general enrichment of HS near or within known genes, within CpG islands, and within regions of human–mouse conservation. An enrichment of HS was also reported (Crawford et al., 2004) in regions of the genome with high GC levels, and it was suggested that this preference was due to the presence of CpG islands.

The high GC level of the regions enriched in HS goes, however, well beyond the increased presence of CpG islands. Indeed, differences in GC level partition vertebrate genomes into well-defined compartments. It is well established that the human genome is a mosaic of isochores that not only show different GC levels but also have many correlated

Abbreviations: HS; hypersensitive sites; MPSS; massive parallel signature sequencing; ENCODE; Encyclopedia of DNA Elements.

* Coresponding author. Tel.: +39 081 5833402; fax: +39 081 2455807.
E-mail addresses: miriam@szn.it (M. Di Filippo), bernardi@szn.it (G. Bernardi).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.02.012

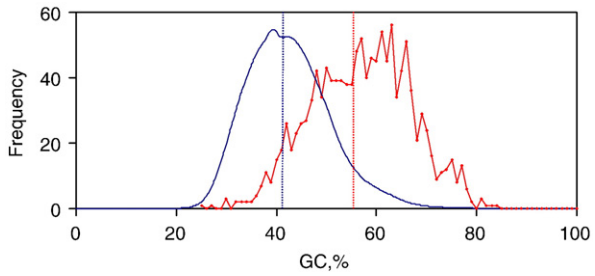

Fig. 1. Compositional distribution of human DNA and of HS. The GC level of the sequences (average length ~500 bp) corresponding to HS (red line), downloaded from http://research.nhgri.nih.gov/DNaseHS/chip_2006, was superimposed on the GC profile of the human genome (in blue), as calculated at a window size of 500 bp. The superposition shows a striking GC enrichment for the sequences containing HS. The average GC level of the sequences analyzed was found to be ~56% (red dashed line), whereas that of the human genome was ~41% (blue line).

structural, functional and evolutionary features (for recent reviews see Bernardi, 2004, 2007) and represent therefore a fundamental level of genome organization (Eyre-Walker and Hurst, 2001). Five families of isochores have been identified in the human genome (Costantini et al., 2006): the GC-poor L1 (GC <37%), L2 (GC 37–41%) and H1 families (GC 41–46%), and the GC-rich H2 (GC 46–53%) and H3 families (GC >53%).

The strikingly non-uniform gene distribution of the human genome, with gene density increasing with GC level (Bernardi et al., 1985; Mouchiroud et al., 1991; Zoubak and Bernardi, 1996) indicated the existence of two "gene spaces": the gene-rich "genome core" (comprising isochore families H2 and H3) and the gene-poor "genome desert" (comprising isochore families L1, L2 and H1). These two genomic compartments show differences not only in gene density, but also in other properties. Indeed, the "genome core" is characterized by an early replication timing, a higher level of recombination and gene expression, short introns and UTRs (untranslated sequences), higher GC heterogeneity, low methylation level, whereas the "genome desert" is endowed with opposite properties. Very importantly, the chromatin organization of the genome is known to be related to the GC level of isochores. In fact, when the GC-richest and the GC-poorest fractions of the genome were hybridized on interphase nuclei, they were found to be distributed in the centre and in the periphery of the nucleus, respectively, and to be characterized by different chromatin conformations (Saccone et al., 2002; Federico et al., 2006), the GC-richest fractions being endowed with a more open, relaxed chromatin, the GC-poorest fractions with a closed, compact chromatin.

Starting with these observations, we investigated the DNase-I accessibility of chromatin as related to the compositional pattern of the human genome.

2. Materials and methods

The sequences corresponding to the genomic coordinates on hg17 (International Human Genome Sequencing Consortium, 2004) for human CD4+T, GM 06990, HepG2 and HeLaS3 cells HS (HS described in Crawford et al., 2006a,b; and in Sabo et al., 2006) were downloaded from http://research.nhgri.nih.gov/DNaseHS/chip_2006 and from http://genome.ucsc.edu/ENCODE/. The sequences of the HS flanking regions were downloaded from http://genome.ucsc.edu/ (UCSC hg17). The average GC level of the sequences was calculated using a script (available upon request) implemented by us.

The coordinates of human isochores and of Human ENCODE targets, taken from Costantini et al. (2006) and from the preceding paper Costantini et al. 2008 respectively, were used to assign HS to isochores. The density of HS in the isochore families was calculated by dividing the number of HS in each family by the total size of the isochore family. The gene density of a set of 24,346 human genes (retrieved from GeneBank) is that reported in Bernardi (2007) and Costantini et al. (2007).

3. Results

The publicly available genomic coordinates on the finished human genome assembly (UCSC Release hg17; International Human Genome Sequencing Consortium, 2004) of all human HS from the CD4+T cells, described in Crawford et al. (2006b) were downloaded from http://research.nhgri.nih.gov/DNaseHS/chip_2006. We averaged the replicates over the three DNase concentrations used in the DNase-chip assay. The downloaded sequences showed an average length of 500 bp whose average GC level was calculated. The observed value, 56%, is much higher than the average GC level of the human genome, 41%, as calculated from the finished sequence of human genome, hg17 (see Fig. 1).

Starting from this result, the base composition of more extended regions of DNA in which HS were located was analyzed. The flanking genomic regions from http://genome.ucsc.edu/ at the window sizes of 1, 5 and 10 kb were downloaded and their GC levels were calculated. A positive correlation was found between the GC level of HS and that of the flanking regions (R_{1kb}=0.67; R_{5kb}=0.55; R_{10kb}=0.54), the average GC level of which was found to correspond to GC%$_{1kb}$=51.6, GC%$_{5kb}$=49.8, GC%$_{10kb}$=49.3, respectively (see Fig. 2). Therefore, these HS are not only GC-rich but they are also predominantly located in GC-rich regions.

A question could be raised about the contribution of CpG islands to the high GC level of the regions enriched in DNase HS and whether that contribution be mainly responsible for this enrichment. In order to answer this question, we eliminated the CpG islands from the dataset. In so doing we found that the average GC level of the HS corresponds to 52.5% (Supplementary Fig. 1). This value is expectedly lower than 56.4%, which was found before, but yet significantly higher than the average GC level of the total human genome.

Moreover, we calculated the density of HS in the isochore families. Because of the data obtained by chip assay are calculated only for the regions selected by ENCODE, it is necessary to extrapolate these data to whole-genome data. This extrapolation was performed according to Costantini et al. (2008; see preceding paper). In this extrapolation a

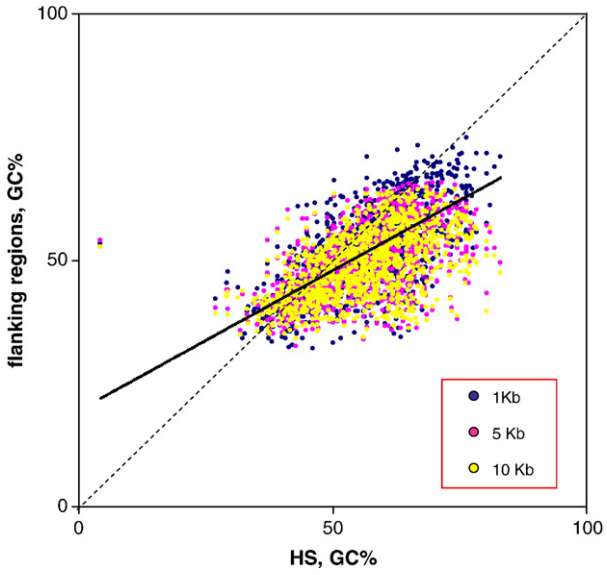

Fig. 2. Compositional correlation with flanking regions. Correlations between GC levels of HS and of 1, 5 and 10 kb flanking regions (see also text). A strong correlation between GC level of HS and GC level of the flanking regions was found at the window size of 1, 5 and 10 kb (GC$_{1kb}$=51.6, GC$_{5kb}$=49.8, GC$_{10kb}$=49.3; R_{1kb}=0.67, shown in the figure, R_{5kb}=0.55, R_{10kb}=0.54), respectively.

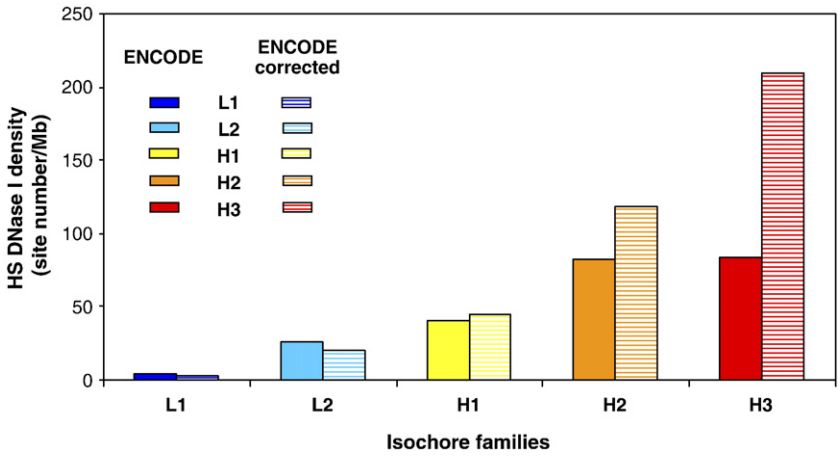

Fig. 3. HS density in human isochore families. The histogram shows the DNase-I hypersensitive sites (downloaded from http://research.nhgri.nih.gov/DNaseHS/chip_2006/) density (site number/megabases) as calculated from ENCODE targets (left-hand set of bars) and as corrected for the whole genome (right-hand set of bars). The correction was performed by multiplying the size in Mb of each isochore family by the G/E (whole genome/ENCODE) ratio (see Table 1) and then the site density was recalculated.

conversion factor was calculated that allows, for each size value (Mb) of the isochore families constructed in the ENCODE DNA targets, to obtain the corresponding value for the whole genome.

We found that the vast majority of the human HS belong to the GC-richest isochore families H2 and H3 (see Fig. 3). We performed the same analysis, also considering the data without the contribution of the CpG islands (Supplementary Fig. 2). In this case, the HS density increases with the increasing of GC level, even if in H3 it shows some decreases.

Table 1 compares the densities of HS and of genes (see Bernardi, 2007) in isochore families. This clearly shows more HS densities compared to gene densities for GC-rich isochores, whereas the reverse is true for GC-poor isochores.

We extended this analysis also to other cell types for which data are publicly available on www.genome.ucsc.edu/ENCODE, such as GM 06990, HepG2 and HeLaS3, analyzed by DNase-chip. We analyzed also the data obtained by MPSS assay on CD4+T cells, and by DNase/array on GM 06990 cell, in this case using the coordinates of human isochores (Costantini et al., 2006). In all cases we found an enrichment of HS in the isochore family H3 (see Supplementary Fig. 3). In agreement with Crawford et al. (2006a) we noted some small differences in the amount of sites for each isochore family in different tissues and for the same tissue using different approaches (see Table 2), but the general pattern of HS density in the isochores is conserved among the data obtained from different tissue and approaches.

4. Discussion

Vertebrates display a mosaic organization of the genome in which the isochores are the structural units. In fact, isochores are not only involved in the structure, but also in the function and evolution of the vertebrate genomes. Several important genomic parameters, such as gene density, chromatin structure, replication timing, gene expression and methylation level were found to be highly related to isochores (for recent reviews see Bernardi, 2004, 2007). Interestingly, an association of GC-rich genes and expression level was previously reported (Arhondakis et al., 2004, 2006) as well as a compositional preference in the accessibility of the genomes to retroviral sequences and in their expression (Bernardi et al., 1985; Mouchiroud et al., 1991; Saccone et al., 2002; Federico et al., 2006; Rynditch et al., 1998; Zoubak et al., 1994; Tsyba et al., 2004; Holman and Coffin, 1992; Muller et al., 1993; Pryciak and Varmus, 1992; Schroeder et al., 2002; Elleder et al., 2002; see for review Bernardi, 2004;). Here we tried to relate chromatin accessibility, as assayed by DNase digestion, with genome structure, as revealed by its compositional pattern.

In our analysis, we found a striking compositional preference in the accessibility of the chromatin to DNase-I. In fact HS showed a much

Table 1
Densities of HS and genes (Bernardi, 2007) in human isochore families

	HS density, %	Gene density, %	HS density/Gene density
L1	0.72	2.0	0.36
L2	5.0	7.2	0.69
H1	10.8	13.5	0.8
H2	30.2	29.1	1.03
H3	53.2	48.2	1.10

The density of the HS sequences in the isochore families, calculated by dividing the number of the sequences by the total size of the isochore family in ENCODE targets, corrected by conversion factor to which they belong, was compared to the gene density (Bernardi, 2007) in the isochore families. The HS values are clearly lower than the gene values for the GC-poor isochores, whereas the contrary was found for the GC-rich isochores.

Table 2
Cell types, total size (in Mb) of isochores in the ENCODE targets (corrected as in Costantini et al., 2008, preceding paper) and in whole genome, number of the HS regions identified in each family, and density (number/Mb) are reported

	L1	L2	H1	H2	H3
(ENCODE) Mb:	5.8	11.4	9.4	3.4	0.9
(Whole genome) Mb:	613.5	1040.1	781.2	352.2	67.2
NHGRI CD4+T (Dnase_Chip)					
Region number	18	226	420	401	197
Density	3.1	19.8	44.7	118	218.8
NHGRI_MPSS_Cd4+T					
Region number	4	14	86	70	55
Density	0.006 [a]	0.01 [a]	0.11 [a]	0.2 [a]	0.82 [a]
NHGRI GM 06990 (Dnase_Chip)					
Region number	36	241	488	367	222
Density	6.2	18.7	51.9	108	246.6
Regulome GM 06990 (Chip)					
Region number	277	817	984	454	235
Density	47.65	71.6	104.6	133.5	261.1
NHGRI HeLaS3 (Dnase_Chip)					
Region number	62	240	352	245	143
Density	10.7	21.1	37.5	272.2	158.8
NHGRI HepG2 (Dnase_Chip)					
Region number	213	502	612	338	220
Density	36.7	44.0	65.1	99.4	244

[a] Region number/(whole genome) Mb.

higher GC level than the average GC level of the total human genome and almost all HS belonged to the isochore family H3, namely the GC-richest part of the genome. The density ratio HS/gene ranged from 0.36 to 1.1, a 3 fold range, from isochore family L1 to isochore family H3. This means that, by far, not all genes in the "genome desert" are associated with HS, whereas in the "genome core", there is a slight excess of HS relative to genes. In connection with this conclusion, it should be mentioned that several authors (Crawford et al., 2006a,b; Sabo et al., 2006) found a large population of cleavage sites far away from genes, and suggested that these clusters of accessibility may reflect a regular feature of chromatin, embedded in large domains of activity or repression occurring on a scale of hundreds of kilobases. Moreover, they found that HS tend to be enriched in regions of the genome with high GC levels. This was suggested to be due to the presence of CpG islands in such regions, but this cannot be the only reason, because CpG island density mimics gene density (Jabbari and Bernardi, 1998).

The compositional preference in the accessibility of the chromatin to DNase-I, suggested by the high GC-rich level showed by HS regions, is an evidence of the existing relationship between chromatin structure and genome organization. The different structure which chromatin can assume in the nucleus, on the basis of the GC level of the genome, has also important implication in term of evolution. In fact, a discussion about the evolution of mammalian and avian genomes, and the compositional compartmentalization of the chromatin has been discussed by Bernardi (2007).

Acknowledgments

We thank Fabio Auletta for his help in computer work, Maria Costantini, Gabriele Amore and Oliver Clay for helpful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.02.012.

References

Arhondakis, S., Auletta, F., Torelli, G., D'Onofrio, G., 2004. Base composition and expression level of human genes. Gene 325, 165–169.
Arhondakis, S., Clay, O., Bernardi, G., 2006. Compositional properties of human cDNA libraries: practical implications. FEBS Lett. 580 (24), 5772–5778.
Bernardi, G., 2004. Structural and Evolutionary Genomics, Natural Selection in Genome Evolution. reprinted in 2005 Elsevier, Amsterdam.
Bernardi, G., 2007. The neoselectionist theory of genome evolution. Proc. Natl. Acad. Sci. USA 104, 8385–8390.
Bernardi, G., et al., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.
Boyes, J., Felsenfeld, G., 1996. Tissue-specific factors additively increase the probability of the all-or-none formation of a hypersensitive site. EMBO J. 15 (10), 2496–2507.
Costantini, M., Clay, O., Auletta, F., Bernardi, G., 2006. An isochore map of human chromosomes. Genome Res. 16, 536–541.
Costantini, M., Di Filippo, M., Auletta, F., Bernardi, G., 2007. Isochore pattern and gene distribution in the chicken genome. Gene 400 (1-2), 9–15.
Costantini, M., Di Filippo, M., Bernardi, G., 2008. Extrapolating ENCODE data to the whole human genome. Gene 419, 66–69. doi:10.1016/j.gene.2008.02.013.
Crawford, G.E., et al., 2004. National Institute of Health Intramural Sequencing Center, Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. U S A 101, 992–997.
Crawford, G.E., et al., 2006a. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 16, 123–131.
Crawford, G.E., et al., 2006b. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods 3 (7), 503–509.
Gross, D.S., Garrard, W.T., 1988. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159–197.
Elgin, S.C., 1988. The formation and function of DNase I hypersensitive sites in the process of gene activation. J. Biol. Chem. 263, 19259–19262.
Elleder, D., Pavlicek, A., Paces, J., Hejnar, J., 2002. Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence. FEBS Lett. 517, 285–286.
Eyre-Walker, A., Hurst, L.D., 2001. The evolution of isochores. Nat. Rev. Genet. 2 (7), 549–555.
Federico, C., Scavo, C., Cantarella, C.D., Motta, S., Saccone, S., Bernardi, G., 2006. Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. Chromosoma 115 (2), 123–128.
Holman, A.G., Coffin, J.M., 1992. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leucosis virus, and murine leukemia virus integration sites. Proc. Natl. Acad. Sci. U S A 102, 6103–6107.
International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431 (7011), 931–945.
Jabbari, K., Bernardi, G., 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene 224, 123–128.
McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., Felsenfeld, G., 1981. A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion. Cell 27 (1 Pt 2), 45–55.
Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G., 1991. The distribution of genes in the human genome. Gene 100, 181–187.
Muller, H.P., Pryciak, P.M., Varmus, H.E., 1993. Retroviral integration machinery as a probe for DNA structure and associated proteins. Cold Spring Harb. Symp. Quant Biol. 58, 533–541.
Pryciak, P.M., Varmus, H.E., 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell 69, 769–780.
Rynditch, A.V., Zoubak, S., Tsyba, L., Tryapitsina-Guley, N., Bernardi, G., 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. Gene 222 (1), 1–16.
Sabo, P.J., et al., 2004. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. Proc. Natl. Acad. Sci. U. S. A. 101 (13), 4537–4542.
Sabo, P.J., et al., 2006. Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. Nat. methods 3 (7), 511–514.
Saccone, S., Federico, C., Bernardi, G., 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 300 (1-2), 169–178.
Schroeder, A.R., Shinn, H., Chen, C.B., Ecker, J.R., Bushman, F., 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110, 521–529.
Szabo, G.J., Damjanovich, S., Sumegi, J., Klein, G., 1987. Overall changes in chromatin sensitivity to DNase I during differentiation. Exp. Cell. Res. 169 (1), 158–168.
The Encode Consortium, 2004. The ENCODE (ENcyclopedia Of DNA Elements) Project. Science 306, 636–640.
Tsyba, L., Rynditch, A., Boeri, E., Jabbari, K., Bernardi, G., 2004. Distribution of HIV-I in the genomes of infected individuals: localization in GC-poor isochores and high viremia correlated. Cell. Mol. Life Sci. 61 (6), 721–726.
Wolfe, S.L., 1993. Molecular and Cellular Biology. Wadsworth Publishing Company, Belmont, California.
Wu, C., 1980. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286, 854–860.
Wu, C., Wong, Y.C., Elgin, S.C., 1979. The chromatin structure of specific genes, II: Disruption of chromatin structure during gene activity. Cell 16, 807–814.
Zoubak, S., Bernardi, G., 1996. The gene distribution of the human genome. Gene 174, 293–307.
Zoubak, S., Rynditch, A., Bernardi, G., 1994. Regional specificity of HTLV-I proviral integration in the human genome. Gene 143, 155–163.