

THE RELATIONSHIP BETWEEN FINE SCALE DNA STRUCTURE, GC CONTENT, AND FUNCTIONAL ELEMENTS IN 1% OF THE HUMAN GENOME

STEPHEN C. J. PARKER¹
parker@bu.edu

ELLIOTT H. MARGULIES²
elliott@nhgri.nih.gov

THOMAS D. TULLIUS^{1,3}
tullius@bu.edu

¹ *Graduate Program in Bioinformatics, Boston University, Boston MA 02215, U.S.A.*

² *National Human Genome Research Institute, National Institutes of Health, Bethesda MD 20892, U.S.A.*

³ *Department of Chemistry, Boston University, Boston MA 02215, U.S.A.*

GC content has been shown to be an important aspect of human genomic function. Extending beyond the scope of GC content alone, there is a class of regions in the genome that have especially high GC content and are enriched for the CG dinucleotide—called CpG islands. CpG islands have been linked to biologically functional genomic elements. DNA structure also contributes to biological function. Recent studies found that some DNA structural properties are correlated with CpG island functionality [5, 14]. Here, we use hydroxyl radical cleavage patterns as a measure of DNA structure, to explore the relationship between GC content and fine-scale DNA structure. We show that there is a positive correlation between GC content and the solvent-accessible structural properties of a DNA sequence, and that the strength of this correlation decreases as genomic resolution increases. We demonstrate that regions of the genome that have highly solvent-accessible DNA structure tend to overlap functional genomic elements. Our results suggest that fine-scale DNA structural properties that are encoded in the genome are important for biological function, and that the highly solvent-accessible nature of high GC content regions and some CpG islands may account for some of their functional properties.

Keywords: DNA structure; GC content; CpG islands; hydroxyl radical cleavage; functional element; human genome

1. Introduction

GC content—the fraction of G or C nucleotides within a given window—is variable across the human genome [17, 36]. This observed heterogeneity in sequence composition has been implicated as a marker for some functional genomic regions. One example of this is CpG islands, which are regions of the genome characterized by high GC content and enrichment of the CG dinucleotide [11]. CpG islands have been linked to many regulatory processes [7, 18, 24, 33, 37-39].

Beyond the primary order of nucleotides in a genome that is used to define GC content and CpG islands, the local structural profile of DNA has been implicated in a number of biological processes. Recent studies suggest that DNA structure is important for some of the same processes as CpG islands: namely DNA-protein interactions [20], promoter function [1, 29], epigenetically controlled gene regulation [4, 23, 32, 34, 40],

and DNase I hypersensitivity [14]. However, the precise relationship between GC content, fine-scale DNA structure, and genome function remains unclear.

A critical first step in assessing this relationship is the ability to predict the local DNA structural profile for genomic sequences. Hydroxyl radical cleavage patterns of DNA have been used to study structural properties for a wide variety of sequences [13, 19, 30]. The cleavage pattern of naked DNA is a reflection of an important structural parameter, the solvent-accessible surface area of the DNA backbone [2]. The cleavage pattern thus provides a high-resolution quantitative measure of the shape of the DNA backbone and how it varies with respect to its sequence. We have recently shown that using a database of experimentally-determined hydroxyl radical cleavage patterns, the cleavage pattern of any DNA sequence can be predicted with a high degree of accuracy [13].

Although GC content has recently been implicated in defining hydroxyl radical cleavage patterns of DNA [35], this analysis was conducted at a relatively low genomic resolution of 333 base pairs. Single-nucleotide, genome-scale DNA structure predictions are feasible [13], which makes exploring the relationship between GC content and fine-scale DNA structure possible. Since different DNA sequences can have similar local structural properties [10, 13], directly correlating GC content with DNA structure is an important experiment. Results from the ENCODE Pilot Project provide a rich resource for functional annotations in 1% of the human genome [3]. These developments facilitate the investigation of the relationship between GC content, DNA structure, and functional elements in this 1% of the human genome.

Here, we compare GC content to DNA structure (measured as hydroxyl radical cleavage patterns) at various genomic resolutions, with an emphasis on fine-scale DNA structure. We then measure the occurrence of significantly over-represented DNA structural motifs with known functional annotations. Our results show that GC content only weakly influences fine-scale DNA structure, and that local structural properties may be important in conferring biological functionality to genomic regions like CpG islands.

2. Materials and Methods

2.1. DNA sequence and functional annotation data sources

The DNA sequence for NCBI build 36 (March 2006), hg18 version of the ENCODE regions within the human genome was downloaded from the UCSC genome browser (<http://genome.ucsc.edu/ENCODE/>) [21, 22].

We used the following functional annotations for comparisons with DNA sequence and structural features. All the annotations are available through the UCSC genome browser (see above), unless otherwise noted. For all analyses, the hg18 version of each annotation track was used.

- DNase I hypersensitive sites (DHSs) represent regions of open chromatin architecture where protein-DNA interactions occur. We used a Union set of DHSs derived from the human GM06990 cell line, as described in [3, 14].
- Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE) is an alternative method used to locate regions of open chromatin. FAIRE sites are enriched for regulatory elements [12].
- Promoters were defined as the region 2.5 kilobases upstream from gene start sites. We used the GENCODE [16] gene track to define genes.
- Ancestral Repeats (ARs) are mobile elements that inserted before the common ancestor of most mammals. They are thought to be neutrally evolving and are therefore typically used to represent nonfunctional regions of the human genome [9, 15, 28, 31, 41]. We used the AR regions defined in [3].
- CpG islands are regions of the human genome with high GC content and higher-than-expected CG dinucleotide density. We used the CpG islands track from the UCSC genome browser, which was constructed using the CpG island definition described in [11].
- Evolutionarily constrained regions are areas of the human genome that are under purifying selection against nucleotide changes. We used the ‘moderate track’ – which is a summary of regions identified by multiple sequence alignment and constraint detection algorithms—described in [3, 25] for this analysis.
- Transcription start sites used here are described in [3, 8].
- As a control, we constructed a ‘random annotation’ by randomly selecting 500 base pair intervals within the ENCODE regions. We repeated this process 1000 times to create the random annotation track used here. Since this annotation set was derived randomly, there should be no association with any given set of functional elements.

2.2. *Local DNA structure prediction and GC content analysis*

We used predicted hydroxyl radical cleavage patterns as a measure of local DNA structure. Hydroxyl radical cleavage patterns were predicted using the Sliding Tetramer Window algorithm described in [13] for all the ENCODE regions. After the cleavage intensity at each base was predicted, we averaged the cleavage values within a window for all possible windows within the ENCODE regions.

For GC content analysis we calculated the fraction of G or C bases within all possible windows of various sizes within the ENCODE regions. To calculate CpG density we counted the observed number of CG dinucleotides within the same windows.

2.3. Annotation proximity and overlap statistics

To calculate the proximity of various windows to functional annotations we computed the distance, in base pairs, from the closest base in a given window to the closest base from the nearest element in the specified annotation.

To calculate the observed overlap statistics between different annotations, for example—comparing the regions in annotation X to the regions in annotation Y , we first computed the fraction of regions in annotation X that overlap any region from annotation Y . We then constructed a null distribution of the fraction of expected overlaps by using the block bootstrap method described in [3]. We calculated the mean and standard deviation from the null distribution to assess the statistical significance of the observed overlap. This allowed us to determine if the regions in annotation X overlap the regions in annotation Y significantly more or less than random expectation.

3. Results

3.1. Correlation between GC content and local DNA structure

Given the data reported in [35] that shows a high correlation between GC content and mean hydroxyl radical cleavage patterns at a window size of 333 base pairs, we first sought to reproduce and supplement these results. We computed the Pearson correlation between GC content and mean hydroxyl radical cleavage for windows of size N , where $N = \{2, 3, 4, 5, 10, 20, 50, 100, 333, 500, 1000, 10000\}$, in the ENCODE regions. We observe a positive correlation between the size of a window and the strength of the correlation between GC content and hydroxyl radical cleavage (Figure 1A). That is, while large windows have a high correlation between GC content and mean hydroxyl radical cleavage, small windows—which are a reflection of the fine-scale structure of DNA—do not.

To determine if the above result is unique to the DNA in the ENCODE regions we randomized all of the ENCODE sequences. We used a first order Markov model trained on the real ENCODE sequences to preserve all dinucleotide frequencies. The random sequences follow the same correlation trend as the real ENCODE sequences (data not shown), which suggests that the observed correlations are an inherent property of DNA and not an artifact of the ENCODE sequences chosen for this analysis.

We next focused on the relationship between CpG density and mean hydroxyl radical cleavage over windows of size N (Figure 1B). For equivalent values of N , the strength of the correlation between DNA structure and CpG density is less than for GC content (compare Figure 1B to Figure 1A).

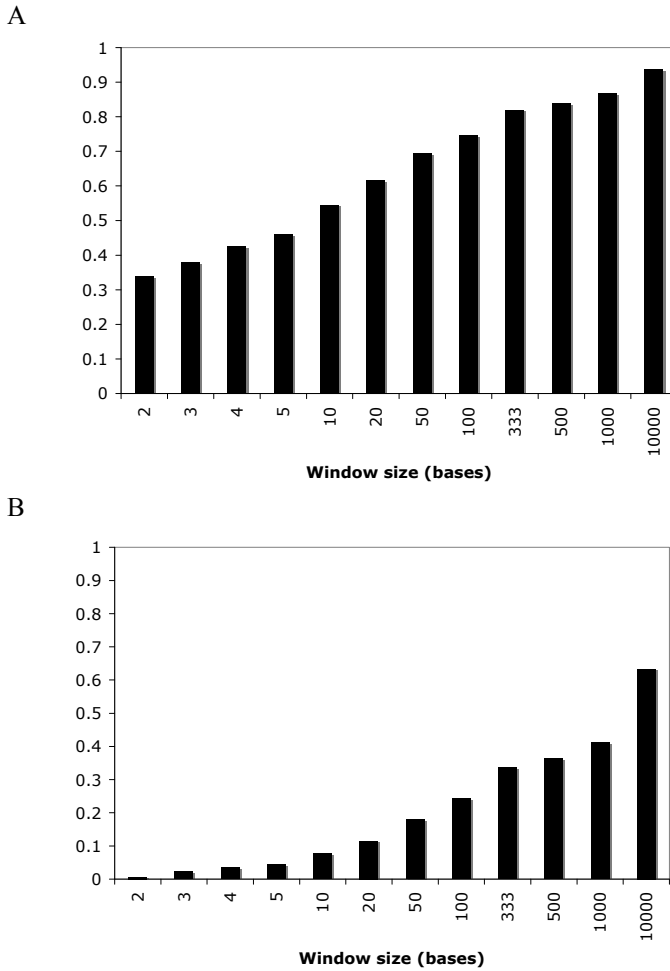


Fig. 1. Pearson correlation coefficient (r^2) between GC content or CpG island density and hydroxyl radical cleavage at various genomic scales. **A.** Correlation between GC content and predicted hydroxyl radical cleavage. **B.** Correlation between CpG density and predicted hydroxyl radical cleavage. Note that all r -values are positive.

The above results demonstrate that the correlation between GC content or CpG density and DNA structure is variable depending on genomic scale. Importantly, when fine-scale DNA structure is considered the correlation with GC content or CpG density is low. Therefore nucleotide composition is not a good predictor of fine-scale DNA structural properties. To further demonstrate this point we focus the rest of this manuscript on an in-depth analysis of $N = 10$. We specifically select this scale because it represents about one turn of the DNA double helix and is the approximate size of a

transcription factor binding site [27], which should allow for biologically relevant interpretations of the results.

Looking at the entire distribution of GC content and mean hydroxyl radical cleavage for a window of size 10 (Figure 2) clearly demonstrates that GC content is not a good predictor of fine-scale DNA structure. For example, it is possible to have windows with 0% GC content and higher mean hydroxyl radical cleavage than some windows with 100% GC content (Figure 2A). Examples of cleavage profiles for windows with different GC content are shown in Figure 2B-C.

3.2. High hydroxyl radical cleavage regions overlap with functional elements

We focused in on the highest and lowest mean cleavage intensity 10 base windows. To do this, we calculated Z-scores for all windows using the observed mean cleavage intensity over the window and the mean and standard deviation for all windows in the random sequence distribution we constructed (as described in section 3.1). We used a Z-score threshold of $|Z| = 3.09$, which is equivalent to a p-value of $p = 0.001$, to extract windows with the highest and lowest mean cleavage values. This process resulted in 43,096 high cleavage windows and 306,089 low cleavage windows. Overlapping windows for each set were merged so that disjoint and contiguous genomic regions are present in the two resulting annotation sets. This merging process resulted in 14,914 high cleavage regions and 57,307 low cleavage regions.

To determine if the resulting high and low hydroxyl radical cleavage regions occur near biologically active areas of the genome we measured their proximity to annotated transcription start sites (Figure 3). We observe that the low cleavage regions cluster near transcription start sites, and the high cleavage regions do so to a more pronounced extent, suggesting that these particular DNA structural features may be associated with some aspect of gene regulation.

To further examine the possibility that low or high cleavage regions are associated with biological function, we employed a more rigorous statistical test. We used the block bootstrap method [3] to measure the statistical confidence associated with how often low or high cleavage regions overlap a number of functional annotations (see section 2.3 for an overview of this method and section 2.1 for an explanation of each annotation). Figure 4 shows the results of this analysis. The observed overlap of low and high cleavage regions with a random annotation (see methods) is the same as random expectation. Low and high cleavage regions overlap ancestral repeats significantly less than random. The fraction of high cleavage regions that overlap with promoters and FAIRE sites is statistically significant ($p < 0.05$). This result suggests that high cleavage regions may have an association with functional regions for the genome. The observation that 72% of CpG islands overlap high cleavage regions is highly statistically significant ($p < 10^{-27}$).

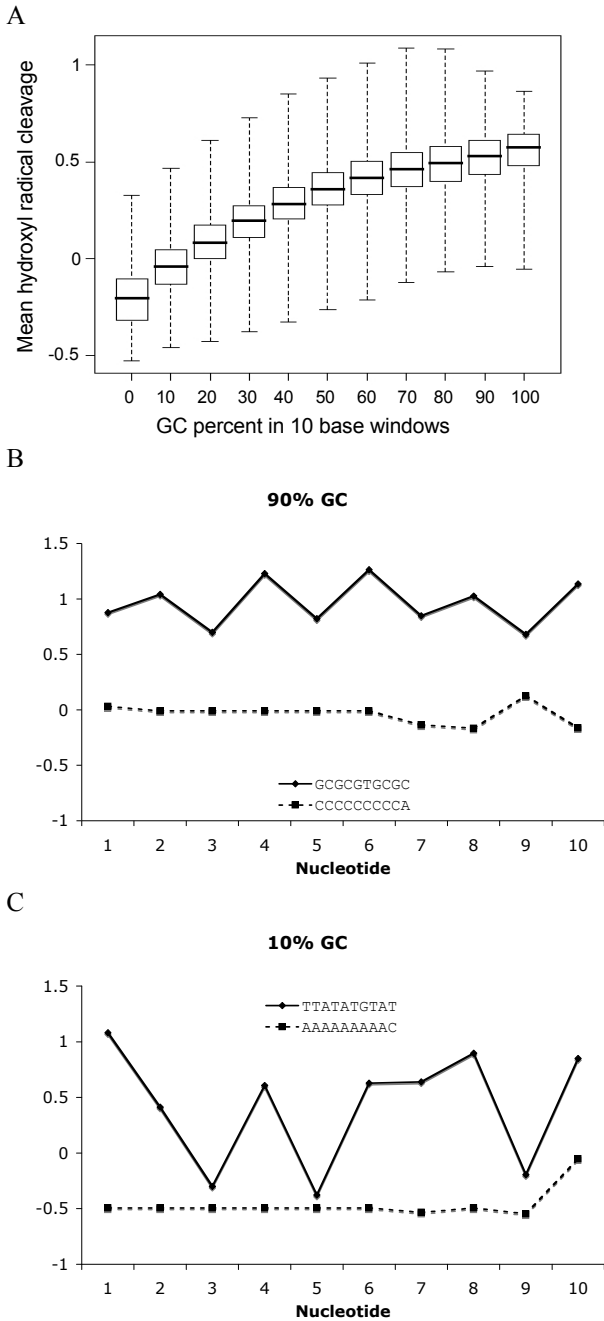


Fig. 2. Hydroxyl radical cleavage in 10 base windows binned by GC content. A. The correlation between GC percent and hydroxyl radical cleavage is not perfect. Windows with high GC content can have low cleavage, and windows with low GC content can have high cleavage. Example hydroxyl radical cleavage profiles for 10 base windows with high (B) and low (C) GC content.

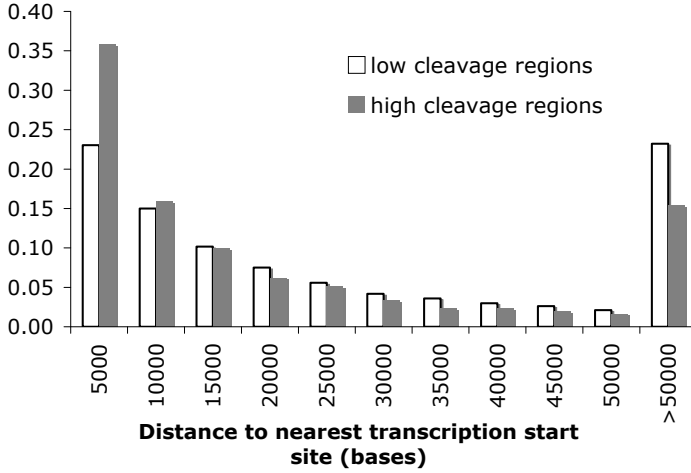


Fig. 3. Proximity analysis of high and low cleavage regions relative to annotated transcription start sites. High and low cleavage regions tend to occur near annotated transcription start sites, and this effect is more pronounced with high cleavage regions.

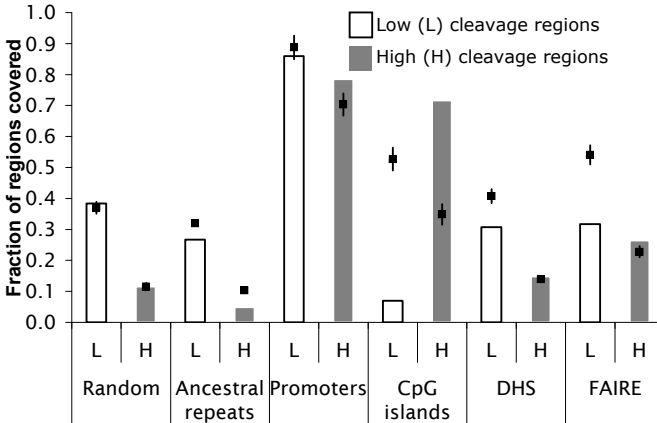


Fig. 4. Low and high cleavage region overlaps with functional annotations. Black points represent the mean of a null distribution estimated using the block-bootstrap method (see section 2.3 for a summary) and error bars represent +/- one standard deviation. DHS = DNase I Hypersensitive Sites; FAIRE = Formaldehyde Assisted Isolation of Regulatory Elements sites.

3.3. CpG islands overlapping high cleavage regions are more likely to be functional

The CpG islands used in this analysis all meet a common criteria that was developed using primary DNA sequence-based metrics. We observe that most, but not all, CpG islands overlap high cleavage windows (Figure 4), suggesting that DNA structural features can be used to partition CpG islands into different groups. Given that high cleavage windows have a statistically significant association with promoters and FAIRE sites (Figure 4), the CpG island set that overlaps these windows may have enhanced functional tendencies compared to their non-high cleavage region overlapping counterparts.

To specifically test the above hypothesis, we first partitioned all CpG islands within ENCODE into two groups: 1) CpG islands that do not overlap high cleavage regions, and 2) CpG islands that overlap high cleavage regions (Figure 5A). We then performed a statistical overlap analysis with these two groups relative to other annotations and compared the overlaps between groups (Figure 5B). All CpG islands have a statistically significant association with promoters, DNase I hypersensitive sites, and evolutionarily constrained regions. However, the group 2 CpG islands overlapped significantly more of each annotation compared to the group 1 CpG islands (compare open bars to grey bars in Figure 5B). These results suggest CpG islands that overlap high cleavage regions are more likely to be functional.

4. Discussion

We have performed a general assessment of the relationship between GC content, DNA structure (as measured by hydroxyl radical cleavage patterns), and genome function. Our results demonstrate that the correlation between GC content and DNA structure varies depending on the scale of the comparison. At low resolution scales the two variables are correlated, but the strength of this correlation decreases as resolution increases. When a biologically meaningful scale is considered—for example 10 bases represent one turn of the DNA double helix and is the approximate size of a transcription factor binding site [27]—GC content is not a strong predictor of local DNA structure. Even at scales greater than 10 bases, CpG density does not predict local overall structure well.

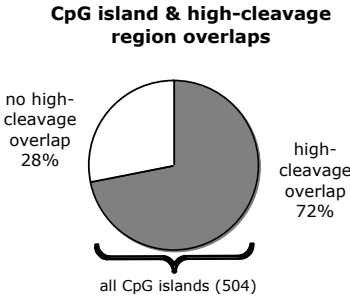
We found more low hydroxyl radical cleavage regions in the ENCODE regions compared to high regions. However, the high cleavage regions seem are more significantly associated with functional genomic elements like promoters, FAIRE sites, and CpG islands.

The finding that despite equality based upon a common primary DNA sequence-based definition, not all CpG islands are the same with respect to fine-scale DNA structure, is particularly interesting. We previously reported a common hydroxyl radical cleavage pattern found among DNase I hypersensitive sites (DHSs) that occurs more often in DHSs overlapping CpG islands compared to DHSs that do not overlap CpG islands [14]. The results presented here suggest that fine-scale local DNA structural

motifs may be associated with differentiating CpG islands that have greater functional potential.

The set of CpG islands overlapping high cleavage regions occur within evolutionarily constrained regions of the human genome significantly more often than do the set of CpG islands that do not overlap high cleavage regions (Figure 5B). It is interesting to speculate that local DNA structural features that distinguish the former set of CpG islands can act as a substrate for natural selection. The above result, along with recent literature perspectives [6, 26], suggests this may be a possibility. Collectively, the results reported here illustrate the importance of considering local DNA structure when investigating the relationship between genomic sequence and the biological functionality encoded therein.

A



B

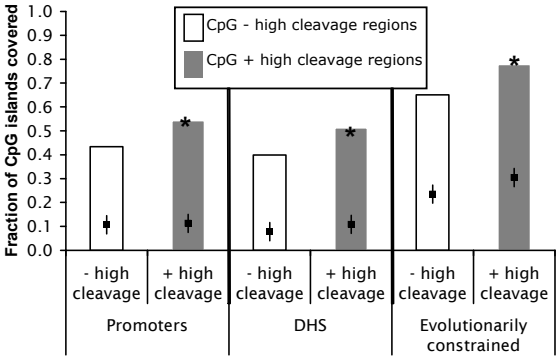


Fig. 5. CpG islands with high-cleavage windows are more likely to be functional. **A.** 72% of annotated CpG islands overlap at least one high-cleavage window. **B.** CpG islands containing high-cleavage windows overlap significantly more promoters, DHSs, and evolutionarily constrained regions compared to CpG islands that do not contain high-cleavage windows (* = $p < 0.05$ for grey bar compared to open bar in the same category; Fisher exact test). Black points and error bars are as described in Figure 4. DHS = DNase I hypersensitive sites.

Acknowledgments

We would like to thank Eric Bishop for providing code to calculate the proximity of regions to known transcription start sites and for critical evaluation of the manuscript. We would like to thank Gayle McEwen for providing code to calculate region overlap statistics using the block-bootstrap method. This work was funded by a grant from the National Human Genome Research Institute of the National Institutes of Health (R01 HG003541) to TDT. EHM was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. SCJP was supported by a Ford Foundation Dissertation Fellowship.

References

- [1] Abeel, T., Saeys, Y., Bonnet, E., *et al.*, Generic eukaryotic core promoter prediction using structural features of DNA, *Genome Res.*, 18(2):310-323, 2008.
- [2] Balasubramanian, B., Pogozelski, W.K. and Tullius, T.D., DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone, *Proc Natl Acad Sci U S A*, 95(17):9738-43, 1998.
- [3] Birney, E., Stamatoyannopoulos, J.A., Dutta, A., *et al.*, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 447(7146):799-816, 2007.
- [4] Bock, C., Paulsen, M., Tierling, S., *et al.*, CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure, *PLoS genetics*, 2(3):e26-e26, 2006.
- [5] Bock, C., Walter, J., Paulsen, M., *et al.*, CpG Island Mapping by Epigenome Prediction, *PLoS Computational Biology*, 3(6):e110-e110, 2007.
- [6] Cooper, G.M. and Brown, C.D., Qualifying the relationship between sequence conservation and molecular function, *Genome Res.*, 18(2):201-205, 2008.
- [7] Davuluri, R.V., Grosse, I., and Zhang, M.Q., Computational identification of promoters and first exons in the human genome, *Nat Genet*, 29(4):412-417, 2001.
- [8] Denoeud, F., Kapranov, P., Ucla, C., *et al.*, Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions, *Genome Res.*, 17(6):746-759, 2007.
- [9] Ellegren, H., Smith, N.G.C., and Webster, M.T., Mutation rate variation in the mammalian genome, *Current Opinion in Genetics & Development*, 13(6):562-568, 2003.
- [10] Gardiner, E.J., Hunter, C.A., Lu, X.J., *et al.*, A structural similarity analysis of double-helical DNA, *J Mol Biol*, 343(4):879-89, 2004.
- [11] Gardiner-Garden, M. and Frommer, M., CpG Islands in vertebrate genomes, *Journal of Molecular Biology*, 196(2):261-282, 1987.
- [12] Giresi, P.G., Kim, J., McDaniell, R.M., *et al.*, FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin, *Genome research*, 17(6):877-85, 2007.

- [13] Greenbaum, J.A., Pang, B., and Tullius, T.D., Construction of a genome-scale structural map at single-nucleotide resolution, *Genome research*, 17(6):947-53, 2007.
- [14] Greenbaum, J.A., Parker, S.C.J., and Tullius, T.D., Detection of DNA structural motifs in functional genomic elements, *Genome research*, 17(6):940-6, 2007.
- [15] Hardison, R.C., Roskin, K.M., Yang, S., *et al.*, Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution, *Genome Res.*, 13(1):13-26, 2003.
- [16] Harrow, J., Denoeud, F., Frankish, A., *et al.*, GENCODE: producing a reference annotation for ENCODE, *Genome Biology*, 7(Suppl 1):S4-S4, 2006.
- [17] International Human Genome Sequencing, C., Initial sequencing and analysis of the human genome, *Nature*, 409(6822):860-921, 2001.
- [18] Ioshikhes, I.P. and Zhang, M.Q., Large-scale human promoter mapping using CpG islands, *Nat Genet*, 26(1):61-63, 2000.
- [19] Jain, S.S. and Tullius, T.D., Footprinting protein-DNA complexes using the hydroxyl radical, *Nat. Protocols*, 3(6):1092-1100, 2008.
- [20] Joshi, R., Passner, J.M., Rohs, R., *et al.*, Functional specificity of a Hox protein mediated by the recognition of minor groove structure, *Cell*, 131(3):530-43, 2007.
- [21] Karolchik, D., Baertsch, R., Diekhans, M., *et al.*, The UCSC genome browser database, *Nucl. Acids Res.*, 31(1):51-54, 2003.
- [22] Kent, W.J., Sugnet, C.W., Furey, T.S., *et al.*, The human genome browser at UCSC, *Genome Res.*, 12(6):996-1006, 2002.
- [23] Kogan, S.B., Kato, M., Kiyama, R., *et al.*, Sequence structure of human nucleosome DNA, *Journal of Biomolecular Structure & Dynamics*, 24(1):43-8, 2006.
- [24] Kudla, G., Lipinski, L., Caffin, F., *et al.*, High guanine and cytosine content increases mRNA levels in mammalian cells, *PLoS Biology*, 4(6):e180 EP --e180 EP -, 2006.
- [25] Margulies, E.H., Cooper, G.M., Asimenos, G., *et al.*, Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome, *Genome Res.*, 17(6):760-774, 2007.
- [26] Margulies, E.H. and Birney, E., Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes, *Nat Rev Genet*, 9(4):303-313, 2008.
- [27] Maston, G.A., Evans, S.K., and Green, M.R., Transcriptional regulatory elements in the human genome, *Annu Rev Genomics Hum Genet*, 2006.
- [28] Mouse Genome Sequencing, C., Initial sequencing and comparative analysis of the mouse genome, *Nature*, 420(6915):520-562, 2002.
- [29] Pedersen, A.G., Baldi, P., Chauvin, Y., *et al.*, DNA structure in human RNA polymerase II promoters, *J Mol Biol*, 281(4):663-73, 1998.
- [30] Price, M.A. and Tullius, T.D., Using hydroxyl radical to probe DNA structure, *Methods in Enzymology*, 212(194-219), 1992.
- [31] Rat Genome Sequencing Project, C., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, *Nature*, 428(6982):493-521, 2004.

- [32] Salih, F., Salih, B., Kogan, S., *et al.*, Epigenetic Nucleosomes: Alu Sequences and CG as Nucleosome Positioning Element, *Journal of Biomolecular Structure & Dynamics*, 26(1):9-16, 2008.
- [33] Sandelin, A., Carninci, P., Lenhard, B., *et al.*, Mammalian RNA polymerase II core promoters: insights from genome-wide studies, *Nat Rev Genet*, 8(6):424-436, 2007.
- [34] Segal, E., Fondufe-Mittendorf, Y., Chen, L., *et al.*, A genomic code for nucleosome positioning, *Nature*, 442(7104):772-778, 2006.
- [35] Thomas, D.J., Rosenbloom, K.R., Clawson, H., *et al.*, The ENCODE Project at UC Santa Cruz, *Nucl. Acids Res.*, 35(suppl_1):D663-667-D663-667, 2007.
- [36] Venter, J.C., Adams, M.D., Myers, E.W., *et al.*, The Sequence of the Human Genome, *Science*, 291(5507):1304-1351, 2001.
- [37] Vinogradov, A.E., Isochores and tissue-specificity, *Nucl. Acids Res.*, 31(17):5212-5220, 2003.
- [38] Vinogradov, A.E., Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth, *Trends in Genetics*, 21(12):639-643, 2005.
- [39] Vinogradov, A.E., Noncoding DNA, isochores and gene expression: nucleosome formation potential, *Nucl. Acids Res.*, 33(2):559-563, 2005.
- [40] Wanapirak, C., Kato, M., Onishi, Y., *et al.*, Evolutionary conservation and functional synergism of curved DNA at the mouse epsilon- and other globin-gene promoters, *J Mol Evol*, 56(6):649-57, 2003.
- [41] Yang, S., Smit, A.F., Schwartz, S., *et al.*, Patterns of Insertions and Their Covariation With Substitutions in the Rat, Mouse, and Human Genomes, *Genome Res.*, 14(4):517-527, 2004.