Genome Wide Analysis Reveals Strong Correlation between CpG Islands and Tissue-Specificity

Riu Yamashita[1,2] ryamasi@hgc.jp
Yutaka Suzuki[1] ysuzuki@ims.u-tokyo.ac.jp
Toshihisa Takagi[3] tt@k.u-tokyo.ac.jp
Sumio Sugano[1] ssugano@ims.u-tokyo.ac.jp
Kenta Nakai[1] knakai@ims.u-tokyo.ac.jp

[1] HumanGenome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
[2] Undergraduate Program for Bioinformatics and Systems Biology, Faculty of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3] Graduate school of Frontier Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8562, Japan

Keywords: gene expression, CpG island, transcription, DBTSS, promoter

1 Introduction

Vertebrate genomes contain GC-rich or poor compartments, called isochores, It has been reported that heavy isochores (GC-rich) contain housekeeping genes while light ischores (GC-rich) have tissue-specific genes [1]. On the other hand, it was mentioned oppositely that there was no correlation between GC contents and tissue specificity [4, 5]. One of the major reasons for this conflicting idea is probably caused by the inaccurate information of transcription start site (TSS) in each gene. Recently, we constructed DBTSS: DataBase of Transcription Start Sites [6]. In this database, the clones which were constructed by the oligo-capping method show TSSs of a gene with high accuracy. Here we used data in DBTSS to determine precise TSSs and studied correlation between CpG islands and tissue-specificity using them.

2 Methods and Results

2.1 TSSs and Promoter Sequence from DBTSS

We obtained 9,219 promoter sequences for human genes and 4,778 for mouse genes from DBTSS. Although there are several definitions for CpG islands, we used a simple definition that is used by Gardiner-Garden *et al.* [3]: the averages of value for GC contents and for CpG score were calculated for every sequence by using a 200 bp window at TSS position. If the GC content was grater than 0.5, and the CpG score was grater than 0.6, we regarded the gene CpG islands-positive. We obtained 6,600 CpG positive genes and 2,619 negatives in the human genes. We also obtained 2,948 CpG positive genes and 1,830 negatives in the mouse genes.

2.2 Detection of Expression Profile from UniGene

We used NCBI UniGene for obtaining gene expression profiles of above genes. Each UniGene entry contains the information of the source of libraries where the EST clones were taken. We counted the number of the library sources, and regarded it as an index representing the degree of tissue-specificity. In LocusLink of NCBI, there are tables to connect NM IDs, LocusLink IDs, and UniGene IDs; therefore, we could obtain the correspondence between NM ID and UniGene expression degree (Figure 1). Student's t-test showed clear difference between CpG+ genes and CpG- genes ($p < 10^{-193}$ for human and $p < 10^{-279}$ for mouse).

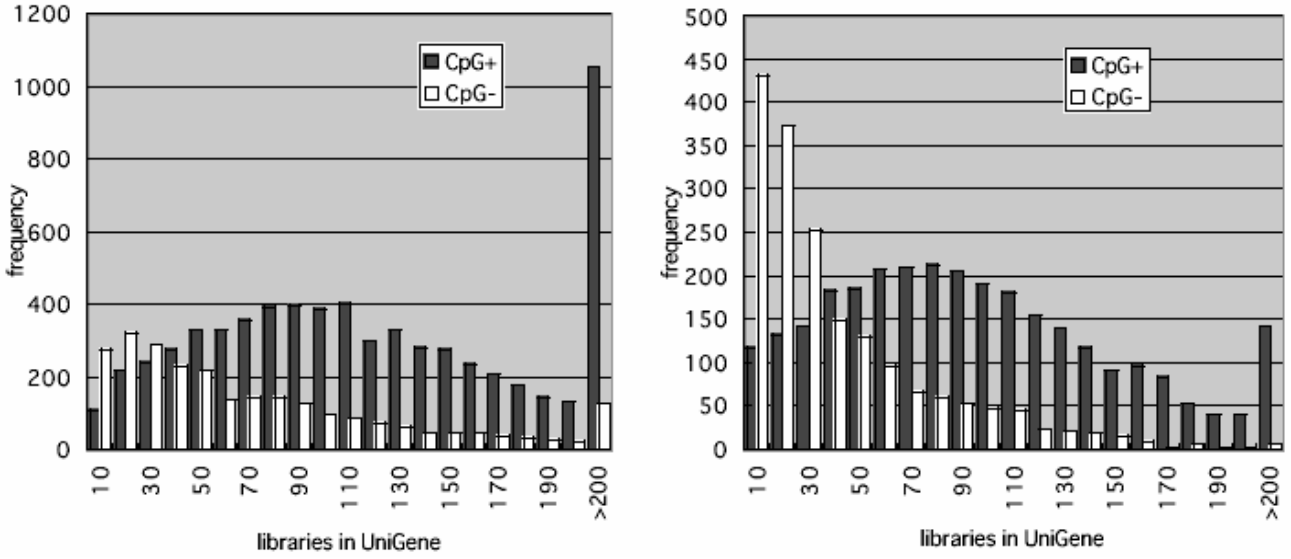

Figure 1: Correlation between CpG islands and tissue-specificity majored by the number of source libraries. The left figure indicates the case for human genes (N=9219) and the right one does mouse genes (N=4,778).

3 Discussion

Even though there are still conflicting ideas for the correlation between the GC content and the tissue-specificity of genes [2, 7]. Here we showed in figure 3 that there is a clear difference between the expressions of CpG+ genes and CpG- genes. Therefore, we suggest that there is a strong correlation with CpG island and tissue specificity in human and mouse genes. It is notable, however, that several CpG negative genes express without any tissue specificity. That is probably because that housekeeping genes are regulated by not only CpG island, but also other factors, which might be unknown. Interestingly, Vinogradov showed a clear difference between human and mouse cases in the correlation with GC contents of 3^{rd} codon position and gene expression profiles [7]. However, we could not detect any difference about it in the promoter region.

References

[1] Bernardi, G., The human genome: organization and evolutionary history, *Annu. Rev. Genet.*, 29:445–476, 1995.

[2] D'Onofrio, G., Expression patterns and gene distribution in the human genome, *Gene*, 300(1-2):155–160, 2002.

[3] Gardiner-Garden, M. and Frommer, M., CpG islands in vertebrate genomes, *J. Mol. Biol.*, 196:261–282, 1987.

[4] Goncalves, I., Duret, L., and Mouchiroud, D., Nature and structure of human genes that generate retropseudogenes, *Genome Res.*, 10(5):672–678, 2000.

[5] Ponger, L., Duret, L., and Mouchiroud, D., Determinants of CpG islands: expression in early embryo and isochore structure, *Genome Res.*, 11(11):1854–1860, 2001.

[6] Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S., DBTSS: database of human transcriptional start sites and full-length cDNAs, *Nucleic Acids Res.*, 30(1):328–331, 2002.

[7] Vinogradov, A.E., Isochores and tissue-specificity, *Nucleic Acids Res.*, 31(17):5212–5220, 2003.