In this work we did not consider the distant regions from the TSSs, more than 1 kb upstream and 0.2 kb downstream, in order to maintain the fidelity of the search results. Therefore, the current dataset does not cover the TF binding sites located very far from the TSSs. However, these are the regions where the sequence conservations were most significant throughout the neighboring 10 kb (data not shown) and where the transcriptional initiation events actually take place. Thus, it should be important to start the characterization of the promoters by investigating the nature of these regions.

Genome sequencing and full-length cDNA sequencing projects are underway for various kinds of model organisms, such as chimpanzee, macaque and zebrafish as well as many other microbes (http://www.nih.gov/science/models/). The progress of these projects should shortly accumulate genomic sequences and a large number of full-length cDNA data, from which promoter sequences could be retrieved and analyzed in a similar manner as described here. Also, very recently, new technologies named the CAGE and the 5'TSS library were developed. Using these technologies, accumulation of the TSS data in even higher throughput manner will be enabled without degrading the data quality [Shiroki *et al.*, 2003; Hashimoto *et al.*, 2004]. These data should be presented in DBTSS, which enable further accurate and versatile analyses of the promoters. Comprehensive analyses of the conservation/divergence of the promoters between human and monkeys, mouse, fish, flies, worms and other model organisms should identify which populations of promoters and what kinds of promoter elements therein play the roles for modulating the transcriptional network for each of the organisms. These analyses should clarify what features of the transcriptional network of human genes allow human cells to function as the cells of a human, a primate, a mammal and a multi-cellular organism and so on. To this end, our data resource together with newly developed database, DBTSS, should for the first time lay the firm foundation for this, as well as providing an invaluable platform for genome-wide comparative analysis of promoters. From this, the achievements of the genome projects, which would otherwise be no more than a meaningless DNA sequence, should truly come alive.

Acknowledgements

We are grateful to H. Hata and members of the HGC sequencing team for their excellent sequencing work. We are thankful to E. Nakajima and C. Gough for helpful discussion and critical reading of the manuscript. This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas and by special coordination funds for promoting science and technology (SCF), both from the Ministry of Education, Culture, Sports, Science and Technology in Japan. This study was also supported by Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.
- Baeuerle, P. A. and Baltimore, D. (1996). NF-kappa B: ten years after. Cell **87**, 13-20.
- Boguski, M. S. (2002). Comparative genomics: the mouse that roared. Nature **420**, 515-516.
- Carninci, P. and Hayashizaki, Y. (1999). High-efficiency full-length cDNA cloning. Methods Enzymol. **303**, 19-44.

- Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Birney, E. (2003). Ensembl 2002: accommodating comparative genomics. Nucleic Acids Res. **31**, 38-42.

- Fickett, J. W. and Wasserman, W. W. (2000). Discovery and modeling of transcriptional regulatory regions. Curr. Opin. Biotechnol. **11**, 19-24.

- Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8**, 967-974.

- Giardine, B., Elnitski, L., Riemer, C., Makalowska, I., Schwartz, S., Miller, W. and Hardison, R. C. (2003). GALA, a database for genomic sequence alignments and annotations. Genome Res. **13**, 732-741.

- Hannenhalli, S. and Levy, S. (2002). Predicting transcription factor synergism. Nucleic Acids Res. **30**, 4278-4284

- Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet. **16**, 369-372.

- Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. (2004). SAGE for 5'-ends transcriptome. Nat. Biotechnol., in press.

- Ho, I. C. and Glimcher, L. H. (2002). Transcription: tantalizing times for T cells. Cell **109 Suppl.**, S109-S120.

- Huang, X., Miller, W., Schwartz, S. and Hardison, R. C. (1992). Parallelization of a local similarity algorithm. Comput. Appl. Biosci. **8**, 155-165.

- Jareborg, N., Birney, E. and Durbin, R. (1999). Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. **9**, 815-824.

- Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D. and Kent, W. J. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. **31**, 51-54.

- Kawai, J. *et al.*; RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium (2001). Functional annotation of a full-length mouse cDNA collection. Nature **409**, 685-690.

- Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. **31**, 3576-3579.

- Lander, E. S. *et al.*; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature **409**, 860-921.

- Liu, R., McEachin, R. C. and States, D. J. (2003). Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. Genome Res. **13**, 654-661.

- Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E. M. (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. **12**, 832-839.
- Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. **31**, 374-378.
- Mitchell, P. J. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science **245**, 371-378.
- Novina, C. D. and Roy, A. L. (1996). Core promoters and transcriptional control. Trends Genet. **12**, 351-355.
- Okazaki, Y. *et al.*; FANTOM Consortium; RIKEN Genome Exploration Research Group Phase I & II Team (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature **420**, 563-573.
- Osada, N., Kusuda, J., Suzuki, Y., Sugano, S. and Hashimoto, K. (2000). Sequence analysis, gene expression, and chromosomal assignment of mouse Borg4 gene and its human orthologue. J. Hum. Genet. **45**, 374-377.
- Praz, V., Perier, R., Bonnard, C. and Bucher, P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. **30**, 322-324.
- Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2003). NCBI Reference Sequence project: update and current status. Nucleic Acids Res. **31**, 34-37.
- Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA **100**, 15776-15781.
- Suzuki, Y. and Sugano, S. (2003). Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. Methods Mol. Biol. **221**, 73-91.
- Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A. and Sugano, S. (2000). Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. Genomics **64**, 286-297.
- Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A. and Sugano, S. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. **11**, 677-684.
- Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. **30**, 328-331.
- Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat. Rev. Genet. **4**, 251-262.
- Venter, J. C. *et al.* (2001). The sequence of the human genome. Science **291**, 1304-1351.

- Waterston, R. H. *et al.*; Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. Nature **420**, 520-562.
- Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A. and Wagner, L. (2003). Database resources of the National Center for Biotechnology. Nucleic Acids Res. **31**, 28-33.
- Yamashita, R., Suzuki, Y., Nakai, K. and Sugano, S. (2003). Small open reading frames in 5' untranslated regions of mRnas. C. R. Biol. **326**, 987-991.

Footnotes:

[a] TRANSFAC is a registered trademark, Match is a trademark of BIOBASE GmbH, Germany