DNA Research, 16: 371-383 (2009)

Identification and functional analyses of 11,769 full-length human cDNAs focused on alternative splicing

Ai Wakamatsu[1], Kouichi Kimura[2], Jun-ichi Yamamoto[3], Tetsuo Nishikawa[3], Nobuo Nomura[4], Sumio Sugano[5], Takao Isogai[1, 3]

Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan[1]; Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo 185-8601, Japan[2]; Reverse Proteomics Research Institute, 1-9-11 Kaji, Chiyoda-ku, Tokyo 101-0044, Japan[3]; National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan[4]; Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 4-6-1 Shiroganedai, Minato-ku, Tokyo 108-8639, Japan[5]

Abstract

We analyzed diversity of mRNA produced as a result of alternative splicing (AS) in order to evaluate gene function. First, we predicted the number of human genes transcribed into protein-coding mRNAs by using the sequence information of full-length cDNAs and 5'-ESTs, and obtained 23,241 of such human genes. Next, using these genes, we analyzed the mRNA diversity and consequently sequenced and identified 11,769 human full-length cDNAs whose predicted ORFs were different from other known full-length cDNAs. Especially, 30% of the cDNAs we identified contained variation in the transcriptional start site (TSS). Our analysis, which particularly focused on multiple variable first exons (FEVs) formed due to the alternative utilization of TSSs, led to the identification of 261 FEVs expressed in tissue-specific manner. Quantification of the expression profiles of 13 genes by real-time PCR analysis further confirmed the tissue-specific expression of FEVs, for example OXR1 had specific TSS in brain and tumor tissues, and so on. Finally, based on the results of our mRNA diversity analysis, we have created the FLJ Human cDNA Database. From our result, it has been understood mechanisms that one gene produces suitable protein coding transcripts responding to the situation and the environment.

Key words: full-length cDNA, alternative splicing, alternative transcription start site, mRNA diversity, tissue-specific expression