Construction of Human Gene Catalogue by mRNA Using 1.4 Million of Full-Length cDNAs by Oligo-Capping Method

Ai Wakamatsu¹  Jun-ichi Yamamoto¹  Kouichi Kimura²
wakamatsu-a@reprori.jp  yamamoto-j@reprori.jp  kokimura@crl.hitachi.co.jp

Keisuke Tsuchiya¹  Tetsuo Nishikawa¹,²
tsuchiya-k2@reprori.jp  nisikawa-t@reprori.jp

Sumio Sugano³  Takao Isogai¹
ssugano@ims.u-tokyo.ac.jp  isogai-t@reprori.jp

¹ Reverse Proteomics Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan
² Biosystems Research Department, Central Research Laboratory, Hitachi, Ltd.,
1-280 Higashi-Koigakubo, Kokubunji, Tokyo, 185-8601, Japan
³ Graduate School of Frontier Sciences, The University of Tokyo,
4-6-1 Shiroganedai, Minato-ku, Tokyo, 108-8639, Japan

Keywords: full-length cDNA, genome mapping, human gene, mRNA

1 Introduction

Human gene number was estimated to be 20,000-25,000 by human genome sequencing project. However number of human mRNA including varieties was predicted to be about 100,000. The varieties are thought to be caused by variations of transcription start site (TSS) and splicing. In our human cDNA project, about 30,000 of human full-length sequenced cDNAs (FLJ cDNAs) were deposited to DDBJ/GenBank/EMBL, and we obtained about 1.4 million of 5’-onepass sequences (5’-onepass) of full-length cDNAs from about 100 kinds of cDNA libraries consist of human tissues and cells constructed by oligo-capping method. About 500 bp of all 5’-end of the cDNAs were obtained. The majority of the insert cDNA sizes were over 2 kb and the full-length rate of 5’-end was 90%. And about 19,000 of full-length sequenced FLJ cDNAs was newly obtained, major part of which were selected focusing on TSS and splicing variations. In this study we constructed human gene catalogue by mRNA using 1.4 million of human full-length cDNAs constructed by oligo-capping method and human full-length cDNA sequences in public domains

2 Methods

For construction of human gene catalogue by mRNA, about 1.4 million of our 5’-onepass of full-length cDNAs, 19,265 of unpublished full-length human cDNAs, 76,935 of full-length human cDNAs composed by 30,754 cDNAs of our FLJ project and 46,181 cDNAs of KIAA, MGC, DKFZ and others, 28,931 of human RefSeq (2005.1.31), 33,666 of human Ensembl gene transcripts (human 26.35/UCSC hg17), and about 110 thousand of our 3’-onepass of full-length cDNAs were mapped to human genome. Clustering analysis was performed (Figure 1). Then the 32,856 clusters of Fig.1 were classified by estimated ORF regions of each cluster and we estimated number of human genes (Fig.2).

Figure 1: Clustering of human cDNA sequences
3 Results and Discussion

From 32,856 clusters, 27,745 clusters were selected by ORF prediction. As a result, we primary estimated human gene number for 27,745 (Fig.2). Then we compared clusters of I, II and III shown in Fig.2. In core 16,784 clusters, category 1, our FLJ cDNAs were covered about 14,100 clusters (84%). In category 2, outside of core clusters, our FLJ cDNAs were covered about 3,800 clusters (77%). In category 3, about 160 clusters were selected as FLJ cDNAs from our 5'-onepass. Total number of our human gene catalogue is about 18,000 clusters. For estimation of human gene number, we calculated coverage of our 5'-onepass for each category. Our 5'-onepass were covered 15,219 clusters (91%), 4,246 clusters (86%) and 1,094 clusters (18%) in category 1, 2 and 3, respectively. Therefore genes category 3 estimated only by RefSeq and/or Ensembl need to be required proof by cDNAs. From these results, we estimated number of human genes for 23,237 and our FLJ cDNAs were covered about 18,000 genes (77%) of those. Furthermore, we analyzed variations of the splicing patterns that affected ORFs of human full-length cDNAs. Some TSS and splicing variations which had specific expression or motif were found in newly obtained FLJ cDNAs by focusing on TSS and splicing variations from about 1.4 million of full-length cDNAs.

This work was supported by a grant from New Energy and Industrial Technology Developmental Organization (NEDO) project of the Ministry of Economy, Trade and Industry of Japan.

References


[3] DDBJ/GenBank/EMBL Accession No.: DA000001 - DA999999, DB000001 - DB384947, AU116788 - AU160826, AU279383 - AU280837