Characterization of One-Pass and Full-Length Sequences of Oligo-Capping cDNA Clones by Genome Mapping

Tetsuo Nishikawa[1,2]
nisikawa@crl.hitachi.co.jp

Kouichi Kimura[1]
kokimura@crl.hitachi.co.jp

Tomohiro Yasuda[1]
tyasuda@crl.hitachi.co.jp

Jun-ichi Yamamoto[2]
yamamoto@reprori.jp

Ai Wakamatsu[2]
wakamatsu@reprori.jp

Shizuko Ishii[2]
ishii@reprori.jp

Masashi Nemoto[2]
nemoto@reprori.jp

Jun-ichi Uechi[1]
j-uechi@rd.hitachi.co.jp

Yutaka Suzuki[3]
ysuzuki@manage.ims.u-tokyo.ac.jp

Keiichi Nagai[1]
k-nagai@crl.hitachi.co.jp

Sumio Sugano[3]
ssugano@ims.u-tokyo.ac.jp

Nobuo Nomura[4]
nnomura@jbirc.aist.go.jp

Takao Isogai[2]
isogai-t@reprori.jp

[1] Central Research Laboratory, Hitachi, Ltd., 1-280 Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan
[2] Reverse Proteomics Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu-si, Chiba 292-0818, Japan
[3] Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan
[4] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

Keywords: clustering, oligo-capping, cDNA, genome, mapping

1 Introduction

Whole human genome sequence determination was completed this year. To obtain accurate gene structures and promoter information, full-length cDNA sequence information as well as the genome sequences is indispensable. Moreover, the full-length cDNA clones and sequences are valuable for functional analysis of genes. The oligo-capping cDNA library developed by Maruyama and Sugano is an effective source of the full-length cDNA clones. The full-length human cDNA sequencing project [2] supported by New Energy and Industrial Technology Developmental Organization (NEDO) determined 30,000 full-length and more than one million 5' one pass cDNA sequences of oligo-capping clones obtained at HRI and Tokyo University. Consequently, the NEDO human splicing variant cDNA project started in 2002, taking advantage of these oligo-capping cDNAs which are regarded as a valuable source of splicing variant cDNAs.

Several large scale splicing variant studies by mapping EST sequences to genome sequences have recently been published. These studies merely detect splicing variants from EST sequences. Also information of ORFs near translation start site is insufficient in ESTs. In the NEDO splicing variant project, differing from these studies, full-length sequences of the detected splicing clones are newly determined. It is expected that there are new splicing variants which are not found by analyzing ESTs, in 5' one-pass sequences. These 5' one pass sequences contain information about transcription start site variations and about expression specificity of the clones. In this study we, therefore, clustered one pass and full-length sequences of the clones by genome mapping and characterized those clusters.

2 Methods

As shown in Figure 1, clustering was done firstly by mapping cDNA sequences to genome sequences, and secondly by comparing remaining cDNA sequences each other, which were not clustered by the first step, mapping. For the first step, mapping was done by a method that we developed using dynamic programming for optimizing alignments of segments detected by MegaBLAST. For the second step, a fast algorithm, ESC algorithm that we developed, was used [1].

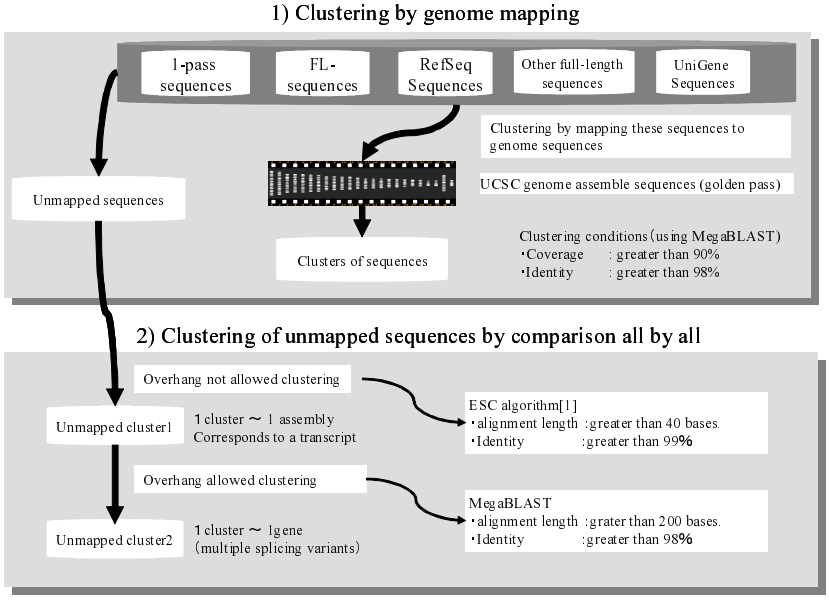

Figure 1: Clusterting method.

3 Results

By first and second clustering, 53,000 and 40,000 clusters were obtained, respectively. The first clustering results are shown in Figure 2. Total clusters are classified into full-length sequence clusters, Amsemble sequence clusters, and FL5'one-pass sequence clusters. The overlaps of the clusters were counted. The number of FL5'one-pass sequence clusters that does not overlap other kind of clusters is 31,000. This means that many unknown genes are contained in the FL5'one-pass sequences.

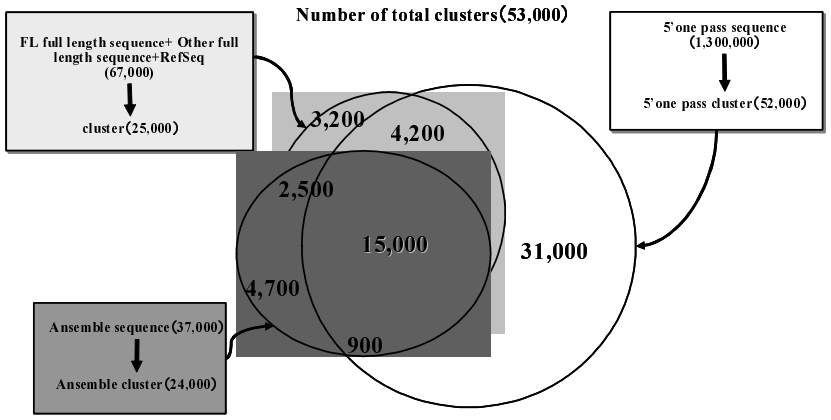

Figure 2: Clustering results.

Expression profile of each cluster was obtained by using cDNA library information. Tissue specific expression profiles were detected by using a measure of expression specificity based on a probabilistic model. We also are now studying splicing and transcription start site variations from clusters. An integrated database system that stores extracted results will be developed. This work was supported by a grant from NEDO Project of the Ministry of Economy, Trade and Industry of Japan.

References

[1] Yasuda, T. and Nishikawa, T., A fast clustering system for a huge number of nucleotide sequences, *Genome Informatics*, 13:388–389, 2002.

[2] `http://www.nedo.go.jp/bio-e/`