ORIGINAL ARTICLE

Large-scale correlation of DNA accession numbers to the cDNAs in the FANTOM full-length mouse cDNA clone set

Itsuki Ajioka,[1,2] Takuya Maeda,[1,2] and Kazunori Nakajima[2,3]

[1]*Equal contributors in this work*
[2]*Department of Anatomy, School of Medicine, Keio University*
Tokyo, Japan;
[3]*Department of Molecular Neurobiology, Institute of DNA Medicine, Jikei University*
School of Medicine, Tokyo, Japan

(Received for publication on February 2, 2006)
(Revised for publication on June 1, 2006)
(Accepted for publication on June 1, 2006)

Abstract. Oligonucleotide-based microarrays, such as GeneChip, are widely used to determine the large-scale gene expression profiles. However, GeneChip only provides information on the identity of the molecules, and the investigator must obtain each cDNA clone for further analyses. In this study, we devised a program which enables us to correlate a large number of DNA accession numbers to the FANTOM (functional annotation of the mouse) full-length mouse cDNA clone set, and made a correlative table between mouse GeneChip clones and FANTOM clones. This allows easy identification of the corresponding FANTOM clone for each GeneChip clone, even if the sequence of the GeneChip clone does not directly match the FANTOM clone. Using this table, for example, a large number of *in situ* hybridization probes can be synthesized easily, because the FANTOM clones are flanked by T3/T7 promoters on both ends. In addition, we further developed a program which retrieves the amino acid sequence (AA Seq) for each clone, even for the FANTOM clones that lack the AA Seq description, and classifies the proteins automatically. As an example, we devised a correlation table with predictions of the secretory or transmembrane molecules. The correlation table is useful for a large-scale screening of molecules involved in cell-cell communication in various biological processes. The full correlation table for the GeneChip clones is available at http://www.kjm.keio.ac.jp/past/55/3/correlation_table1.html (Keio J Med 55 (3): 107–110, September 2006)

Key words: GeneChip, FANTOM, screening, secretory protein, transmembrane protein

Introduction

As various databases of gene expression profiles have become available, it has become increasingly necessary for biologists to have to deal with a long list of DNA accession numbers. Oligonucleotide-based microarrays, such as GeneChip, for example, are widely used to determine the large-scale expression profiles of numerous genes. However, GeneChip analysis only provides information on the identity of the expressed molecules, and the investigator must obtain each cDNA clone for further analyses, including *in situ* hybridization and analysis by ectopic overexpression. Thus, from this point of view, although the GeneChip is very powerful for large-scale analyses of gene expression, use of this database is rather impractical for determining the expression profiles of genes at the cellular level or performing functional assays of genes at a large scale. The emerging strategy of *in silico* screening on the internet, that is, the digital differential display, has the same problem. The mouse cDNAs clone set, namely, FANTOM clones, comprised of 60,770 full-length cDNAs (FANTOM2[1]), is com-

Reprint Requests to: Kazunori Nakajima, M.D., Ph.D., Department of Anatomy, School of Medicine, Keio University, 35 Shinanomachi, Shinjuku-ku, Tokyo160-8582, Japan, E-mail: kazunori@sc.itc.keio.ac.jp

mercially available (although we used the FANTOM2 clones[1] in this study, the updated FANTOM3 clones have been reported recently[2]). After the GeneChip assay, investigators often search for the cDNAs such as FANTOM clones by using world wide web services, and also use other web services to obtain additional information for further assays. However, manual operation of the internet services by means of a GUI-based web browser practically confines the investigators to analysis of only a relatively small number of molecules. Thus, in this study, we report on our development of a Perl script program which enables us to correlate a large number of DNA accession numbers to the FANTOM clones. The program accesses various web-based databases and services imitating the manual operation, so the investigators need not to construct and maintain enormous local databases. As an example, we devised a correlative table between mouse GeneChip clones and FANTOM clones, which allows easy identification of the corresponding FANTOM clone for each GeneChip clone, even if the sequence of the GeneChip clone does not directly match the FANTOM clone. In addition, since most GeneChip experiments are being performed for screening molecules, it would be useful if the clones could be further classified by the type of proteins encoded. Thus, we further developed a program which retrieves the amino acid sequence (AA Seq) for each clone, even for the FANTOM clones that lack the AA Seq description in the flat files, and classifies the proteins automatically. As an example, we devised a correlation table with predictions of the secretory or transmembrane molecules (S/TMs).

Methods and Results

Hyper text transfer protocol (HTTP) was applied to the network program, and a Perl module, LWP, was downloaded from the Comprehensive Perl Archive Network (CPAN) to enable the program to communicate at HTTP level (http://www.perl.com/CPAN/). First, we collected HTML-formatted files that were displayed while performing manual operations through websites, and researched the interconnection by investigating the HTML files. Since the query data themselves change according to the series of HTML-formatted files as responses from each web service, our programs extract proper data that match patterns predetermined by our investigations of the server's responses, and communicate with the server by submitting proper query data. The first program prepared flat files of the Affymetrix GeneChip U74A, B, and C clones by submitting their accession numbers to the website of DDBJ (http://getentry.ddbj.nig.ac.jp/getstart-e.html)[3] and receiving all the flat files via FTP. The program executed an FTP program, wget, and 'zgip' program to expand a compressed file. The second program converted the flat files into a multi-FASTA formatted file of DNA Seqs. Then, we executed the stand-alone BLAST program downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Ftp/index.html)[4] by using the FASTA file as queries and by setting the FANTOM database (http://genome.gsc.riken.jp)[5] as a targeted database. Then, the third program correlated each accession number of the GeneChip clones to that of the highest-scored FANTOM clones in the obtained BLAST search results (Fig. 1). As a result, while 27,533 of the 35,813 clones (76.9%) were found to match with the FANTOM clones, 8,280 clones did not match. Some of the non-matching clones of GeneChip were due to low BLAST scores, because we set the threshold of E-value at 0.0001. The other non-matching ones might have corresponded to a part of the untranslated region (UTR) that was not contained in the FANTOM clones. To obtain the longer cDNA Seq corresponding to the GeneChip clones, the fourth program submitted the multi-FASTA formatted file of the 8,280 DNA Seqs to the DDBJ BLAST site with a 'Rodents' database as the targeted database, and 7,589 of the 8,280 clones were identified as the rodent cDNA clones. Then, we correlated them to the FANTOM clones again and found that 2,458 of the 7,589 clones (32.3%) matched with the FANTOM clones, although they did not have any overlapping sequences with the GeneChip clones. Finally, we combined these 2,458 clones with the above 27,533 clones and obtained a correlation table of 29,991 of the 35,813 clones (83.7%) of GeneChip U74. The full list is available as Supplementary table 1 at http://www.kjm.keio.ac.jp/past/55/3/correlation_table1.html This "2-step correlation program" allows easy development of correlation tables between independent cDNA clone sets in general, even if the two

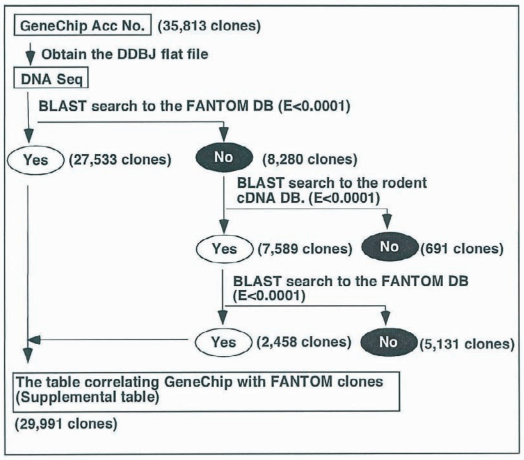
GeneChip Acc No. (35,813 clones)
Obtain the DDBJ flat file
DNA Seq
BLAST search to the FANTOM DB (E<0.0001)
Yes (27,533 clones)
No (8,280 clones)
BLAST search to the rodent cDNA DB. (E<0.0001)
Yes (7,589 clones)
No (691 clones)
BLAST search to the FANTOM DB (E<0.0001)
Yes (2,458 clones)
No (5,131 clones)
The table correlating GeneChip with FANTOM clones (Supplemental table)
(29,991 clones)

Fig. 1 Correlation of each GeneChip clone with FANTOM clones

We developed the four Perl programs to correlate each GeneChip Clone with FANTOM clones as described in the main text. Acc: accession. E: E-value.

clones in each set derived from the same mRNA and contained only partial sequences without any overlap.

To predict S/TMs in the FANTOM clones, we applied the first program to the accession numbers of the FANTOM clones and obtained their flat files. Our second program generated not only a multi-FASTA file of DNA Seqs but also that of AA Seqs from flat files. By examining the multi-FASTA file of the AA seqs, we found that while 17,487 clones had the AA Seq description, 12,504 clones did not (Fig. 2). We carefully read the flat files without the AA Seq description and found that the FANTOM clones occasionally had sequence errors that prevented translation into an AA, as reported previously.[6] To obtain the AA Seq of such clones, we applied the "2-step correlating program" to the 12,504 clones. As a result, 11,359 clones were found to match with the rodent cDNA clones, and 5,187 clones were identified as clones with an AA Seq description (Fig. 2). We combined these 5,187 clones with the 17,487 clones that had the AA Seq description in the FANTOM flat files, and obtained the AA Seqs of a total of 22,674 clones (Fig. 2). To determine whether or not they were S/TMs, we used 2 independent prediction programs. One was a simple program predicting a signal peptide in the N-terminus of the proteins based on an argorithm of the iPSORT program[7] and the other was a web client program that communicates with a web service of TMHMM that accurately predicts transmembrane helices in proteins more accurately as compared with the other prediction programs.[8,9] Using iPSORT-based programs, 4,485 clones were identified as having a signal sequence, while the TMHMM programs identified 4,841 clones which had a transmembrane region (Fig. 2, Supplementary table 1 at http://www.kjm.keio.ac.jp/past/55/3/correlation_table1.html Thus, the "2-step correlation program" used herein is also useful for generating correlation tables of clones for any types of proteins by using the respective prediction programs.

Discussion

The correlation table devised in this study can be generally applied to the analyses of the stage-, region- and cell type-specific molecules involved in both normal and pathological events. Since the FANTOM cDNA clones are inserted between T3/T7 promoters on the vector plasmids, a large number of *in situ* hybridization probes can be synthesized at once easily, for example.[10] Thus, the correlation table, which combines the large-scale microarray analysis and the full-length cDNA clone set, will effectively facilitate the large-scale analyses of the molecules involved in various biological events.

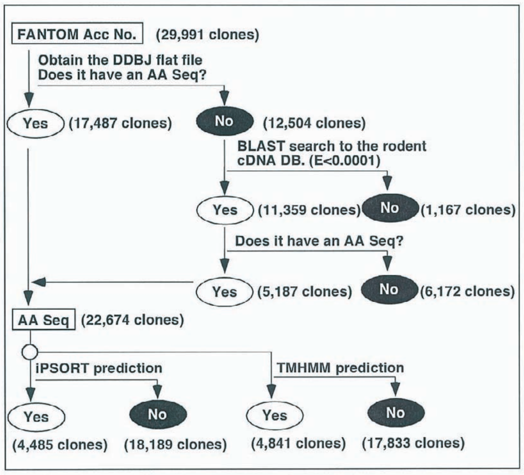

Fig. 2 Screening of predicted secretory or transmembrane molecules

By using the "2-step correlation program", we obtained the AA Seq of a total of 22,674 clones as described in the main text. To determine whether or not they were S/TMs, we developed a simple program predicting a signal peptide in the N-terminus of the proteins based on an algorithm of the iPSORT program[7] and a web client program that communicates with a web service of the TMHMM that accurately predicts transmembrane helices.[9]

Acknowledgement: We thank members of DDBJ for their valuable advice on the utilization of the DDBJ web services via FTP, and members of the Nakajima laboratory for the valuable discussions. We also thank Genome Exploration Research Group of RIKEN Genome Sciences Center for the use of FANTOM database in this study. This work was supported by grants from the Ministry of Education, Culture, Sports, and Science and Technology of Japan, Japan Society for the Promotion Science, a Grant-in-Aid for the 21st Century Center of Excellence Program entitled "Understanding and Control of Life's Function via Systems Biology", a Keio University Special Grant-in-Aid for Innovative Collaborative Research Projects, the Novartis Foundation, the Casio Science Promotion Foundation, the Okawa Foundation, and the Japan Brain Foundation.

References

1. FANTOM Consortium, RIKEN Genome Exploration Research Group Phase I & II Team: Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 2002; 420: 563−573
2. FANTOM Consortium; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group): The transcriptional landscape of the mammalian genome. Science 2005; 309: 1559−1563
3. Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T: DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res 2005; 33: D25−28
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: GenBank. Nucleic Acids Res 2005; 33: D34−38
5. Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y: FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones. Nucleic Acids Res 2002; 30: 116−118
6. Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, Hayashizaki Y, Okazaki Y: CDS annotation in full-length cDNA sequence. Genome Res 2003;13: 1478−1487

7. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 2002; 18: 298–305
8. Moller S, Croning MD, Apweiler R: Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 2001; 17: 646–653
9. Sonnhammer ELL, von Heijne G, Krogh, A: A hidden Markov model for predicting transmembrane helices in protein sequences. Porc Int Conf Intell Syst Mol Biol 1998; 6: 175–182
10. Ajioka I, Maeda T, Nakajima K: Identification of ventricular-side-enriched molecules regulated in a stage-dependent manner during cerebral cortical development. Eur J Neurosci 2006; 23: 296–308