BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression

Osamu Ogasawara¹, Makiko Otsuji¹, Kouji Watanabe¹, Takayasu Iizuka¹, Takuro Tamura¹, Teruyoshi Hishiki², Shoko Kawamoto³ and Kousaku Okubo¹,²,*

¹Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan, ²Biological Information Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto, Tokyo 135-0064, Japan and ³National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Received August 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

BodyMap-Xs (http://bodymap.jp) is a database for cross-species gene expression comparison. It was created by the anatomical breakdown of 17 million animal expressed sequence tag (EST) records in DDBJ using a sorting program tailored for this purpose. In BodyMap-Xs, users are allowed to compare the expression patterns of orthologous and paralogous genes in a coherent manner. This will provide valuable insights for the evolutionary study of gene expression and identification of a responsive motif for a particular expression pattern. In addition, starting from a concise overview of the taxonomical and anatomical breakdown of all animal ESTs, users can navigate to obtain gene expression ranking of a particular tissue in a particular animal. This method may lead to the understanding of the similarities and differences between the homologous tissues across animal species. BodyMap-Xs will be automatically updated in synchronization with the major update in DDBJ, which occurs periodically.

INTRODUCTION

Do homologous genes have similar expression patterns? On the one hand, evolutionary theories predict that paralogous genes have complementary spatio-temporal expression patterns that are based on a model of the consequences of a gene duplication event (1,2). On the other hand, molecular biologists sometimes assume almost similar expression patterns among structurally similar genes because the cis-regulatory element is a part of a gene located next to the coding sequence on the genome (3). The fragmented picture emerging from a limited number of recent genome-wide comparisons (4,5) will become more coherent if a wider range of species are more accurately compared. Expression sequence tag (EST) data are the best resources to carry out such a comparison because they cover a sufficiently wide range of species (currently, 54 species with >40 million ESTs). Further, it has a sufficient resolution in anatomy and nucleotide sequence.

For many years, EST databases have served to provide expression information of individual species (6–12). However, no attempt has been made to integrate expression information across species in a coherent manner despite its potential importance. There are two major obstacles in achieving this goal: (i) In the INSD format—the universal format of EST data available to the public—the RNA source is described in free text in various fields and (ii) biologically sound integration of anatomies across different species is difficult, if at all possible. We have overcome these obstacles by developing a program for identification and subsequent sorting of anatomical names at the organ level across species. By applying this to the entire set of animal EST data in DDBJ, we have developed a database for the comparison of gene expressions across different animal species. This database is automatically updated in synchronization with major updates in DDBJ.

METHODOLOGY

Resources

The data unique to BodyMap-Xs are the anatomical breakdowns of 17 million ESTs across animal species generated by
a sorting program tailored for this purpose. Other data used include the EST data in DDBJ, UniGene (13,14) for EST–gene relationship and RefSeq (15) for the conversion of a UniGene entry to representative amino acid sequences. In addition, a gene cluster from the InParanoid database (16,17) was used to connect genes based on orthologous relationships. The BLASTP program was employed to collect similar genes while the libraries are divided into normal and tumor/cell lines. The resultant program was then applied to the remaining animal ESTs and the same application-modification cycles were repeated. Although the rules based on human anatomy worked properly in most vertebrates as expected, some names of body parts or cells were unique to particular species. These names were added to the dictionary at the most homologous human library groups based on their function. For example, ‘head’ in insects was sorted as ‘cerebrum’ and ‘hemocyte’ in Ciona intestinalis was classified as ‘blood’. Using the present version of the program, 76–100% of the ESTs of each vertebrate species were anatomically classified. Even in C.intestinalis—the simplest chordate in the EST division—five different organs were identified (Supplementary Table 1).

In addition to the anatomical categorization, the libraries were further categorized with respect to two independent aspects—the condition of tissue and distortion in population prior to sampling—by the same sorting program based on the pattern dictionary. Based on the condition of the tissues, the libraries are divided into normal and tumor/cell lines. The distortion aspect discriminates normalization and other processes that distort the population in conventional libraries, which is sometimes employed to avoid redundant isolation of clones for the same gene.

**EST–gene relation and gene–gene relation**

In this database, an individual UniGene cluster is tentatively regarded as the smallest unit that is responsible for an expression pattern. According to the EST_ID:UniGeneID correspondence in NCBI, the data were organized in the UniGene ID × 40-organ matrix of EST frequency. In addition, orthologous genes were interconnected using InParanoid data (16,17) after translating an Ensembl ID (20) to an UniGene ID via LocusLink.
DATA PRESENTATION

Similar to the old BodyMap DB (6), users can navigate to observe the activity ranking of genes in a particular animal and tissue with their anatomical expression patterns at a selectable resolution in a concise tabular format (Figure 2). For each gene in the ranking, users can compare the expression patterns of its orthologous partners across species in the same format.

Genes can be retrieved in the same tabular representation of expression patterns by using an ID or a keyword and by similarity to the query sequence. Similarities are measured against the RefSeq peptides corresponding to the representatives for UniGene clusters using the BLASTP program. The results show the expression patterns of homologous genes. List of libraries can be retrieved either with exact matching of the keyword or after generalization of user’s keyword as a tissue category name by the library sorting program.

Last but not least, we would like to emphasize that the automatic identification process is still error prone even after rounds of careful inspection and correction. This is mainly because the program does not consider the context where the relevant patterns are embedded. Further, a grouping scheme may not be obvious to some users or even inappropriate for some purposes. Therefore, we devised a method that allows the users to backtrack the automatic sorting process. Every number in the tables (either library count or EST count) can be broken down to the library level where the reasons as to why they were categorized in a particular manner are shown by the ‘keyword in context (KWIC)’ representation showing key patterns with short flanking strings in the original EST record. Using such information, users can verify the data or modify them according to their purpose.

FUTURE DEVELOPMENT

In light of the fact that as much as 40–60% of genes have alternatively spliced transcripts in mammals (21) and that a
substantial portion of this splicing may be anatomically controlled (22), the resolution of expression should be enhanced to the transcript level from the gene level. Integration across different measurement platforms is planned through the incorporation of H-ANGEL function (23) to BodyMap-Xs. In the downstream of the expression comparison table for homologous genes, a tree view format that shows mutual evolutionary relations among the homologs is under construction. The BodyMap-Xs data will be updated immediately after major updates in the DDBJ are released; these updates take place four times a year. Further, the pattern dictionary will be updated after every update based on novel patterns found in the new release.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work is supported by a grant-in-aid for scientific research in priority areas (genome information science, grant ID 17017037) from the Ministry of Education, Culture, Sports, Science and Technology in Japan (MEXT) and grants from New Energy and Industrial Technology Development Organization of Japan (NEDO) in the Ministry of Economy, Trade and Industry of Japan (METI). Funding to pay the Open Access publication charges for this article was provided by MEXT grant ID 17017037.

Conflict of interest statement. None declared.

REFERENCES

1. Ohono,S. (1970) Evolution by gene duplication. Springer-Verlag, NY.
2. Force,A., Lynch,M., Pickett,F.B., Amores,A., Yan,Y.L. and Postlethwait,J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics, 151, 1531–1545.
3. Lercher,M.J., Urrutia,A.O. and Hurst,L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat. Genet., 31, 180–183.
4. Huminiecki,L. and Wolfe,K.H. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res., 14, 1870–1879.
5. Castillo-Davis,C.I., Hartl,D.L. and Achaz,G. (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res., 14, 1530–1536.
6. Okubo,K., Hori,N., Matoba,R., Niiyama,T., Fukushima,A., Kojima,Y. and Matsubara,K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat. Genet., 2, 173–179.
7. Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. et al. (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature, 377, 3–174.
8. Hishiki,T., Kawamoto,S., Morishita,S. and Okubo,K. (2000) BodyMap: a human and mouse gene expression database. Nucleic Acids Res., 28, 136–138.
9. Kawashima,T., Kawashima,S., Kohara,Y., Kanehisa,M. and Makabe,K.W. (2002) Update of MAGEST: Maboya Gene Expression patterns and Sequence Tags. Nucleic Acids Res., 30, 119–120.
10. Uenishi,H., Eguchi,T., Suzuki,K., Sawazaki,T., Toki,D., Shinkai,H., Okamura,N., Hamasima,N. and Awata,T. (2004) PEDE (Pig EST Data Explorer): construction of a database for ESTs derived from porcine full-length cDNA libraries. Nucleic Acids Res., 32, D484–D488.
11. Boardman,P.E., Sanz-Ezquerro,J., Overton,I.M., Burt,D.W., Bosch,E., Fong,W.T., Pickle,C., Brown,R.W., Wiln,S.A. and Hubbard,S.J. (2002) A comprehensive collection of chicken cDNAs. Curr. Biol., 12, 1965–1969.
12. Clark,M.S., Edwards,Y.J., Peterson,D., Clifton,S.W., Thompson,A.J., Sasaki,M., Suzuki,Y., Kikuchi,K., Watabe,S., Kawakami,K. et al. (2003) Fugu ESTs: new resources for transcription analysis and genome annotation. Genome Res., 13, 2747–2753.
13. Boguski,M.S. and Schuler,G.D. (1995) ESTablishing a human transcript map. Nat. Genet., 10, 369–371.
14. 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. et al. (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res., 31, 28–33.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res., 33, D501–D504.
16. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol., 314, 1041–1052.
17. O’Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res., 33, D476–D480.
18. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Ogmaa,D., Greiling,G., Jongeneel,C.V., McCarthy,M.I. et al. (2003) eVOC: a controlled vocabulary for unifying gene expression data. Genome Res., 13, 1222–1230.
19. Bard,J., Rhee,S.Y. and Ashburner,M. (2005) An ontology for cell types. Genome Biol., 6, R21.
20. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. et al. (2005) Ensembl 2005. Nucleic Acids Res., 33, D447–D453.
21. Modrek,B. and Lee,C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat. Genet., 34, 177–180.
22. Yeo,G., Holste,D., Kreiman,G. and Burge,C.B. (2004) Variation in alternative splicing across human tissues. Genome Biol., 5, R74.
23. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasaawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. et al. (2005) The Human Anatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. Nucleic Acids Res., 33, D567–D572.