D628–D631 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj137

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

Osamu Ogasawara[1], Makiko Otsuji[1], Kouji Watanabe[1], Takayasu Iizuka[1], Takuro Tamura[1], Teruyoshi Hishiki[2], Shoko Kawamoto[3] and Kousaku Okubo[1,2,*]

[1]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan, [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto, Tokyo 135-0064, Japan and [3]National 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. Expressed 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

*To whom correspondence should be addressed. Tel: +81 559 815 838; Fax: +81 559 815 837; E-mail: kokubo@genes.nig.ac.jp

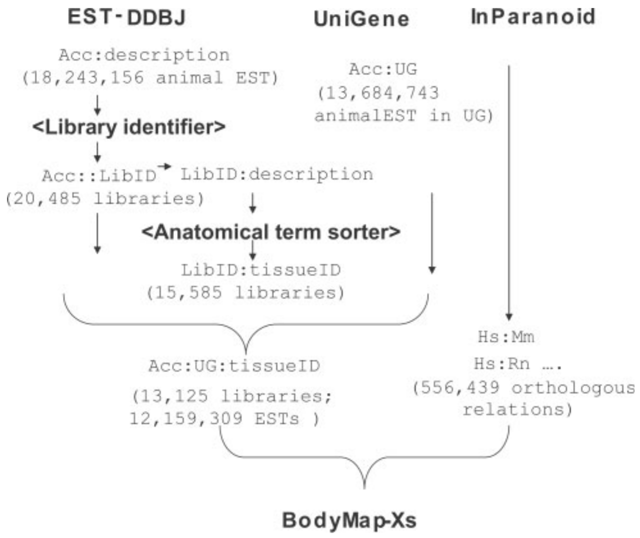


Figure 1. Outline of BodyMap-Xs construction with some relevant figures for the data. Starting from three public resources, construction of database is structured as a pipe-line.

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 within this database. The scheme for EST breakdown and database construction is outlined in Figure 1.

Grouping of EST by the source library

Because INSD formats do not have an official identifier for the source library, EST records are first clustered based on the identity present in the descriptions in the fields for authors, citations and sources. Highly variable strings that are sometimes embedded in source field, such as clone ID, were suppressed in advance. From each cluster, hereafter referred to as the library, we arbitrarily selected one record to be processed during anatomical sorting.

Grouping of libraries at the organ level

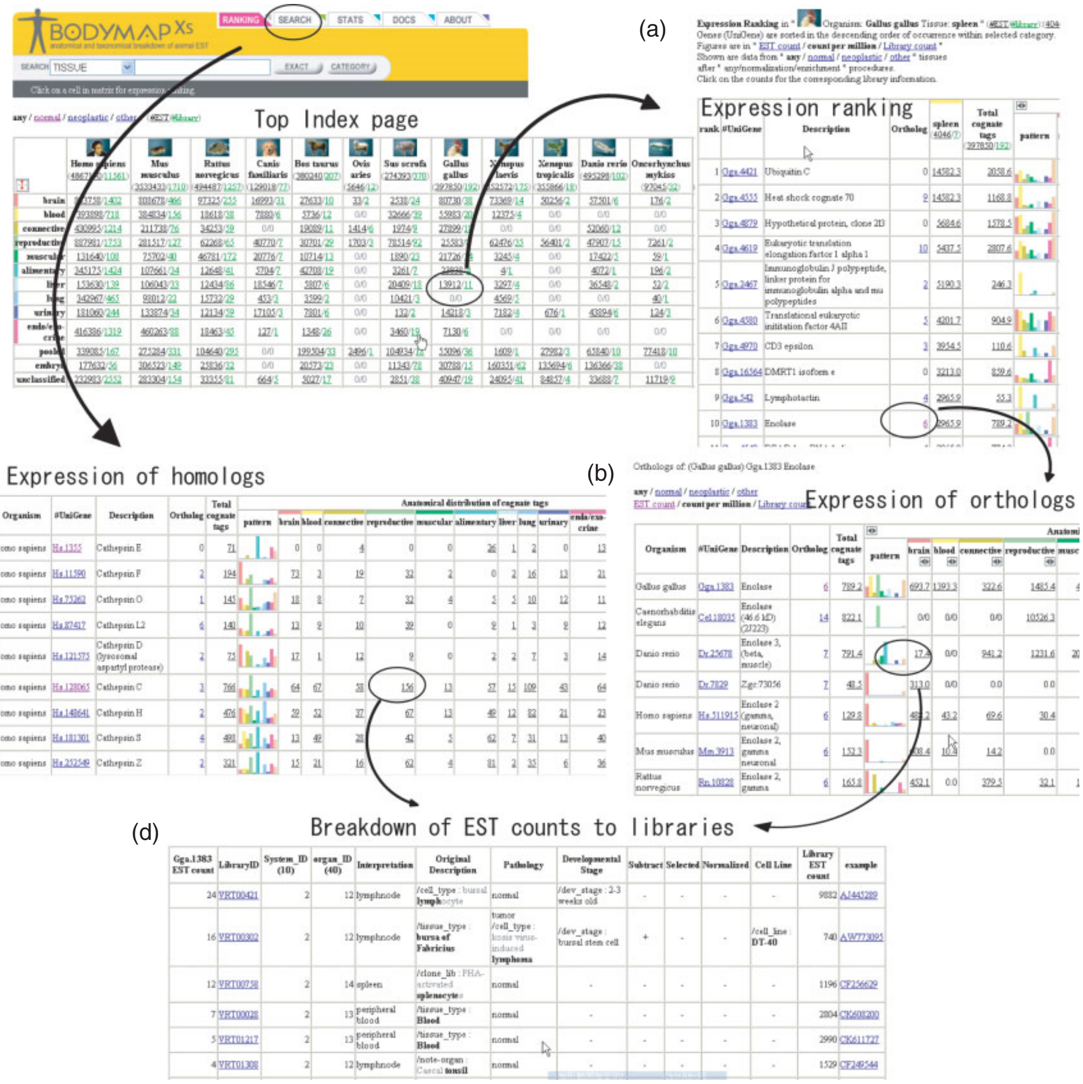
In INSD, the description logic and terminology of the RNA used for library construction are not controlled. Therefore, it is advised that the procedure for material preparation can be written in free text as a 'note'. However, in reality, the descriptions are scattered across multiple fields of the records. For the coherent categorization of libraries across species, we need to classify those libraries in some manner based on those free and scattered descriptions. Structured terminology or ontology for anatomical concepts have been proposed for some species (18,19); however, assignment of these concepts in ontologies, having mutually different description logics to the tissue descriptions of continuously growing EST records across species is a daunting task. Further, regarding the source information of gene expression studies, we valued the reproducibility and clarity of the assignment process more than the soundness of the resulting classification in a purely anatomical sense.

Based on this concept, we developed a sorting program that detects anatomical names in the EST records across species and sorts them according to explicit rules. First, we selected 200 most populated libraries and manually grouped them in a bottom-up manner by joining the less populated libraries to their closest neighbors until we obtained 40 groups of almost homogeneous libraries. These 40 groups, mostly representing individual organs, were labeled with the most concise organ name. Subsequently, the key patterns within the anatomical terms responsible for the grouping process were extracted and compiled into a pattern dictionary that maps each pattern to 1 of the 40 organ names. Patterns in this dictionary were then expanded manually to cover variations and synonyms to the key patterns and were applied back to the same 200 libraries. Based on the erroneous identification and conflicts found in this application, several rules that resolve these problems were added to the program. Subsequently, the sorting program was applied to rest of the human libraries. From the libraries that lacked matching patterns, 100 most populated libraries were selected, manually assigned to 1 of the 40 groups and the pattern dictionaries were expanded accordingly. By repeating this cycle several times, we obtained a program that can sort 95% of the human EST records including those that were positively identified as pooled tissue or whole body. 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.



any / normal / neoplastic / other / (#EST/#library)

	Homo sapiens (4867[illegible]/11561)	Mus musculus (3533433/1710)	Rattus norvegicus (494487/1257)	Canis familiaris (129018/72)	Bos taurus (380240/207)	Ovis aries (5646/12)	Sus scrofa (274393/370)	Gallus gallus (397850/192)	Xenopus laevis ([illegible]52572/175)	Xenopus tropicalis (355866/18)	Danio rerio (495298/102)	Oncorhynchus mykiss (97045/32)
brain	[illegible]3758/1402	808678/466	97325/255	16993/31	27633/10	33/2	2538/24	80730/38	73369/14	50256/2	57501/6	176/2
blood	393898/718	384834/156	18618/38	7880/6	5736/12	0/0	32666/39	55983/20	12375/4	0/0	0/0	0/0
connective	430995/1214	211738/76	34253/59	0/0	19089/11	1414/6	1974/9	27899/1[illegible]	0/0	0/0	52060/12	0/0
reproductive	887981/1753	281517/127	62268/65	40770/7	30701/29	1703/3	78514/92	25583/[illegible]	62476/35	56401/2	47907/15	7261/2
muscular	131640/108	75702/40	46781/172	20776/7	10714/13	0/0	1890/23	21726/[illegible]	3245/4	0/0	17422/5	59/1
alimentary	345175/1424	107661/34	12648/41	5704/7	42708/19	0/0	3261/7	23838/[illegible]	4/1	0/0	4072/1	196/2
liver	153630/139	106043/33	12434/86	18546/7	5807/6	0/0	20409/18	13912/11	3297/4	0/0	36548/2	52/2
lung	342967/465	98012/22	15732/29	453/3	3599/2	0/0	10421/3	0/0	4569/5	0/0	0/0	40/1
urinary	181060/244	133874/34	12134/59	17105/3	7801/6	0/0	132/2	14218/3	7182/4	676/1	43894/6	124/3
endo/exo-crine	416386/1319	460263/88	18463/45	127/1	1348/26	0/0	3460/19	7130/6	0/0	0/0	0/0	0/0
pooled	339085/167	275284/331	104640/295	0/0	199304/33	2496/1	104934/1[illegible]	55096/36	1609/1	27982/3	65840/10	77418/10
embryo	177632/56	306523/149	25836/32	0/0	20573/23	0/0	11343/78	30788/15	160351/62	135694/6	136366/38	0/0
unclassified	232983/2552	283304/154	33355/81	664/5	5027/17	0/0	2851/38	40947/19	24095/41	84857/4	33688/7	11719/9

Expression Ranking in " Organism: **Gallus gallus** Tissue: **spleen** " (#EST/#library) (404
Genes (UniGene) are sorted in the descending order of occurrence within selected category.
Figures are in " EST count / **count per million** / Library count "
Shown are data from " **any** / normal / neoplastic / other " tissues
after " any/normalization/enrichment " procedures.
Click on the counts for the corresponding library information.

rank	#UniGene	Description	Ortholog	spleen (4046/7)	Total cognate tags (397850/192)	pattern
1	Gga.4421	Ubiquitin C	0	14582.3	2058.6	
2	Gga.4555	Heat shock cognate 70	9	14582.3	1168.8	
3	Gga.4879	Hypothetical protein, clone 2I3	0	5684.6	1578.5	
4	Gga.4619	Eukaryotic translation elongation factor 1 alpha 1	10	5437.5	2807.6	
5	Gga.2467	Immunoglobulin J polypeptide, linker protein for immunoglobulin alpha and mu polypeptides	2	5190.3	246.3	
6	Gga.4380	Translational eukaryotic initiation factor 4AII	5	4201.7	904.9	
7	Gga.4970	CD3 epsilon	3	3954.5	110.6	
8	Gga.16564	DMRT1 isoform e	0	3213.0	859.6	
9	Gga.542	Lymphotactin	4	2965.9	55.3	
10	Gga.1383	Enolase	6	2965.9	789.2	

Orthologs of: (Gallus gallus) Gga.1383 Enolase

any / normal / neoplastic / other
EST count / **count per million** / Library count

Organism	#UniGene	Description	Ortholog	Total cognate tags	pattern	brain	blood	connective	reproductive	musc
Gallus gallus	Gga.1383	Enolase	6	789.2		693.7	1393.3	322.6	1485.4	4
Caenorhabditis elegans	Cel.18035	Enolase (46.6 kD) (2J223)	14	822.1		0/0	0/0	0/0	10526.3	
Danio rerio	Dr.25678	Enolase 3, (beta, muscle)	7	791.4		17.4	0/0	941.2	1231.6	20
Danio rerio	Dr.7829	Zgc:73056	7	48.5		313.0	0/0	0.0	0.0	
Homo sapiens	Hs.511915	Enolase 2 (gamma, neuronal)	6	129.8		48[illegible].2	43.2	69.6	30.4	
Mus musculus	Mm.3913	Enolase 2, gamma neuronal	6	152.3		[illegible]08.4	10.4	14.2	0.0	
Rattus norvegicus	Rn.10828	Enolase 2, gamma	6	165.8		452.1	0.0	379.5	32.1	1

Organism	#UniGene	Description	Ortholog	Total cognate tags	Anatomical distribution of cognate tags: pattern	brain	blood	connective	reproductive	muscular	alimentary	liver	lung	urinary	endo/exo-crine
Homo sapiens	Hs.1355	Cathepsin E	0	71		0	0	4	0	0	26	1	2	0	13
Homo sapiens	Hs.11590	Cathepsin F	2	194		73	3	19	32	2	0	2	16	13	21
Homo sapiens	Hs.75262	Cathepsin O	1	145		18	8	7	32	4	5	5	10	12	11
Homo sapiens	Hs.87417	Cathepsin L2	6	140		13	9	10	39	0	9	1	3	9	12
Homo sapiens	Hs.121575	Cathepsin D (lysosomal aspartyl protease)	2	75		17	1	12	9	0	2	2	7	3	14
Homo sapiens	Hs.128065	Cathepsin C	3	766		64	67	58	156	13	57	15	109	43	64
Homo sapiens	Hs.148641	Cathepsin H	2	476		59	52	37	67	13	49	12	82	21	23
Homo sapiens	Hs.181301	Cathepsin S	4	493		13	49	28	42	5	62	7	31	13	40
Homo sapiens	Hs.252549	Cathepsin Z	2	321		15	21	16	62	4	81	2	35	6	36

Gga.1383 EST count	LibraryID	System_ID (10)	organ_ID (40)	Interpretation	Original Description	Pathology	Developmental Stage	Subtract	Selected	Normalized	Cell Line	Library EST count	example
24	VRT00421	2	12	lymphnode	/cell_type : bursal **lymph**ocyte	normal	/dev_stage : 2-3 weeks old	-	-	-	-	9882	AJ445289
16	VRT00302	2	12	lymphnode	/tissue_type : **bursa of Fabricius**	tumor /cell_type : leosis virus-induced **lymphoma**	/dev_stage : bursal stem cell	+	-	-	/cell_line : **DT-40**	740	AW773095
12	VRT00758	2	14	spleen	/clone_lib : PHA-activated **splenocyte**s	normal	-	-	-	-	-	1196	CF256629
7	VRT00028	2	13	peripheral blood	/tissue_type : **Blood**	normal	-	-	-	-	-	2804	CK608200
5	VRT01217	2	13	peripheral blood	/tissue_type : **Blood**	normal	-	-	-	-	-	2990	CK611727
4	VRT01308	2	12	lymphnode	/note-organ : Caecal **tonsil**	normal	-	-	-	-	-	1529	CF249544

Figure 2. Browsing data in BodyMap-Xs. Starting from the index page, the expression ranking of relevant organs is shown (**a**). For genes with expression patterns of interest, the expression of orthologous partners can be shown (**b**). Similarity search allows users to compare expression patterns of genes structurally similar to the query (**c**). For validation of expression patterns, every figure in the table can be broken down to the library level. The library list shows why these libraries were categorized in a particular manner with KWIC format (**d**).

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., Okumura,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., Tickle,C., Brown,W.R., Wilson,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., Otgaar,D., Greyling,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., Ogasawara,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.