Theor Chem Acc (2010) 125:651–658
DOI 10.1007/s00214-009-0670-5

REGULAR ARTICLE

TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles

Sunitha Kogenaru · Coral del Val · Agnes Hotz-Wagenblatt · Karl-Heinz Glatting

Received: 31 July 2009 / Accepted: 21 October 2009 / Published online: 10 November 2009
© Springer-Verlag 2009

Abstract Tissue-distribution profiles are crucial for understanding the characteristics of cells and tissues in terms of their differential expression of genes. Most of the currently available resources for tissue-distribution profiles are either specialized for a few particular organisms, tissue types and disease stages or do not consider the "tissue ontology" levels for the calculation of the tissue-distribution profiles. Therefore, we have developed "TissueDistributionDBs", a repository of tissue-distribution profiles based on the expressed sequence tags (ESTs) data extracted from the UniGene database by employing "Tissue Ontology" available at BRENDA. To overcome the occurrence of the natural language variations in the EST's source tissue-type terms, we have generated a "tissue synonym library" and standardized these tissue-type terms by cross-referencing to the controlled vocabulary for tissue-type terms available at BRENDA "Tissue Ontology". Furthermore, we have provided a quantitative expression for genes among the tissue types at various anatomical levels by constructing "tissue slims". Concurrently, the expression among tissue types is used for tissue-distribution calculations. The resulting output profiles can be queried by the Sequence Retrieval System (SRS) and are currently available for 20 different model organisms. We benchmarked our database system against the Swissprot database using a set of 40 different tissue types. This database system is useful for the understanding of the tissue-specific expression patterns of genes, which have implications for the identification of possible new therapeutic drug targets, in gene discovery, and in the design and analysis of micro-arrays. TissueDistributionDBs can be accessed via the World Wide Web (www) at http://genius.embnet.dkfz-heidelberg.de/menu/tissue_db/.

Keywords Tissue · Tissue type · Tissue slims · Tissue ontology · Tissue synonym library · Tissue-distribution pattern · Tissue-distribution profiles · Biomarker · Gene

1 Introduction

Tissue-distribution profiles are the representation of the quantitative expression of thousands of genes in a tissue or an organ of a given organism. These organism-specific tissue-distribution profiles play an important role in the identification of novel drug targets, in gene discovery, in the customization of micro-arrays and in the discovery of novel biomarkers for disease screening [1]. The pattern of gene expression determines the characteristics of the tissue types [2]. The transcripts that are expressed in a tissue type at a certain point in time are captured by constructing cDNA libraries by reverse transcription of mRNAs [3–7]. Expressed sequence tags (ESTs), which are partially sequenced cDNAs, provide a snapshot of the transcripts

Dedicated to Professor Sandor Suhai on the occasion of his 65th birthday and published as part of the Suhai Festschrift Issue.

Electronic supplementary material The online version of this article (doi:10.1007/s00214-009-0670-5) contains supplementary material, which is available to authorized users.

S. Kogenaru (✉) · C. del Val · A. Hotz-Wagenblatt · K.-H. Glatting
Department of Molecular Biophysics,
Deutsches Krebsforschungszentrum (DKFZ),
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany
e-mail: skogenaru@yahoo.com

Present Address:
C. del Val
Department of Computer Science and Artificial Intelligence,
University of Granada, Granada, Spain

[8]. The ever increasing numbers of ESTs obtained from the cDNA libraries are stored in several repositories [9–14]. A careful analysis of ESTs not only provides significant functional, structural and evolutionary information, but also provides in silico analysis of tissue-specific transcriptional profiles [13, 15, 16]. Because the number of non-redundant ESTs representing a particular gene indicates the expression level of that gene in a given tissue type [16], the tissue-distribution profile of the gene can be calculated by counting the number of ESTs representing a gene in different tissue types to the total number of ESTs from that particular tissue. Since the tissue-type source of the ESTs and relationships among these tissue types at the anatomical level play a major role, the quality of the tissue-distribution profiles greatly depends on the extent of the standardized tissue-type terms used and also the depth of the anatomical relationships among the tissue types considered for the tissue-distribution calculations.

Currently, several resources such as SOURCE [17], UniGene, EST Expression Profile Viewer [18], Digital Differential Display (DDD) [19], BODYMAP [20, 21] and TIGR gene indices [13] provide the tissue-distribution profiles for several organisms. However, all these resources neither consider the natural language variations of the source tissue-type terms nor the relationship among the tissue types, which may result in inconsistency in the tissue-distribution calculations. The next generation of resources such as TissueInfo [22] and ExQuest [23] consider the natural language variation in tissue-type terms and define the relationship among the tissue types using "Tissue hierarchy", where anatomical relationships among the tissue types are described in a parent–child (one child one parent) hierarchical manner. "Tissue Ontology" on the other hand is the next generation concept for establishing the relationship among tissue types. It provides well-structured controlled vocabularies by establishing standard anatomical relationships among the tissue types in a structure known as a directed acyclic graph (DAG). This is a graph in which tissue types can have multiple parents (one child multiple parents).

We have created "TissueDistributionDBs", a comprehensive catalog of genes and their tissue-distribution profiles based on the "Tissue Ontology" for 20 different model organisms by taking into account the natural language variation in the source tissue-type terms. The UniGene database which automatically sorts the ESTs into a non-redundant set of gene-oriented clusters is an ideal repository of ESTs for the tissue-distribution calculations. Therefore, our approach is to rank the genes defined by UniGene cluster in a spectrum of tissue-specificity as in several other resources [2, 16, 17, 22–24]. Additionally, we use the "Tissue Ontology", a standard Open Biomedical Ontology (OBO) [25] platform available at BRENDA [26], where the tissue types and their anatomical relationships are described in a species-independent manner (uni- and multicellular organisms). Finally, we generate the tissue-distribution profiles by calculating the relative expression measure of genes to rank the tissue types at four different anatomical levels of the "Tissue Ontology".

2 Database constructions

The algorithm behind TissueDistributionDBs retrieves the tissue-type source for all the ESTs present in each UniGene cluster from GenBank [27], as schematically shown in Fig. 1. The retrieved tissue-type source information contains natural language variations, aliases and typographical errors [22]. We therefore, manually curated and standardized these raw tissue types by referring to the corresponding controlled vocabulary tissue-type terms available at BRENDA "Tissue Ontology" and generated an organism-specific "Tissue Synonym Library" (Supplementary material 1), which is a collection of raw tissue-type terms found in GenBank along with their corresponding curated tissue-type terms.

"Tissue Ontology" available at BRENDA provides tissue-type terms and their ontological hierarchy describing the anatomical relationships among the tissue types in an organized structure called DAG, wherein each node is considered as one tissue type. We use the BrendaTissue.obo.txt file, an OBO flat file because it not only offers well-structured controlled vocabularies that enable human readability, minimal redundancy, and ease of parsing, but also can be shared for use across different biological and medical domains. The controlled vocabularies of the tissue types are structured in a way that, we can query them at different levels of the tissue types: for example, we can use the "Tissue Ontology" to find all the tissue types which are part of the nervous system, or we can retrieve all the sub-anatomical tissue types of the brain or head. The tissue types are assigned to a particular level depending upon how far a tissue-type is from the root tissue-type node.

The root node is the most generalized tissue type and corresponds to the whole body. Depending upon the number of tissue types present between a particular tissue type and the root, by tracing back from the target tissue type to the root node, a "Level" is assigned to a tissue type. For instance, if a tissue type is adjacent to the root node, then it is said to be at Level-1 and at Level-2 if there exists another tissue type in between itself and the root node. This process is called "mapping" of the tissue types to definite levels of the "Tissue Ontology". Thus, the farther a tissue type is from the root node, the more "specialized" or specific is the tissue type, while on the other hand, the closer a tissue type is to the root node, the more

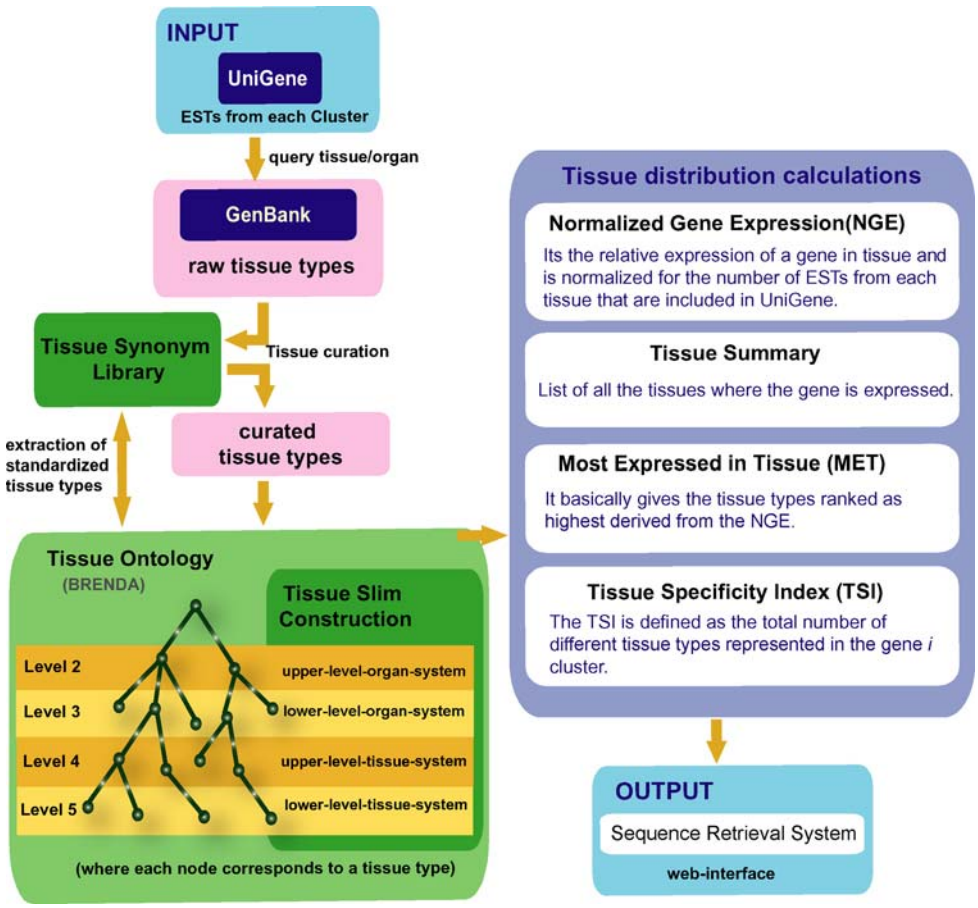

Fig. 1 TissueDistributionDBs generation pipeline provides an outline of the steps followed in the generation of the tissue-distribution profiles, where the data flow is indicated by the *arrows*

"generalized" it is [26]. "Tissue Ontology" basically is an advanced generation concept to handle relationships among the tissue types in a definitive manner [26]. A simplified example to illustrate this would be that the tissue-type brain has two parents, head and central nervous system. This is because the brain is anatomically present within the head structure and as well as forms the basis for the central nervous system. Therefore, if any EST is from the tissue-type brain, then automatically both head and central nervous system are considered as its parents or generalized tissue types. We mapped the curated tissue types at four different anatomical levels of BRENDA "Tissue Ontology". This mapping of the tissue types to the "Tissue Ontology" is called construction of the "Tissue Slim" (Supplementary material 2), which basically provides a uniform view of all the tissue types present at a particular level. We manually curated and generated the organism-specific "Tissue Slims" at four different anatomical levels (Table 1) using the BrendaTissue.obo.txt file. To further validate the relationship among the tissue types we used the ontology-editing program, DAG-Edit (http://www.godatabase.org/dev/java/dagedit/docs/) for better visualization.

Table 1 Categorization of the "Tissue Slim"

Tissue slims	Description of "Tissue Slim" levels	Examples
Level-2	Upper-organ system level	Skeletal system, hematopoietic system, visceral system
Level-3	Lower-organ system level	Blood, lymph, hemolymph tissue, lymphoid tissue
Level-4	Upper-tissue-system level	Blood plasma, blood cancer cell, blood clot
Level-5	Lower-tissue-system level	Blood platelet, blood serum

An overview of the four different "Tissue Slim" levels available in TissueDistributionDBs, illustrated with examples at each level

As tabulated in Table 1, "Tissue Slim" at Level-2 specifies the "Upper-organ system level", and, for example, has one of the tissue type, hematopoietic system, branching from the root node whole body. The "Lower-organ system level" which constitutes the "Tissue Slim" at Level-3 includes the tissue-type blood, a sub-level from the hematopoietic system.

The "Tissue Slim" at Level-4 or "Upper-tissue-system level" which branches from the tissue-type blood includes blood plasma. This further divides into sub-levels defined as "Lower-tissue-system level" at Level-5 of the "Tissue Slim" into blood platelet and blood serum. Even though, the tissue types from all the organisms are grouped under the single "Tissue Ontology" at BRENDA with animal and plant, under two separate nodes, the organism-specific "Tissue Slims" are constructed purely for programming purposes.

2.1 Tissue-distribution calculations

TissueDistributionDBs employs normalization in the tissue-distribution calculations as in SOURCE [17]. However, we further consider the tissue types at all the four "Tissue Slims" levels to concurrently score and rank the differential expression of genes in the tissue types.

The "Tissue Expression Frequency" $f_i(t)$ of a gene i defined by a UniGene cluster in a tissue type t can be calculated as the number of ESTs in tissue type t for gene i represented as $N_i(t)$ to the total number of ESTs for the tissue type t in UniGene database denoted as $N(t)$ [17].

$$f_i(t) := \frac{N_i(t)}{N(t)}$$

The "Total Expression Level" of a gene i within each tissue type $t(\lambda_i)$, is then calculated by summing up the expression of gene i in the total number of tissue types T for a given organism represented by UniGene database [17].

$$\lambda_i := \sum_{t \in T} f_i(t)$$

"Normalized gene expression" presents the relative expression level of a gene in different tissue types, which are normalized for the number of ESTs from each tissue type that are included in UniGene database. Therefore, the normalized gene expression v_i of gene i in tissue type t is calculated by dividing the frequency of gene i in the tissue type t by the total expression for the overall organism as followed in SOURCE [17]. However, we further extend this normalization to all the four "Tissue Slims" levels.

$$v_i := \frac{f_i(t)}{\lambda_i}$$

"Most Expressed In Tissue" retrieves all the tissue types which have the highest weightage score assigned to them in each gene defined by UniGene cluster like in TissueInfo [22]. However, we further extend it to provide the tissue types in which the gene is most expressed at all the four different levels of the "Tissue Slims".

The Tissue-Specificity Index is defined as the count of the total number of different tissue types represented in the gene i UniGene cluster [16].

2.2 Access to the database repository

We use PERL scripts for performing the queries and generating the organism-specific outputs in the flat file format, which are then subsequently integrated into the sequence retrieval system (SRS) [28] installed at the German Cancer Research Center (DKFZ). SRS indexes the data from flat files and renders it as HTML pages suitable for viewing using any modern web browser. Each gene defined by a UniGene cluster has a unique identifier corresponding to the UniGene cluster identifier. Additionally gene-based, ESTs-based, cluster-based and tissue-based information are also available for each gene. The user can start a query with TissueDistributionDBs by selecting one or more organisms. A detailed user guide manual is available on the home page at http://genome.dkfz-heidelberg.de/menu/tissue_db/examples.html, to help in navigating and making queries with TissueDistributionDBs. Additionally, information about the levels at which the tissue types are assigned in the "Tissue Ontology" is also available. The results from TissueDistributionDBs can be viewed using different types of display options (Table 2) depending on what the user is looking for in the tissue-distribution profiles. For instance, the user can avail the "BarChartView" option to get a better graphical visualization of the tissue types and their distribution in a particular gene defined by UniGene cluster ID. The user can also personalize the views by selecting the desired data fields from the available fields in the database and creating a personalized view (Table 2). TissueDistributionDBs repository is automatically updated with every new release of the corresponding organism-specific UniGene database which insures up-to-date access.

3 Evaluation of the database system

In order to scrutinize TissueDistributionDBs, we considered the human dataset (ftp://ftp.expasy.org/databases/swiss-prot/special_selections/human.seq.gz) available at Swissprot database [29]. We parsed this data file and retrieved only the tissue-type information if available at the TISSUE field in the reference comment (RC) line which provides information about the source of the tissue type as cited in the literature. This information occurs optionally and specifies the expression of a gene in a particular tissue type. Even though it does not provide any quantitative information on gene expression in a particular tissue type, the source of tissue type is curated and more reliable. We used our "Tissue Synonym Library" to curate the tissue types from the RC field, so that they are compatible with the tissue-type nomenclature used in our database system.

Table 2 Views for displaying the output

An outline of the different output view's along with the contents available in each view

Sl. no.	Tissue type	Contents available
1	Tissue simple	Cluster ID, gene symbol, description, tissue-specificity index
2	Tissue simple summary	Cluster ID, gene symbol, description, tissue summary
3	Tissue level-2	Cluster ID, gene symbol, most expressed in, tissue types at level-2
4	Tissue level-3	Cluster ID, gene symbol, most expressed in, tissue types at level-3
5	Tissue level-4	Cluster ID, gene symbol, most expressed in, tissue types at level-4
6	Tissue level-5	Cluster ID, gene symbol, most expressed in, tissue types at level-5
7	BarChartView-Tissuelevel-2	Gene-wise, tissue types at level-2, percentage of expression
8	BarChartView-Tissuelevel-3	Gene-wise, tissue types at level-3, percentage of expression
9	BarChartView-Tissuelevel-4	Gene-wise, tissue types at level-4, percentage of expression
10	BarChartView-Tissuelevel-5	Gene-wise, tissue types at level-5, percentage of expression
11	Complete entry	Gene-based information, ESTs-based information, Cluster-based information, Tissue-based information
12	Name only	Cluster ID
13	Personalized view	User specified details

We randomly selected 40 different tissue types from different levels of the "Tissue Ontology" and queried Swissprot and our database system for the genes which are expressed in one or more of these 40 different tissue types. If any tissue type for a gene from Swissprot is also found in our database system irrespective of the level, we consider it as a hit for that particular tissue type, otherwise we consider that the gene is not expressed in that particular tissue type. We calculate the percentage of hits $P_{\text{hits}}(t)$ for a particular tissue type t as the number of gene hits in TissueDistributionDBs $n(t)$ divided by the total number of genes $N(t)$ from Swissprot database, which are expressed in the tissue type t multiplied by 100.

$$P_{\text{hits}(t)} := \left[\frac{n_{(t)}}{N_{(t)}}\right] \times 100$$

The results obtained showed that out of the total 19,272 entries from the organism *Homo sapiens* section of Swissprot, 17,164 number of genes were found to be expressed in at least one of the 40 randomly selected tissue types. By querying these genes against our database system, we found that on an average 86.38% of the tissue types from our database system are consistent with the tissue-type information from Swissprot database as tabulated in Table 3. The percentage of inconsistency between these two database systems may be attributed to the unavailability of the tissue-type source information for some of the genes in Swissprot or due to the incompatibility of the tissue types terms between the two databases, in spite of curating them using the "Tissue Synonym library". Furthermore, the number of hits for any particular tissue type for the 40 randomly selected tissue types is always higher than the number of tissue types which do not have any hits. This evaluation qualitatively shows that the tissue-specific expression indicated in TissueDistributionDBs reflects to a large extent what is already known about the gene expression in specific tissue system, but does not necessarily reflect its accuracy. The consistency percentage between TissueDistributionDBs and Swissprot database is tabulated in detail for all the 40 different tissue types in Table 3.

4 Implications

The quantitative gene expression profiles provided by TissueDistributionDBs at all the "Tissue Slim" levels of the "Tissue Ontology" have implications for the identification of novel therapeutic drug targets, in gene regulation, in gene discovery, as biomarkers for disease screening, and in the design and analysis of micro-arrays [30–35]. Furthermore, TissueDistributionDBs can be used to understand better the gene of interest by knowing the tissue-distribution pattern in any of the 20 model organisms.

Drugs exert their therapeutic effect by binding to the targets and modulating their activity. A large number of these targets are continuingly being explored for potential drug binding targets. These drug targets are considered to

Table 3 Validation of TissueDistributionDBs

Sl. no.	Tissue type	% of hits
1	Adipose tissue	55.33
2	Adrenal gland	45.96
3	Aorta	80.51
4	Bladder	92.89
5	Blood	73.58
6	Blood vessel	66.66
7	Bone	87.75
8	Bone marrow	64.55
9	Brain	97.69
10	Cartilage	60.52
11	Colon	89.91
12	Cornea	84.61
13	Decidua	100
14	Embryo	87.87
15	Eye	98.4
16	Head	100
17	Hematopoietic stem cell	74.07
18	Hippocampus	95.13
19	Kidney	96.21
20	Liver	86.2
21	Lung	94.12
22	Lung cancer cell	94.59
23	Muscle	92.37
24	Ovary	89.69
25	Pancreas	86.57
26	Pericardium	100
27	Peripheral nervous system	78.28
28	Placenta	92.19
29	Renal cortex	84
30	Skeletal muscle	90.17
31	Skin	96.59
32	Skin cancer cell	88.09
33	Small intestine	91.32
34	Spleen	74.14
35	Testis	92.64
36	Thalamus	96.7
37	Thymus	88.11
38	Umbilical cord	92.16
39	Urinary bladder	97.15
40	Uterus	98.54
Average	All	86.3815

The results obtained for validation of TissueDistributionDBs by performing quality check (as indexed on 30 July 2009) against Swissprot database. The column "Tissue type" specifies the randomly selected tissue types. The "% of hits" provides the percentage of tissue types that are consistent between our database system and Swissprot. On an average 86.38% of the tissue types from our database system are consistent with Swissprot. Moreover, the consistency between the two database systems is always greater than the inconsistency between them for any given tissue type

have specific tissue-expression levels that significantly distinguish them from other genes [30, 31]. The sequence-derived physicochemical properties of the currently available targets are used to predict novel drug targets. Toward this goal, TissueDistributionDBs are queried for the tissue-specificity at "Tissue Slim" Level-4. Many of the successful drug targets are distributed in less than three tissue types [30, 31]. Side effects can be avoided if drugs against a target gene expressed in only one tissue type can be used. This seems to indicate that tissue-specificity might be one of the important factors for the successful exploration of new drug targets [30, 31]. The drugs that are currently available on the market are targeted to less than 500 genes which represent only a small fraction when compared to the overall estimated number of $\sim$5,000 drug targets [32]. Therefore, there is an increasing demand for the identification of novel drug targets based on the specific properties like tissue-specificity by computational methods. In this regard, the tissue-distribution pattern of several targets is compared using the data from TissueDistributionDBs [30, 31]. Here, the quantitative differences in the tissue expression of the genes are considered at the "Tissue Slim" levels from Level-3 to Level-5 and this data is used to train the support vector machine model along with other sequence information. This model is able to accurately distinguish targets from non-targets [33]. Another similar study associated with exploring specific properties of successful drug targets also identified several quantitative measures that distinguish them from non-targets based on their tissue-specificity and many other properties. Here, the tissue-distribution pattern data from TissueDistributionDBs for the successful drug targets is useful in deriving quantitative guidelines that could aid in the computational screening of new drug targets [34]. These studies together provide a new perspective for pursuing new drug targets [33, 34].

TissueDistributionDBs is also useful in the study associated with the DNA methylation, which is an epigenetic modification important for regulating the gene expression and suppressing the spurious transcription. Often in mammalian genomes methylation occurs in the CpG dinucleotides regions, but most CpG islands are resistant to such epigenetic modification. Currently, the mechanism underlying the methylation resistance of CpG islands is poorly understood. The flanking sequences of the unmethylated CpG islands are explored using the in vivo DNA methylation data from the human brain. This has lead to the discovery of the enrichment of putative transcription factor binding sites in the flanking regions. These sites may block the spreading of methylation into these islands as they are occupied by the transcription factors which further promote transcription. This hypothesis has been tested by looking at the tissue expression patterns of the transcription factors corresponding to the enriched transcription factor binding

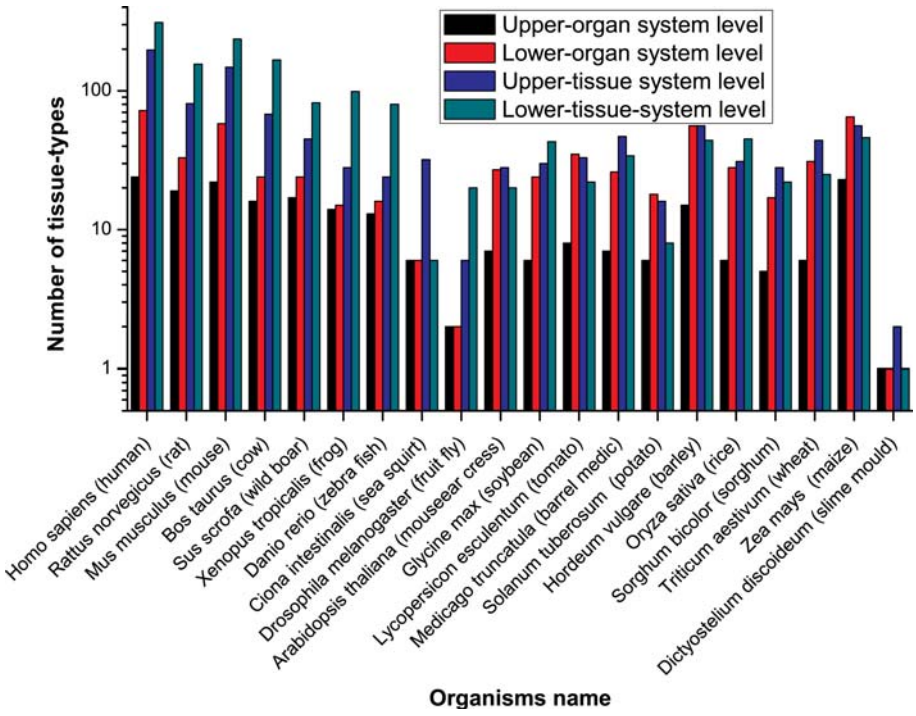

Fig. 2 The histogram of the organism-wise distribution of the tissue types shows the distribution of tissue types in 20 different model organisms on a log scale, where the tissue types at four different "Tissue Slim" levels are considered for each organism

sites in TissueDistributionDBs at the "Tissue Slim" Level-4 which contains the profiles specific to brain. Consistent with the hypothesis, the majority of the transcription factors are indeed over-expressed in brain [35].

5 Discussion

TissueDistributionDBs extends the framework for a systematic survey and analysis of the EST-based tissue-distribution profiles. The raw tissue-type terms are not curated in UniGene, EST Expression Profile Viewer, DDD and TIGER Gene Indices [36]; therefore, the natural language variations occurring in the source tissue-type terms may jeopardize the quantitative tissue-distribution calculations. TissueInfo and ExQuest like TissueDistributionDBs curate the raw tissue-type terms, however, these two resources define the relationship among the tissue type using "Tissue hierarchy" unlike TissueDistributionDBs which uses the "Tissue Ontology" to handle such relationships. The use of "Tissue Ontology" also makes TissueDistributionDBs unique from the EST Expression Profile Viewer even though both use the same source data. DDD provides statistical significance for the gene clusters which is limited by the number of ESTs being more than 1,000, therefore, does not provide information in cases where the number of ESTs per cluster are usually low as in the case of plants. Some of the tissue types of the diseased conditions are not covered by BRENDA "Tissue Ontology", we assigned the ontological levels by considering their corresponding normal tissue types and ontology levels available at BRENDA. TissueDistributionDBs extends the tissue-distribution calculations to four different anatomical levels of the "Tissue Ontology" by construction of the "Tissue Slims" while TissueInfo focuses on only one available hierarchical level. Since "Tissue Ontology" is primarily based on distinct anatomical structures, there is less need for defining the levels on the fly. Therefore, we provide only pre-defined anatomical levels of the "Tissue Slims" in TissueDistributionDBs. The benchmarking of the database system shows that the information provided by TissueDistributionDBs is consistent to a large extent (86.38%) to that of Swissprot database. Any anomalies between the two database systems can be attributed to the incompatibility between the tissue-type terms of the two databases. TissueDistributionDBs provides tissue-distribution profiles at a wider coverage range with $\sim$1,000 different tissue types at four different levels of the "Tissue Ontology" in 20 different model organisms at one platform (Fig. 2).

Acknowledgments We express our gratitude to Prof. Dr. Sándor Suhai for his ample support. We thank Vikram Alva for the technical assistance in home page setup and Andrea Mclntosh-Suhr for proofreading. We also thank all the five anonymous referees for their very helpful comments. We acknowledge the editors for their invitation to the Suhai Festschrift issue, and for their helpful comments, corrections and advice. Financial support by the German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ) is gratefully acknowledged.

References

1. Klee EW (2008) Clin Lab Med 28:127–143, viii
2. Stanton JA, Macgregor AB, Green DP (2003) Appl Bioinformatics 2:S65–S73
3. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Proc Natl Acad Sci USA 95:14863–14868
4. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Nat Biotechnol 14:1675–1680
5. Ramsay G (1998) Nat Biotechnol 16:40–44
6. Schena M, Shalon D, Davis RW, Brown PO (1995) Science 270:467–470
7. Shalon D, Smith SJ, Brown PO (1996) Genome Res 6:639–645
8. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF (1991) Science 252:1651–1656
9. Ball CA, Dolinski K, Dwight SS, Harris MA, Issel-Tarver L, Kasarskis A, Scafe CR, Sherlock G, Binkley G, Jin H, Kaloper M, Orr SD, Schroeder M, Weng S, Zhu Y, Botstein D, Cherry JM (2000) Nucleic Acids Res 28:77–80
10. Blackshear PJ, Lai WS, Thorn JM, Kennington EA, Staffa NG, Moore DT, Bouffard GG, Beckstrom-Sternberg SM, Touchman JW, Bonaldo MF, Soares MB (2001) Gene 267:71–87
11. Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WR, Wilson SA, Hubbard SJ (2002) Curr Biol 12:1965–1969
12. Boguski MS, Lowe TM, Tolstoshev CM (1993) Nat Genet 4:332–333
13. Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) Nucleic Acids Res 28:141–145
14. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J (2001) Nucleic Acids Res 29:82–86
15. Bortoluzzi S, Danieli GA (1999) Trends Genet 15:118–119
16. Vasmatzis G, Essand M, Brinkmann U, Lee B, Pastan I (1998) Proc Natl Acad Sci USA 95:300–304
17. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA (2003) Nucleic Acids Res 31:219–223
18. Schuler GD (1997) J Mol Med 75:694–698
19. Strausberg RL, Dahl CA, Klausner RD (1997) Nat Genet 15(Spec No):415–416
20. Kawamoto S, Matsumoto Y, Mizuno K, Okubo K, Matsubara K (1996) Gene 174:151–158
21. Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K (1992) Nat Genet 2:173–179
22. Skrabanek L, Campagne F (2001) Nucleic Acids Res 29:E102
23. Brown AC, Kai K, May ME, Brown DC, Roopenian DC (2004) Genomics 83:528–539
24. Christoffels A, van Gelder A, Greyling G, Miller R, Hide T, Hide W (2001) Nucleic Acids Res 29:234–238
25. Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA, Smith B, Day-Richter J, Noy NF, Musen MA (2006) OMICS 10:185–198
26. Schomburg I, Chang A, Schomburg D (2002) Nucleic Acids Res 30:47–49
27. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) Nucleic Acids Res 34:D16–D20
28. Etzold T, Ulyanov A, Argos P (1996) Methods Enzymol 266:114–128
29. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) Nucleic Acids Res 31:365–370
30. Zheng CJ, Han LY, Yap CW, Ji ZL, Cao ZW, Chen YZ (2006) Pharmacol Rev 58:259–279
31. Zhu F, Han L, Zheng C, Xie B, Tammi MT, Yang S, Wei Y, Chen Y (2009) J Pharmacol Exp Ther 330:304–315
32. Drews J (2000) Science 287:1960–1964
33. Xu H, Xu H, Lin M, Wang W, Li Z, Huang J, Chen Y, Chen X (2007) Proteomics 7:4255–4263
34. Yao L, Rzhetsky A (2008) Genome Res 18:206–213
35. Fan S, Fang F, Zhang X, Zhang MQ (2007) PLoS One 2:e1184
36. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) Nucleic Acids Res 29:159–164