BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2328–2333
doi:10.1093/bioinformatics/btq417

Databases and ontologies Advance Access publication July 13, 2010

# C-It: a knowledge database for tissue-enriched genes

Pascal Gellert, Katharina Jenniches, Thomas Braun* and Shizuka Uchida*

Max-Planck-Institute for Heart and Lung Research, Ludwigstr. 43, 61231 Bad Nauheim, Germany

Associate Editor: Jonathan Wren

**ABSTRACT**

**Motivation:** Due to the development of high-throughput technologies such as microarrays, it has become possible to determine genome-wide expression changes in a single experiment. Although much attention has been paid to identify differentially expressed genes, the functions of tens of thousands of genes in different species still remain unknown.

**Results:** C-It is a knowledge database that has its focus on 'uncharacterized genes'. C-It contains expression profiles of various tissues from human, mouse, rat, chicken and zebrafish. By applying our previously introduced algorithm DGSA (Database-Dependent Gene Selection and Analysis), it is possible to screen for uncharacterized, tissue-enriched genes in the species mentioned above. C-It is designed to include further expression studies, which might provide more comprehensive coverage of gene expression patterns and tissue-enriched splicing isoforms. We propose that C-It will be an excellent starting point to study uncharacterized genes.

**Availability:** C-It is freely available online without registration at http://C-It.mpi-bn.mpg.de

**Contact:** thomas.braun@mpi-bn.mpg.de; shizuka.uchida@mpi-bn.mpg.de

Received on May 14, 2010; revised on July 2, 2010; accepted on July 8, 2010

## 1 INTRODUCTION

The development of high-throughput technologies such as DNA microarrays and Serial Analysis of Gene Expression (SAGE) has revolutionized profiling of complex transcriptomes, enabling researchers to determine the transcription level of all known, unknown or putative genes within one experiment. During the last years, numerous genome-wide expression studies were performed to analyze the regulation of genes in different developmental stages, diseases or tissues. Due to the rapidly decreasing costs of next-generation sequencing, it is expected that the number of such expression studies will dramatically increase within the next years.

Differentially expressed genes identified by these experiments became a starting point to characterize putative functions, interactions and biological roles of identified genes. Recent studies reveal that the number of genes encoded in higher organism (e.g. human) is not as high as initially expected; the current number of protein-coding genes is predicted to be between 20 000 and 25 000 (Collins *et al.*, 2004). Yet, a considerable number of putative genes is still not characterized (Uchida *et al.*, 2009), which is in part due to the current mode of data processing. Researchers tend to exclude genes with unknown annotations from further analysis even if they show a specific expression profile suggesting a distinct function in the process under examination (Pawłowski, 2008). Pawłowski (2008) assumes that one of the reasons for this bias is caused by the tendency to focus on certain well-known signaling pathways. Such a focus facilitates further evaluation and manipulation of processes under study because of available knowledge and tools. However, there is no doubt that characterization of unknown genes is needed to obtain a comprehensive understanding of the underlying biology in a system-wide manner (Ideker *et al.*, 2001; Kitano, 2002).

C-It is a knowledge database focusing on uncharacterized genes to build a starting point for biologists to study genes with unknown functions. The database implements literature information from the PubMed database to identify genes that lack publication records. Based on the assumption that genes are likely to fulfill important functions when their expression is enriched in a certain tissue, C-It uses the tissue expression information of UniGene (Wheeler *et al.*, 2003) expressed sequence tags (EST) profiles to identify tissue-enriched genes. Since evolutionary conservation of gene expression profiles usually indicates important functions of genes in a specific tissue, we implemented our algorithm 'Database-dependent Gene Selection and Analysis' (DGSA; Uchida *et al.*, 2009). This algorithm identifies tissue-enriched genes by using EST profiles in all available tissues of organisms that biologists typically use for their research [*Mus musculus* (mouse), *Rattus norvegicuis* (rat), *Gallus gallus* (chicken) and *Danio rerio* (zebrafish)] as well as human (*Homo sapiens*). C-It combines microarray and SAGE data to give users integrated access to comprehensive transcriptional profiles. Furthermore, C-It is linked with a custom version of the exon array analyzer (EAA; Gellert *et al.*, 2009) to allow tissue-enriched alternative splicing analysis.

## 2 METHODS

### 2.1 Identification of tissue-enriched genes

The UniGene profiles from human, mouse, rat, chicken and zebrafish were downloaded from the NCBI ftp-server (UniGene Build #222, #182, #181, #42, #117, respectively). To take into account differences in the sampling depth (the total number of ESTs) of each tissue, the EST counts of each UniGene cluster were normalized by the total EST count of the tissue to transcripts per million (TPM) and imported into MySQL tables for displaying the gene expression across all tissues. Our DGSA algorithm was applied to these normalized TPM values. In brief, the DGSA defines a gene as enriched in a tissue if the tissue is in the top 20% compared with all UniGene tissues.

*To whom correspondence should be addressed.

 

In other words, the TPM of the tissues of each UniGene cluster were sorted and the highest 20% tissues were considered as enriched for the cluster. Additionally, our 'further rules' which are part of the DGSA were applied to all UniGene clusters. The additional rules exclude genes that are ubiquitously expressed in multiple organs: (i) genes must be expressed in <50% of total tissues; (ii) sum of gene EST must be <400; and (iii) at least 25% of gene EST must be expressed in the target tissue. The DGSA is less conservative than other approaches, for example, the one developed by Audic and Claverie (1997) identifies tissue-'specific' rather than tissue-'enriched' genes. This is more applicable to C-It because its intention is to provide genes that are highly expressed in a tissue of interest across several species.

### 2.2 Homolog gene clusters

To identify homologs across the selected five organisms, the HomoloGene database (http://www.ncbi.nlm.nih.gov/homologene, Release 64) from NCBI was utilized. Because HomoloGene does not include UniGene cluster IDs, we employed the 'gene2unigene' dataset from NCBI to map the clusters to their corresponding Entrez Gene ID (GI). In the case of ambiguous mapping from GI to UniGene, C-It uses the rank of the 'best' UniGene cluster. This is, depending on the user-defined search option, the cluster with the highest or lowest rank (searching for enriched or depleted genes, respectively).

### 2.3 Additional expression studies

For microarray data, raw data were downloaded from the BioGPS project (http://biogps.gnf.org/; Lattin *et al.*, 2008; Su *et al.*, 2002; Walker *et al.*, 2004) and normalized by RMA provided by the R package 'affy' (Gautier *et al.*, 2004). The SAGE libraries from the Mouse Atlas of Gene Expression (Siddiqui *et al.*, 2005) were downloaded and normalized to TPM as described above. SAGE tags were mapped to their corresponding genes by using SAGEgenie (Boon *et al.*, 2002). To identify tissue-enriched isoforms, C-It was linked to a custom version of the EAA with datasets from various tissues downloaded from the Affymetrix web site (http://www.affymetrix.com) and the BioGPS project (Wu *et al.*, 2009).

### 2.4 Literature data

To determine the characterization level of a gene, literature information based on the PubMed database was obtained through the 'gene2pubmed' dataset provided by NCBI. As it has been defined previously (Uchida *et al.*, 2009), publications that include more than 100 GIs were considered as articles that report screening results (e.g. microarrays, sequencing) and excluded from the publication counts. To identify publications that deal with specific tissues, datasets for the linkage between the PubMed IDs and a collection of tissue terms based on the MeSH were downloaded from the MeSH database.

### 2.5 Implementation

The web server itself runs on Ubuntu 8.04 with an Apache 2 server using PHP and MySQL. An automated script has been set up to update UniGene expression profiles and literature information every 2 months.

### 2.6 Reverse transcription polymerase chain reaction (RT-PCR)

RT-PCR experiments are described elsewhere (Uchida *et al.*, 2009).

## 3 RESULTS

Despite major attempts to understand the cellular function of genes in the heart a large number of heart-enriched genes exist, for which no specific function has been assigned (Uchida *et al.*, 2009). To determine the total number of uncharacterized genes, we utilized the 'gene2pubmed' dataset and found that >26% of the human protein-coding genes in the NCBI Entrez database have no publication in PubMed. Another recently published study yielded a slightly higher number of genes without publication using a similar approach (Wren, 2009). Reduction of stringency criteria by three publications per gene increases the proportion of uncharacterized genes to over 41% (Table 1). In other words: a significant number of all protein-coding genes are only poorly characterized, which emphasizes the need to develop new tools for further analysis.

**Table 1.** Uncharacterized genes

| Organism | Number of genes | Without publication | Less than three publications |
|---|---|---|---|
| Human | 26 992 | 7212 (26.7%) | 11 115 (41.2%) |
| Mouse | 36 268 | 8461 (23.3%) | 11 751 (32.4%) |
| Rat | 26 136 | 12 884 (49.2%) | 22 156 (84.8%) |
| Chicken | 18 691 | 13 906 (74.4%) | 18 265 (97.7%) |
| Zebrafish | 33 359 | 18 760 (56.2%) | 31 280 (93.8%) |

The number of protein-coding genes for each organism based on the NCBI database.

Construction of C-It was based on EST expression profiles from the UniGene database, which were used to identify tissue-enriched genes employing our previously introduced DGSA (Uchida *et al.*, 2009) to all available tissues. We included the expression profiles of human, mouse and rat as well as the more distantly related organisms, chicken and zebrafish, into C-It. By utilizing HomoloGene, C-It displays evolutionary-conserved, tissue-enriched genes in up to four organisms. Expression patterns of genes in functionally similar tissues are often conserved between different organisms, which often indicates that these genes fulfill important biological functions in the respective tissues. Hence, the identification of evolutionary-conserved, tissue-enriched genes serves as an additional filter to exclude false positive (i.e. non-expressed) genes. In addition, C-It provides the possibility to exclude genes that are expressed in other tissues. This feature can be used, for example, to filter genes that are enriched in two similar tissues.

Literature information from the PubMed database provided a measure for uncharacterized genes. A gene with no or only few published articles was defined as 'uncharacterized', while we assume that the function of a gene mentioned in many articles is known. Of course, this approach cannot rule out that a gene is mentioned in articles without further investigation or has been studied under a synonym not listed in NCBI. However, we reason that this approach usually gives a good estimation of a gene at the level of characterization. In addition to PubMed, we used the Medical Subject Headings (MeSH) classification to identify articles related to a specific tissue. This approach allows a user to filter for genes that are uncharacterized with respect to the tissue of interest.

To gain further information on gene expression patterns, large-scale transcriptomics studies can be added. Furthermore, pre-calculated exon array data can be viewed to identify tissue-enriched alternative splicing isoforms. Using this information, it is possible to obtain a comprehensive overview of the expression patterns of uncharacterized genes. All data mentioned above were integrated into C-It for querying by the web interface (see Table 2 for all

**Table 2.** Data sources included in C-It

| Resource | Type | Organism | Reference/URL |
|---|---|---|---|
| UniGene | EST profiles | Human, Mouse, Rat, Chicken, Zebrafish | Wheeler *et al.* (2003) http://www.ncbi.nlm.nih.gov/unigene |
| PubMed | Literature | Human, Mouse, Rat, Chicken, Zebrafish | http://www.ncbi.nlm.nih.gov/pubmed |
| BioGPS | Micorarrays | Human, Mouse, Rat | Lattin *et al.* (2008); Su *et al.* (2002); Walker *et al.* (2004) http://biogps.gnf.org/ |
| Mouse Atlas of Gene Expression | SAGE libraries | Mouse | Siddiqui *et al.* (2005) http://www.mouseatlas.org |
| Affymetrix | Exon Arrays | Human, Mouse, Rat | http://www.affymetrix.com |
| BioGPS | Exon Arrays | Mouse | Wu *et al.* (2009) http://biogps.gnf.org/ |

To build C-It, we acquired gene expression studies and literature information from various sources.

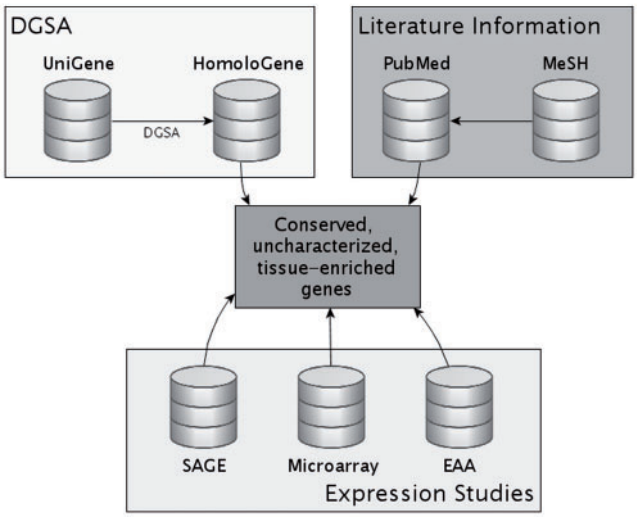


**Fig. 1.** Database scheme of C-It. Our DGSA was applied to the UniGene profiles of five organisms by using information from the HomoloGene database to identify homologs; this step allows identification of evolutionary-conserved genes. PubMed articles and MeSH terms were used to classify uncharacterized genes. Additional expression studies from microarrays, SAGE and exon arrays were included to build a comprehensive source of expression analyses.

acquired data sources). A simplified overview of the database scheme is shown in Figure 1.

### 3.1 Web interface

The web interface of C-It enables screening for uncharacterized genes in various different tissues. A list of uncharacterized, tissue-enriched and species-conserved genes can be generated in four steps: first, a user must select a tissue of interest from the top page of C-It. Currently, 78 tissues are available in the UniGene profiles of five organisms used in this study. Depending on the availability of the selected tissues in five organisms, it is possible to choose up to four organisms for further analysis. This step also requires to set thresholds for the maximum number of articles published about the gene. The default settings are less than five publications and less than three publications listed in the Medical Subject Headings (MeSH) category of the selected tissue.

After all parameters are specified, C-It generates a Venn diagram showing the number of uncharacterized, tissue-enriched genes in user-defined organisms (Fig. 2). The user can deselect previously chosen organism(s) to define which overlap of the Venn diagram will be shown in the list. If no organism is deselected, genes enriched in all selected organism will be displayed. To gain more evidence about the tissue enrichment, other transcriptomic data from exon arrays, microarrays and SAGE data can be added to the list. The final list can be sorted by each column and is linked to expression diagrams for all five organisms. These diagrams show all expression profiles of the selected gene for a direct comparison between different tissues and, if available, across SAGE and microarrays. If a gene maps to several UniGene clusters, SAGE tags or micorarray probe sets, the expression of all datasets is shown in the diagrams. In the case of exon array data, the number of differentially expressed exons, if any, for the selected gene is linked to a custom version of EAA, which enables identification of tissue-enriched isoforms. For better handling, the final list can be downloaded as a comma-separated text file or opened in a new window without surrounding boxes. The work flow of C-It is schematically shown in Figure 3.

In addition to the above described work flow, C-It can also be queried by single gene or a set of genes. To quickly obtain the expression profiles of a gene of interest, we implemented an option to search by GI or Symbol. C-It displays the number of publications and all available expression studies of the queried gene and its homologs, if available. Another possibility is to query C-It for a set of genes by uploading a list of GIs. C-It uses these genes as starting point. In the following steps, the list can be filtered for tissue-enrichment and publication records, similar to the work flow described above.

### 3.2 Validation

We previously validated the power of DGSA for heart-enriched genes (Uchida *et al.*, 2009). In this study, we searched for genes that are uncharacterized in mouse and enriched in brain or liver using C-It. The parameters for literature information were set to the maximum stringency, therefore only genes without any publication published, both with and without the MeSH filter, were considered.

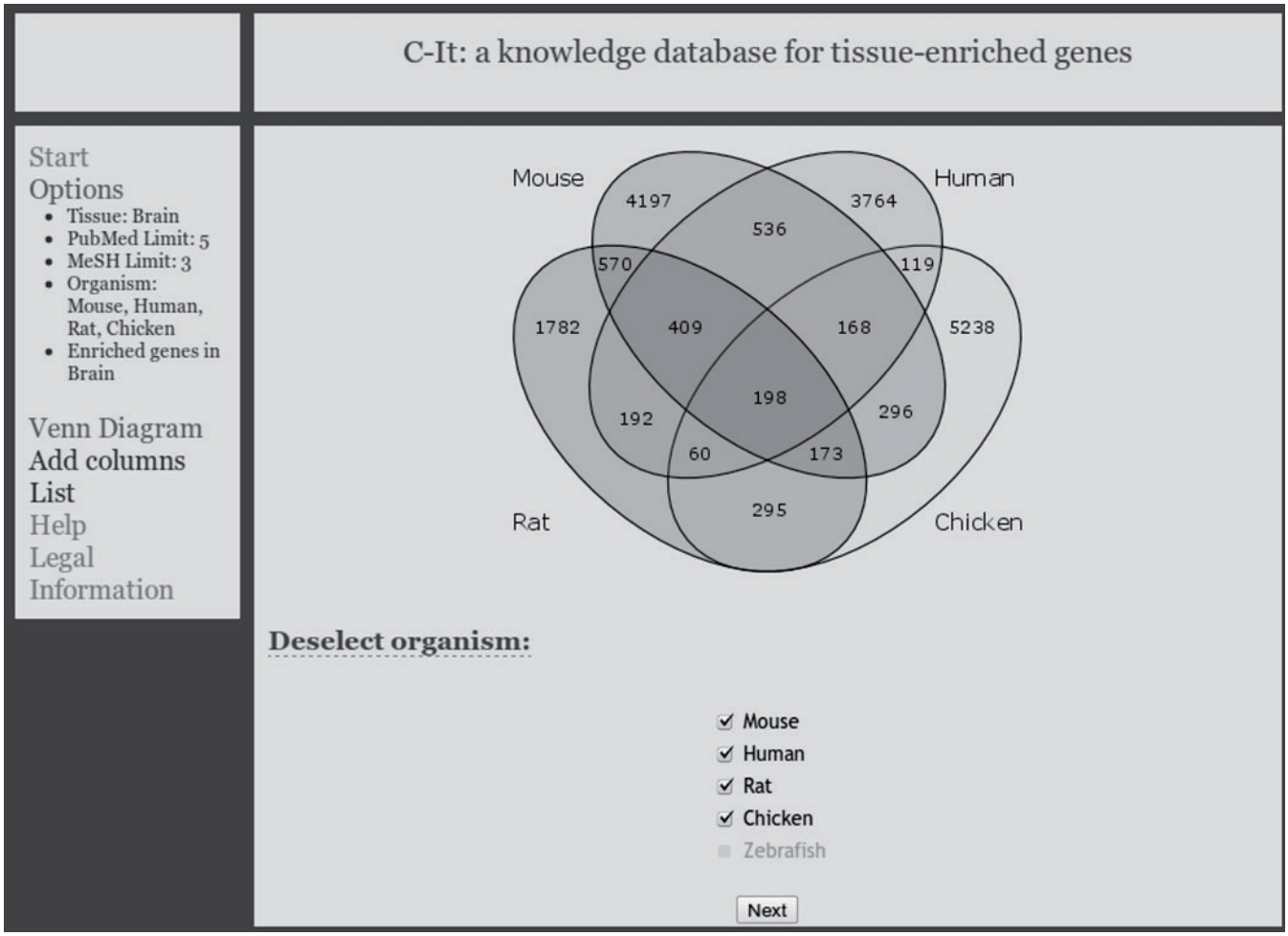


**Fig. 2.** Screenshot of C-It showing a Venn diagram of genes that are enriched in the brain of human, mouse, rat and chicken. Default settings for published articles for each gene were used. This step allows modifications of the list of genes by deselecting individual organism.

RT-PCR analysis of the eight randomly selected genes revealed strong enrichments in selected tissues compared with 14 other murine adult tissues (Fig. 4).

### 3.3 Comparison to other publicly available databases

There are a number of publicly available databases that allow identification of tissue-specific expression of individual genes. Some of these databases, such as TiGER (Liu *et al.*, 2008), BodyMap (Hishiki *et al.*, 2000) and TissueDistributionDBs (Kogenaru *et al.*, 2009), utilize EST profiles, while others such as GeneSpeed (Kutchma *et al.*, 2007) use micorarray data. These databases use different approaches to identify the tissue-specificity or tissue-enrichment of genes. C-It uses datasets from EST profiles, SAGE libraries and microarrays to eliminate biases arising from different technologies, thus providing a comprehensive overview of the transcriptomic data. The currently available databases, e.g FastDB (de La Grange *et al.*, 2007) and ASTD (Koscielny *et al.*, 2009), which list alternative splicing isoforms, are mostly based on sequencing data. In addition, the USCS Genome Browser (Kent *et al.*, 2002) and the BioGPS project include heatmaps showing differences in exon array signals across several tissues. However, these databases do not apply appropriate algorithms to identify differentially expressed exons; thus, it remains difficult to locate alternatively spliced transcripts. Our C-It integrates exon array data calculated by the EAA (Gellert *et al.*, 2009) to identify differentially expressed exons in a tissue-enriched manner. Furthermore, C-It is the first knowledge database focusing on uncharacterized genes. To the best of our knowledge, no other database includes literature information to identify genes without known functions.

## 4 CONCLUSIONS

In this study, we describe the construction of the new knowledge database C-It, which focuses on genes without known functions. We employed expression profiles of the UniGene database of human, mouse, rat, chicken and zebrafish and applied our DGSA algorithm to each profile to identify genes that show tissue enrichment in various species. Literature information was used to identify genes that lack assigned functions. We included different types

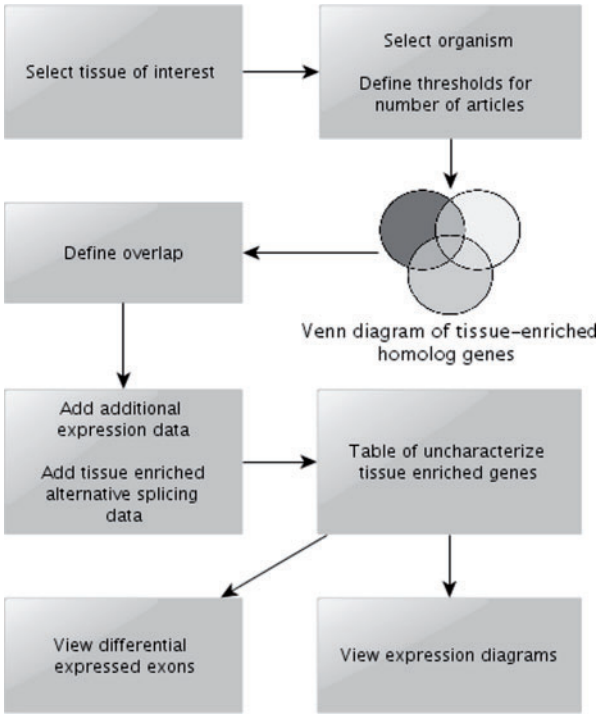


**Fig. 3.** Flow chart of C-It. After selecting the tissue of interest, the organisms and a threshold for uncharacterized genes, C-It generates a Venn diagram showing overlaps between homologous genes. Additional information (e.g. microarray expression and splicing isoforms) can be added to this list of overlapping genes. The final list is linked to a custom version of EAA to allow tissue-enriched isoform analyses and generation of expression diagrams.

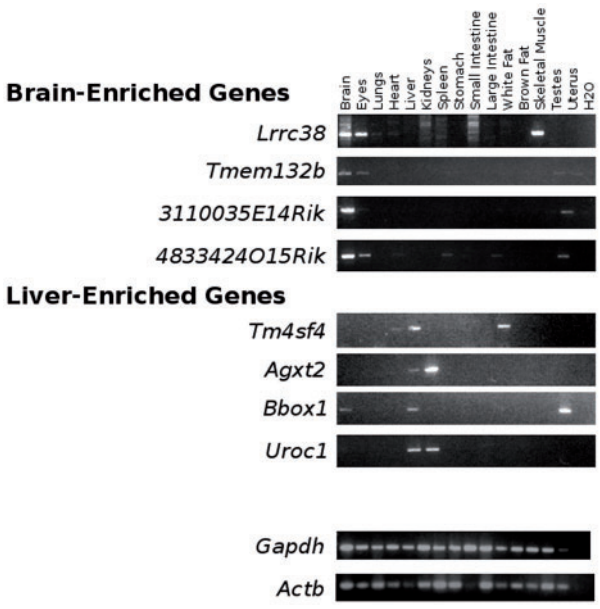


**Fig. 4.** RT-PCR results for eight randomly selected brain- or liver-enriched, uncharacterized genes using 15 adult organs from the mouse.

of transcriptional studies into C-It to provide a comprehensive overview of gene expression profiles. In addition, exon array data were added to identify tissue-enriched splicing isoforms. C-It provides an easy-to-use, one-stop-shop for biologists to study uncharacterized genes. C-It is freely available without registration at http://C-It.mpi-bn.mpg.de.

## ACKNOWLEDGEMENTS

The authors wish to thank Dr Jose C. Clemente and Dr Petra Uchida for their valuable advice and comments on this article.

*Funding*: Start-up-grant of the Excellence Cluster Cardio-Pulmonary System (ECCPS) (to S.U.); a fellowship of the International Max Planck Research School for Heart and Lung Research (IMPRS-HLR) (to K.J.); the Max Planck Society, the DFG (Br1416); the EU Commission (MYORES network of excellence); the Kerckhoff-Foundation and the Excellence Initiative 'Cardiopulmonary System' (to T.B.).

*Conflict of Interest*: none declared.

## REFERENCES

Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.

Boon,K. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287.

Collins,F. *et al.*; International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

de La Grange,P. *et al.* (2007) A new advance in alternative splicing databases: from catalogue to detailed analysis of regulation of expression and function of human alternative splicing variants. *BMC Bioinformatics*, **8**, 180.

Gautier,L. *et al.* (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Gellert,P. *et al.* (2009) Exon array analyzer: a web interface for affymetrix exon array analysis. *Bioinformatics*, **25**, 3323–3324.

Hishiki,T. *et al.* (2000) BodyMap: a human and mouse gene expression database. *Nucleic Acids Res.*, **28**, 136–138.

Ideker,T. *et al.* (2001) A new approach to decoding life: systems biology. *Ann. Rev. Genomics Hum. Genet.*, **2**, 343–372.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Kogenaru,S. *et al.* (2009) TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles. *Theor. Chem. Acc.*, **125**, 651–658.

Koscielny,G. *et al.* (2009) ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, **93**, 213–220.

Kutchma,A. *et al.* (2007) GeneSpeed: protein domain organization of the transcriptome. *Nucleic Acids Res.*, **35**, D674–D679.

Lattin,J.E. *et al.* (2008) Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res.*, **4**, 5.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Pawłowski,K. (2008) Uncharacterized/hypothetical proteins in biomedical 'omics' experiments: is novelty being swept under the carpet? *Brief. Funct. Genomics Proteomics*, **7**, 283–290.

Siddiqui,A.S. *et al.* (2005) A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing c57bl/6j mouse tissues and cells. *Proc. Natl Acad. Sci. USA*, **102**, 18485–18490.

Su,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465.

Uchida,S. *et al.* (2009) An integrated approach for the systematic identification and characterization of heart-enriched genes with unknown functions. *BMC Genomics*, **10**, 100.

Walker,J.R. *et al.* (2004) Applications of a rat multiple tissue gene expression data set. *Genome Res.*, **14**, 742–749.

Wheeler,D.L. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.

Wren,J.D. (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics*, **25**, 1694–1701.

Wu,C. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.