FEBS Letters 559 (2004) 22–26 FEBS 28044

Absolute expression values for mouse transcripts: re-annotation of the READ expression database by the use of CAGE and EST sequence tags

Rimantas Kodzius[a], Yonehiro Matsumura[b,c], Takeya Kasukawa[c,d], Kazuro Shimokawa[c], Shiro Fukuda[c], Toshiyuki Shiraki[c], Mari Nakamura[c], Takahiro Arakawa[c], Daisuke Sasaki[c], Jun Kawai[a,c], Matthias Harbers[e], Piero Carninci[a,c,*], Yoshihide Hayashizaki[a,c]

[a] *Genome Science Laboratory, RIKEN, Wako Main Campus, Hirosawa 2-1, Wako, Saitama 351-0198, Japan*
[b] *Division of Genomic Information Resource Exploration, Science of Biological Supramolecular Systems, Graduate School of Integrated Science, Yokohama City University, 1-7-29 Suehiro-Cho, Tsurumi-Ku, Yokohama, Kanagawa 230-0045, Japan*
[c] *Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan*
[d] *Network Service Solution Business Group, NTT Software Corporation, 209 Yamashita-cho, Naka-ku, Yokohama, Kanagawa 231-8551, Japan*
[e] *K.K. Dnaform, Tsukuba Branch, 3-1 Chuo 8-chome, Ami-machi, Inashiki-gun, Ibaraki 300-0332, Japan*

Received 30 October 2003; revised 26 December 2003; accepted 5 January 2004

First published online 15 January 2004

Edited by Takashi Gojobori

Abstract The RIKEN expression array database (READ) provides comprehensive gene expression data for the mouse, which were obtained as relative values from microarray double-staining experiments with E17.5 mRNA as common reference. To assign absolute expression values for mouse transcripts within READ, we applied the E17.5 reference sample to CAGE (cap analysis of gene expression) and expressed sequence tag (EST) high-throughput tag sequencing. Newly assigned values within the READ database were validated by comparison to expression data from serial analysis of gene expression, CAGE and EST experiments. These experiments confirmed the great significance of the absolute expression values within the improved READ database. The new Absolute READ database on absolute expression data is available under http://genome.gsc.riken.jp/absolute.

Key words: Gene expression; Cap analysis of gene expression; Serial analysis of gene expression; Microarray; Absolute value; Expressed sequence tag count

1. Introduction

The RIKEN mouse encyclopedia project led to the cloning and full-length sequencing of 60 770 cDNA clones [1]. As most of those cDNA clones comprised uncharacterized transcripts at the time of their discovery, further experiments were undertaken to confirm the expression of most of the RIKEN clones in high-throughput expression profiling studies. The cDNA-based microarray experiments provided one of the most comprehensive datasets on genes expressed in mouse and are available as part of the FANTOM (functional annotation of mouse) and READ (RIKEN expression array database) databases [2].

Commonly microarrays are used in double-labeling experiments providing relative expression levels between two samples [3,4]. The use of two probes having distinct labels overcomes major problems caused by different printing and hybridization efficiency, where the reference should comprise all transcripts present on the microarray to offer reliable data. References commonly in use comprise oligonucleotides [5], genomic DNA [6,7], or exogenous controls from a different organism [8].

Besides double-labeling approaches, single-labeling methods are in use [6,9], which aim at direct measurements of absolute expression values in combination with special software solutions [10,11]. Alternatively, several attempts have been undertaken to measure absolute expression values [5] or to calibrate absolute expression data from double-labeling experiments [12]. However, those efforts are often limited by the insufficient information available on the reference data, and comparison of absolute and relative numbers derived from different approaches is problematic, as computational analysis of the expression values depends on the consistency of the input data [13].

Due to the limitations of microarray studies, other methods for expression profiling are in use, which focus on the collection of sequence tags to allow further for the discovery of new transcripts while providing at the same time absolute expression values. Such approaches include EST (expressed sequence tag) sequencing [14,15], SAGE (serial analysis of gene expression) [16] or the novel CAGE method (cap analysis of gene expression) [17]. CAGE was developed to deliver a large number of expression data points and to allow for the identification of transcriptional start sites. In addition, CAGE offers all the advantages of SAGE and EST sequencing including the

*Corresponding author. Fax: (81)-48-4624686.
E-mail address: rgscerg@gsc.riken.jp (P. Carninci).

Abbreviations: CAGE, cap analysis of gene expression; E17.5, mouse embryo whole body mixed sex E17.5 RNA; EST, expressed sequence tag; FANTOM, functional annotation of mouse; READ, RIKEN expression array database; RTS, representative transcript set; SAGE, serial analysis of gene expression; tpm, transcripts per million; TU, transcriptional unit

0014-5793/04/$30.00
doi:10.1016/S0014-5793(04)00018-3

detection of rare and novel transcripts and providing absolute expression values.

The READ database contains information on 50 tissues, where relative gene expression levels are provided as compared to expression in an E17.5 mRNA (mouse embryo whole body mixed sex embryonic day 17.5 RNA) [2]. For better compatibility to other expression profiling data, however, the READ values have to be converted into absolute values on a common basis. Therefore, we undertook additional experiments to re-annotate the READ dataset by converting the relative expression values into absolute expression data based on high-throughput sequencing of the reference sample. Here we describe the thorough characterization of reference E17.5 mRNA as used in the READ expression project and the assignment of absolute values to the READ dataset. The newly assigned expression values were further confirmed by comparison to expression data obtained by other experimental means.

2. Materials and methods

All tissues were obtained from mouse strain C57BL/6J [2]. The cDNA library from E17.5 mRNA was prepared according to [18], and all other RIKEN cDNA libraries and their sequencing have been described in [19]. cDNA libraries used for statistical analyses were non-normalized, non-subtracted and non-fractionated libraries. CAGE libraries from E17.5, whole brain, and cerebellum mRNA were prepared according to [17]. To minimize polymerase chain reaction (PCR) bias in GC-rich regions of 5′ untranslated regions, we used dimethylsulfoxide as additive, supplied reaction mixtures with excess of dNTP and Taq polymerase and kept PCR cycles to a minimum. Additional expression data were obtained from different public databases including a SAGE kidney library [20,21], together with E17.5 [22] and placenta [23] libraries from EST databases [24,25]. Accession numbers and counts for all clones in EST and SAGE libraries can be downloaded from http://genome.gsc.riken.jp/absolute. A tpm (transcripts per million) value for a specified transcriptional unit (TU) was calculated by counting appearance of CAGE tags (or corresponding SAGE tags, ESTs) for a given TU, divided by the total number of TU counts obtained from a particular tissue and normalized per million.

The representative transcript set (RTS) [1,26] consists of mouse transcripts representing TUs that were defined as regions within the genome based on public databases (GenBank, RefSeq and Ensembl). RTS was used as reference to cluster sequence tags and transcripts to TU (T. Kasukawa, in preparation). This approach was applied to all transcripts in READ and tags from EST, SAGE and CAGE libraries. CAGE tags were processed as described in [17], assigned to TUs by mapping to the mouse genome version UCSC mm3 (University of California Santa Cruz, *Mus musculus* 2003) and annotated by searching for the nearest TU within a 10 kb window within the mouse genome. EST sequence tags were assigned to corresponding TUs as described for CAGE, whereas in case of the SAGE library, a list of GenBank accession numbers and corresponding counts was downloaded from the public domain. Correlation to RTS was achieved using 'RTS correspondence tables to GenBank' (T. Kasukawa, in preparation). It should be noted that SAGE data could be mis-assigned due to possible multiple affiliations, which may lower the correlation to other expression data.

To convert relative into absolute values we used normalized data from the READ database, as obtained by PRIM (preprocessing implementation for microarray) [27]. Genes found lowly expressed in CAGE or EST experiments (fewer than three annotated tags) were not included in statistical analysis. Statistical analysis was performed both considering TU and tpm counts in the libraries as independent variables assuming a linear dependence between datasets describing the same sample. Covariance and correlation (Pearson) coefficients were obtained. Absolute values for TUs using E17.5 sequence tags and READ values were calculated according to: $G = \mathrm{E17.5} \times 2^A$, where G stands for absolute value for specific TU, A for relative logarithmic expression value in READ, and E17.5 for tpm values with the E17.5 standard. Data from different sources were compared to the newly assigned READ data on the basis of tpm values for the same TU as depicted in Fig. 1.

3. Results

3.1. High-throughput analysis of E17.5 transcripts

To determine absolute expression values on transcripts expressed in the reference E17.5 sample, a cDNA library was prepared from which 49 806 5′-EST reads could be obtained, and grouped into 7164 unique TUs by RTS. Here we decided to apply the TU/RTS system as a common reference system for our studies on the mouse transcriptome, as it is today the most comprehensive transcript set based on genomic data and incorporating information from GenBank, RefSeq and Ensembl (T. Kasukawa, in preparation). To achieve a much deeper sequence coverage than possible by conventional EST sequencing, a CAGE E17.5 library [17] was obtained, which provided an additional 86 555 CAGE tags equivalent to 7507 different TUs. In combination with the EST and CAGE tags, 10 675 unique TUs could be identified as transcribed in an E17.5 embryo at this level of sequencing coverage. Sequence tags related to E17.5 transcripts were further analyzed by statistical means to obtain numerical values describing their distribution within the dataset. Out of this analysis we defined tpm values for each transcript to describe their absolute expression values. As for E17.5, tpm values were used throughout this study including new values calculated from the READ dataset, as we feel that they could become a general unit for measuring absolute expression values. Tpm values ranked from 7 tpm for rarely expressed genes, e.g. cyclin K (READ ID A330093M23) to 81 299 tpm for hemoglobin as the most abundant transcript (READ ID 1020007M19) in E17.5. To further confirm data consistency, we compared TU frequency using tpm values as obtained from EST and CAGE E17.5 libraries, where a high linear correlation between the two data sets was observed (correlation coefficient: 0.709), indicating that the additional manipulations during CAGE library preparation did not affect the quantitative values derived from it.

3.2. Assignment of absolute expression values to READ database

The READ database contains expression data for transcripts related to 57 931 cDNA clones used for microarray printing. Sequence information derived from those clones was subjected to the same grouping procedure as used for the sequencing tags from E17.5 resulting in 22 406 unique TUs. TUs represented by READ were compared to TUs found by EST and CAGE sequencing of E17.5, where overlapping TUs were identified in 7164 and 7507 cases respectively. The combination of both datasets covered 8845 unique TUs within READ (~40% of the READ TUs), out of which only about 20% of the TUs within READ were covered by more than three tags found in E17.5. The 'E17.5 standard' reference set was defined combining tags from E17.5 EST and CAGE libraries, and then applied to convert individual relative logarithmic values for a given TU within READ into absolute tpm values by applying the formula $G = \mathrm{E17.5} \times 2^A$. This formula was derived from the formula used in the past to generate READ microarray data $A = \log_2(G/\mathrm{E17.5})$ [27]. As an example for the range covered by the approach, in the case of lung, newly assigned tpm values in

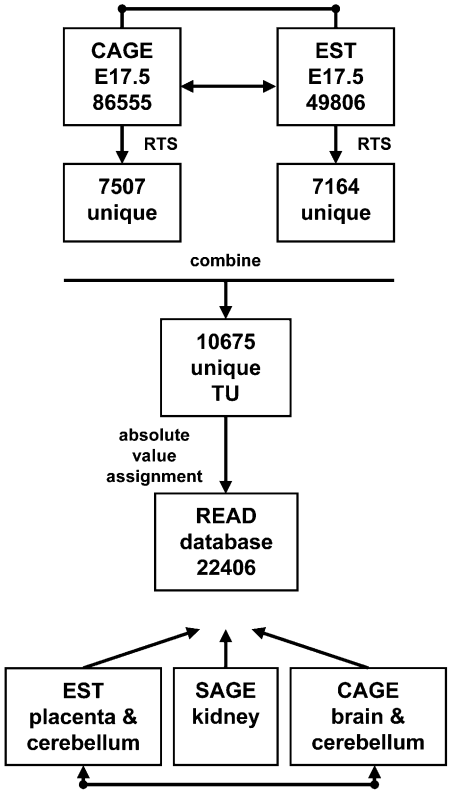

Fig. 1. Schematic representation for absolute value assignment and validation. CAGE and EST data for E17.5 and cerebellum libraries were compared to ensure data consistency. Two E17.5 libraries were merged into one reference library. After absolute values assignment, new READ values were validated by comparison to EST, SAGE and CAGE libraries.

Absolute READ ranked from 0.26 tpm to 510 203 tpm (see Web_Table 1) with hemoglobin as by far the most abundant gene found.

3.3. Confirmation of newly assigned READ values

The newly assigned expression values within READ were further analyzed by correlating a selected subset of data points with the absolute expression data from SAGE and EST experiments in the public domain, and additional CAGE and EST libraries prepared in-house (Fig. 1). All sequence tags included in this analysis were processed, and grouped into TUs as described above, and tpm values were assigned as summarized in Table 1. As the newly assigned expression values within READ as well as the absolute expression values obtained from SAGE, CAGE and EST libraries are given in the same tpm units, and are based on the same TU/RTS system, the data sets could be directly compared. To allow for a quantitative analysis, we excluded lowly expressed TUs for which only three or fewer tags were obtained. As summarized in Table 1, correlation coefficients ranked from 0.562 to 0.699 in this analysis, where as a trend larger numbers of tags improved the data quality.

As an example, a more general analysis of data obtained from the CAGE cerebellum library is shown in Fig. 2, which we selected here as it provided the highest number of CAGE tags in this study. Similar to the analysis of the E17.5 data, tpm values as obtained from CAGE and EST tags were directly correlated as shown in Fig. 2a. Though a linear correlation was observed, the overall correlation of the data was lower than in the case of E17.5 because of the rather small number of EST tags available. Next, tpm values derived from the new Absolute READ database and the CAGE cerebellum library were directly correlated as shown in Fig. 2b. Based on 10 874 TUs, a high linear correlation was observed demonstrating the potential of our new approach for the definition of absolute expression values based on deep CAGE tag sequencing. As we observed an uneven distribution for TUs with low tpm values, we further analyzed the data from the CAGE E17.5 and cerebellum libraries for the abundance of individual TUs (Fig. 2c,d). For most TUs in both CAGE libraries, only a small number of tags were found at this level of sequencing, indicating that deep sequencing of CAGE libraries is desirable. However, already with the limited number of tags presently available, we measured 821 tpm for tubulin (READ ID 5730555P04) in the 'E17.5 standard'. After applying our formula, the relative READ value of -0.93 in cerebellum was converted to the absolute value of 431 tpm ($821 \times 2^{-0.93} = 431$). This is in good agreement with the 382 tpm found for tubulin in the CAGE cerebellum library (Web_Table 4), underlining again the potential of our approach.

3.4. New READ database on absolute expression values for mouse transcripts

The newly assigned expression values were used to create the 'Absolute READ database'. The database holding four tables as delimited text files, along with additional user instructions, can be downloaded from: http://genome.gsc.riken.

Table 1
TU distribution between datasets

Source	Total number of tags/seq	Assigned to RTS TUs	Number of unique TUs	Conformance with READ	Correlation coefficient to READ absolute values
READ database	57 931	57 931	22 406	22 406	–
CAGE E17.5	86 555	36 284	7 507	6 229	–
CAGE brain	42 349	10 273	3 617	3 023	0.689
CAGE cerebellum	327 178	123 387	14 227	10 874	0.699
SAGE kidney	12 154	1 168	205	172	0.562
EST E17.5	49 806	47 493	7 164	6 284	–
EST placenta	5 347	3 727	1 049	936	0.592
EST cerebellum	6 409	4 667	2 266	2 062	0.529
E17.5 standard	136 361	83 777	10 675	8 845	–
RTS	42 690	42 690	42 690	22 406	–

Source: library or database used; total numbers indicate total sequence tags; assigned indicates number of tags mapped to genome; unique TUs indicates the number after grouping within RTS; conformance indicates how many sequence tags are shared with READ TU; correlation indicates the correlation coefficient which is calculated based on READ absolute values and corresponding library used for validation.

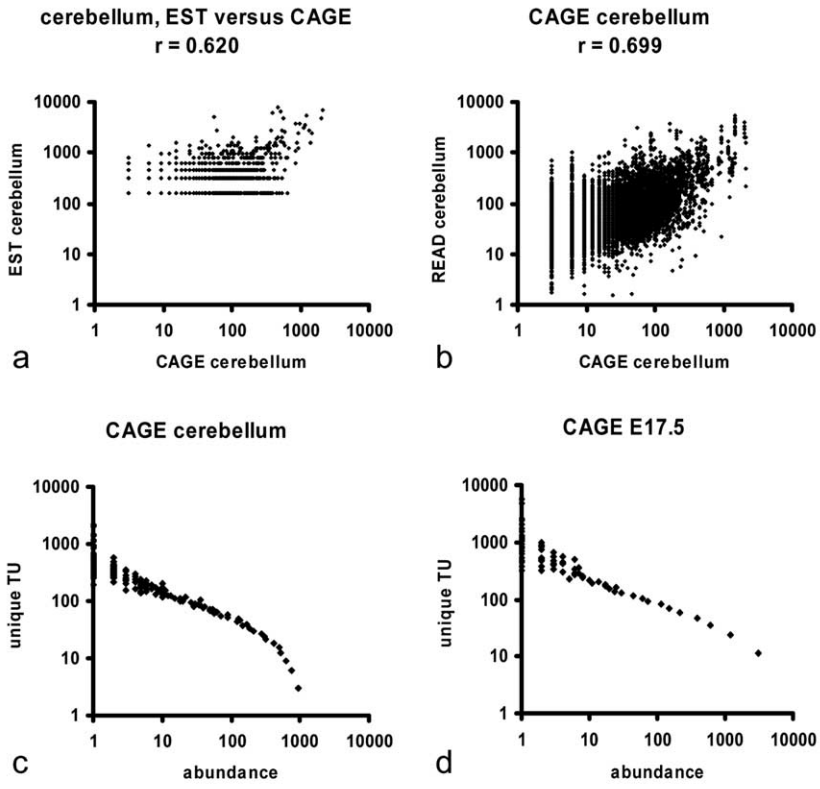

Fig. 2. Scatter plots and unique transcript distributions for E17.5 and cerebellum libraries. a: Correlation of EST and CAGE tags from cerebellum EST and CAGE libraries. b: Correlation of tpm values from CAGE library and absolute READ values for cerebellum. c: Unique TU distribution according to their abundance in CAGE cerebellum library. d: Unique TU distribution according to their abundance in CAGE E17.5 library.

jp/absolute. Web_Table 1 (absolute_count_20k_50tissues.txt) contains READ absolute values for FANTOM1 clones in 50 tissues. The relative logarithmic READ data set does not contain expression information on E17.5 itself. However, absolute values on expression in E17.5 are given as out of the analysis of the ESTs and CAGE tags ('E17.5 standard'). These data are also included in Web_Table 2 (absolute_count_60k_21tissues.txt) which contains READ absolute values for FANTOM2 clones in 21 tissues [26]. In Web_Table 3 (validation_CAGE_SAGE_EST.txt), the user can find libraries with TUs and counts used for validation of Absolute READ values. Web_Table 4 (tpm_count_libr_READ.txt) contains additional tpm expression information for libraries used in bioinformatics analysis during this project. All tables as presented here cover the present status of the project as of the publication date. However, the Absolute READ database will be subject to continuous updates and improvements in the future, when more sequence information will be available.

4. Discussion

We propose here an approach for the standardization of expression studies by assigning quantitative absolute expression values given in tpm. Expression profiling by cDNA microarrays has major limitations as the experiments usually provide only relative expression values, and genes expressed at very low levels are commonly out of the detection range [28]. Similarly, the READ database encompasses only relative expression values as compared to genes expressed in E17.5. As relative expression values are of limited value for a distinction between high, medium, or rarely expressed genes in a given tissue, we aimed at providing absolute values for transcripts expressed in the mouse based on high-throughput sequencing information on the common reference sample.

To achieve a high coverage on the genes expressed in E17.5, we applied the novel CAGE approach along with classical EST sequencing, where in total 136 361 sequence tags covering 10 675 unique TUs were included in our analysis. Although the number of sequence tags included in this study is much higher than commonly used, it cannot be excluded that rare transcripts present in the libraries were still missed. In case of rare transcripts like the hormone-sensitive lipase (READ ID A830014N15), or ubiquitin-specific protease (READ ID A630093H23), we failed to detect them in the E17.5 reference, whereas they showed low expression in the CAGE cerebellum library having nearly three times as many tags. Thus 136 361 tags from 'E17.5 standard' allowed only for the evaluation of 8845 transcripts (40%) out of the 22 406 TUs covered by the READ database. Sequencing of additional CAGE tags from the E17.5 reference along with the integration of absolute expression data from other tissues will help to more comprehensively validate the expression levels, and will include genes for which we could not yet obtain experimental proof for their expression in the reference set.

Currently, after covering 10 675 unique TUs from E17.5, the detection limit within Absolute READ is 7 tpm. Statistical analysis on the sequence tags obtained from the reference sample suggests a dynamic range of expression within the

dataset covering at least four orders of magnitude ranking from hemoglobin with 11 086 tags to many TUs, for which only a single tag could be found. For individual transcripts within Absolute READ, tpm values for a given TU can fluctuate as sometimes several different although unique clones may have represented the same TU on the microarrays used in the READ experiments. Here alternative splicing, promoter usage, or polyadenylation may have caused some of the variability, although full-length cDNAs as used for microarray printing should be less sensitive to alternative exon usage.

The number of high-quality tags obtained for E17.5 within this study was sufficient to prove the concept of our approach for establishing absolute expression values for READ data. However, an even higher number of sequence tags will be necessary to achieve a full coverage of the genes expressed in E17.5 and to have statistics that are more reliable on their absolute expression values. Thus we are targeting in the future for a much deeper sequencing of the reference sample and other CAGE libraries from additional mouse tissues to cover also genes not yet found expressed in E17.5.

To further confirm the absolute expression levels within Absolute READ, we compared our newly assigned values with data obtained by other experimental means including EST sequencing, CAGE and SAGE. Using a cutoff of at least three confirmed tags per TU, we could already within the present dataset observe a good correlation between those data, which surely will further be improved in the future with larger numbers of tags available in the public domain. As shown in our study, even the 327 178 tags sequenced from the cerebellum CAGE library could cover only 14 227 unique TUs, which most likely do not represent the complete cerebellum transcriptome, suggesting that as many as 1 000 000 tags per library would be a desirable goal.

Beside defining absolute expression values for transcripts using tpm expression values as a 'standard unit', it is furthermore of high importance for the establishment of a reference system that a common clustering and anthology system will be used for all transcripts to enable a direct comparison of expression data from different sources. Here we propose the RTS/TU system as a common standard, which we have applied here to cluster transcripts to TUs described within Absolute READ. TUs are regions within the mouse genome, which relate to RTS sets holding transcripts derived from a given TU. As a genome-based approach including information from GenBank, RefSeq and Ensembl it is presently the most comprehensive system to describe the mouse transcriptome.

We have here undertaken efforts to integrate the information from available gene expression datasets within READ to establish a unified system describing gene expression values in the mouse. Global attribution of absolute expression values in tpm for a given TU is a first step towards unification and standardization of expression databases including READ, SAGE, CAGE and ESTs. In extension of our initial studies, expression profiling of more specific samples of pathological interest is awaiting the same quantitative treatment.

Acknowledgements: We thank A. Hasegawa for database management and bioinformatic analysis, also Y. Mitsuiki for secretarial support. This work was supported by Research Grants for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H., Research Grant for Advanced and Innovational Research Program in Life Science from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to J.K. and Presidential Research Grant for Intersystem Collaboration of RIKEN to J.K. R.K. is the recipient of an INCO-JAPAN fellowship from the European Union.

References

[1] Okazaki, Y. et al. (2002) Nature 420, 563–573.
[2] Miki, R. et al. (2001) Proc. Natl. Acad. Sci. USA 98, 2199–2204.
[3] Frederiksen, C.M., Aaboe, M., Dyrskjot, L., Laurberg, S., Wolf, H., Orntoft, T.F. and Kruhoffer, M. (2003) APMIS Suppl. 96–101.
[4] Sterrenburg, E., Turk, R., Boer, J.M., van Ommen, G.B. and den Dunnen, J.T. (2002) Nucleic Acids Res. 30, e116.
[5] Dudley, A.M., Aach, J., Steffen, M.A. and Church, G.M. (2002) Proc. Natl. Acad. Sci. USA 99, 7554–7559.
[6] Kerr, M.K. and Churchill, G.A. (2001) Proc. Natl. Acad. Sci. USA 98, 8961–8965.
[7] Kim, H., Zhao, B., Snesrud, E.C., Haas, B.J., Town, C.D. and Quackenbush, J. (2002) BioTechniques 33, 924–930.
[8] Hill, A.A., Brown, E.L., Whitley, M.Z., Tucker-Kellogg, G., Hunter, C.P. and Slonim, D.K. (2001) Genome Biol. 2, RESEARCH0055.
[9] Nimgaonkar, A., Sanoudou, D., Butte, A.J., Haslett, J.N., Kunkel, L.M., Beggs, A.H. and Kohane, I.S. (2003) BMC Bioinformatics 4, 27.
[10] Rao, J.S. and Li, J. (2003) Respir. Physiol. Neurobiol. 135, 109–119.
[11] Hekstra, D., Taussig, A.R., Magnasco, M. and Naef, F. (2003) Nucleic Acids Res. 31, 1962–1968.
[12] Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. and Sealfon, S.C. (2002) Nucleic Acids Res. 30, e48.
[13] Stoeckert, C. et al. (2001) Bioinformatics 17, 300–308.
[14] Banfi, S., Guffanti, A. and Borsani, G. (1998) Trends Genet. 14, 80–81.
[15] Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) Nat. Genet. 4, 332–333.
[16] Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Science 270, 484–487.
[17] Shiraki, T. et al. (2003) Proc. Natl. Acad. Sci. USA 100, 15776–15781.
[18] Shibata, Y., Carninci, P., Watahiki, A., Shiraki, T., Konno, H., Muramatsu, M. and Hayashizaki, Y. (2001) BioTechniques 30, 1250–1254.
[19] Carninci, P. et al. (2003) Genome Res. 13, 1273–1289.
[20] Virlon, B., Cheval, L., Buhler, J.M., Billon, E., Doucet, A. and Elalouf, J.M. (1999) Proc. Natl. Acad. Sci. USA 96, 15286–15291.
[21] http://www-dsv.cea.fr/thema/get/datasets.html.
[22] http://www.ncbi.nlm.nih.gov/UniGene/library.cgi?ORG=Mm&LID=12267.
[23] http://www.ncbi.nlm.nih.gov/UniGene/library.cgi?ORG=Mm&LID=1783.
[24] Wheeler, D.L. et al. (2003) Nucleic Acids Res. 31, 28–33.
[25] http://www.ncbi.nlm.nih.gov/UniGene/.
[26] Bono, H. et al. (2003) Genome Res. 13, 1318–1323.
[27] Kadota, K., Miki, R., Bono, H., Shimizu, K., Okazaki, Y. and Hayashizaki, Y. (2001) Physiol. Genomics 4, 183–188.
[28] Kothapalli, R., Yoder, S.J., Mane, S. and Loughran Jr., T.P. (2002) BMC Bioinformatics 3, 22.