Available online at www.sciencedirect.com

SCIENCE DIRECT®

C. R. Biologies 326 (2003) 923–929

COMPTES RENDUS BIOLOGIES

Molecular biology and genetics

The Riken mouse genome encyclopedia project

Yoshihide Hayashizaki

Gene Exploration Research Group, Riken Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Received 16 September 2003; accepted 23 September 2003

Presented by François Gros

Abstract

The Riken mouse genome encyclopedia a comprehensive full-length cDNA collection and sequence database. High-level functional annotation is based on sequence homology search, expression profiling, mapping and protein–protein interactions. More than 1 000 000 clones prepared from 163 tissues were end-sequenced and classified into 128 000 clusters, and 60 000 representative clones were fully sequenced representing 24 000 clear protein-encoding genes. The application of the mouse genome database for positional cloning and gene network regulation analysis is reported. ***To cite this article: Y. Hayashizaki, C. R. Biologies 326 (2003).***

Résumé

Le projet d'encyclopédie du génome murin de Riken. L'encyclopédie du génome de la souris de Riken est une collection extensive d'ADNc de pleine longueur et une base de données de leurs séquences. Une annotation fonctionnelle de haut niveau est fondée sur la recherche d'homologies de séquence, les profils d'expression, la cartographie et les interactions protéine–protéine. Plus d'un million de clones préparés à partir de 163 tissus ont été séquencés à leur extrémité et classés en 128 000 groupes, et 60 000 clones représentatifs entièrement séquencés qui codent 24 000 protéines distinctes. Nous rapportons l'application de cette base de données pour le clonage positionnel et l'analyse de la régulation des réseaux génétiques. ***Pour citer cet article : Y. Hayashizaki, C. R. Biologies 326 (2003).***

Keywords: full-length cDNA; functional annotation; integrated database; mouse genome encyclopedia

Mots-clés : ADNc complet ; annotation fonctionnelle ; base de données intégrée ; encyclopédie du génome de la souris

1. Goal of Riken mouse genome encyclopedia project

The goal of our project is to establish a mouse genome encyclopedia and to develop a series of original technologies to achieve it. Riken mouse encyclopedia is the platform for the second goal to draw genome-wide pictures of gene cascades that can account for the mechanism to connect genes to phenotypes. The encyclopedia will consist of five components; the non-redundant mouse full-length cDNA clone bank, the mouse full-length cDNA sequence bank, the chromosomal locations, expression profiles and protein–protein interactions covering as many

E-mail address: yosihide@gsc.riken.go.jp (Y. Hayashizaki).

1631-0691/$ – see front matter © 2003 Published by Elsevier SAS on behalf of Académie des sciences.
doi:10.1016/j.crvi.2003.09.018

genes as can be collected. We have been developing the full-length cDNA technologies and high-speed sequencing technologies to analyze these materials. The purpose of our project is not only the analysis of the sequence of full-length cDNAs but also the development of a new approach for research based on the encyclopedia. We would like to develop new post-sequence technologies and systems which can maximize the usefulness of the encyclopedia by utilizing full-length cDNA clones.

2. History and current status of Riken mouse genome encyclopedia project

Since 1995, the Riken genome exploration research group has been developing a series of new systems to construct full-length cDNA libraries [1–3], high-speed sequencing system named RISA [4]. Using this system, approximately 1 000 000 mouse full-length cDNA clones were isolated from 163 independent tissues at different stages from different organs, and 3′ end sequence data could classify these clones into 128 000 clusters. Full stretches of sequences of 60 000 representative clones were sequenced, resulting in 36 000 unique sequences. We are also producing the data of the expression profiles and protein–protein interaction [5–8]. The integrated database, including not only full-length cDNA sequences but also mapping, expression profiles and protein–protein interaction data of all these genes were very useful for the analysis of gene functions, to support a positional candidate approach, and for gene network analysis to connect genes to phenotypes.

3. Achievements

3.1. Development of a system to prepare the mouse genome encyclopedia

To establish the mouse genome encyclopedia, two series of technologies are needed: A method for constructing high quality full-length cDNAs, and RISA (Riken Integrated Sequence Analysis) system (high-speed sequencing system).

At present, our Riken full-length cDNA method consists of four key technologies: an elongation method for first strand cDNA synthesis [3], a selection method for eliminating partial cDNA [1,2], a normalization and subtraction method [9] to avoid redundancy in subsequent sequencing efforts, and a new cloning vector [10]. Also, the following technologies play an important role: protocol for making library from small amount of tissues, removal of poly-A stretch, mixed- and tagged-cDNA libraries.

We also have developed a large-scale plasmid preparation system [11,12], a transcriptional sequencing reaction system [13–15], and a high-speed 384-capillary sequencer that should enable analysis of 40 000 samples/day [4]. The most important point is that all of these technologies have been incorporated into a single system to achieve our goal.

3.2. Full-length cDNA library technology

Our group is approaching this project with originally developed technologies. The Riken full-length cDNA method consists of four key technologies as described below.

3.2.1. Elongation method

The basic principle for the elongation method is that the first strand cDNA synthesis should be undertaken at high temperature because the main cause of partial cDNA synthesis is the secondary structure of the mRNA. The elongation method is based on trehalose-mediated thermostabilization, thermoactivation and thermoprotection of reverse transcriptase [3]. In addition to this development, we also discovered that trehalose has some effect, which may be large, on half of the enzymes generally used.

3.2.2. Selection method

The selection method is named ‘the Cap trapper method’ [1,2]. This method employs chemical modification of the Cap site to oxidize the diol structure, which is specific for non-redundant cDNA. The product dialdehyde structure is connected to the hydrazide. Thus, the diol group can be biotinylated by biotin hydrazide. After synthesis of the first strand synthesis, the Cap site is labeled by the above-mentioned chemical modification, and RNaseI is used to cleave the single strand RNA but not the DNA-RNA hybrid. In the mRNA with partial cDNA, the single-strand RNA is exposed and can be attached by RNaseI. Therefore, RNaseI treatment removes the biotin group from mRNA with partial cDNA, and only the biotin label at the Cap site of mRNA with the full-length first strand cDNA remains. Subsequently, full-length cDNA can be collected using avidin beads.

3.2.3. Normalization-subtraction method

Highly expressed genes and already-collected genes reduce the efficiency of collection of novel cDNAs by one-pass sequencing. To eliminate them, we developed a reiterative normalization and subtraction method utilizing biotinylated RNAs as drivers [9]. Currently an amplified library subtraction system is being introduced in order to rescue from cDNA libraries the cDNAs that have not yet been classified by one-pass sequencing from existing cDNA libraries.

3.2.4. Removal of poly-A tails from FL-cDNAs

Poly-stretches of nucleotides, such as poly (A) tail in cDNAs are known to interfere with the processibility of DNA/RNA polymerase in the sequencing reaction, resulting in reduced read-lengths and rates of successful sequences. To overcome this problem, we developed a new method to remove poly (A) tails from cDNAs using Type II restriction enzymes [16]. We also removed the G-stretch previously used for priming the second strand by a new linker adapter strategy that induces with high efficiency a sequence to prime the second strand cDNA [17].

3.2.5. Development of host/vectors for FL-cDNA libraries

Two kinds of cloning vectors were developed [10]: lambda FLC1 and FLC2. Lambda FLC1 can clone a wide range of cDNAs from 0.5 to over 13 kb long and shows slight size preference for long cDNA clones [18]. We could routinely prepare a cDNA library that once bulk-excised into plasmid, showed average sixes of 2 to 3 kbp. Lambda FLC2 is a modified lambda FLC1, which contains att sites flanking the cloning sites. The att sites allow easy transfer of a cDNA insert into other vectors for expression and other functional studies with lambda recombinase. This system should facilitate the functional analyses of cDNAs in post-sequence research.

3.3. High throughput sequencing system

We established and expanded our large-scale sequencing system. It comprises of FL-cDNA library construction, an E. coli picking system, a plasmid preparation system, a sequencing reaction system, a sequencing system, and the management of samples and data. The current capacity of sequencing is 40 000 samples per day. All samples are well-tracked to avoid confusion of IDs (ID errors), and quality control checking is routinely done.

3.3.1. Plasmid preparation system

We have designed and introduced three instruments: an instrument for medium distribution and E. coli inoculation, a harvester of E. coli culture solution, and a plasmid extractor [13]. These instruments can process 40 000 plasmid extractions per day and are now being optimized to achieve constant yield and quality for sequencing templates.

3.3.2. Transcriptional sequencing (TS)

To develop the TS system [14–16], we originally developed a mutated RNA polymerase, which can incorporate the $3'$ dNTP preferentially and uniformly, and fluorescent $3'$dNTP dye terminator.

3.3.3. Development of 384-capillary sequencer (RISA sequencer)

We completed the development of the first version of a 384-capillary sequencer (RISA 2) [4] at the end of 1996. Shimadzu from November 1999 has commercially marketed it.

3.4. Data management system

We have developed many programs to analyze sequence data produced in our high-throughput sequencing system [19], such as a set of tools automatically classifying cDNA clones based on the $3'$-end sequences and tools for automatically registering sequence data in the encyclopedia database [20].

We have also established an assembly and primer design system. The assembly system can handle three kinds of sequence data produced by different kinds of sequencers in a uniform style. If a gap remains, primers for the primer walking sequencing can be easily designed. Public available sequences, such as EST data, can also be utilized to fill the gaps.

We have a database system based on Sybase DBMS to manage most information derived from our sequencing system, tissue sources for FL-cDNA libraries, clone ID, 3′-end sequences, full-length sequences, and clustering information. This database is updated daily. The summarized information can be viewed with Web browsers in a user-friendly manner.

3.5. Data accumulation

The mouse genome encyclopedia are prepared in five phases. In Phase I, we have constructed the mouse full-length cDNA using as many tissues as can be collected and clustered these clones by end sequencing, to produce non-redundant full-length cDNA. In Phase II, the full-sequence of the rearrayed clones from the non-redundant cDNA library have been determined. In Phase III, the chromosomal location of all full-length cDNA have been identified by *in-silico* hybridization to the human and mouse genome sequences [21, 22]. Phase IV is producing a basic database of gene expression in the body and during the development from embryo to adult [5,6,8]. Finally, Phase V is the step to produce the protein–protein interactions, based on the biggest advantage of full-length cDNA which can express the whole structure of protein [7,23], although partial cDNA (expression sequence tag; EST) and genome can not.

3.6. Data production

3.6.1. cDNA libraries

Almost all cDNAs were normalized and/or subtracted, constructed from over 163 tissues and cells for the first volume of the Riken Mouse Genome Encyclopedia.

3.6.2. One million cDNA clones and their 3′-end sequences

By clustering of the 3′-end sequences, a total of 128 000 clusters of cDNAs have been obtained. However, this classification includes some overlap or redundancy, because of various forms of splicing, alternative poly adenylation sites, some internal priming, and clustering limitation due to sequencing fidelity and other factors.

3.6.3. Full sequences of cDNA clones derived from non-redundant FL-cDNA set

Representative clones of clusters that are based on 3′-end sequences of cDNAs were rearrayed and used for the full-sequencing phase. Apparently novel genes estimated by comparison of 3′-end sequences and public DNA databases were given high priorities for full-sequencing. So far, about 60 000 full-sequences from 128 000 clusters were determined. The 60 000 sequences still contained redundancy, resulting in being clustered into 36 000 completely independent unique sequences.

3.7. Functional annotation of FL-cDNAs

In order to annotate the function of 60177 full-length Riken mouse cDNAs sequenced in RIKEN, we held the FANTOM (Functional Annotation Of Mouse) meeting [24] through 28th August to 8th September to establish the international standard of annotations. This annotation activity covers not only functional information itself, but also many other informative data, such as supplemental descriptions of gene function (gene symbol and its synonym), the functional classification (Gene Ontology, and TIGR EGAD), chromosomal localization (from genetic mapping and physical mapping if available), expression specificity (organ localization and sub-cellular localization of cDNA), mutation information (disease and knockout mouse information) [18,25].

The Riken mouse genome encyclopedia is with the human the most detailed transcriptome described in any organism to date. Analysis of these cDNAs extends known gene families and identifies new ones.

3.8. DNA microarrays

3.8.1. Construction of high-throughput arrayer

A new arrayer has been constructed, having two arms, each of which holds a pin head, and a large stage on which 96 microarrays can be prepared simultaneously. When a 16-pin head is adopted, 96 microarrays of 30K cDNAs can be done in 100 h. A 48-pin head device allowing faster mode is also available. The Maximum performance is expected to be 96 microarrays of 30K cDNAs in 16 h.

3.8.2. 19K RIKEN microarray to 40K RIKEN microarray

We have established 19K microarrays of Riken cDNAs. Expression profiles of various tissues and several developmental stages have been investigated [6]. Microarray data are analyzed and stored in our expression database, READ [8]. 19 000 genes were analyzed in 20 tissues. This database serves as a fundamental resource for the functional researches of each cDNA and each gene cascade. The size of the database is expanded to 40K cDNAs in 20 tissues.

3.9. Development of a protein–protein interaction analysis system

To uncover the function of each gene as a systematic genome-wide approach, the protein–protein interaction (PPI) panel covering all genes [7], is very important. PPIs play pivotal roles in the network of cellular biological processes and also they should be potential targets for drugs developments. However, it seems not to be so easy to establish entire PPI panel in mouse, because the estimated total number of mouse genes of (100 000) is far larger than those of budding yeast ($\sim$ 6000) and *C. elegans* ($\sim$ 20 000).

To address this difficulty, we have developed a high-throughput PPI assay system that consists of a PCR-mediated sample preparation and a modified mammalian two-hybrid method. In the pilot study, the system achieved the examination of more than 10^6 combinations per day. We have also developed a selection method of assay samples allowing us to find significantly interacting combinations efficiently, based on the demonstration that two genes co-expressed in the same tissues at the same stages preferentially interact with each other. These two key developments paved a way to enable us completion of a rough draft of an entire protein–protein interaction panel in mouse within a few years.

4. Application of cDNA system to other projects

Arabidopsis thaliana full-length cDNA Project

To determine the chromosomal locations, and expression profiles of this plant-model organism our group is collaborating with the Plant Molecular Biology Laboratory with our cDNA cloning technique [26,27]. At the moment 115 000 cDNA clones were constructed based on $3'$-end sequences of cDNAs [28]. 15 000 full, sequences at 99.99% accuracy will be finished sequences very soon as an international collaboration. All of these clones were mapped onto the *Arabidopsis thaliana* genome sequence [29].

5. Future plan

Our final goal is to establish a system for genome-wide understanding of biological phenomena at the molecular level, particularly in the medical field. In order to achieve this goal, the first step is to collect data on all full-length cDNAs, their primary structures, and expression sites. Also important is the chromosomal mapping of the cDNAs at sequencing level in order to connect the gene and the phenotype. We have started developing a system to establish such as encyclopedia using a full-length cDNA system and the RISA sequencing system. To overcome possible resource problems, we chose mouse materials from inbred, congenic and knockout strains, which are available with no limitation for the preparation of tissues such as samples at very early embryonic stages and fertilized eggs. Predictions of human full-length cDNA sequences *in-silico* can be done by homology search in comparison with our mouse full-length cDNA [21,29]. This enhances the significance of our strategy of choosing mouse cDNA as a target.

We have begun collecting mouse full-length cDNAs and are finding our approach an extremely powerful one for analyzing and explaining why certain genes cause a phenotype. To connect gene(s) and phenotype(s), our encyclopedia is very useful for identifying candidate gene(s) responsible for the phenotype(s) in the positional candidate approach. The cDNA microarrays are also useful for selecting a set of genes, which are transcriptionally regulated downstream of the target gene, using mutant and normal tissues. We plan to continue development of the mouse genome encyclopedia and the technologies to establish it and make it widely useful in order to enable the depiction of genome-wide maps from gene(s) to phenotype.

Acknowledgements

This study has been supported by Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government to Y.H.

References

[1] P. Carninci, C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki, M. Itoh, M. Kamiya, K. Shibata, N. Sasaki, M. Izawa, M. Muramatsu, Y. Hayashizaki1, C. Schneider, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, Genomics 37 (1996) 327–336.

[2] P. Carninci, A. Westover, Y. Nishiyama, T. Ohsumi, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki, M. Muramatsu, C. Schneider, Y. Hayashizaki, High efficiency selection of full-length cDNA by improved biotinylated cap trapper, DNA Res. 4 (1997) 61–66.

[3] P. Carninci, Y. Nishiyama, A. Westover, M. Itoh, S. Nagaoka, N. Sasaki, Y. Okazaki, M. Muramatsu, Y. Hayashizaki, Thermostabilization and thermoactivation of thermolabile enzymes by trehalose and its application for the synthesis of full length cDNA, Proc. Natl Acad. Sci. USA 95 (1998) 520–524.

[4] K. Shibata, M. Itoh, K. Aizawa, S. Nagaoka, N. Sasaki, P. Carninci, H. Konno, J. Akiyama, K. Nishi, T. Kitsunai, H. Tashiro, M. Itoh, N. Sumi-Kikuchi, Y. Ishii, S. Nakamura, M. Hazama, T. Nishine, A. Harada, R. Yamamoto, H. Matsumoto, S. Sakaguchi, T. Ikegami, K. Kashiwagi, S. Fujiwake, K. Inoue, Y. Togawa, M. Izawa, E. Ohara, M. Watahiki, Y. Yoneda, T. Ishikawa, K. Ozawa, T. Tanaka, S. Matsuura, J. Kawai, Y. Okazaki, M. Muramatsu, Y. Inoue, Y. Hayashizaki, RIKEN integrated sequence analysis (RISA) system – 384-format sequencing pipeline with 384 multicapillary sequencer, Genome Res. 10 (2000) 1757–1771.

[5] K. Kadota, R. Miki, H. Bono, K. Shimizu, Y. Okazaki, Y. Hayashizaki, Preprocessing Implementation for Microarray (PIRM): an efficient method for processing cDNA microarray data, Physiol. Genomics 4 (2001) 183–188.

[6] R. Miki, K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J.L. DeRisi, M.B. Eisen, W.R. Iyer, P.O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki, Y. Hayashizaki, Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18 816 full-length enriched mouse cDNA arrays, Proc. Natl Acad. Sci. USA 98 (2001) 2199–2204.

[7] H. Suzuki, Y. Fukunishi, I. Kagawa, H. Bono, R. Saito, H. Oda, T. Endo, S. Kondo, Y. Okazaki, Y. Hayashizaki, Protein–protein interaction panel using mouse full-length cDNAs, Genome Res. 11 (2001) 1758–1765.

[8] H. Bono, T. Kasukawa, M. Furuno, Y. Hayashizaki, Y. Okazaki, READ: RIKEN expression array database, Nucleic Acids Res. 30 (2002) 211–213.

[9] P. Carninci, Y. Shibata, N. Hayatsu, Y. Sugahara, K. Shibata, M. Itoh, H. Konno, M. Okazaki, Y. Muramatsu, Y. Hayashizaki, Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes, Genome Res. 10 (2000) 1617–1630.

[10] P. Carninci, Y. Shibata, N. Hayatsu, M. Itoh, T. Shiraki, T. Hirozane, A. Watahiki, K. Shibata, H. Konno, M. Muramatsu, Y. Hayashizaki, Balanced-size and long-size cloning of full-length, Cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis, Genomics 77 (2001) 79–90.

[11] M. Ito, P. Carninci, S. Nagaoka, N. Sasaki, Y. Okazaki, T. Ohsumi, M. Muramatsu, Y. Hayashizaki, Simple and rapid preparation of plasmid template by a filtration method using microtiter filter plates, Nucleic Acids Res. 25 (1997) 1315–1316.

[12] M. Itoh, T. Kitsunai, J. Akiyama, K. Shibata, M. Izawa, J. Kawai, Y. Tomaru, P. Carninci, Y. Shibata, Y. Ozawa, M. Muramatsu, Y. Okazaki, Y. Hayashizaki, Automated filtration-based high-throughput plasmid preparation system, Genome Res. 9 (1999) 463–470.

[13] N. Sasaki, M. Izawa, M. Watahiki, K. Ozawa, T. Tanaka, S. Yoneda, Y. Matsuura, P. Carninci, M. Muramatsu, Y. Okazaki, Y. Hayashizaki, Transcriptional sequencing: A method for DNA sequencing using RNA polymerase, Proc. Natl Acad. Sci. USA 95 (1998) 3455–3460.

[14] N. Izawa, M. Sasaki, M. Watahiki, E. Ohara, Y. Yoneda, M. Muramatsu, Y. Okazaki, Y. Hayashizaki, Recognition sites of 3′-OH group by T7 RNA polymerase and its application to transcriptional sequencing, J. Biol. Chem. 273 (1998) 14242–14246.

[15] N. Sasaki, M. Izawa, Y. Sugahara, T. Tanaka, M. Watahiki, K. Ozawa, E. Ohrara, H. Funaki, Y. Yoneda, S. Matsuura, M. Muramatsu, Y. Okazaki, Y. Hayashizaki, Identification of stable RNA hairpins causing band compression in transcriptional sequencing and their elimination by use of inosine triphosphate, Gene 222 (1998) 17–24.

[16] Y. Shibata, P. Carninci, K. Sato, N. Hayatsu, T. Shiraki, Y. Ishii, T. Arakawa, A. Hara, N. Ohsato, M. Izawa, K. Aizawa, M. Itoh, K. Shibata, A. Shinagawa, J. Kawai, Y. Ota, S. Kikuchi, N. Kishimoto, M. Muramatsu, Y. Hayashizaki, Removal of polyA tails from full-length cDNA libraries for high efficiency sequencing, BioTechniques 31 (2001) 1042–1049.

[17] Y. Shibata, P. Carninci, A. Watahiki, T. Shiraki, H. Konno, M. Muramatsu, Y. Hayashizaki, Cloning full-length, cap-trapper-selected cDNAs by using the Single-Strand Linker Ligation Method, BioTechniques 30 (2001) 1250–1254.

[18] P. Carninci, T. Shiraki, Y. Mizuno, M. Muramatsu, Y. Hayashizaki, Extra-long first-strand cDNA synthesis, BioTechniques 32 (2002) 984–985.

[19] Y. Sugahara, P. Carninci, M. Itoh, K. Shibata, H. Konno, T. Endo, M. Muramatsu, Y. Hayashizaki, Comparative evaluation of 5′-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries, Gene 263 (2001) 93–102.

[20] H. Konno, Y. Fukunishi, K. Shibata, M. Itoh, P. Carninci, Y. Sugahara, Y. Hayahizaki, Computer-based methods for the

mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library, Genome Res. 11 (2001) 281–289.

[21] S. Kondo, A. Shinagawa, T. Saito, H. Kiyosawa, I. Yamanaka, K. Aizawa, S. Fukuda, A. Hara, M. Itoh, J. Kawai, K. Shibata, Y. Hayashizaki, Computational analysis of full-length mouse cDNA compared with human genome sequences, Mamm. Genome 12 (2001) 673–677.

[22] I. Yamanaka, H. Kiyosawa, S. Kondo, T. Saito, P. Carninci, A. Shinagawa, K. Aizawa, S. Fukuda, A. Hara, M. Itoh, J. Kawai, K. Shibata, T. Arakawa, Y. Ishii, Y. Hayashizaki, Mapping of 19032 mouse cDNAs on the mouse chromosomes, Journal of Structural and Functional Genomics 2 (2001) 23–28.

[23] R. Saito, H. Suzuki, Y. Hayashizaki, Interaction generality, a measurement to assess reliability of protein–protein interaction, Nucleic Acids Res. 30 (2002) 1163–1168.

[24] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L.M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli, G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M.F. de Bonaldo, M.J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D.A. Hume, M. Kamiya, N.H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schonbach, T. Seya, Y. Shibata, K.F. Storch, H. Suzuki, K. Toyo-oka, K.H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y. Hasegawa, H. Kawaji, Y. Kohtsuki, S. Hayashizaki, Functional annotation of 21 076 sequenced mouse cDNAs prepared from full-length enriched libraries, Nature 409 (2001) 685–690.

[25] H. Bono, T. Kasukawa, M. Furuno, Y. Hayashizaki, Y. Okazaki, FANTOM DB: database of Functional Annotation of RIKEN Mouse cDNA Clones, Nucleic Acids Res. 30 (2002) 116–118.

[26] M. Seki, P. Carninci, Y. Nishiyama, Y. Hayashizaki, K. Shinozaki, High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated CAP trapper, Plant J. 15 (1998) 707–720.

[27] M. Seki, M. Narusaka, H. Abe, M. Kasuga, K. Yamaguchi-Shinozaki, P. Carninci, Y. Hayashizaki, K. Shinozaki, Monitoring the Expression pattern of 1300 Arabidopsis genes under drought and cold stresses using full-length cDNA microarray, The Plant Cell 13 (2001) 61–72.

[28] M. Seki, M. Naramura, A. Kamiya, J. Ishida, M. Satou, T. Sakurai, M. Nakajima, A. Enju, K. Akiyama, Y. Oono, M. Muramatsu, Y. Hayashizaki, T. Arakawa, K. Shibata, A. Shinagawa, K. Shinozaki, Functional annotation of a full-length Arabidopsis cDNA collection, Science 296 (2002) 141–145.

[29] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, J. Szustakowki, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.