doi:10.1093/nar/gkj034

CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis

Hideya Kawaji[1], Takeya Kasukawa[1,2], Shiro Fukuda[2], Shintaro Katayama[2,*], Chikatoshi Kai[2], Jun Kawai[2,3], Piero Carninci[2,3] and Yoshihide Hayashizaki[2,3]

[1]NTT Software Corporation, Teisan Kannai Building 209, Yamashita-cho Naka-ku, Yokohama, Kanagawa, 231-8551, Japan, [2]Genome Exploration Research Group (Genome Network Project Core Group), RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan and [3]Genome Science Laboratory, Discovery Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan

Received August 15, 2005; Revised and Accepted September 26, 2005

ABSTRACT

Cap-analysis gene expression (CAGE) Basic and Analysis Databases store an original resource produced by CAGE, which measures expression levels of transcription starting sites by sequencing large amounts of transcript 5′ ends, termed CAGE tags. Millions of human and mouse high-quality CAGE tags derived from different conditions in >20 tissues consisting of >250 RNA samples are essential for identification of novel promoters and promoter characterization in the aspect of expression profile. CAGE Basic Database is a primary database of the CAGE resource, RNA samples, CAGE libraries, CAGE clone and tag sequences and so on. CAGE Analysis Database stores promoter related information, such as counts of related transcripts, CpG islands and conserved genome region. It also provides expression profiles at base pair and promoter levels. Both databases are based on the same framework, CAGE tag starting sites, tag clusters for defining promoters and transcriptional units (TUs). Their associations and TU attributes are available to find promoters of interest. These databases were provided for Functional Annotation Of Mouse 3 (FANTOM3), an international collaboration research project focusing on expanding the transcriptome and subsequent analyses. Now access is free for all users through the World Wide Web at http://fantom3.gsc.riken.jp/.

INTRODUCTION

Cap-analysis gene expression (CAGE) is a high-throughput method to measure expression levels by counting large amounts of sequenced capped 5′ ends of transcripts, termed CAGE tags (1). A similar approach is proposed as 5′ end SAGE (2). The average length of these 5′ end tags of transcripts is 20 bp and the tags are aligned with the genome directly, although original SAGE (3) tags are aligned with 3′ ends of transcripts (4). CAGE tags are an essential resource for profiling transcriptional starting sites and can be used for profiling gene expressions by counting CAGE tags associated with genes. Millions of mouse and human high-quality CAGE tags derived from different conditions in >20 tissues consisting of >250 RNA samples are subjected for analysis in the international collaboration research project, Functional Annotation Of Mouse 3 (FANTOM3). The CAGE tags are used for the analysis of the transcriptional landscape of mammalian genome (5), antisense transcription in the mammalian transcriptome (6), comprehensive promoter analysis (P.Carninci, A.Sandelin, B.Lenhard, S.Katayama, K.Shimokawa, J.Ponjavic, C.A.Semple, M.S.Taylor, P.Engstrom, M.C.Frith, A.R.Forrest, W.B.Alkema, S.L.Tan, C.Plessy, R.Kodzius, T.Ravasi, T.Kasukawa, S.Fukuda, M.Kanamori-Katayama, Y.Kitazume, H.Kawaji, C.Kai, H.Konno, K.Nakano, S.Mottagui-Tabar, P.Arner, A.Chesi, S.Gustincich, F.Persichetti, H.Suzuki, S.M.Grimmond, C.Wells, V.Orlando, C.Wahlestedt, E.T.Liu, M.Harbers, J.Kawai, V.B.Bajic, D.A.Hume and Y.Hayashizaki, manuscript submitted) and subsequent analyses.

We constructed two database systems to utilize the CAGE resource, CAGE Basic and Analysis Databases. Their aims are (i) to manage and trace the CAGE data consistently and (ii) to demonstrate the promoter usage (using CAGE and other data). The former is required to support the novel experimental processes of CAGE and to manage the large amount of RNA samples provided in the FANTOM3 collaboration. The latter is to support subsequent analyses using all of the required data, without influence of our management of the CAGE data. Additionally, we constructed CAGE Expression 3D Viewer for

*To whom corresponding should be addressed. Tel: +81 45 503 9222; Fax: +81 45 503 9216; Email: rgscerg@gsc.riken.jp novel type of expression view (K.Shimokawa, Y.Okamura-Oho, C.Kai, P.Carninci and Y.Hayashizaki, manuscript in preparation). The database systems described here were used in FANTOM3 and are now publicly accessible. Here, we present the systems' overview and functions to facilitate the use of the CAGE resource.

DATA BASIS

A consistent and comprehensive dataset is crucial to allow biological analyses in different kinds of viewpoints. Our two database systems are built on the same basis: CAGE tag starting site (CTSS), tag cluster (TC) and transcriptional unit (TU).

CTSS is a nucleotide position on the genome from which an alignment of CAGE tag starts. Counts of CAGE tags sharing the same starting sites represent expression profiles in base pairs level. TC is an operationally defined unit to characterize promoters. It is constructed by clustering 5′ end overlapped region of transcripts (P.Carninci, A.Sandelin, B.Lenhard, S.Katayama, K.Shimokawa, J.Ponjavic, C.A.Semple, M.S.Taylor, P.Engstrom, M.C.Frith, A.R.Forrest, W.B.Alkema, S.L.Tan, C.Plessy, R.Kodzius, T.Ravasi, T.Kasukawa, S.Fukuda, M.Kanamori-Katayama, Y.Kitazume, H.Kawaji, C.Kai, H.Konno, K.Nakano, S.Mottagui-Tabar, P.Arner, A.Chesi, S.Gustincich, F.Persichetti, H.Suzuki, S.M.Grimmond, C.Wells, V.Orlando, C.Wahlestedt, E.T.Liu, M.Harbers, J.Kawai, V.B.Bajic, D.A.Hume and Y.Hayashizaki, manuscript submitted), such as 5′ end 20 bp long of RIKEN full-length cDNA and RIKEN-5′-expressed sequence tag (EST), 5′ end tags of GIS (7) and GSC (4) ditags, DBTSS (8), 5′ end SAGE and CAGE. Of these, overlapping sequences on the genome with at least 1 bp are clustered, and define a TC. Counts of CAGE tags within TCs represent expression profiles on promoter level. TU is also an operationally defined unit proposed in FANTOM2 (9), defined as a region or a set of discontinuous regions in the genome from where all exons of a mature full-length mRNA are derived (10). Counts of CAGE tags within TUs represent expression profiles on gene level. TUs are associated with Entrez Gene (11) and gene ontology term (12) by means of transcripts belonging to them, if possible. CTSS are associated with TCs, and TCs are associated with TUs. Users can access the CAGE resource of interest by searching TUs with their own keywords.

SYSTEM OVERVIEW

Figure 1 is an overview of the CAGE database systems. CAGE Basic Database is a primary database of the CAGE resource, and provides a central view of CAGE resources. CAGE Analysis Database stores TC related information, and provides a central view of promoters. As a complementary system, Genomic Elements Database is constructed to provide a central view of genome positions. Their main contents are described in Table 1. CAGE Analysis Database would be the most convenient gateway for users, especially new to the CAGE data. Hyperlinks from the database to the others are available depending on their interests, CAGE Basic Database for CAGE sequences themselves and Genomic Elements Database for a conventional genome view.

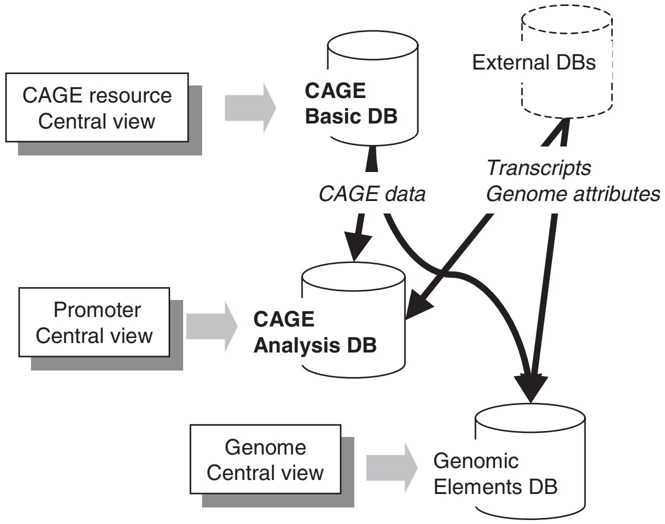


Figure 1. An overview of the CAGE supporting systems and data flow among them.

Table 1. Main contents of the database systems

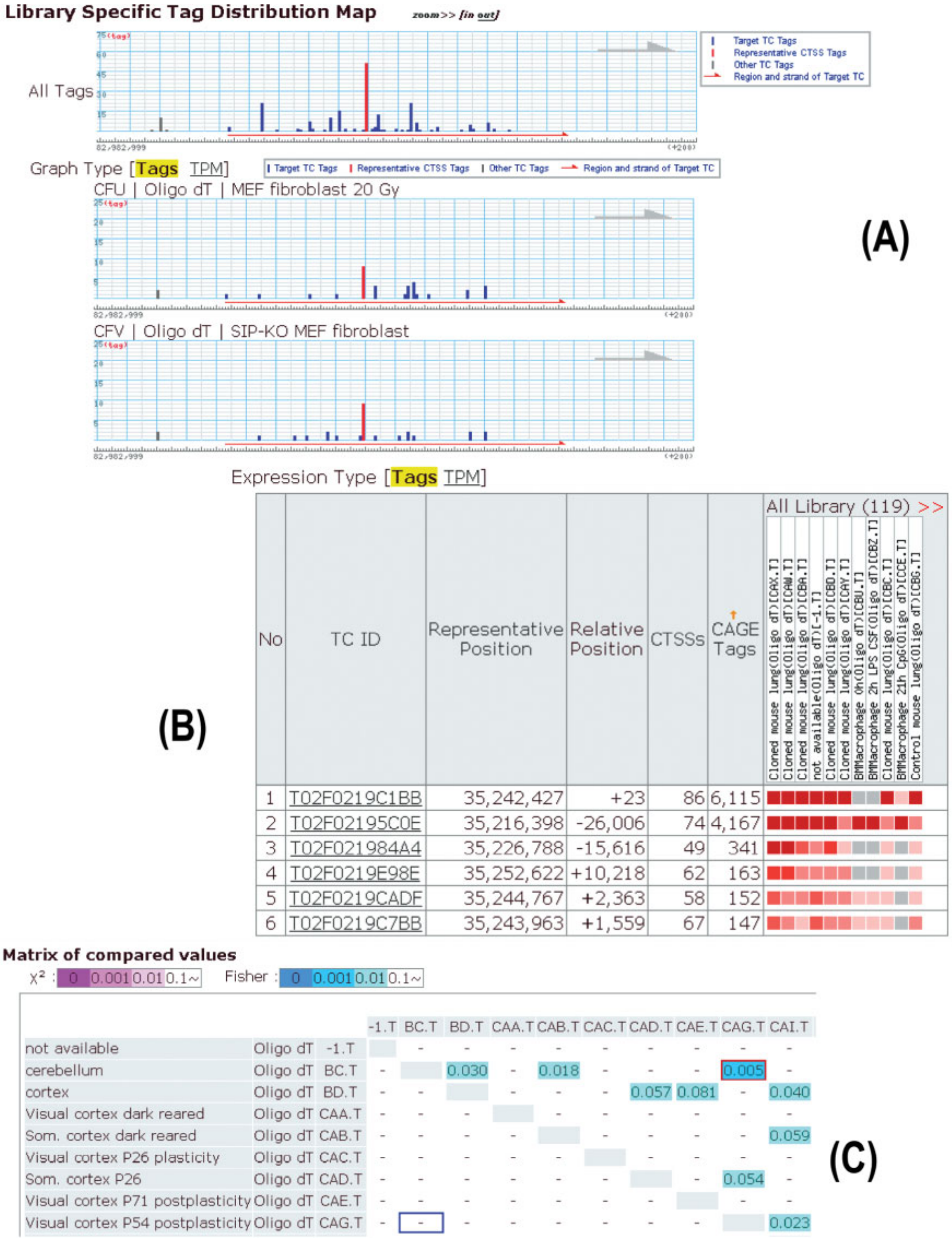
Database	Contents
CAGE Basic Database	RNA sample information
	CAGE library information
	CAGE clone plate/spot
	CAGE clone sequence
	CAGE clone sequence quality
	CAGE tag sequence
	CAGE tag mapping status
	Associations of CAGE tags with CTSS
	Associations of CTSS with TCs
	Associations of TCs with transcript and TUs
CAGE Analysis Database	Base pair level expression profile
	TC expression profile within TUs
	Statistical significance expression fluctuations
	Presence of predicted core promoter elements[a] in upstream region
	Presence of conserved genome region between human and mouse (axtNet)
	Presence of CpG islands
	Counts of TC related mRNA, 5′-EST, GIS/GSC ditags
Genomic Elements Database	TC
	Predicted core promoter elements[a]
	mRNA
	GIS/GSC ditag
	5′- and 3′-ESTs
	Candidates of imprited transcripts in EICO DB
	Transcription factors listed in TFdb
	Gene prediction[b]
	CpG islands[b]
	Repeat detected by repeatmasker and tandem repeats finder[b]
	Assemble gap[b]
	Conserved genome region between human and mouse (axtNet)[b]

[a] TATA box, CCAAT box, GC box and initiator.
[b] Retrieved from the UCSC Genome Browser Database.

CAGE BASIC DATABASE

In the CAGE protocol, 5′ ends of full-length cDNA synthesized from RNA samples are cleaved with MmeI, a class II restriction enzyme, which cleaves 20/18 bp outside the recognition sequence. The cleaved 5′ end cDNA tags (CAGE tags) are ligated to form concatemers and cloned as CAGE clones in CAGE library. After sequencing the CAGE clone, CAGE tag sequences are extracted and mapped computationally onto the genome.

CAGE clone sequence, CAGE tag location on the clone and its genome mapping information are stored to facilitate their traceability. To manage a broad range of RNA samples provided in the FANTOM3 collaboration, RNA sample ID, tissue name, developmental stage, sample treatment,


Library Specific Tag Distribution Map
zoom>> [in out]
All Tags
75(tag)
60
45
30
15
82,982,999
(+200)
Target TC Tags
Representative CTSS Tags
Other TC Tags
Region and strand of Target TC
Graph Type [Tags TPM]
CFU | Oligo dT | MEF fibroblast 20 Gy
25(tag)
20
15
10
5
CFV | Oligo dT | SIP-KO MEF fibroblast
(A)
Expression Type [Tags TPM]
(B)
(C)

No	TC ID	Representative Position	Relative Position	CTSSs	CAGE Tags	Cloned mouse lung(Oligo dT)[CAX.T]	Cloned mouse lung(Oligo dT)[CAW.T]	Cloned mouse lung(Oligo dT)[CBA.T]	not available(Oligo dT)[-1.T]	Cloned mouse lung(Oligo dT)[CBD.T]	Cloned mouse lung(Oligo dT)[CAY.T]	BMMacrophage 0h(Oligo dT)[CBU.T]	BMMacrophage 2h LPS CSF(Oligo dT)[CBZ.T]	Cloned mouse lung(Oligo dT)[CBC.T]	BMMacrophage 21h CpG(Oligo dT)[CCE.T]	Control mouse lung(Oligo dT)[CBG.T]
1	T02F0219C1BB	35,242,427	+23	86	6,115	🟥	🟥	🟥	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥
2	T02F02195C0E	35,216,398	-26,006	74	4,167	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
3	T02F021984A4	35,226,788	-15,616	49	341	🟥	🟥	🟥	🟥	🟥	🟥	⬜	⬜	🟥	⬜	🟥
4	T02F0219E98E	35,252,622	+10,218	62	163	🟥	🟥	🟥	🟥	🟥	🟥	⬜	⬜	🟥	⬜	🟥
5	T02F0219CADF	35,244,767	+2,363	58	152	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬜	🟥
6	T02F0219C7BB	35,243,963	+1,559	67	147	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬜	🟥

All Library (119) >>

Matrix of compared values

χ² : 0 0.001 0.01 0.1~ Fisher : 0 0.001 0.01 0.1~

			-1.T	BC.T	BD.T	CAA.T	CAB.T	CAC.T	CAD.T	CAE.T	CAG.T	CAI.T
not available	Oligo dT	-1.T		-	-	-	-	-	-	-	-	-
cerebellum	Oligo dT	BC.T	-		0.030	-	0.018	-	-	-	0.005	-
cortex	Oligo dT	BD.T	-	-		-	-	-	0.057	0.081	-	0.040
Visual cortex dark reared	Oligo dT	CAA.T	-	-	-		-	-	-	-	-	-
Som. cortex dark reared	Oligo dT	CAB.T	-	-	-	-		-	-	-	-	0.059
Visual cortex P26 plasticity	Oligo dT	CAC.T	-	-	-	-	-		-	-	-	-
Som. cortex P26	Oligo dT	CAD.T	-	-	-	-	-	-		-	0.054	-
Visual cortex P71 postplasticity	Oligo dT	CAE.T	-	-	-	-	-	-	-		-	-
Visual cortex P54 postplasticity	Oligo dT	CAG.T	-	-	-	-	-	-	-	-		0.023

Figure 2. Screenshots of CAGE Analysis Database: (**A**) a view of base pair scale expression within a TC, where CAGE tag count of each genome position is displayed in histogram, (**B**) a view of TC expressions within a TU, in which expression levels are represented by a heat map like representation, (**C**) a view of statistical significances of expression fluctuations between RNA samples, and their *E*-values are displayed in a matrix.

cell type and collaboration name are stored. The amount of the CAGE data derived from each RNA sample is presented to examine if targeted samples are analyzed with CAGE and to which extent CAGE tags in the samples were sequenced.

CAGE ANALYSIS DATABASE

Expression levels are measured by counting associated CAGE tags, and they can be used to measure different levels of expression profiles from base pair to chromosomal band level. Two levels of expression profiles are presented in the CAGE Analysis Database for each RNA sample: base pair scale expressions inside a TC are displayed in histogram (Figure 2A), and TC expressions within a TU are represented by a heat map like representation (Figure 2B). CAGE tag counts and transcripts per million, (tag counts)/(total mapped tag counts in the sample) × 1 000 000, are used as units of expression level. Additionally, statistical significances of expression fluctuations between RNA samples are also accessible in a matrix (Figure 2C). They provide users with graphical views of transcriptional start variation, promoter variation and expression fluctuation of promoters.

Rarely expressed promoters contain only a few tags. Although our RACE experiment using an oligo-capping method supported 91% of the tested cases (5), some CAGE tags could be artifacts caused by some errors in library preparation, sequencing and genome mapping. To provide some evidences for promoters, associations of TCs with (genome) conserved regions (13), CpG islands (14), predicted core promoter elements (15–17) and different transcript counts are stored. Users can search TCs with different reliability levels by specifying search conditions.

GENOMIC ELEMENTS DATABASE

Genomic Elements Database is a supplementary database to the two CAGE databases. The aim is to integrate TCs and other data onto the genome and display them in a conventional way. Generic Genome Browser (18) with MySQL DBMS is used to present a genome view. Candidates of imprinted transcripts in EICO DB (19,20), transcription factors in TFdb (21) and other data in the UCSC Genome Browser Database (22) are stored in addition to the utilized data above. This system is also utilized in full-length cDNA annotation in FANTOM3 (5).

CONCLUSION

The CAGE database systems have successfully provided a large amount of mouse and human CAGE tags derived from various RNA samples for the FANTOM3 project, resulting in biological analyses in various viewpoints. The systems have supported these analyses by providing central views of CAGE resource, promoter and genome position depending on the aspects of interests to researchers. They are publicly available now, and are expected to promote subsequent analyses by using the CAGE resource in scientific research community.

AVAILABILITY

The database systems described here are hyperlinked from http://fantom3.gsc.riken.jp/. Their user's guide, glossary and/or database schema are available from their help pages, and their raw data files, table definitions in SQL and tab-delimited data files, are also available for download from http://fantom3.gsc.riken.jp/download.html.

ACKNOWLEDGEMENTS

We would like to thank A. Hasegawa for data processing; A. Karlsson for English editing; K. Nakano and H. Murakami for building computational systems; K. Yoshida for support; and all members of the FANTOM consortium. This study was 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., a grant of the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan. to Y.H., Grant for the Strategic Programs for R&D of RIKEN to Y.H. Funding to pay the Open Access publication charges for this article was provided by RIKEN.

Conflict of interest statement. None declared.

REFERENCES

1. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al*. (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.
2. Hashimoto,S., Suzuki,Y., Kasai,Y., Morohoshi,K., Yamada,T., Sese,J., Morishita,S., Sugano,S. and Matsushima,K. (2004) 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol*., **22**, 1146–1149.
3. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
4. Harbers,M. and Carninci,P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*, **2**, 495–502.
5. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
6. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J. *et al*. (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564–1566.
7. Ng,P., Wei,C.L., Sung,W.K., Chiu,K.P., Lipovich,L., Ang,C.C., Gupta,S., Shahab,A., Ridwan,A., Wong,C.H. *et al*. (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods*, **2**, 105–111.
8. Suzuki,Y., Yamashita,R., Sugano,S. and Nakai,K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*., **32**, D78–D81.
9. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60 770 full-length cDNAs. *Nature*, **420**, 563–573.
10. Kasukawa,T., Katayama,S., Kawaji,H., Suzuki,H., Hume,D.A. and Hayashizaki,Y. (2004) Construction of representative transcript and protein sets of human, mouse, and rat as a platform for their transcriptome and proteome analysis. *Genomics*, **84**, 913–921.
11. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **33**, D54–D58.

12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
13. Kent,W.J., Baertsch,R., Hinrichs,A., Miller,W. and Haussler,D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA*, **100**, 11484–11489.
14. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol*., **196**, 261–282.
15. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem*., **72**, 449–479.
16. Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*., **212**, 563–578.
17. Schmid,C.D., Praz,V., Delorenzi,M., Perier,R. and Bucher,P. (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res*., **32**, D82–D85.
18. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
19. Nikaido,I., Saito,C., Wakamoto,A., Tomaru,Y., Arakawa,T., Hayashizaki,Y. and Okazaki,Y. (2004) EICO (Expression-based Imprint Candidate Organizer): finding disease-related imprinted genes. *Nucleic Acids Res*., **32**, D548–D551.
20. Nikaido,I., Saito,C., Mizuno,Y., Meguro,M., Bono,H., Kadomura,M., Kono,T., Morris,G.A., Lyons,P.A., Oshimura,M. *et al*. (2003) Discovery of imprinted transcripts in te mouse transcriptome using large-scale expression profiling. *Genome Res*., **13**, 1402–1409.
21. Kanamori,M., Konno,H., Osato,N., Kawai,J., Hayashizaki,Y. and Suzuki,H. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun*., **322**, 787–793.
22. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.