Nucleic Acids Research, 2005, Vol. 33, Database issue ***D25–D28***
doi:10.1093/nar/gki020

DDBJ in collaboration with mass-sequencing teams on annotation

Y. Tateno*, N. Saitou[1], K. Okubo, H. Sugawara and T. Gojobori

Center for Information Biology and DNA Data Bank of Japan and [1]Laboratory of Evolutionary Genetics, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima, 411-8540, Japan

Received September 15, 2004; Revised and Accepted September 17, 2004

ABSTRACT

In the past year, we at DDBJ (DNA Data Bank of Japan; http://www.ddbj.nig.ac.jp) collected and released 1 066 084 entries or 718 072 425 bases including the whole chromosome 22 of chimpanzee, the whole-genome shotgun sequences of silkworm and various others. On the other hand, we hosted workshops for human full-length cDNA annotation and participated in jamborees of mouse full-length cDNA annotation. The annotated data are made public at DDBJ. We are also in collaboration with a RIKEN team to accept and release the CAGE (Cap Analysis Gene Expression) data under a new category, MGA (Mass Sequences for Genome Annotation). The data will be useful for studying gene expression control in many aspects.

INTRODUCTION

As one of the International Nucleotide Sequence Databases (INSD), DDBJ's primary mission is undoubtedly to collect, annotate and release the original and authentic DNA sequence data. In fact, the amount of data collected and released by DDBJ has continued to grow. However, annotation has not caught up with these activities. The reason for that may be attributed to our limited labor power and knowledge that cannot meet biological examination and verification of all submitted data, though this may not be our task. At any rate, this problem will unfortunately remain to be resolved.

Fortunately, however, there have been other approaches to conducting annotation at DDBJ. One of them is to hold or participate in a workshop at which a number of experts systematically annotate some specified data. We actively participated in the annotation jamborees, FANTOM I and II, to annotate and release mouse full-length cDNAs (1). We also hosted the workshops, H-Invitational I and II, to annotate 21 037 human full-length cDNAs (2). The data annotated at the workshops have been made public on http://hinv.ddbj.nig.ac.jp/index.html at DDBJ and on another site at the Japan Biological Information Research Center (JBIRC). Accordingly, we have updated our home page as shown in Figure 1. The numbers of accesses to the H-Invitational data and our paper (2) indicate that both have apparently made a great impact on research communities of biology, medicine and others. A similar activity is in the process for rice genome annotation. As another approach, we are in collaboration with a research team of RIKEN to process and release their data that will be useful for annotation of genome data from different angles.

In this paper, we report our primary activity of data collection and release in the past year and our approach to the collaboration in genome annotation.

RECENT TRENDS IN DATA SUBMISSION

In the past year, we at DDBJ collected and released 1 066 084 entries or 718 072 425 bases. This includes the complete chromosome 22 sequence data of chimpanzee, *Pan troglodyte*. The chimpanzee chromosome was sequenced and submitted by two Japanese, three German, one Chinese, one Korean and one Taiwanese groups (3). This is the first case of data submission for the entire chromosome of non-human primates. Contrary to our expectation for the genetic similarity between man and chimpanzee (4), the new data revealed that 83% of the coding sequences on the corresponding chromosomes differed at the amino acid level between the two species. This would make the relationship between the two species more complicated and interesting than previously considered. Also included is the whole-genome shotgun (WGS) data of silkworm, *Bombyx mori*, submitted by the Silkworm Genome Research Program of the National Institute of Agrobiological Sciences, Japan (5). It covers about 80% of the whole genome of the species.

In addition to the data collection and release, we edit all the data made public by INSD (DDBJ/EMBL/GenBank) and publish as a DDBJ release four times a year. The main reason for editing and publishing the release is to regularly take statistics of the submitted data to INSD in several aspects. For example, the most recent one published as release 58 in June 2004

*To whom correspondence should be addressed. Tel: +81 55 981 6857; Fax: +81 55 981 6858; Email: ytateno@genes.nig.ac.jp

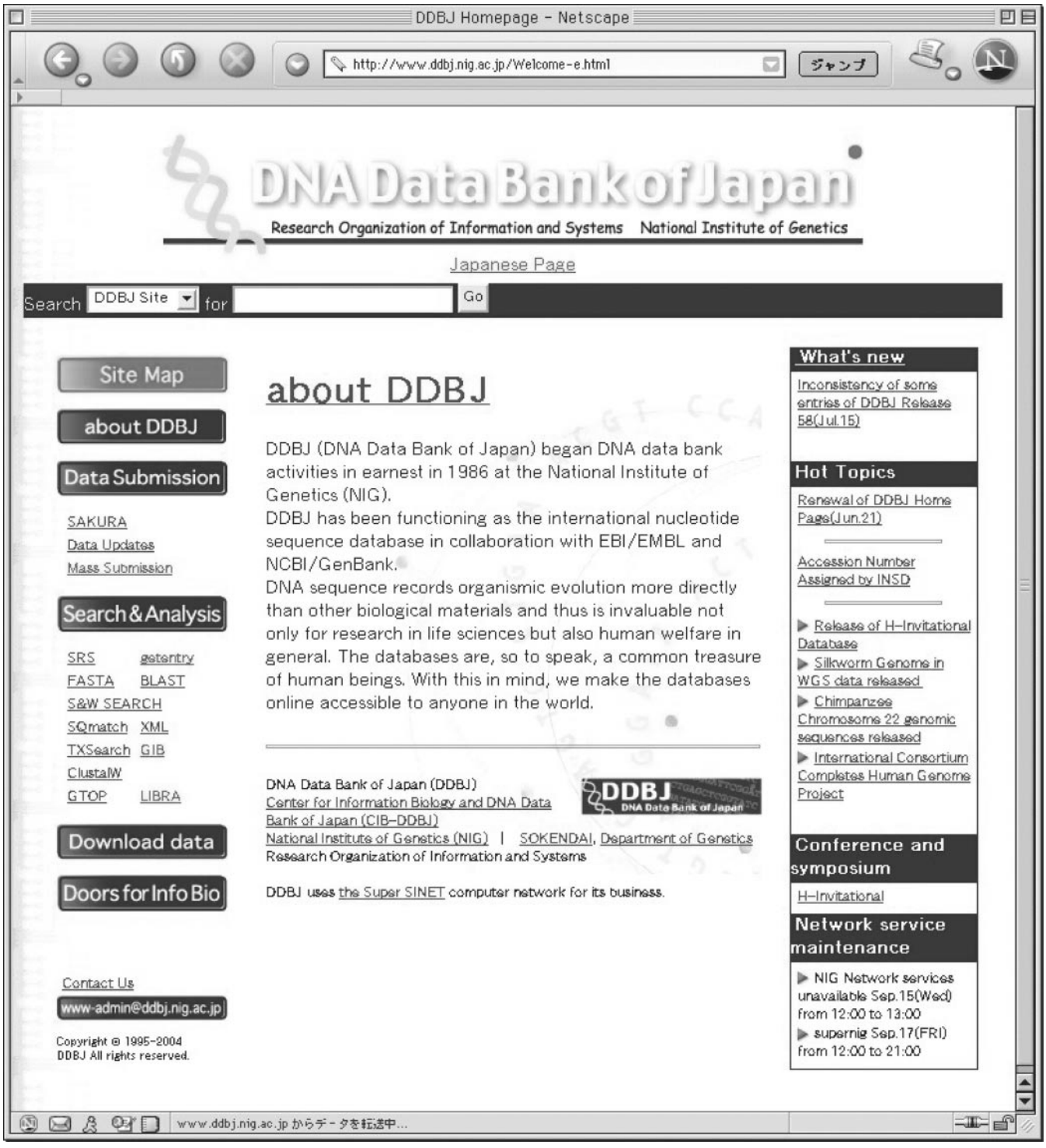

Figure 1. New home page of DDBJ.

contains 34 917 581 entries or 39 812 635 108 bases. Release 58 also shows that the total number of bases increased by 73 billion bases in the past year or 1.23 times as large as the number of the last year.

To indicate the recent trends in data submissions, we extracted and obtained the statistics focusing on the top nine species in the past four years, from 2001 to 2004. The result is given in Figure 2. It is clear from the figure that *Homo sapiens* have been ranked top in the past 4 years. Human genes and genomic regions have been extensively sequenced and submitted even after the completion of human genome sequencing in 2001 (6). The H-Invitational I and II workshops mentioned above apparently contributed to maintaining the human data highest. With the accumulation of

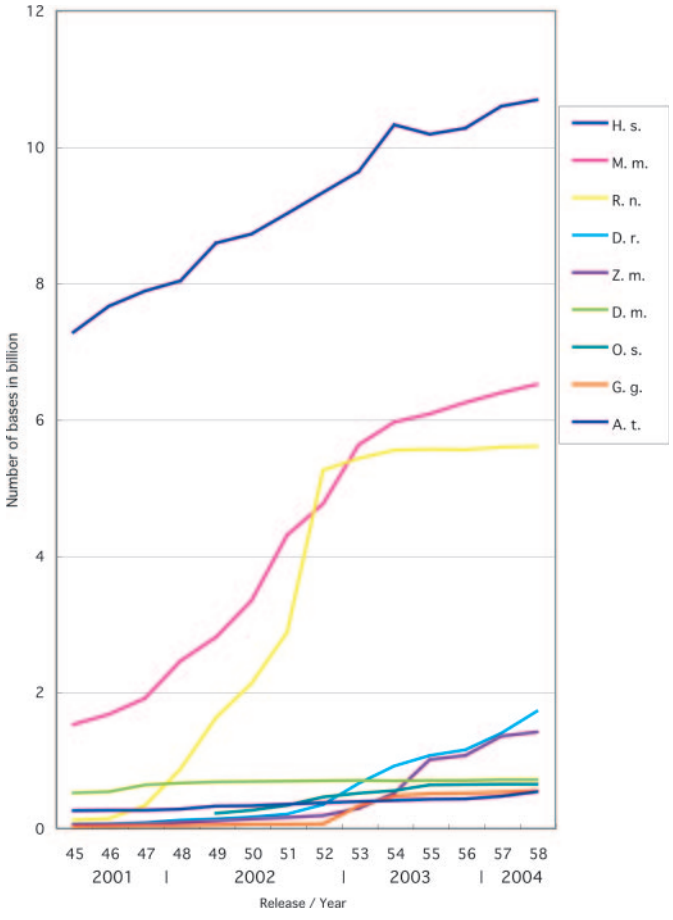

Figure 2. Recent trends in data submission. Successions of data submissions in the past four years are shown for the top nine species. H. s., *Homo sapiens*; M. m., *Mus musculus*; R. n., *Rattus norvegicus*; D. r., *Danio rerio*; Z. m., *Zea mays*; D. m., *Drosophila melanogaster*; O. s., *Oryza sativa*; G. g., *Gallus gallus*; A. t., *Arabidopsis thaliana*.

chimpanzee genome data, human genes and genomic regions will further be sequenced and studied to clarify what makes man a man at the genome level.

The mouse, *Mus musculus*, and rat, *Rattus norvegicus*, data steeply increased in the past years. The two rodent species have been recognized and used as the model experimental animals closest to man. This tendency has particularly been strengthened with some recent genome-scale works for the species (7,8). Below the three major species, several species are clustered in Figure 2. Among them zebrafish, *Danio rerio*, is slightly ahead, indicating that the fish has recently been recognized as the most suitable model system for genetic study of vertebrate development (9). In the cluster, there are three plant species in which maize, *Zea mays*, has been more extensively sequenced than the other two species, *Oryza sativa* and *Arabidopsis thaliana*.

The trends in the data submissions in Figure 2 clearly demonstrate that as research for a particular species progresses, sequence data of the species are produced and accumulate accordingly, which further stimulates the research to advance, and so forth. Data sharing certainly promotes the progress of research. This should be truly understood by those who are not willing to share their data with others.

COLLECTION OF DATA FOR GENOME ANNOTATION

With the accumulation of genome sequence data at INSD, genome research has turned also on non-coding regions such as 5′ UTRs and microRNA regions. Those regions are known to be responsible for regulation of gene expression. However, their roles have not exactly been understood. For example, no one knows completely about how gene expression is regulated at the promoter region. The regulation of gene expression is unquestionably important for understanding many aspects in biology, including development, metabolism, aging and speciation for closely related species. With this in mind, a RIKEN team sequenced a huge number of expressed sequences in 5′ UTR, CAGE (Cap Analysis Gene Expression) sequences, for mouse (10) and plans to submit the data to DDBJ. A CAGE sequence more specifically is the initial 20 bases from a 5′ end mRNA. CAGE is expected to produce 10 000 to 500 000 sequences in a tissue of a species, which makes it possible to conduct high-throughput analysis of gene expression, profiling of transcriptional start points and others.

At the collaborative meeting of INSD in 2004, we thus proposed a new division to accept and release the CAGE data and those similar to them, because we understood and expected that the data would be crucially important for studying comprehensive aspects of promoter usage. The new division was finally accepted and named MGA (Mass sequences for Genome Annotation). The definition of MGA is the sequences that are produced in large quantity in view of genome annotation. MGA thus includes sets of short sequences that are meaningful in the genome context, such as sequences from libraries of CpG islands (11–12) and DNase hypersensitive sites (13–14).

CONCLUDING REMARKS

As gene expression research rapidly advances, microarray data have accumulated at many individual laboratories. Some of the data were submitted to the public databases such as ArrayExpress, GEO and CIBEX. The open letter (15) sent by the MGED (Microarray Gene Expression Data) Society to the editors of relevant journals will certainly accelerate data submission to the public databases. Then the user of the public gene expression databases will demand the link between the gene expression databases and INSD, because they naturally need information not only on gene expression but also on the relevant DNA sequences. Perhaps, discussion about gene expression will not be settled until information on the pertinent DNA sequences is provided. Currently, the only vehicle connecting these two types of the databases is the accession number issued by INSD.

REFERENCES

1. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
2. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y. *et al.*

(2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*., **2**, 856–875.
3. Watanabe,H., Fujiyama,A., Hattori,M., Taylor,T.D., Toyoda,A., Kuroki,Y., Noguchi,H., BenKahla,A., Lehrach,H., Sudbrak,R. *et al*. (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, **429**, 382–388.
4. Chen,F.C. and Li,W.-H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet*., **68**, 444–456.
5. Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H., Yasukochi,Y. *et al*. (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res*., **11**, 27–35.
6. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
7. Gibbs,R.A., Weinstock,G.M., Metzker,M.L., Muzny,D.M., Sodergren,E.J., Scherer,S., Scott,G., Steffen,D., Worley,K.C., Burch,P.E. *et al*. (2004) Rat Genome Sequencing Project Consortium, Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
8. 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.
9. Detrich,H.W.III, Westerfield,M. and Zon,L.I.(eds) (1999) *The Zebrafish*. Academic Press, San Diego, CA.
10. 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.
11. Cross,S.H., Clark,V.H. and Bird,A.P. (1999) Isolation of CpG islands from large genomic clones. *Nucleic Acids Res*., **27**, 2099–2107.
12. Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet*., **6**, 236–244.
13. Sabo,P.J., Humbert,R., Hawrylycz,M., Wallace,J.C., Dorschner,M.O., McArthur,M. and Stamatoyannopoulos,J.A. (2004) Genome-wide identification of DNase I hypersensitive sites using active chromatin sequence libraries. *Proc. Natl Acad. Sci. USA*, **101**, 4537–4542.
14. Crawford,G.E., Holt,I.E., Mullikin,J.C., Tai,D., Blakesley,R., Bouffard,G., Young,A., Masiello,C., Green,E.D., Wolfsberg,T.G. and Collins,F.S. (2004) Identifying gene regulatory elements by genome-wide recovery of Dnase hypersensitive sites. *Proc. Natl Acad. Sci. USA*, **101**, 992–997.
15. Ball,C.A., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H., Quackenbush,J., Ringwald,M. *et al*. (2004) Submission of microarray data to public repositories. *PLoS Biol*., **2**, 1276–1277.