D16–D18 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 16 October 2008*
doi:10.1093/nar/gkn724

DDBJ dealing with mass data produced by the second generation sequencer

Hideaki Sugawara, Kazuho Ikeo, Satoshi Fukuchi, Takashi Gojobori and Yoshio Tateno*

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

Received September 18, 2008; Accepted September 30, 2008

ABSTRACT

DNA Data Bank of Japan (DDBJ) (http://www.ddbj.nig.ac.jp) collected and released 2 368 110 entries or 1 415 106 598 bases in the period from July 2007 to June 2008. The releases in this period include genome scale data of *Bombyx mori*, *Oryzas latipes*, *Drosophila* and *Lotus japonicus*. In addition, from this year we collected and released trace archive data in collaboration with National Center for Biotechnology Information (NCBI). The first release contains those of *O. latipes* and bacterial meta genomes in human gut. To cope with the current progress of sequencing technology, we also accepted and released more than 100 million of short reads of parasitic protozoa and their hosts that were produced by using a Solexa sequencer.

INTRODUCTION

As a member of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/), DDBJ has steadily collected, annotated, released and exchanged the original DNA sequence data, which, for example, is shown by a growth curve of the data submissions in the past years (visit http://www.ddbj.nig.ac.jp/images/breakdown_stats/percentage-e.gif). However, the current situation of data submissions is dramatically changing due to the emergence of ultra high speed or the 2nd generation sequencers (2GS), such as 454 (by 454 Life Sciences, Branford, USA), Solexa (by Illumina, Inc., San Diego, USA), SOLiD (by Applied Biosystems, Foster City, USA) and Helicos (by Helicos BioSciences Corporation, Cambridge, USA). With those machines the whole human genome could now be sequenced at one-thousandth or less speed of the first cases in 2001 (1,2). Recently, two reports announced that the whole genome was sequenced for two well-known persons (3,4), which was perhaps the beginning of personal genomics. Also known is the 1000 human genomes project that is underway in USA, Europe and China to obtain a complete and detailed catalogue of genetic variations of humans (http://www.1000genomes.org/page.php). Those activities warn us that the above growth curve will drastically be steepen. At present, INSDC release about 100 billion bases in total. This is the outcome of the collaboration among the three member banks for >20 years. However, this number will easily be surpassed when the 1000 human genomes project is completed and the result is submitted to INSDC in a few years, or even before that.

To cope with those activities INSDC collaborators discussed in 2008 the attitude towards handling mass submissions produced by 2GS. The common fear among the collaborators was limited computer storages that will sooner or later be filled with continuously coming mass submissions. Nevertheless, the collaborators agreed to collect, distribute and exchange mass data of transcriptomes, such as trace archives (TRA) and short reads (SR), upon the condition that the sequences are assembled. DDBJ has also started to accept and release such mass sequence data. In the following, DDBJ's activity is reported focusing mainly on mass data submissions from Japanese universities and institutes.

COLLECTION OF ORDINARY DATA IN THE PAST YEAR

In the period from July 2007 to June 2008, DDBJ collected, annotated and released the original data of 2 368 110 entries or 1 415 106 598 bases. More than 90% of the data came from Japanese researchers and Japan Patent Office (JPO), and the rest were mainly from researchers in China, Korea and Taiwan.

The released data newly include 282 117 entries of patent data from Korean Industrial Property Office (KIPO) that will continue to send their data to DDBJ for public release. The other portion of the released data contains WGS, GSS (fosmid ends and BAC ends) and HTG (BAC clones) of silkworm (*Bombyx mori*) submitted by National Institute of Agrobiological Sciences; EST entries of medaka (*Oryzas latipes*) submitted by National Institute of Basic Biology; EST entries of *Drosophila simulans*, *D. sechellia* and *D. auraria* submitted

*To whom correspondence should be addressed. Tel: +81 55 981 6857; Fax: +81 55 981 6858; Email: ytateno@genes.nig.ac.jp
The authors wish it to be known that, in their opinion, the all authors should be regarded as joint First Authors.

by Kyoto Institute of Technology and WGS and PLN of *Lotus japonicus* by Kazusa DNA Research Institute. Those data can be obtained at the DDBJ ftp site (http://www.ddbj.nig.ac.jp/ftp_soap-e.html).

It may be worthwhile to refer to the data on *L. japonicus* among them. This plant is widely used as a model organism to study symbiotic nitrogen fixation. This species experienced whole-genome duplication in evolution, and the genome is now composed of six linkage groups that together contain about 30 000 genes (5). The number of the genes is in agreement with that of *Arabidopsis thaliana* for which the number was estimated as 29 500 (6). These results may suggest that the number of genes for an angiosperm species is about 30 000, unless the species has experienced further genome duplication in evolution.

COLLECTION AND RELEASE OF TRA DATA

TRA is a repository of DNA sequence chromatograms (traces), base calls and quality estimates for a single-pass reads from a large-scale sequencing project. TRA data could be useful for confirming SNP sites in question, and, once assembled, provide information for finding new ORFs or genes. With the support by National Project of Integrating Life Science Databases in Japan (ILSD, http://dbcls.rois.ac.jp/en/), we are now able to collect and release TRA data at DDBJ. The released data are as follows.

(1) TRA data of *O. latipes* WGS sequences: The data were submitted by National Institute of Genetics and released at the DDBJ ftp site mentioned above. The data were also sent to National Center for Biotechnology Information (NCBI) TRA Repository (NTR, http://www.ncbi.nlm.nih.gov/Traces/home) and their TI numbers were given by NTR. The total number of entries is about 1.5 millions and the TI numbers without the first three digits (209) are 5 022 956–5 389 675, 5 396 176–6 435 759 and 6 858 496–6 933 759. The length of each entry is several thousand bases. Using any of these numbers one can retrieve at NTR and observe the chromatogram of the entry with the number. The data were also assembled to 24 entries with accession numbers, DG000001–DG00024, (see http://medaka.utgenome.org/ for more details).

(2) TRA data of meta bacterial-genomes in human gut: The data were submitted by University of Tokyo, RIKEN and other universities and institutes (7) and released at the DDBJ ftp site. The samples taken from 13 healthy individuals revealed 237 gene families in the adults and 136 gene families for the infants, though the names of the bacteria in the samples were not identified (7). Another interesting finding is the existence of a conjugative transposon family that could mediate gene transfer between bacteria in the samples (7). Similarly, TI numbers given by NTR without the first three digits (209) are 7 946 941–9 007 079.

COLLECTION OF DATA PRODUCED BY 2GS

2GS, Solexa for example, can produce more than 1 billion sequences per run with the accuracy of 99.9% in several days, though the length of each sequence is very short and thus called SR. However, SR could be valuable if the reference genome sequence to them is available, and assembled against it. In this sense, 2GS is quite powerful for the study of personal (or individual) genomics, population genetics and diagnostic medicine among others. SR data could also be useful for studying the gene expression patterns of a species. Therefore, INSDC set up an archive for SR data as Short Reads Archive (SRA). The participation of DDBJ in SRA is also supported by ILSD.

DDBJ received a tremendous amount of sequence data from Genome Sequence Center of Tokyo University. The submitters used a Solexa machine to sequence full-length cDNAs of eight species, *Plasmodium faliparum*, *P. vivax*, *P. yoelii*, *P. berghei*, *Toxoplasma gondii*, *Cryplosporidium* sp., *Anopheles stephensi* and *Glossina* sp. The first six are parasitic pathogens and the last two are host species. In particular, the first four and the seventh are known to be malarial pathogens and their host, respectively. The length of each entry is 36 or 48 bases due to the specification of Solexa, and the total number of entries is more than 100 millions in the present submission (Table 1). As long as the

Table 1. Species and amounts of submitted short reads

Species	Block	Read Length
Toxoplasma_v2	200	36
Toxoplasma_2nd	300	36
Toxoplasma_v1	300	36
Cryptosporidium_ref	300	36
Cryptosporidium_nref	300	36
Cryptosporidium_2nd	300	36
Plasmodium yoelii_ref	300	36
Plasmodium yoelii	300	36
Plasmodium yoelii_xzl_nref	300	36
Plasmodium yoelii_xzl_ref	300	36
Plasmodium yoelii_xzn_nref	300	36
Plasmodium yoelii_2nd1	300	36
Plasmodium yoelii_2nd2	300	36
P. falciparum_v1	300	36
P. falciparum_2nd1	300	36
P. falciparum_2nd2	300	36
P. falciparum_v1	300	36
P. falciparum_v2	300	36
P. vivax	200	36
P. vivax_ref1	100	36
P. vivax_ref2	100	36
P. vivax_nref	100	36
P. vivax_2nd2	100	36
P. vivax_2nd1	100	36
P. vivax_2nd3	100	36
Babesia bovis_2nd1	100	36
Babesia bovis_2nd2	100	36
P. berghei_2nd	300	36
P. berghei	200	36
Anopheles stephensi_tss	100	48
Anopheles stephensi2nd_1	100	48
Anopheles stephensi2nd_2	100	48
Anopheles stephensi2nd_3	100	48
Glossina_pup_tss	100	36
Glossina_pup_2nd_1	100	48
Glossina_pup_2nd_2	100	48
Glossina_lar_tss	100	36
Glossina_lar_2nd_1	100	48
Glossina_lar2nd_2	100	48

1 block contains 20 000–30 000 SR each of which is 38 or 48 bases in length.

number of entries is concerned, the present submission alone exceeds the total number of ordinary entries that INSDC together have collected and released since 1980. This implies something; the new sequencing technology will perhaps change biology considerably. Individualized biology could emerge in the near future. Namely, biologists would focus intensively on individual genomic characters and the difference between them to elucidate what life really is.

The SR data were released from DDBJ and the SRA repository at NCBI. We have been informed that more SR data will soon be submitted to DDBJ from Japanese universities and institutes. One problem with sending such a tremendous amount of data through Internet would be traffic congestion and an extremely slow rate, even if transmission is possible. We have learned that as long as the data amount is <50 GB the transmission can be done within a few hours. However, we have to resolve two problems to realize and promote individualized biology in the future, capacities of computer and Internet.

REMARKS

As personal genomes can be scrutinized now by the state-of-the-art sequencing technology, one problem emerges. One's genome is not only one's property but also one's ancestors' and descendants'. We are products of evolution. We will not be able to freely publicize the contents of our genomes. The genome of a person hides many recessive inferior genes that are shared with his parents and children (3). In general, children would oppose to sequencing the genome of their parents or *vice versa*. It is thus necessary to pay great care and attention in handling or dealing with person's genome contents.

ACKNOWLEDGEMENTS

We thank all staff of DDBJ for the data collection, annotation, release, management and software development. In particular, we are grateful to Tomohiro Koike and Makoto Yamamoto for their engagements in the collection and release of TRA and SR data.

FUNDING

DDBJ is funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) with the management expenses grant for national university cooperation. DDBJ is also supported by a grant from National Project of Integrating Life Science Databases. Funding to pay the open access publication charges for this article was provided by the Japan Society for the Promotion of Science.

Conflict of interest statement. None declared.

REFERENCES

1. 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) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.-J., Makhijani,V., Roth,G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
4. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2008) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, 2113–2144.
5. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, *Lotus japonicus, DNA Res.* [Epub ahead of print; doi:10.1093/dnares/dsn008].
6. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
7. Kurosawa,K., Itoh,T., Kuwahara,T., Oshima,K., Toh,H., Toyoda,A., Takami,H., Morita,H., Sharme,V.K., Srivastava,T.P. *et al.* (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* [Epub ahead of print; doi: 10.1093/dnares/dsm018].