DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data

Ayako Suzuki¹, Hiroyuki Wakaguri², Riu Yamashita³, Shin Kawano⁴, Katsuya Tsuchihara⁵, Sumio Sugano¹, Yutaka Suzuki²,* and Kenta Nakai⁶,*

¹Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan, ²Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan, ³Tohoku Medical Megabank Organization, Tohoku University, Miyagi, Japan, ⁴Database Center for Life Science, Research Organization of Information and Systems, Chiba, Japan, ⁵Division of TR, The Exploratory Oncology Research and Clinical Trial Center, National Cancer Center, Chiba, Japan and ⁶Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan

Received September 15, 2014; Revised October 16, 2014; Accepted October 16, 2014

ABSTRACT

DBTSS (http://dbtss.hgc.jp/) was originally constructed as a collection of uniquely determined transcriptional start sites (TSSs) in humans and some other species in 2002. Since then, it has been regularly updated and in recent updates epigenetic information has also been incorporated because such information is useful for characterizing the biological relevance of these TSSs/downstream genes. In the newest release, Release 9, we further integrated public and original single nucleotide variation (SNV) data into our database. For our original data, we generated SNV data from genomic analyses of various cancer types, including 97 lung adenocarcinomas and 57 lung small cell carcinomas from Japanese patients as well as 26 cell lines of lung cancer origin. In addition, we obtained publically available SNV data from other cancer types and germline variations in total of 11,322 individuals. With these updates, users can examine the association between sequence variation pattern in clinical lung cancers with its corresponding TSS-seq, RNA-seq, ChIP-seq and BS-seq data. Consequently, DBTSS is no longer a mere storage site for TSS information but has evolved into an integrative platform of a variety of genome activity data.

INTRODUCTION

With our unique oligo-capping technique (1), it has become apparent that the transcriptional start site (TSS) position is not a single fixed point but is observed as a peak with various distribution widths in many genes (2), which was later confirmed by the FANTOM project in a larger scale (3). To know accurate TSS positions is valuable as it could lead to more accurate characterization of its upstream transcriptional regulatory region. Thus, we constructed a database containing such information of mostly human genes in 2002. Since then, its updates have been regularly reported in the Nucleic Acids Research database issues (2004, 2006, 2008, 2010 and 2012 (4)). With the advances in sequencing technologies, we have developed TSS-seq, where the oligo-capping technique is applied to next generation sequencing (NGS), allowing even more accurate genome-wide determination of TSSs (5). The NGS sequencers are not only suited for determining genomic DNA sequences but also for transcriptome analysis (RNA-seq (6)) and epigenome analysis (ChIP-seq (7) and bisulfite sequencing (BS-seq; (8))). Since such additional data enable further biological characterization of transcriptional regulatory regions, we have also incorporated transcriptomic/epigenomic data of various tissues/cell cultures in DBTSS. In the latest update, Release 9, it contains 1257 million TSS tag sequences collected from 24 tissues and 33 cell cultures (see Table 1). It also contains the data of subcellularly fractionated RNAs as well as the ChIP-seq data of various histone modifications, binding sites of RNA polymerase II and several transcription factors, mainly, in cultured cell lines.

In this report, we introduce Release 9 of DBTSS, where we significantly enlarged the number of incorporated single nucleotide variation (SNV) data, which were both collected from publically available databases and generated from our own experiments (see below). The association of such large-scale SNV data and the multi-omics data should be useful for finding regulatory SNVs that play important roles in diseases, especially lung cancers, the data of which we have extensively collected.
Figure 1. Basic usage. (A) Top page of DBTSS. A simple search for ‘TSS Viewer’ and ‘Genome Viewer’ can be made by specifying a keyword, such as a gene name ‘BRAF’ in the Database Search at the left frame (red box). Search by ‘SNV Summary in Cancers’ and ‘Pathway Map’ can be made from the positions indicated by orange and purple boxes, respectively. (B) A part of the TSS Viewer display for the BRAF gene. The overview and the detailed positions of the TSSs are shown in the upper and lower panels, respectively. Many of the fields are expandable. (C) The default display of Genome Viewer for the BRAF gene. Displayed items are as indicated in the margin. The displayed items can be controlled from the panels located under the ‘Select track items’ headline. (D) A sample output of SNV information for the BRAF gene. Surrounding region of a previously reported cancer driver mutation (V600E of the BRAF gene; highlighted in red box), is displayed.
Table 1. Statistics of the data sets

| Data set     | Number of samples | Number of used tags (Average per data set) |
|--------------|-------------------|--------------------------------------------|
| TSS-seq      | 73                | 16,620,753                                  |
| RNA-seq      | 42                | 31,880,393                                  |
| ChIP-seq     | 255               | 19,493,875                                  |
| RIP-seq      | 12                | 1,386,112                                   |
| BS-seq       | 26                | 113,946,186                                 |
| ChromHMM     | 36                | n.d.                                        |
| SNV          | 49                | 1,022,073,467                               |

n.d., not determined.

Figure 2. Other useful information. (A) Upper panel: A part of the Mutation frequency table for the BRAF gene. Enriched fields are as highlighted; lower panel: Summary of multi-omics data mainly collected from cell lines. (B) Pathway Map representation of characteristic genes. In this example, gene expression level (in RPKM) of node genes in a lung adenocarcinoma cell line, LC2/ad, in the ErbB/HER signaling pathway is shown. Further links will appear when the users click the circle corresponding to each gene.

Table 2. SNVs registered in the database

| Data source          | Definition                          | Number of samples | Reference |
|----------------------|-------------------------------------|-------------------|-----------|
| NCCE, Japan          | Lung adenocarcinoma                 | 97                | PLoS One. 2013 8(9):e73484. |
| NCCE, Japan          | Small cell lung cancers             | 57                | J Thorac Oncol. 2014 9(9):1324–1331. |
| ICGC                 | 43 of ICGC DCC Project Codes       | 6590              | [https://dcc.icgc.org/](https://dcc.icgc.org/) |
| Meyerson’s Lab.      | Lung adenocarcinoma                 | 183               | Cell. 2012 150(6):1107–1120. |
| Ogawa’s Lab.         | Myelodysplasia                      | 29                | Nature. 2011 478(7367):64–69 |
|                     | Clear-cell renal cell carcinoma     | 106               | Nat Genet. 2013 45(8):860–867 |
| TCGA                 | Gastric adenocarcinoma              | 295               | Nature. 2014 513(7517):202-209 |
|                     | Urothelial bladder carcinoma        | 131               | Nature. 2014 507(7492):315–322. |
|                     | Glioblastoma                        | 291               | Cell. 2013 155(2):462–477. |
|                     | Clear cell renal cell carcinoma     | 446               | Nature. 2013 499(7456):43–49. |
|                     | Endometrial carcinoma               | 373               | Nature. 2013 497(7447):67–73. |
|                     | Acute myeloid leukemia              | 200               | NEJM. 2013 368(22):2059–2074. |
|                     | Breast tumors                       | 507               | Nature. 2012 390(7418):61–70. |
|                     | Squamous cell lung cancers          | 178               | Nature. 2012 489(7417):519–525. |
|                     | Colon and rectal cancer             | 224               | Nature. 2012 487(7407):330–337. |
|                     | Ovarian carcinoma                   | 316               | Nature. 2011 474(7353):609–615. |
|                     | Glioblastoma                        | 91                | Nature. 2008 455(7216):1061–1068. |
| HGVD                 | Normal tissues (Japanese)           | 1208              | [http://www.genome.med.kyoto-u.ac.jp/SnpDB](http://www.genome.med.kyoto-u.ac.jp/SnpDB) |

Total 11,322
NEWLY INCORPORATED SNV DATA
It is now widely accepted that a number of SNVs in transcriptional regulatory regions are related to many human diseases, including cancers. However, it remains difficult to identify such functionally relevant SNVs from many other neutral SNVs. To overcome this difficulty, intensive efforts should be made in (i) collecting SNVs systematically from a variety of diseases and/or sources (i.e. different cells/tissues/cell cultures) and in (ii) associating such SNV information at each locus with any related functional information, such as the expression profiles of neighboring genes and their surrounding epigenetic profiles. For the first point, we have identified SNVs from genome analyses of various types of cancers, including somatic mutations of 97 lung adenocarcinoma (9) and 57 lung small cell carcinoma (10) (Table 2). In addition, we have collected SNVs from various public resources, such as ICGC (https://icgc.org/) (11) and TCGA (http://cancergenome.nih.gov/) (12) as well as the germline variation data of 1000 Japanese individuals (Table 2; also see our website for the data content and references). We also considered representative cancer mutations in COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/). For the second point, we have generated a series of data sets of ChIP-seq, BS-seq, RNA-seq and TSS-seq in 26 cell lines of lung cancer or sources (i.e. different databases) and in (ii) associating such SNV information with any related functional information, such as the expression profiles of neighboring genes and their surrounding epigenetic profiles. For the first point, we have identified SNVs from genome analyses of various types of cancers, including somatic mutations of 97 lung adenocarcinoma (9) and 57 lung small cell carcinoma (10) (Table 2). In addition, we have collected SNVs from various public resources, such as ICGC (https://icgc.org/) (11) and TCGA (http://cancergenome.nih.gov/) (12) as well as the germline variation data of 1000 Japanese individuals (Table 2; also see our website for the data content and references). We also considered representative cancer mutations in COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/). For the second point, we have generated a series of data sets of ChIP-seq, BS-seq, RNA-seq and TSS-seq in 26 cell lines of lung cancer origin when determining the SNVs in lung adenocarcinoma cell lines (13). Now the database contains a total of 73 TSS-seq data sets (1,266,782,562 TSS tags in total), 420 ChIP-seq and other next generation sequence data sets, which are associated with the SNV/SNP data of 11,322 individuals (Table 1).

SAMPLE USAGES
(i) Basic Tour
At the top page (http://dbtss.hgc.jp/; Figure 1A), users can specify any human gene name in the Keyword field on the top left (keyword is case insensitive). For example, Figure 1B shows the result when a keyword ‘BRAF’, a frequently mutated gene in various types of cancers, especially in melanomas, is searched. The page contains the TSS information in normal adult and fetal tissues as well as eight cell lines, such as DLD-1. The information of ChIP-seq, RNA-seq and RIP (RNA-immunoprecipitation)-seq, which is for characterizing miRNA-mediated regulation (14), also appears if available. At the bottom, detailed TSS distribution patterns at single-base resolution are shown. When available, SNP information from dbSNP is also shown. For more detailed multi-omics information, users can employ our newly implemented Genome Viewer from the link, which is on the top of the page. In this viewer, users can control whether to show or to hide the display of all available information, such as cells and histone marks. The default display of BRAF is shown in Figure 1C (please consult the help file in the database to know how to reproduce this display). In this figure, the users can confirm that the TSS region is rich in active marks, such as the H3K4me3 and H3K9/14ac marks, while repressive marks, such as H3K27me3 and H3K9me3, are poor and DNA methylation level is low.

SNV information, which is stored in recent cancer genomics data sets, such as TCGA and ICGC, as well as in our uniquely generated data sets of Japanese cancer patients, is also integrated and can be browsed. For example, users can double click to magnify the genomic position of chr7:140753336 to examine the prevalence of this previously well-characterized mutation in cancers (BRAF V600E; (11,12)) in other cell types (Figure 1D). At the same time, the users can browse the epigenome and transcriptome information in the surrounding region by clicking the buttons in the indicated sections, which are at the bottom of the main viewer.

(ii) Mutation frequency information
As another example, users can specify any gene name in the ‘SNV Summary in Cancers’ field at the left column of the top window. When the users search by ‘BRAF’, two tables appear as shown in Figure 2A. The first table shows the mutation frequency of this gene at distal upstream region (from −50 to −1 kb; when TSS is designated as 0), proximal promoter region (−1 to +1 kb) and genomic regions (TSS to the 3′-end of the gene model) in a variety of cancer types. In this example, users will find that mutations in its coding region are as frequent in head and neck thyroid carcinomas (ICGC) as in melanoma (TCGA). Users may also find that BRAF mutations are observed in its proximal region of the TSS in many other cancers as well, although their functional consequence remains to be characterized (upper panel; Figure 2A). To further investigate the biological relevance of the mutations, the second table may give a clue by showing the summary of RNA-seq and ChIP-seq data for this gene, which were collected from a series of cell lines. In the example of the BRAF gene, the users can obtain the information that the RNA is expressed relatively ubiquitously among cell lines despite the intensities of histone marks are rather variable (lower panel; Figure 2A).

(iii) The Pathway Map
The third unique feature of the database is the Pathway Map representation of the information. When users click the Search button of the Pathway Map field, which is found at the left bottom of the top page, lists of pathways, including our original pathway maps, which were constructed from the CST pathways of Cell Signaling Technology, Inc. (http://www.cellsignal.com/), and the KEGG pathways (http://www.genome.jp/kegg/), are shown. In the pathway diagram, gene products that show some chosen characteristics, such as higher level in either its expression or any histone mark, are highlighted (Figure 2B). The users can choose a cancer type from a variety of sources and the cells/cell lines as they wish to check the mutation patterns of genes belonging to the pathway.

AVAILABILITY
A detailed user manual and a document on experimental procedures are also available on the website (http://dbtss.hgc.jp/docs/help_2014.html). Statistics for the current database are also presented in the statistics section.
CONCLUDING REMARKS

In this release of DBTSS, there has been significant advance in its capability for suggesting potentially important SNVs especially in cancers. With these enhancements, DBTSS will continue to be a useful resource in the age of clinical genomics.

ACKNOWLEDGEMENTS

The database is maintained on the supercomputer system at Human Genome Center, the Institute of Medical Science, the University of Tokyo. We thank Minoru Kanehisa and Cell Signaling Technology, Inc, for letting us use the information of KEGG and CST pathways respectively, in our database.

FUNDING

DBTSS is financially supported with a Grant-in-Aid for Publication of Scientific Research Results (Databases) by Japan Society for the Promotion of Science, a Grant-in-aid for Scientific Research on Innovative Areas ‘Genome Science’ [221S0002] from the Ministry of Education, Culture, Sports, Science and Technology of Japan and Togo Database Project from Japanese Agency for Science and Technology. Funding for open access charge: Publication charges are covered by Togo Database Project from Japanese Agency for Science and Technology.

REFERENCES

1. Suzuki, Y. and Sugano, S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. Methods Mol. Biol., 221, 73–91.

2. Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S. et al. (2001) Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep., 2, 388–393.

3. Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Lassmann, T., Itoh, M., Summers, K.M., Suzuki, H., Daub, C.O. et al. (2014) A promoter-level mammalian expression atlas. Nature, 507, 462–470.

4. Yamashita, R., Sugano, S., Suzuki, Y. and Nakai, K. (2012) DBTSS: Database of Transcriptional Start Sites progress report in 2012. Nucleic Acids Res., 40, D150–D154.

5. Tsuchihara, K., Suzuki, Y., Wakaguri, H., Irie, T., Tanimoto, K., Hashimoto, S., Matsuhashi, K., Mizushima-Sugano, J., Yamashita, R., Nakai, K. et al. (2009) Massive transcriptional start site analysis of human genes in hypoxia cells. Nucleic Acids Res., 37, 2249–2263.

6. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621–628.

7. Furey, T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet., 13, 840–852.

8. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462, 315–322.

9. Suzuki, A., Mimaki, S., Yamane, Y., Kawase, A., Matsushima, K., Suzuki, M., Goto, K., Sugano, S., Esumi, H., Suzuki, Y. et al. (2013) Identification and characterization of cancer mutations in Japanese lung adenocarcinoma without sequencing of normal tissue counterparts. PLoS ONE, 8, e73484.

10. Unemura, S., Mimaki, S., Makinoshima, H., Tada, S., Ishii, G., Ohmatsu, H., Niño, S., Yoh, K., Matsumoto, S., Takahashi, A. et al. (2014) Therapeutic priority of the PI3K/AKT/mTOR pathway in small cell lung cancers as revealed by a comprehensive genomic analysis. J. Thorac. Oncol., 9, 1324–1331.

11. International Cancer Genome Consortium. (2010) International network of cancer genome projects. Nature, 464, 993–998.

12. The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet., 45, 1113–1120.

13. Suzuki, A., Makinoshima, H., Wakaguri, H., Esumi, H., Sugano, S., Kohno, T., Tsuchihara, K. and Suzuki, Y. (2014) Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res., doi:10.1093/nar/gku885.

14. Kanematsu, S., Tanimoto, K., Suzuki, Y. and Sugano, S. (2014) Screening for possible miRNA-mRNA associations in a colon cancer cell line. Gene, 533, 520–531.