D150–D154 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 15 November 2011*
doi:10.1093/nar/gkr1005

DBTSS: DataBase of Transcriptional Start Sites progress report in 2012

Riu Yamashita[1,2], Sumio Sugano[3], Yutaka Suzuki[3] and Kenta Nakai[2,*]

[1]Frontier Research Initiative, Institute of Medical Science, [2]Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and [3]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, 277-8568, Japan

Received September 15, 2011; Accepted October 19, 2011

ABSTRACT

To support transcriptional regulation studies, we have constructed DBTSS (DataBase of Transcriptional Start Sites), which contains exact positions of transcriptional start sites (TSSs), determined with our own technique named TSS-seq, in the genomes of various species. In its latest version, DBTSS covers the data of the majority of human adult and embryonic tissues: it now contains 418 million TSS tag sequences from 28 tissues/cell cultures. Moreover, we integrated a series of our own transcriptomic data, such as the RNA-seq data of subcellular-fractionated RNAs as well as the ChIP-seq data of histone modifications and the binding of RNA polymerase II/several transcription factors in cultured cell lines into our original TSS information. We also included several external epigenomic data, such as the chromatin map of the ENCODE project. We further associated our TSS information with public or original single-nucleotide variation (SNV) data, in order to identify SNVs in the regulatory regions. These data can be browsed in our new viewer, which supports versatile search conditions of users. We believe that our new DBTSS will be an invaluable resource for interpreting the differential uses of TSSs and for identifying human genetic variations that are associated with disordered transcriptional regulation. DBTSS can be accessed at http://dbtss.hgc.jp.

INTRODUCTION

To understand the precise mechanism of transcriptional regulation of a gene, it is essential to identify and analyze its transcriptional start sites (TSSs), which are located in the vicinity of its potential promoter regions. Several genome-wide studies, including ours, have identified multiple TSSs or their corresponding alternative promoters of each gene (1–3). However, it still remains mostly elusive what biological roles those TSSs play. It is even unknown whether they represent no more than intrinsic biological errors or cloning artifacts. To understand biological relevance of those divergent TSSs, we have developed DBTSS, DateBase of Transcriptional Start Sites, and continued its updates since 2001 (4). DBTSS provides TSS information, which was determined using our cap-site detection method, oligo-capping (5). To obtain the genome-wide data of TSS information, RNA sequences immediately downstream of the TSSs are sequenced with an Illumina massively parallel sequencing platform as TSS tags. We call this technology ‘TSS-seq’ (6,7).

On the other hand, a large number of systematic epigenomic studies are underway, aiming at comprehensive understanding of chromatin conditions in the genome. Recently, a group of the ENCODE project published a chromatic map of nine representative cell types from the ChIP-seq analyses of nine chromatin markers for various types of histone modifications (8). Integration of such kind of data with the TSS data would be useful for both studies. In fact, we too have generated similar types of ChIP-seq data for the studies of individual cells as listed in our web site. Similarly, RNA-seq data from several subcellular components, such as nucleus, cytoplasm, and polysome fractions, are useful to further characterize the TSS data. Meanwhile, human genome re-sequencing projects including exome sequencing projects have been also massively accumulating single-nucleotide variation (SNV) data (9,10) although understanding their biological consequences is still difficult. Since we believe that many of genetic disorders should be attributed to malfunctions in transcriptional regulations, the link between SNPs and the transcriptional information based on TSSs, histone modification status, and transcripts must be established.

In this report, we report three main progresses in DBTSS. First, we expanded our TSS-seq data by 3-fold,

*To whom correspondence should be addressed. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: knakai@ims.u-tokyo.ac.jp so that a major part of human adult and embryonic tissues are covered. Second, we added various types of transcriptomics data that can be useful to further interpret TSS information. For example, we integrated our original RNA-seq data of subcellular-fractionated RNAs (11) as well as the ChIP-seq data of histone modifications, RNA polymerase II and several transcriptional regulatory factors (12–14) in cultured cell lines to collectively understand the relationship between TSS, transcript dynamics and epigenetic factors. Some of the external epigenomic data, such as those by the ENCODE project (8), are also added. Third, we associated our TSS data with SNV data to identify SNV candidates that may be responsible for disorders of transcriptional regulation. We believe that such integration provides in-depth biological insight of divergent TSSs *in vivo*.

New TSS-seq data

In this update, we added new TSS data that were derived from our TSS-seq experiment. Now DBTSS contains 418 146 632 TSS tags, collected from 28 tissues or cell types, including 16 kinds of human adult tissues, 5 kinds of human fetal tissues and 7 kinds of cultured cells (Table 1). These TSS tags were clustered into subgroups with 500-bp bins to define TSS clusters (TSCs) as putative promoter unit (see Ref. 15 for more detail). As a default viewer setting, we adopted TSCs with the expression levels higher than 5 ppm (particles or tags per million; >5 ppm TSCs), which exists from 11 300 to 36 718 in varying cell types (Table 2). In several cell lines, we collected TSS-seq information under different experimental conditions such as before or after the stimulation by cytokines or hypoxic shocks. Also, a total of 70 386 438 TSS-seq tag data from several developmental stages and cultured cells are included in mice. Users can search TSSs where different expressions are observed between different tissue types or culture conditions (Figure 1B).

RNA-seq and ChIP-seq data

In addition to the TSS-seq data, we also included RNA-seq tags in the data of cultured cells. For example, in DLD-1 cell data, RNA-seq data from total RNA, nuclear, cytoplasm, polysome, and the Argonaute complex-associated fractions are stored (15). Users can search and browse the transcriptional levels of each gene in each fraction to understand the dynamics of its transcripts. To further understand the relationship between TSSs and chromatin states of the surrounding regions, we integrated ChIP-seq data of histone modifications: H3K4Me3 and H3Ac for active chromatin markers and H3K27Me3 for a silent chromatin marker. Data of the binding sites of RNA polymerase II and several transcription factors, such as STAT6 in Ramos and BEADS2B cells and HIF1A in DLD-1 cells, are also included. Details of experimental conditions are described on our web site. We also associated our data with external epigenomic data. Particularly, we added the above-mentioned data of chromatin states in nine human cell lines (8).

Table 1. Statistics of TSS-seq data

Category	No. of cell types or tissues	No. of total condition	No. of TSS-seq
human adult Tissues	16	20	138 864 978
human fetal tisssues	5	5	41 744 136
human cell lines	7	23	237 537 518
human total	**28**	**48**	**418 146 632**
mouse embrio	1	4	38 897 846
mouse cell lines	3	3	31 488 592
mouse total	**4**	**7**	**70 386 438**
total	**32**	**55**	**488 533 070**

Connection to SNV information

We linked our TSS information with various SNV data. In addition to the dbSNP data, we obtained exome data for four ethnic groups from the NCBI site (ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp) and generated SNVs uniquely called for these ethnic groups. We called 687 881 JPT (Japanese in Tokyo), 646 522 CEU (Utah residents with northern and western European ancestry from the CEPH collection) 578 299 CHS (Chinese in Singapore) and 1 120 302 YRI (Yoruba in Ibadan) SNVs from 73, 64, 90 and 82 individual's exome data, respectively. Also, we added genome-wide association studies (GWAS) data which connect 5800 SNVs with clinical information (16). The GWAS information was obtained from the UCSC browser 'gwasCatalog.txt' (17).

Data integration and search examples

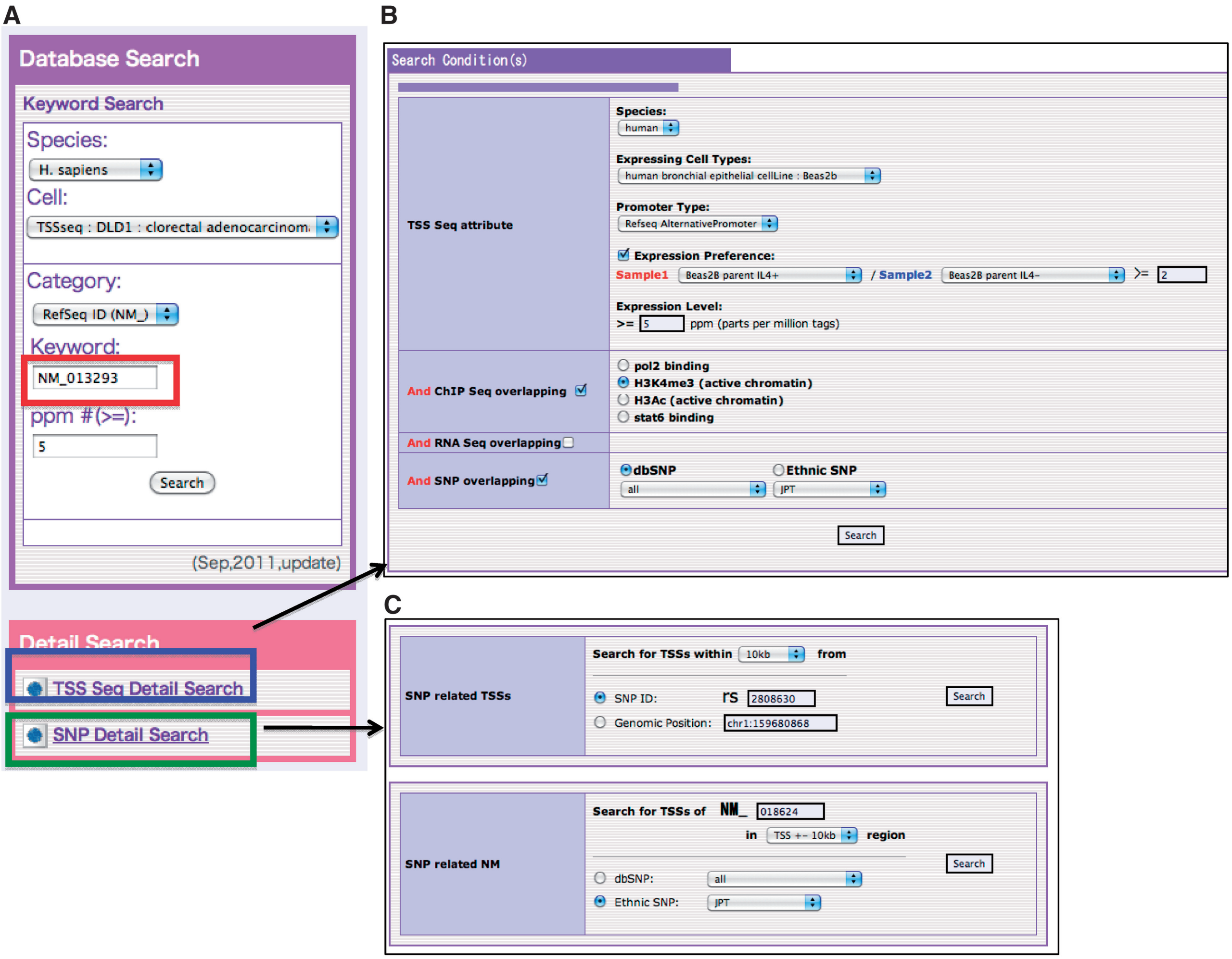
We integrated the internal and external TSS-seq, RNA-seq, ChIP-seq and SNV data with each other. Table 2 (and the 'statistics' page of our web site) represents the statistics of each data and the overlap between them. It should be noted that all of the 'active markers' do not always overlap. In some cases, active chromatin markers are observed neither accompanying RNA polymerase II binding, TSS nor RNA products. These cases may be interesting targets for future analysis of potentially stepwise transcriptional regulations. From a population genetic point of view, it may be also intriguing to further characterize TSSs which are observed in an ethnic group-specific manner with solid supports from ChIP-seq and RNA-seq data in cultured cells.

Users can search TSSs using several search conditions. For example, users can use a RefSeq ID for the simplest search (Figure 1A). Users can narrow down the targets by considering expression levels within a particular cell types or fold expression changes between different cell types (Figure 2A). In the main viewer, users can recalculate tag counts by varying genomic regions (Figure 2B and C). Users can also consider overlap of the ChIP-seq data and RNA-seq data with the TSS-seq data (Figure 2A). Particularly, using the 'SNP search' function, users can input an SNV in a particular category and search TSSs in its surrounding region. Conversely, users can search whether there are any TSSs that have SNVs of various categories in its surrounding regions (Figures 1B and 2D as search results).

Table 2. TSCs corresponding to NCBI RefSeq genes and SNP information

	TSS-seq tags	Total TSCs	TSCs in RefSeq	TSC >5 ppm	overlap RefSeq (>5 ppm)	db_SNP (>5 ppm)	JPT/YRI/CHS/CEU (>5 ppm)
HEK293	20 686 169	193 140	137 518	11 338	10 234	11 285	3930/5471/4391/3476
Ramos	31 022 974	371 759	239 308	11 455	9227	11 418	4036/5433/4340/3513
BEAS2B	98 761 770	708 912	440 302	27 628	20 386	7220	2993/3979/3194/2167
DLD1	48 580 850	462 724	272 171	19 941	16 965	19 878	7084/9735/7731/6211
MCF7	15 785 949	172 834	120 695	11 790	10 383	11 743	4094/5645/4432/3584
TIG3	18 780 087	198 129	144 622	11 893	10 512	11 847	4186/5857/4674/3711
HeLa	3 919 719	99 241	74 719	11 300	9710	11 262	4187/5787/4705/3734
Adult tissue	138 864 978	1 496 409	911 872	36 718	26 023	50 674	16 874/22 996/18 163/14 927
Fetal tissue	41 744 136	822 577	572 941	32 533	26 773	32 386	10 400/14 192/11 088/9260

'Samples': category of samples, 'TSS-seq tags': tag number in each category, 'total TSCs': observed TSC number, 'TSCs in RefSeq': TSCs overlapping with the Refseq transcribed region (including their 50 k bp upstream region), 'TSC > 5 ppm': number of TSCs whose expression level is higher than 5 ppm, 'overlap Refseq (5 ppm)': > 5ppm TSCs which overlap with the RefSeq transcribed region, 'db_SNP (>5 ppm)': number of TSCs which contain SNPs in dbSNP, 'JPT/YRI/CHS/CEU (>5 ppm): number of TSCs which contain ethnic SNPs (JPT: Japanese in Tokyo, CEU: Utah residents with northern and western European ancestry from the CEPH collection, CHS: Chinese in Singapore, and YRI: Yoruba in Ibadan).



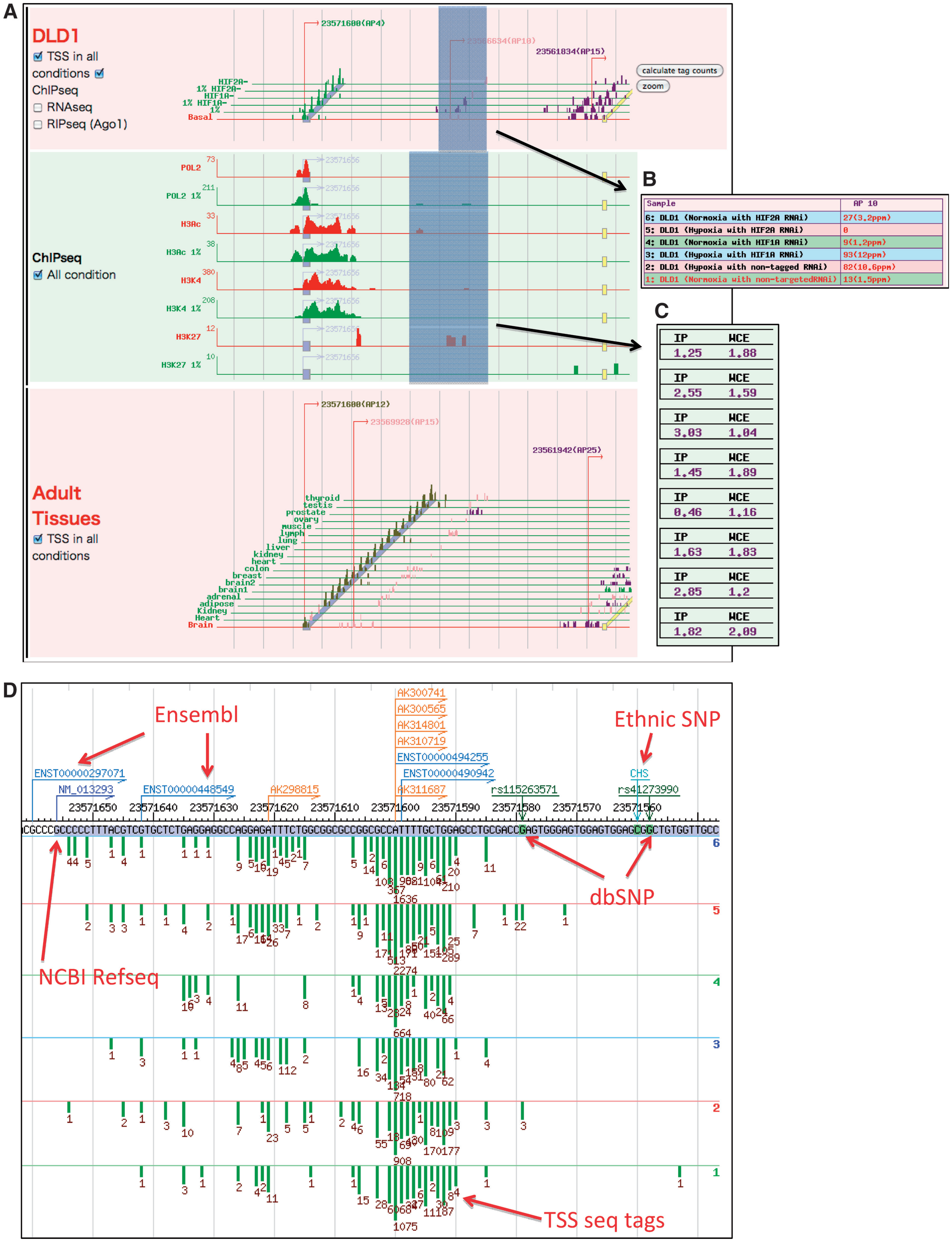
Species: H. sapiens
Cell: TSSseq : DLD1 : clorectal adenocarcinom
Category: RefSeq ID (NM_)
Keyword: NM_013293
ppm #(>=): 5
(Sep,2011,update)

Species: human
Expressing Cell Types: human bronchial epithelial cellLine : Beas2b
Promoter Type: Refseq AlternativePromoter
☑ Expression Preference:
Sample1 Beas2B parent IL4+ / Sample2 Beas2B parent IL4− >= 2
Expression Level:
>= 5 ppm (parts per million tags)
And ChIP Seq overlapping ☑ ○ pol2 binding ● H3K4me3 (active chromatin) ○ H3Ac (active chromatin) ○ stat6 binding
And RNA Seq overlapping □
And SNP overlapping ☑ ● dbSNP all ○ Ethnic SNP JPT

Search for TSSs within 10kb from
● SNP ID: rs 2808630
○ Genomic Position: chr1:159680868

Search for TSSs of NM_ 018624 in TSS +- 10kb region
○ dbSNP: all
● Ethnic SNP: JPT

Figure 1. DBTSS input windows. (**A**) Users can use a RefSeq ID for the simplest search (red box in the figure). (**B**) After clicking 'TSS-seq Detailed Search', users will obtain the 'search condition' window. In this case, users can search TSSs that are overexpressed after IL4 stimulation by 2-folds, with their expression level higher than 5 ppm, showing H3k4me3 signals, and having nearby dbSNP data. (**C**) Users can search TSSs around a given SNP or any genomic position (upper window). SNPs that are neighboring with known genes can be sought, too (bottom window).



Sample	AP 10
6: DLD1 (Normoxia with HIF2A RNAi)	27(3.2ppm)
5: DLD1 (Hypoxia with HIF2A RNAi)	0
4: DLD1 (Normoxia with HIF1A RNAi)	9(1.2ppm)
3: DLD1 (Hypoxia with HIF1A RNAi)	93(12ppm)
2: DLD1 (Hypoxia with non-tagged RNAi)	82(10.6ppm)
1: DLD1 (Normoxia with non-targetedRNAi)	13(1.5ppm)

IP	WCE
1.25	1.88

IP	WCE
2.55	1.59

IP	WCE
3.03	1.04

IP	WCE
1.45	1.89

IP	WCE
0.46	1.16

IP	WCE
1.63	1.83

IP	WCE
2.85	1.2

IP	WCE
1.82	2.09

Figure 2. Example of search results (NM_013293: transformer 2 alpha homolog). (**A**) Overview of TSS-seq and ChIP-seq for NM_013293, transformer 2 alpha homolog. There are three major putative alternative promoters (AP4, AP10 and AP15) in DLD1 cells. The expression of AP10 under the normoxia condition (21%) is relatively low compared with that under the hypoxia condition (1%). Using check boxes, users can also check the TSS-seq and ChIP-seq results in other tissues. (**B**) Function of recalculating tag counts by specifying desired genomic regions. In this case, 1.5 ppm TSC specific to normoxia and 10.6 ppm TSC to hypoxia are observed. (**C**) Users can also recalculate tag counts for ChIP-seq tags. There is a clear difference in the H3K27 states between normoxia and hypoxia. (**D**) Detailed information of AP4. The green bars indicate our TSS-seq data. The start positions of known genes are displayed with arrows. An Ethnic SNP (CHS) and two dbSNPs (rs11523571 and rs41273990) are also found in this region. Searching 'rs11523571' or 'chr1:159680868' based on the input window (Figure 1C) also leads to a similar result.

All of the raw data are freely and anonymously available from our download site at 'ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver8'.

Future perspective

We expect more genome-wide data to become available not only for TSS-seq data but also for RNA-seq and ChIP-seq data from humans, mice and other model organisms in the near future. Also, tens of thousands of SNV data for different ethnic groups are being generated without further biological information except their potential disease associations. We will update DBTSS accordingly. We believe that, by the integration of the data, our DBTSS will continue to serve as a unique database for a wide variety of researchers: both for researchers analyzing transcriptional regulations of individual genes as well as for those analyzing comprehensive transcriptional regulatory networks from a systems biological point of view.

ACKNOWLEDGEMENTS

The authors thank T. Horiuchi, F. Todokoro, H. Wakaguri and J. Zhang for their technical support in database construction. They also thank Dr A. Patil for critically reading the article.

FUNDING

Funding for open access charge: Grant-in-Aid for 'Publication of Scientific Research Results'; 'Program for Improvement of Research Environment for Young Researchers' from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T., Thurman,R.E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
2. 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.
3. Kimura,K., Wakamatsu,A., Suzuki,Y., Ota,T., Nishikawa,T., Yamashita,R., Yamamoto,J., Sekine,M., Tsuritani,K., Wakaguri,H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
4. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
5. Maruyama,K. and Sugano,S. (1994) Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, **138**, 171–174.
6. Tsuchihara,K., Suzuki,Y., Wakaguri,H., Irie,T., Tanimoto,K., Hashimoto,S., Matsushima,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.
7. Wakaguri,H., Yamashita,R., Suzuki,Y., Sugano,S. and Nakai,K. (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res.*, **36**, D97–D101.
8. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
9. Sudmant,P.H., Kitzman,J.O., Antonacci,F., Alkan,C., Malig,M., Tsalenko,A., Sampas,N., Bruhn,L. Shendure,J.1000 genome project and Eichler,E.E (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.
10. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K., Cheetham,R.K. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.
11. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
12. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
13. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
14. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
15. Yamashita,R., Sathira,N.P., Kanai,A., Tanimoto,K., Arauchi,T., Tanaka,Y., Hashimoto,S., Sugano,S., Nakai,K. and Suzuki,Y. (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.*, **21**, 775–789.
16. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
17. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.