Transcriptional Start Site Database for Analyzing Transcriptional Consequences of SNVs

Yutaka Suzuki[1], Riu Yamashita[2], Kenta Nakai[2], Sumio Sugano[1]
[1]Sch. of Frontier Sciences and [2]Human Genome Center,Institute of Medical Sci., University of Tokyo
Correspondence: ysuzuki@hgc.jp: 5-1-5 Kashiwanoha, Kashiwashi, Chiba, 277-8562, JAPAN

ABSTRACT

Although recent human whole genome and exome resequencing studies have identified a large number of single nucleotide variations (SNVs), which of them and to what extent the identified SNVs influence transcriptional levels of the genes still remain elusive. We have developed a method to identify the positions and expression levels of transcriptional start sites (TSSs) on the Illumina Platform. Using this TSS Seq method, we generated a catalogue of TSSs in representative twelve normal tissues and other cultured cell types. All of the TSS tag sequences, their mapped genomic positions and tag counts are freely and publicly available from our database, DBTSS (http://dbtss.hgc.jp). Recently, we have merged the TSS data with publicly available SNV data which were generated by 1000 genome and other international projects and with domestically generated Japanese SNV data. In the current version of DBTSS, a total of SNVs of >200 people are represented, of which approximately 7,907, 12,759 and 5,904 SNVs were located near TSS of canonical promoters, alternative promoters and promoters of intergenic transcripts, respectively. Of these, 15, 21 and 8 SNVs were identified as disease-associated SNVs in previous genome-wide association studies, respectively. To further enrich the information of chromatin statuses surrounding the TSSs, we generated ChIP Seq data of histone modifications and RNA polymerase II in six representative cultured cell types. Also, we incorporated the same kind of information in wider cell types published by ENCODE project. With the expanded data contents, further in-depth evaluation of transcriptional consequences of massively identified SNVs has been firstly enabled. We believe that such integrative information is essential to understand molecular basis underlying disease-related SNVs and other genetic variations in different individuals.

References:
Tsuchihara et al, NAR, 37, 2246-63 (2009)
Yamashita et al, NAR DB issue, D98-104 (2010)
Yamashita et al, Genome Res, 5, 775-89 (2011).

TSS-Seq: Massive TSS Sequencing by "Oligo-cap ping"+ Illumina GA

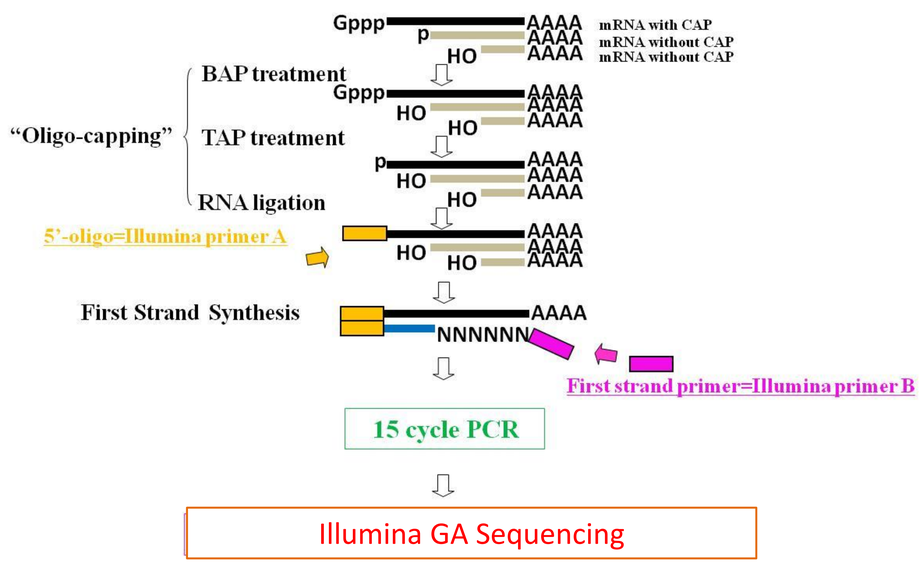


Figure 1 Scheme of TSS Seq (cap-site sequencing) using the Illumina GA Sequencer
Adaptors containing necessary sequence for the Illumina GA sequencer are represented as yellow and pink boxes. For further information, see Supporting Online Material. Gppp: cap structure. AAA: polyA.

Data Contents of the DBTSS (#TSS Seq tags)

	cell	*#cell types or tissues*	*#total culture conditions*	*#total tags*
human	Adult tissues	16	20	138,864,978
human	Fetal tissues	5	5	41,744,136
human	cell lines	7	23	225,574,964
human	**total**	**28**	**48**	**415,760,355**
mouse	embryo	1	4	38,897,846
mouse	cell lines	3	3	31,488,592
mouse	**total**	**4**	**7**	**70,386,438**
total		**32**	**55**	**486,146,793**

Data Contents of the DBTSS (#Promoters/Alternative Promoters)

TSS, TSCs and RefSeq — TSC: TSS cluster; putative promoter unit

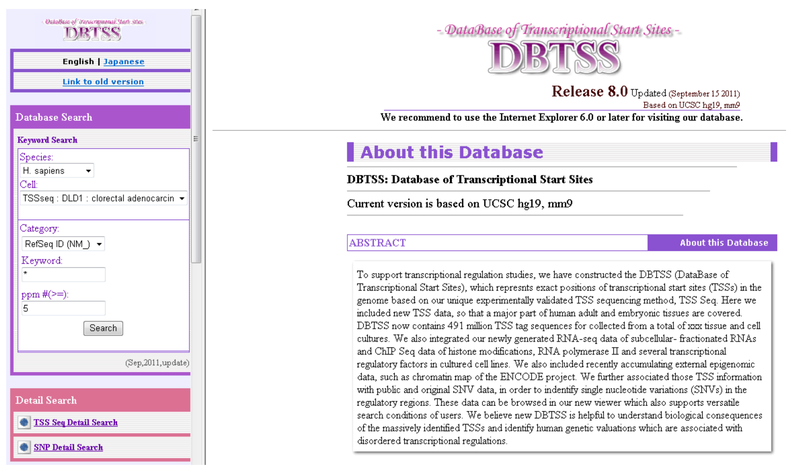
	TSS Seq tags	*Total TSCs*	*TSCs in RefSeq*	*Refseq promoter*	*Refseq promoter > 5ppm*
HEK293	20,686,169	193,140	137,518	11,136	5,133
Ramos	31,022,974	371,759	239,308	9,718	4,507
BEAS2B	98,761,770	708,912	440,302	12,844	5,694
DLD1	48,580,850	462,724	272,171	11,968	6,411
MCF7	15,785,949	172,834	120,695	10,380	4,671
TIG3	18,780,087	198,129	144,622	10,933	4,919
Hela	3,919,719	99,241	74,719	8,804	4,712
Adult tissue	138,864,978	1,496,409	911,872	18,351	8,196
Fetal tissue	41,744,136	822,577	572,941	15,443	6,906

Data Contents of the DBTSS (#Putative Promoters of lncRNAs)

Alternative and Intergenic

	Alternative promoter	*Alternative promoter > 5ppm*	*Intergenic promoter*	*Intergenic promoter > 5ppm*
HEK293	126,382	5,101	55,622	1,104
Ramos	229,590	4,720	132,451	2,228
BEAS2B	427,458	14,692	268,010	7,242
DLD1	260,203	10,554	190,553	2,976
MCF7	110,315	5,712	52,139	1,407
TIG3	133,689	5,593	53,507	1,381
Hela	65,915	4,998	24,522	1,590
Adult tissue	893,521	17,827	584,537	10,695
Fetal tissue	557,498	19,867	249,636	5,760

DataBase of Transcriptional Start Sites (DBTSS@http://dbtss.hgc.jp/)



Integration of the TSS data with ChIP Seq and RNA Seq data

Human

		DLD-1		MCF7		TIG3		HEK293		Beas2B		Ramos		Adult tissue	Fetal tissue
		21% O2	1% O2	21% O2	1% O2	21% O2	1% O2	21% O2	1% O2	IL4+	IL4-	IL4+	IL4-		
TSS Seq		♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦
RNA Seq	total RNA	♦	♦	♦	♦	♦	♦	♦	♦	-	-	-	-	-	-
	polysome	♦	♦	-	-	-	-	-	-	-	-	-	-	-	-
	nucleosome	♦	♦	-	-	-	-	-	-	-	-	-	-	-	-
	cytoplasmic	♦	♦	-	-	-	-	-	-	-	-	-	-	-	-
ChIP Seq	pol2	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	♦	-	-
transcription	HIF1a	-	-	♦	♦	♦	♦	-	-	-	-	-	-	-	-
factor	STAT6	-	-	-	-	-	-	-	-	♦	♦	♦	♦	-	-
ChIP Seq	H3Ac	♦	♦	-	-	♦	♦	-	-	♦	♦	♦	♦	-	-
histone	H3K4Me3	♦	♦	-	-	♦	♦	-	-	♦	♦	♦	♦	-	-
modification	H3K27Me3	♦	♦	-	-	♦	♦	-	-	-	-	-	-	-	-

Mouse

		3T3	10T1/2	embryo				ATDC5
				7d	11d	15d	17d	
TSS Seq		♦	♦	♦	♦	♦	♦	♦

Figure 2 Database contents of DBTSS associated with TSS
ChIP Seq and RNA Seq data associated with the TSS data are shown for the indicated cell types.

dbSNP/1000 genome ethnic SNP idenitified in the neighboring regions of TSSs

	dbSNP (> 5ppm)	*JPT/YRI/CHS/CEU (> 5ppm)*
HEK293	11,285	3,930 / 5,471 / 4,391 / 3,476
Ramos	11,418	4,036 / 5,433 / 4,340 / 3,513
BEAS2B	7,220	2,993 / 3,979 / 3,194 / 2,167
DLD1	19,878	7,084 / 9,735 / 7,731 / 6,211
MCF7	11,743	4,094 / 5,645 / 4,432 / 3,584
TIG3	11,847	4,186 / 5,857 / 4,674 / 3,711
Hela	11,262	4,187 / 5,787 / 4,705 / 3,734
Adult tissue	50,674	16,874 / 22,996 / 18,163 / 14,927
Fetal tissue	32,386	10,400 / 14,192 / 11,088 / 9,260

JPT: Japanese in Tokyo, Japan, YRI: Yoruba in Ibadan, Nigeria, CHS: Han Chinese South
CEU: Utah residents with Northern and Western European ancestry

Figure 3 Genetic Variations associated with TSS
TSSs corresponding to NCBI RefSeq genes and SNP information. '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 50k bp upstream region), 'TSC>5ppm': number of TSCs whose expression level is higher than 5 ppm, 'overlap Refseq (5 ppm)': > 5 ppm TSCs which overlap with the RefSeq transcribed region, 'db_SNP (> 5ppm)': number of TSCs which contain SNPs in dbSNP, 'JPT/YRI/CHS/CEU (>5ppm): 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).

Search Example of DBTSS (Input)

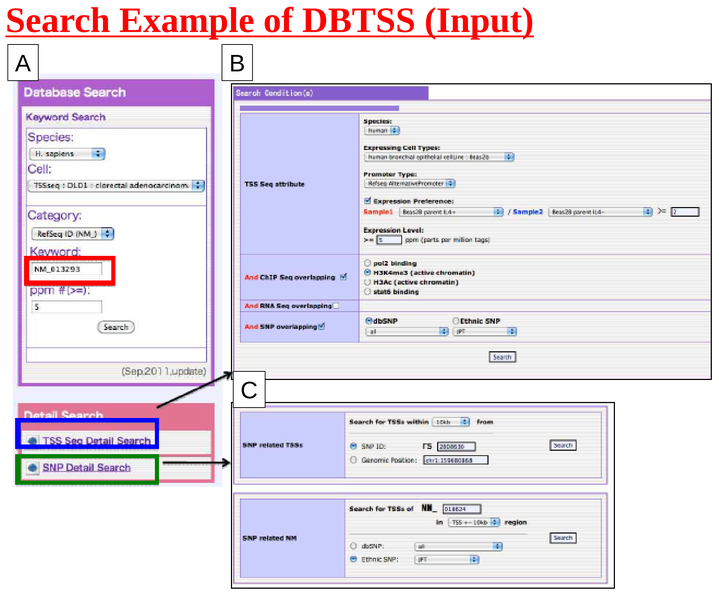


Figure 4 DBTSS search input window
(A) Users can use a RefSeq ID for the simplest search (red box in the figure). (B) After clicking 'TSS-seq Detailed Search', users will obtain the 'search condition' window. In this case, users can search TSSs that are overexpressed after IL4 stimulation by 2 folds, 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).

Search Example of DBTSS (Results)

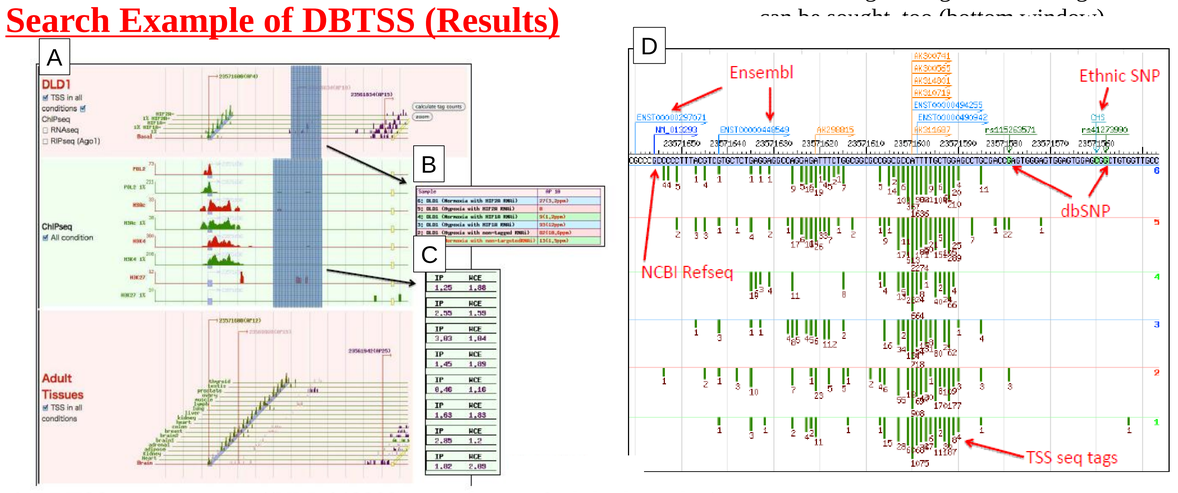


Figure 4 DBTSS search results ((NM_013293: transformer 2 alpha homolog)
(A) Overview of TSS-seq and ChIP-seq for NM_013293, transformer 2 alpha homolog. There are three major putative alternative promoters (AP4, AP10, and AP15) in DLD1 cells. The expression of AP10 under the normoxia condition (21%) 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 also leads to a similar result.

All contents and raw data freely downloadable from DBTSS at: http://dbtss.hgc.jp