DBKERO: Regulatory Omics Database for Analyzing Transcriptional Consequences of Human SNVs

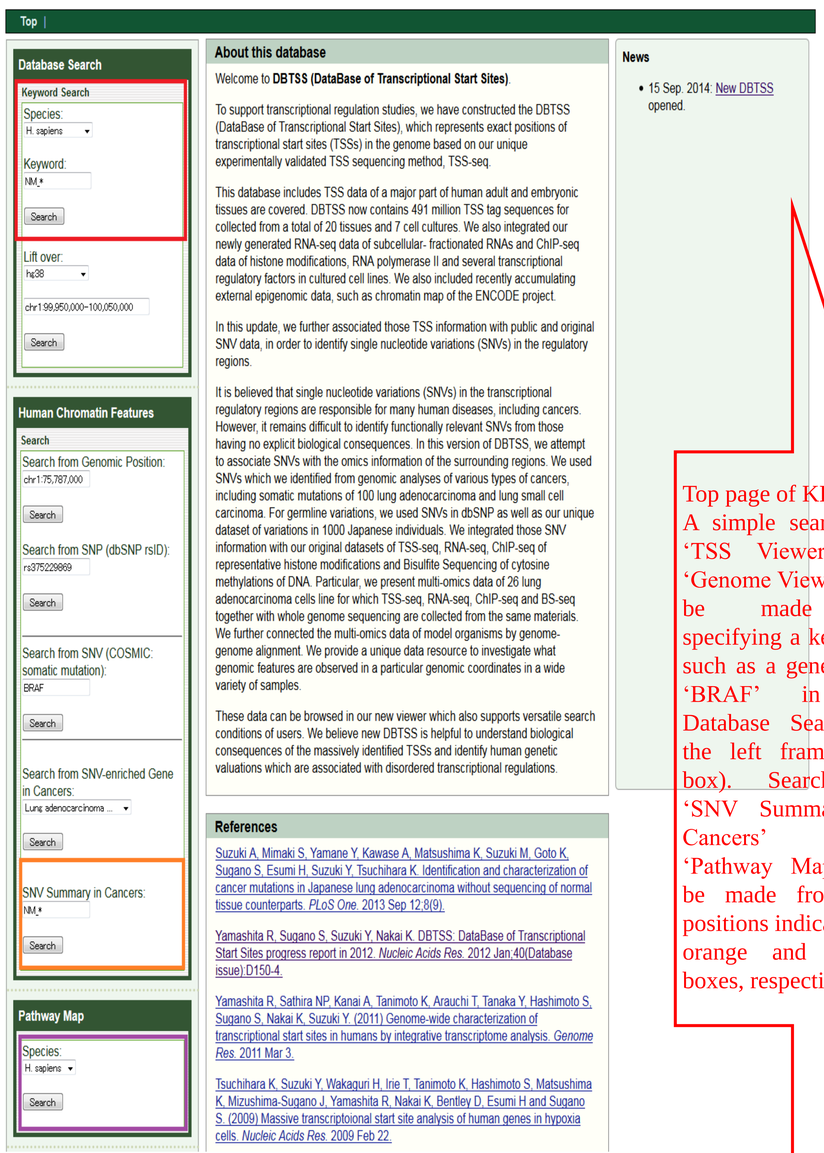
○Yutaka Suzuki[1], Hiroyuki Wakaguri[2], Riu Yamashita[3], Shin Kawano[4], Katsuya Tsuchihara[5], Kenta Nakai[6] and Sumio Sugano[1]

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

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.

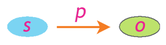
Kashiwa Encyclopedia of Regulatory Omics (KERO@http://dbtss.hgc.jp/)



Top page of KERO. 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.

RDF for Integration of the Databases

What and Why RDF?



- RDF: Resource Description Framework
- Represented as "Subject-Predicate-Object" triple
- All resources except literals have URIs (URLs)
- Fundamental technology of Semantic Web together with OWL ontologies
- Standardized by W3C
- Query language is available (SPARQL)
- Federated query
- RDF resources can join in LOD network

ChIP RDF Schema

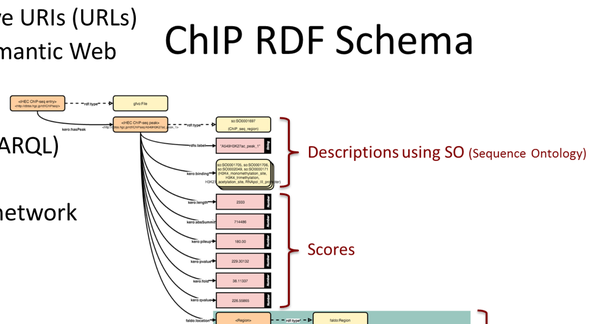
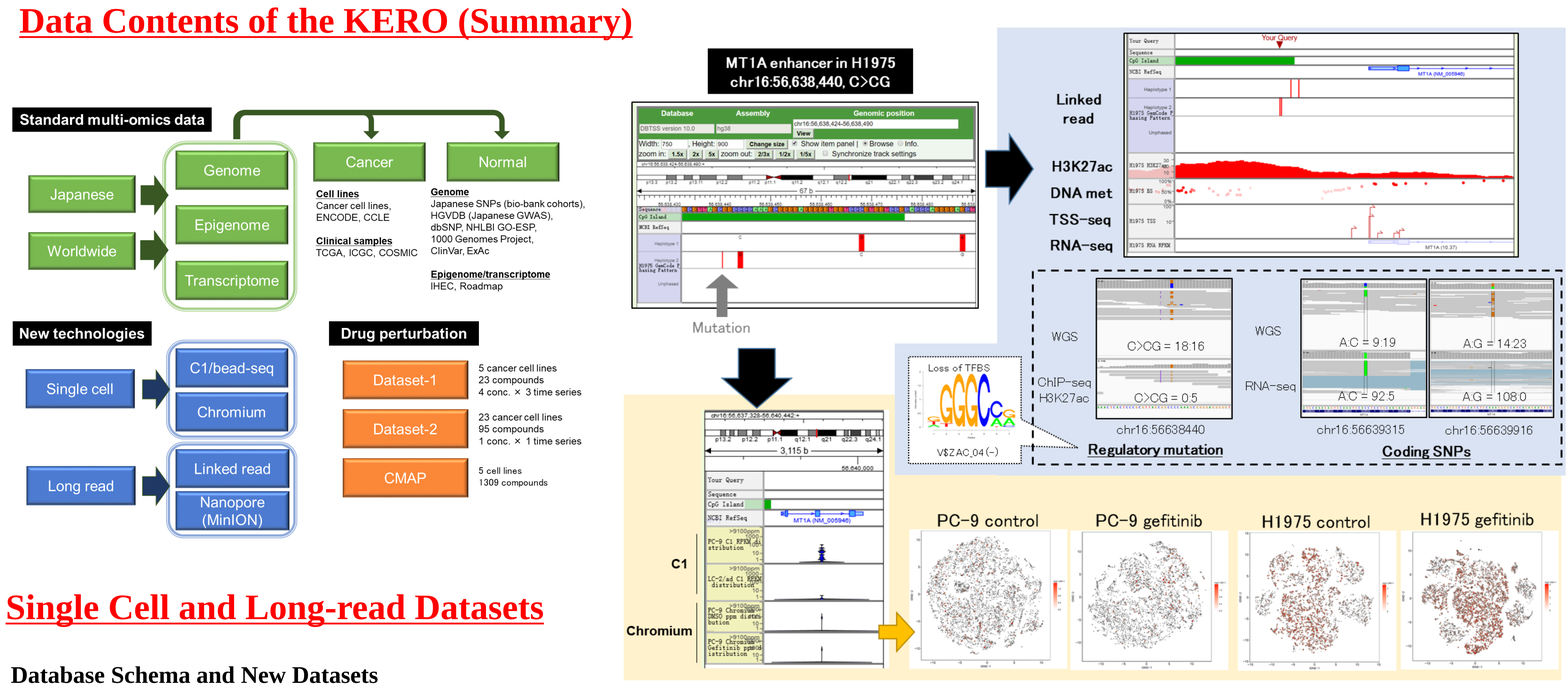


Figure 3 Schema for RDF of ChIP data

Data Contents of the KERO (Summary)



Single Cell and Long-read Datasets

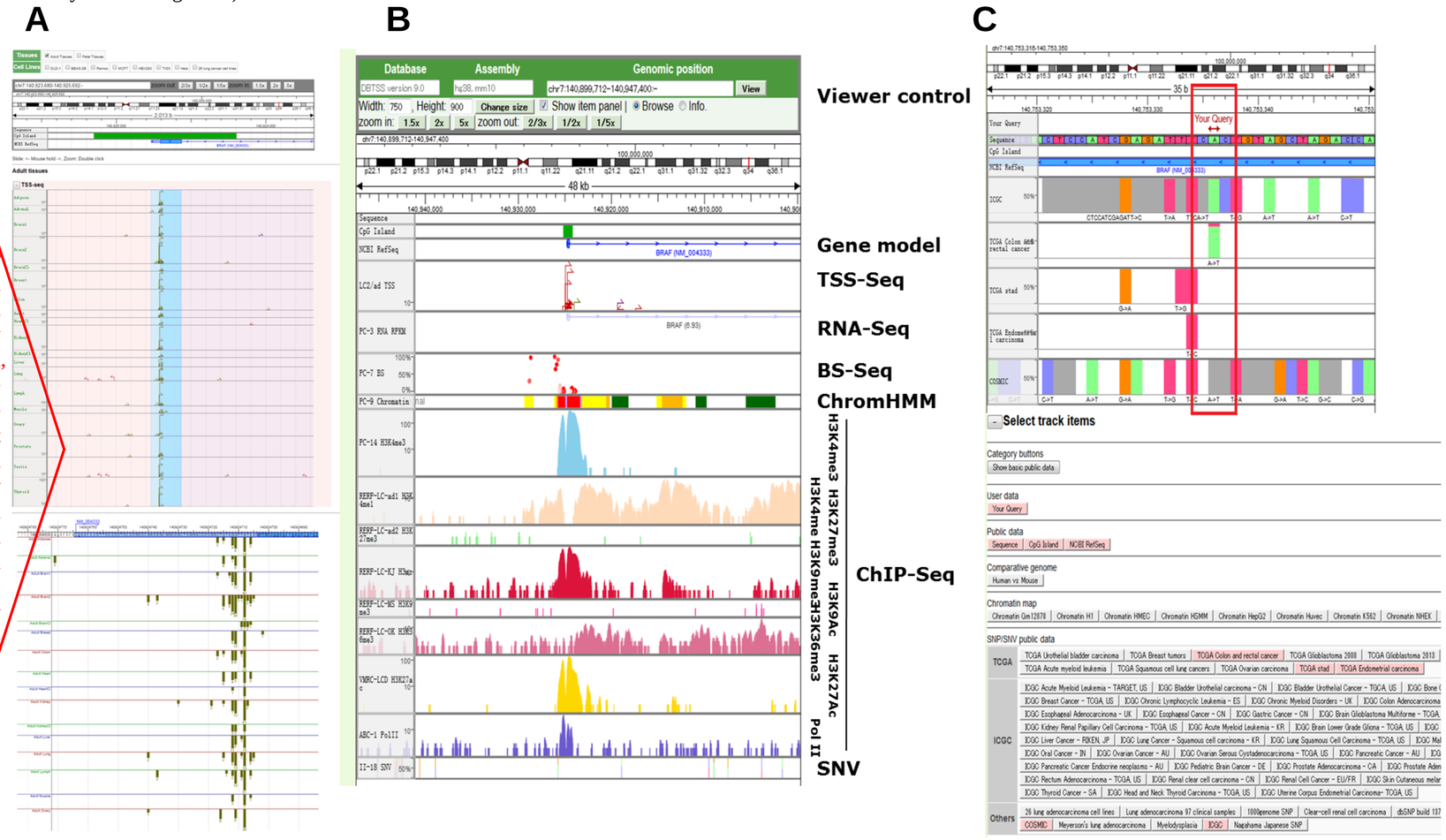
Database Schema and New Datasets

(A) Overall structure of the database is illustrated. How the Japanese clinical omics information is associated with comprehensive omics information from the model systems is shown. This database also included information newly available from single cell and long read technologies and multi-omics perturbation by chemical compounds. Different categories of the datasets are shown in different colors.

(B) The tours of the MT1A gene. The guide how to show the similar results in the database are illustrated. For more details see the web (http://dbtss.hgc.jp/docs/help_2017.html).

Follow the link as illustrated in Supplementary Figure 3:

1. Input 'MT1A' to the keyword field box at the top left part of the top page.
2. Select the 'GemCode Phasing Patterns' of the H1975 cell line. Find a mutation (chr16:56,638,440, C>CG) in the haplotype 2 of the MT1A upstream region (upper left panel).
3. Add epigenome and transcriptome information of the H1975 cell line around the mutation (upper right panel on blue background). Select H3K27ac ChIP-seq and DNA methylation of BS-seq for epigenome patterns, and TSS and rpkm of RNA-seq for transcriptome patterns.
4. To view the data of expression variation in individual single cells, display the 'rpkm distribution' of the C1, bead-seq and Chromium single cell platforms. To view the distribution of the expression levels of the MT1A gene in each cell, select 'the C1 system' (lower left panel on yellow background). For information of a large number of cells, select 'the Chromium system'. Go to the single cell viewer from 'the summary' link and see the expression variation of MT1A gene on the two dimensional t-SNE plot (lower right panel on yellow background).



DBTSS Browser

(A) 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. (B) 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. (C) 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.

A

TSS Viewer: BRAF

Description	Mutation frequency: Upstream distal: -50k to -1	Upstream proximal: -1k to -1	Gene body
Clear-cell renal cell carcinoma by Dr. Ogawa's Lab.	0/106	0/106	0/106
ICGC: Acute Myeloid Leukemia - KR	1/78	1/78	1/78
ICGC: Acute Myeloid Leukemia - TARGET, US	0/2	0/2	0/2
ICGC: Bladder Urothelial Cancer - TGCA, US	1/128	1/128	1/128
ICGC: Bladder Urothelial carcinoma - CN	1/103	1/103	1/103
ICGC: Bone Cancer - UK	0/66	0/66	0/66
ICGC: Brain Glioblastoma Multiforme - TCGA, US	5/268	5/268	5/268
ICGC: Brain Lower Grade Gliona - TCGA, US	1/268	1/268	1/268
ICGC: Breast Cancer - TCGA, US	1/943	1/943	1/943
ICGC: Breast Triple Negative/Lobular Cancer - UK	1/117	1/117	1/117
ICGC: Chronic Lymphocytic Leukemia - ES	1/109	1/109	1/109
ICGC: Chronic Myeloid Disorders - UK	0/129	0/129	0/129
ICGC: Colon Adenocarcinoma - TCGA, US	25/216	25/216	25/216
ICGC: Early Onset Prostate Cancer - DE	1/11	1/11	1/11
ICGC: Esophageal Adenocarcinoma - UK	1/16	1/16	1/16
ICGC: Esophageal Cancer - CN	0/88	0/88	0/88
ICGC: Gastric Adenocarcinoma - TCGA, US	3/289	3/289	3/289
ICGC: Gastric Cancer - CN	0/9	0/9	0/9
ICGC: Head and Neck Thyroid Carcinoma - TCGA, US	232/393	232/393	232/393
ICGC: Kidney Renal Clear Cell Carcinoma -	1/404	1/404	1/404

Multi-omics data in cell lines

Cell	RNA-Seq (RPKM)	H3K4me3 (proximal)	H3K4me1 (distal)	H3K27ac (distal)	PolII (proximal)	H3K36me (gene body)	H3K27me3 (gene body)	H3K27me3 (distal)
A427	6.8	137.4	5.6	15.1	0	2	0.6	0.5
A549	3.4	34.1	0	6.7	16.8	1.6	0.3	0.4
ABC1	3.5	32.4	0	5.2	0	1.8	0.4	0.6
H1299	2.8	178.5	262.9	71.9	850.3	2.3	0.5	0.8
H1437	3	40.8	0	0	0	1.7	0.2	0.2
H1648	7.5	43.4	0	14.5	6.1	2.9	0.4	0.6
H1650	5.8	66.1	0	25.3	33	2.7	0.5	0.6

B

ErbB / HER Signaling

Cancer type: Lung adenocarcinoma 26 cell lines (each) Cell: LC2/ad | Unit: RF
Tint control: light — deep
Coloring: RPKM: 0 255<

Symbol	Value	Go to
ERBB2	109.49	DBTSS, NCBI Gene

Figure 2 Pathway Search and Promoter Search

(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.

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