DBTSS/DBKERO for integrated analysis of transcriptional regulation

Ayako Suzuki1,†, Shin Kawano2,†, Toutai Mitsuyama3,†, Mikita Suyama4, Yae Kanai5, Katsuhiko Shirahige6, Hiroyuki Sasaki7, Katsushi Tokunaga8, Katsuya Tsuchihara1, Sumio Sugano9, Kenta Nakai10,* and Yutaka Suzuki9,*

1Division of Translational Genomics, Exploratory Oncology Research and Clinical Trial Center, National Cancer Center, Chiba, Japan, 2Database Center for Life Science, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Chiba, Japan, 3Computational Regulatory Genomics Research Group, Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan, 4Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan, 5Department of Pathology, Keio University School of Medicine, Tokyo, Japan, 6Institute of Molecular and Cellular Biosciences, the University of Tokyo, Tokyo, Japan, 7Division of Epigenomics and Development, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Japan, 8Department of Human Genetics, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan, 9Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, the University of Tokyo, Chiba, Japan and 10Human Genome Center, the Institute of Medical Science, the University of Tokyo, Tokyo, Japan

Received September 15, 2017; Revised October 11, 2017; Editorial Decision October 11, 2017; Accepted November 03, 2017

ABSTRACT

DBTSS (Database of Transcriptional Start Sites)/DBKERO (Database of Kashiwa Encyclopedia for human genome mutations in Regulatory regions and their Omics contexts) is the database originally initiated with the information of transcriptional start sites and their upstream transcriptional regulatory regions. In recent years, we updated the database to assist users to elucidate biological relevance of the human genome variations or somatic mutations in cancers which may affect the transcriptional regulation. In this update, we facilitate interpretations of disease associated genomic variation, using the Japanese population as a model case. We enriched the genomic variation dataset consisting of the 13,368 individuals collected for various genome-wide association studies and the reference epigenome information in the surrounding regions using a total of 455 epigenome datasets (four tissue types from 67 healthy individuals) collected for the International Human Epigenome Consortium (IHEC). The data directly obtained from the clinical samples was associated with that obtained from various model systems, such as the drug perturbation datasets using cultured cancer cells. Furthermore, we incorporated the results obtained using the newly developed analytical methods, Nanopore/10x Genomics long-read sequencing of the human genome and single cell analyses. The database is made publicly accessible at the URL (http://dbtss.hgc.jp/).

INTRODUCTION

The human genomic variations or mutations in the transcriptional regulatory regions may play roles in the onset and progression of human diseases. In spite of their potential importance and general interests, little is known about how mutations in regulatory regions alter the epigenome or transcription programs, resulting in aberrant cellular phenotypic consequences. In order to understand the effects of these regulatory mutations, the information on the genomic variations needs to be integrated with other omics information, such as epigenome and transcriptome. Since it is sometimes difficult to collect those multi-omics pieces of information directly from clinical samples, information from model experimental systems, such as mouse models and cultured cells, is often used to obtain important clues.

We initiated a database called DBTSS (1), based on our unique full-length cDNA dataset in 2002 (2–4). At the time, we utilized the 5’-end of the full-length cDNA as the precise information of the transcriptional start sites, which can

*To whom correspondence should be addressed. Tel: +81 4 7136 3607; Fax: +81 4 7136 3607; Email: ysuzuki@k.u-tokyo.ac.jp
Correspondence may also be addressed to Kenta Nakai. Tel: +81 3 5449 5131; Fax: +81 3 5449 5133; Email: knakai@ims.u-tokyo.ac.jp
†These authors contributed equally to this work as first authors.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
be regulated by alternative promoters. We analyzed the up-stream potential promoter regions to outline the genome-wide transcriptional regulations. Since its initial launch, we have made several rounds of updates. In recent years, we associated the transcriptome information with a catalogue of genomic variations, such as public SNP database, as well as the epigenome information, mainly obtained from our own studies, to enable further in-depth analyses on the disease-causing molecular mechanisms. We named this part of the database ‘DBKERO’ (Database of Kashiwa Encyclopedia for human genome mutations in Regulatory regions and their Omics contexts).

In this update, we substantially enhanced both the genomic variation datasets and epigenome variation datasets, particularly focusing on the Japanese population. These datasets were collected as a series of genome-wide association studies and a part of International Human Epigenome Consortium (IHEC) projects (5), respectively. We believe that this should serve as a model case for the genomic, epigenomic and transcriptomic variations occurring in a particular ethnic background that underlie various diseases. Also, we associated the data obtained from clinical samples with that from various model systems, such as the drug perturbation datasets using cultured cancer cells. Furthermore, we incorporated the recent results obtained using the newly developed analytical methods, namely, the long read sequencing analysis using Nanopore (6) and Chromium/GemCode (7), as well as the single cell analyses of cancer cells using the C1 (8,9) and Chromium (10) systems. We believe that these new datasets should be useful for further in-depth analysis of diseased genomes, for which current short read sequencing or bulk sequencing of the materials would give limited information. Statistics of the datasets are shown in Tables 1–3 (for more details, see Supplementary figures or visit our website at http://dbtss.hgc.jp/docs/help_2017.html).
# Table 1. Statistics of the omics datasets

| Category | Data source | Number of individuals | References |
|----------|-------------|-----------------------|------------|
| **Germline variation** | Human Genome Variation Database (HGVDDB) (17 GWAS) | 5737 case / 7631 healthy | [link](https://gwas.biosciencebd.be/cgi-bin/hvdb/bv_top.cgi) |
| | The Human Genetic Variation Database (HGVD) | 1,208 | [link](http://www.hgvd.genome.med.kyoto-u.ac.jp/index.html) |
| | Integrative Japanese Genome Variation (JGVD); ToMMo | 1,070 | [link](https://igvd.megabank.tohoku.ac.jp/) |
| | Japan Pgx Data Science Consortium (JPDSC) | 2,994 | [link](http://www.jpdscc.org/english/) |
| **Somatic mutation** | Lung adenocarcinoma – National Cancer Center (NCC) | 97 | [PLoS One 2013](https://www.ncbi.nlm.nih.gov/pubmed/24041242) 8(9) e73484 |
| | Small cell lung cancer - NCC | 57 | [J Thorac Oncol 2014](https://www.ncbi.nlm.nih.gov/pubmed/24647185) 9(9) 1324-31 |
| | International Cancer Genome Consortium (ICGC) Liver cancer - RIKEN | 258 | [link](https://dcc.icgc.org/projects/LIRI-JP) |
| | ICGC Liver cancer - NCC | 244 | [link](https://dcc.icgc.org/projects/LINC-JP) |
| | ICGC Biliary tract cancer | 239 | [link](https://dcc.icgc.org/projects/BTCA-JP) |
| **Normal epigenome** | International Human Epigenome Consortium (IHEC) Liver (64 datasets) | 8 | [link](http://epigenomesportal.ca/ihec/) |
| | IHEC Colon (88 datasets) | 11 | [link](http://epigenomesportal.ca/ihec/) |
| | IHEC Endometrial (132 datasets) | 15 | [link](http://epigenomesportal.ca/ihec/) |
| | IHEC Vascular endometrial (4 datasets) | 1 | [link](http://epigenomesportal.ca/ihec/) |

**Table 2. Statistics of the drug perturbation datasets**

| Dataset-1 | Dataset-2 |
|-----------|-----------|
| Sample | 5 lung cancer cell lines | 23 lung cancer cell lines |
| Number of compounds | 23 + DMSO control | 95 + DMSO control |
| Condition | 4 concentration points; 24, 48, 72 h | 1 concentration point; 24h |
| Total number of datasets (RNA-seq) | 1299 | 2011 |
| Total number of datasets (ATAC-seq) | 1316 | 2077 |

**Table 3. Statistics of the new technologies datasets**

**A) Single cell analysis**

| C1 | bead-seq | Chromium (> 5k tag) |
|----|----------|---------------------|
| Sample | 4 lung cancer cell lines | 5 lung cancer cell lines | 5 lung cancer cell lines |
| Condition | 1μM vandetanib; 6h / No treatment | 1μM gefitinib; 24h / DMSO control | 1μM gefitinib; 24h / DMSO control |
| Total number of single cells | 336 | 442 | 47,665 |
| Average number of reads per cell | 7,119,082 | 1,069,847 | 10,105 |

**B) Long read analysis**

| Chromium/GemCode (whole-exome & regulatory) | Nanopore/MiniION (whole-genome) |
|--------------------------------------------|---------------------------------|
| Sample | 23 lung cancer cell lines | 4 lung cancer cell lines |
| Average number of raw reads | 45,679,789 | 451,582 |
| Average depth | 53.1 (113.7 Mb) | 0.56 (whole-genome) |
Figure 2. The overview of the genome viewer. (A) A representative view of the genome browser harboring standard omics information from Japanese clinical samples. Data for the indicated layers of the omics analyses is shown. The information around the BRAF gene is represented for a Japanese
**DATABASE DESCRIPTIONS**

**New datasets**

*Genome and epigenome catalogues of the Japanese population.* In this update, we integrated several new datasets (Figure 1). The first batch of datasets is the new genome and epigenome datasets collected from clinical samples particularly focusing on the Japanese population. For the genomic variation datasets, we have now included genomic variations of 13,368 Japanese individuals. The raw data was collected from 14 Genome Wide Association Studies (GWAS), which have been conducted around from 2005 to 2015 in Japan (11). A total of 5737 cases and 7631 healthy control samples were enrolled. For the single nucleotide polymorphism (SNP) typing, the SNP arrays from Affymetrix and Illumina (see the web for more details at [https://gwas.biosciencedbc.jp/index.html](https://gwas.biosciencedbc.jp/index.html)) were used. The frequencies at the corresponding genomic coordinates are represented in the database for each group of the case and control cases. We included genomic variation datasets representing previously published whole-genome/exome sequences (WGS/WES) of a total of 5272 Japanese individuals from three bio-bank cohorts (11–15). These Japanese genomic variations were further associated with external public genomic variation reference datasets. For the germline variations, datasets deposited in the public domains of dbSNP (16) and the Exome Aggregation Consortium (ExAC) (17) databases are used as references. For a reference for the somatic mutations in cancers, the public International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA) and the Catalogue of Somatic Mutations in Cancer (COSMIC) databases are used (18–20). Also within the public reference datasets, parts of the data obtained from the Japanese population are highlighted (21,22).

The genomic variation datasets are associated with the transcriptome and epigenome information datasets (Figure 2A). For the epigenome information, a total of 455 epigenome datasets which were obtained from 67 Japanese individuals were included. These datasets have been collected as the contribution to the IHEC from Japanese groups. Since the Japanese groups were allocated for endometrial, gut, stomach, liver, endothelial, and reproductive organ tissues, in the framework of the IHEC, the data for those organs are specially featured. The Japanese epigenome datasets are associated with various tissues obtained from more than 8000 datasets having different ethnic backgrounds, which have been collected by other IHEC teams worldwide ([http://epigenomesportal.ca/ihec/index.html](http://epigenomesportal.ca/ihec/index.html)) (23). The Encyclopedia of DNA Elements (ENCODE) ([https://www.encodeproject.org/](https://www.encodeproject.org/)) (24) and the NIH Roadmap Epigenomics Mapping Consortium (Roadmap) (25) datasets were also associated as the reference dataset. Those collective epigenome datasets are further associated with a total of 286 genome, epigenome and transcriptome datasets, which have been published in previous rounds of the updates of DBTSS/DBKERO (Table 1). To our knowledge, this update has made our database the richest resource of the multi-omics information focusing on the Japanese population.

**RDF as a further integration clue for other databases.** To further expedite the data integration with outer databases, we employed RDF (Resource Description Framework) as the data model. Even not for all, but a significant part of the data can be described in the RDF format. Therefore, the database search from the above datasets are possible from the outside based on the SPARQL (Simple Protocol and RDF Query Language) search. In Figure 2B, the SPARQL search between the DBTSS/DBKERO and the UniProt database (26) at the Ensembl database (27) is exemplified (also see Supplementary Figure S1). This interface, which enables the data search across the databases worldwide, should be useful to complement the data lacking from this database, such as proteome and metabolome.

**Omics datasets from experimental models.** One of the unique features of our database lies in the fact that we focus on the experimental model systems. We believe that model systems would give an important clue about the biological interpretation of the clinical omics information. The use of model systems is particularly important when clinical samples cannot be directly used for hypothesis-generation or for validation analyses. Genetic or drug perturbations are frequently impossible using clinical samples. In addition to datasets of below human cell lines, we generated a series of multi-omics datasets of the mouse model as a model of particular human diseases. The mouse datasets were connected to the human datasets via our genome-genome alignment (28). The human-mouse inter-species association of the tissues depending on the disease models is also represented in a search engine.

Among a series of datasets from the model systems, we would like to especially draw attention to our newly published datasets on the drug perturbations of the cancer cell lines (Figure 2C). For lung cancer cell lines (half are from Japanese origins and half from other ethnic backgrounds), a comprehensive multi-omics catalogue has been generated and represented (WGS; DNA methylation by bisulfite sequencing; epigenome information represented by eight types of histone modifications; open chromatin by ATAC-seq; transcriptome sequencing of mRNA and miRNA and their transcriptional start sites; long-read sequencing; and single cell transcriptome) (8,29). For these cell lines, 95 rep-
Figure 3. Example tours of the TERT and MT1A genes. (A) The tours of the TERT gene. The guide how to show the similar results in the web interface is illustrated. Further detailed guide of the same tour is also shown at the help page (http://dbtss.hgc.jp/docs/help_2017.html). For the tour, follow the link as
representative drugs were administered at varying concentrations and time points (Supplementary Figure S2). RNA-seq and ATAC-seq were conducted to monitor transcriptional and epigenome responses of the cells, resulting in a total of 3240 RNA-seq and 3393 ATAC-seq datasets. Summary of the data contents for this part of the database is shown in Table 2. We associated these original profiles with the datasets from the Connectivity MAP (CMAP), which are publicly available datasets for the similar purpose (30). These datasets should be helpful when the user wants to know possible consequences of the drugs administration in a given mutation types or epigenome/transcriptome types in in vitro cases.

Dataset obtained from new analytical platforms. In this update, we also attempted to further enrich the multi-omics information using the latest genomic technologies (Figure 2D and E). Recently, several new analytical platforms have been developed and the data production has been started in our laboratories as well as world-wide. We particularly focused on the single cell sequencing technologies and the long-read sequencing technologies. These new methods would shed new light on the previously uncharacterized molecular mechanisms of the diseased genomes. However, these technologies are still incomplete and widely-used current platforms usually have distinct advantages and disadvantages. We believe that it would be extremely useful if the data obtained from the same materials using the different platforms are integrated and simultaneously represented as a database.

Single cell analysis. Recent single-cell sequencing technologies have opened the possibility of analyzing individual cells. A number of reports have demonstrated that single-cell analysis provides pivotal information for elucidating cellular plasticity and diversity within a cellular population in vitro and in vivo. Cancer consists of various cell types, including cancer cells, cancer-associated fibroblasts, tumor-infiltrating leukocytes and vascular cells. When we consider the drug perturbations of cancer, we have to inevitably consider the drug response of each population of the cells and their mutual dependence. Analyses at the single-cell resolution is indispensable in case of disease and targets for the treatments occur in a limited number of cells or when the cellular micro-environments play an essential role.

From a technical perspective, the currently used single cell analytical methods can be separated into two categories. In the first method, cells are physically separated by microfluidics, FACS, or other methods, followed by the reactions of reverse transcriptions and PCR amplifications in an individual reaction chamber manually or with the aid of robotics (9). In the second method, cells are confined to a micro-droplet and individual cells are separately marked using molecular barcoding technology (10). The representative platforms of these methods are the C1 system of Fluidigm and bead-seq system (31), and the Chromium system of 10x Genomic, respectively. Although both platforms are commonly employed for single cell analyses for a wide variety of cells, each of the platforms has several intrinsic advantages and disadvantages. Namely, the C1 system yields sufficient sequence coverage per cell but sparse coverage for the cellular population. The converse is the case for the Chromium system.

In this database, single cell RNA-seq datasets obtained from a series of cancer cells are represented. The data was collected using the C1, bead-seq and Chromium platforms (Table 3A). For the materials, we used the same cancer cell materials as described above. We found it extremely informative to integrate a wide variety of datasets to provide a full spectrum of information for the single-cell. Particularly, as exemplified in Figure 2D, the C1 datasets were used to precisely represent the information of the single cells while the Chromium datasets were used to estimate the frequency
of the cells having the corresponding expression patterns within the population. Also see the ‘Example Tour’ below for further concrete ideas.

**Long read sequencing.** Other datasets obtained from ‘the latest analytical methods’ are the long read sequencing dataset (Table 3B). In the example of cancers, the functional relevance of the mutations occurring in the regulatory regions still remains mostly elusive. One of the largest drawbacks preventing its more efficient characterization lies in the fact that the distance between the transcriptional regulatory region and the regulated gene region is occasionally beyond the reach of the short-read sequencing. Therefore, it is generally difficult to identify the allelic context of ‘regulatory’ mutations, which is how the mutation, usually occurring in one of the loci, affects the target gene on the same chromosome. As exemplified in Figure 2E, by utilizing the long read sequencing technologies, the single nucleotide variants (SNVs) in the regulatory regions can be phased to the downstream heterozygous SNPs/SNVs in the coding regions, if they are present. Once phased, we found it possible to identify the regulatory regions of the SNVs and the RNA-seq variant tags of their target transcripts showed biased frequency between the mutant and reference alleles (also see the below ‘Example Tour’).

In this update, we included the long-read sequencing datasets obtained from the Chromium/GemCode system of 10× Genomics and the MinION sequencer of the Oxford Nanopore Technologies. While the Chromium/GemCode method enables long-read sequencing by intensive use of the barcoding technologies and bioinformatics (so-called ‘synthetic long-read (linked read)’ method), the MinION enables it by the ‘physical long-read’ method. The former method has advantages in the cost, but has a disadvantage in leaving some ambiguity in the resulting phasing information. Particularly, the erroneous phasing is inevitable for aneuploid regions or copy number aberration regions in cancer genomes. On the other hand, the MinION method does not have such a disadvantage, instead, its low base-call quality and high sequencing cost impose considerable barrier preventing its sole use for the WGS (note: we employed the MinION to represent the physical long read method, but PacBio sequencing is an alternative). Within this particular dataset, by mutually complementing the drawbacks in the respective methods, we identified 137 potential regulatory mutations affecting transcriptional regulation. Among them, 84 SNVs could create and/or disrupt potential transcription factor binding sites.

**Unique features in this dataset**

Another important unique feature of this database is that intensive data collection was made by a wide variety of analytical methods focusing on a group of model systems, such as lung cancer cells. We also designed the databases by arranging such information from the model systems to collectively enable interpretation of the information obtained from clinical samples, for which complete suites of omics datasets are mostly unavailable. Particularly, for a series of cancer cells, the datasets, covering whole genome, epigenome, transcriptome data in bulk cells as well as long-read sequencing and single cell data in response to various experimental conditions, have been collected using exactly the same materials. To our knowledge, this is the richest datasets for which most, if not all, of the currently available omics analyses have been intensively conducted. We hope these datasets will serve as a valuable resource to obtain a comprehensive omics view in a given cellular entity from a systems biology viewpoint, namely, how the genomics mutations affect the transcriptional programs or how such transcriptomic perturbation observed in bulk could be dissected down to single cell components. Such cellular information should be further interpreted in the context of the clinical samples having similar genomic, epigenomic or transcriptomic aberrations at least in a sub-module in the genomic system.

**AN EXAMPLE TOUR OF DATA RETRIEVAL**

Figure 3 illustrates the model tour for the usage of the database. A user may start the search of the database with a somatic mutation found in the promoter region of the TERT gene in Japanese liver cancer patients (chr5:1295113, G→A) (https://dcc.icgc.org/projects/LINC-JP) (Figure 3A; see Supplementary Figure S3 for the more detailed guide). This gene is the key regulator of telomere stability. The search would find that this is a highly frequent mutation in cancers but never represented in germline variations. Inspection of the transcriptome and epigenome information in the surrounding region suggests that this mutation is located in the transcriptional regulatory region, which regulates the transcription of the TERT gene. When further information is sought in the datasets of the model cancer cell lines, the user would find that there are some cell lines harboring the mutation at exactly the same genomic locus and the transcription of the gene body is aberrantly enhanced, thus, could be used as a model to analyze its biological functions. Indeed, TERT promoter mutations have been detected in melanoma (32,33), liver cancer (34) and other various types of cancers according to the COSMIC database.

In another example, a user may start with a mutation found in the upstream region of the MT1A gene (chr16:56638440, C→G) in H1975 lung cancer cell line (Figure 3B; see Supplementary Figure S4 for the more detailed guide). This gene is a member of the metallothioneins. It is reported that the mRNA/protein levels of this gene are associated with cellular proliferation and migration, differentiation and chemoresistance of cancers (35) as well as for genes from related families (36,37). The user would further connect the searches to find that this enhancer mutation is phased to its transcript region. The ChIP-seq tags for the active histone marks and the transcript RNA-seq tags show a biased representation. When the user examines the single-cell datasets, he/she will also find that the gene expression is diverse among individual cells, suggesting the transcriptional heterogeneity and diverse response to stimulation among cancer cells, which may give essential information, when drug intervention on this gene is contemplated.
AVAILABILITY
A detailed user manual is available on the website (http://dbtss.hgc.jp/docs/help_2017.html). Documents on data processing and information of quality control are also provided at the page of the experimental procedures (http://dbtss.hgc.jp/?doc:protocol_2017.html). Statistics for the current database are also presented in the statistics section (for more details, visit the statistics page at http://dbtss.hgc.jp/docs/data_contents_2017.html). All of the short read sequences used for the database are those which have been deposited in the Short Read Archives and JGA Database for Control Access in DDBJ (http://www.ddbj.nig.ac.jp/index-e.html). Accession numbers are as appear in the statistics section (left frame in the top page).

CONCLUSIONS
In this database, we attempt to develop a versatile database platform. We expect this database to facilitate the analyses regarding how germline variations or somatic mutations in cancers residing in transcriptional regulatory regions may affect the transcriptional regulation of their target genes in the diseased genome contexts. Since it is still generally difficult to collect the complete omics datasets directly from clinical samples, the surrogate data from model systems, such as cultured cells and mouse models, play important roles. In this database, we arranged the retrieval systems so that information from clinical samples should be easily associated with wider variation of omics information in model systems.

For the eventual diseased genomes, we particularly focused on those from the Japanese population. However, the framework of the database should be similarly useful by substituting the genomic information by other datasets of a different ethnic background. Ethnic backgrounds may potentially have substantial effects on the disease etiology. By providing the solid omics data core and referring them to the clinical samples, we believe our database can help a broad range of users who study clinical relevance of the genomic variations/mutations in genes particularly in the regulatory regions.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS
The database is maintained on the supercomputer system at Human Genome Center, the Institute of Medical Science, the University of Tokyo. We thank the members of the CREST/IHEC and other IHEC teams worldwide. We are also grateful to H. Wakaguri, Y. Kuze and T. Horiguchi at the University of Tokyo and the technical staffs of Database Center for Life Science.

FUNDING
DBTSS/DBKERO 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 ‘Platform for Advanced Genome Science’ [16H06279] from the Ministry of Education, Culture, Sports, Science and Technology of Japan; Database Integration Coordination Program from Japan Science and Technology Agency; CREST/IHEC, Japan Agency for Medical Research and Development. Funding for open access charge: CREST/IHEC, Japan Agency for Medical Research and Development. Conflict of interest statement. None declared.

REFERENCES
1. Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., Suzuki, Y. and Nakai, K. (2015) DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. Nucleic Acids Res., 43, D87–D91.
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. 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.
4. 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.
5. Stunnenberg, H.G. and International Human Epigenome Consortium. International Human Epigenome, C. International Human Epigenome, C. and Hirst, M. (2016) The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell, 167, 1089.
6. Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Patton, B. and Akeson, M. (2015) Improved data analysis for the MiniION nanopore sequencer. Nat. Methods, 12, 351–356.
7. Zheng, G.X., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M. et al. (2016) Haplotypeing germline and cancer genomes with high-throughput linked-read sequencing. Nat. Biotechnol., 34, 303–311.
8. Suzuki, A., Matsushima, K., Makinoshima, H., Sugano, S., Kohno, T., Tsuchihara, K. and Suzuki, Y. (2015) Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol., 16, 66.
9. Wu, A.R., Neff, N.F., Kalisky, T., Dauber, P., Treutlein, B., Rothenberg, M.E., Mbura, F.M., Mantalas, G.L., Sim, S., Clarke, M.F. et al. (2014) Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods, 11, 41–46.
10. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J. et al. (2017) Massively parallel digital transcriptional profiling of single cells. Nat. Commun., 8, 14049.
11. Koike, A., Nishida, N., Inoue, I., Tsuji, S. and Tokunaga, K. (2009) Genome-wide association database developed in the Japanese Integrated Database Project. J. Hum. Genet., 54, 543–546.
12. Higasa, K., Miyake, N., Yoshimura, J., Okamura, K., Niihori, T., Saitsu, H., Doi, K., Shimizu, M., Nakabayashi, K., Aoki, Y. et al. (2016) Human genetic variation database, a reference database of genetic variations in the Japanese population. J. Hum. Genet., 61, 547–553.
13. Narahara, M., Higasa, K., Nakamura, S., Tabara, Y., Kagawuchi, T., Ishii, M., Matsubara, K., Matsuda, F. and Yamada, R. (2014) Large-scale East-Asian eQTL mapping reveals novel candidate genes for LD mapping and the genomic landscape of transcriptional effects of sequence variants. PLoS One, 9, e100924.
14. Nagasaki, M., Yasuda, J., Katsuoka, F., Narui, N., Kojima, K., Kawai, Y., Yamaguchi-Kabata, Y., Yokozawa, J., Danjo, H., Saito, S. et al. (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat. Commun., 6, 8018.
15. Yamaguchi-Kabata, Y., Nariai, N., Kawai, Y., Sato, Y., Kojima, K., Tateno, M., Katsuoka, F., Yasuda, J., Yamamoto, M. and Nagasaki, M. (2015) iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. *Hum. Genom. Var.*, 2, 15050.

16. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, 29, 308–311.

17. Lek, M., Karcewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536, 285–291.

18. International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I. et al. (2017) International network of cancer genome projects. *Nature*, 464, 993–998.

19. The Cancer Genome Atlas Research Network (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, 45, 1113–1120.

20. Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L. et al. (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, 45, D777–D783.

21. Umemura, S., Mimaki, S., Makinoshima, H., Tada, S., Ishii, G., Ohmatsu, H., Niho, S., Yoh, K., Matsunaga, H., Goto, M., Arikawa, K., Shirai, Y., 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.

22. 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.

23. Bugold, D., Morais, D.A., Gauthier, C., Cote, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T. et al. (2016) The international human genome epitope consortium data portal. *Cell Syst.*, 3, 496–499.

24. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M. and Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74.

25. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J. et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 518, 317–330.

26. The UniProt, C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, 45, D158–D169.

27. Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P. et al. (2017) Ensembl 2017. *Nucleic Acids Res.*, 45, D635–D642.

28. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Zuber, S., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, 13, 103–107.

29. Suzuki, A., Makinoshima, H., Wagakuri, 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.*, 42, 13557–13572.

30. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, L.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N. et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313, 1929–1935.

31. Matsunaga, H., Goto, M., Arikawa, K., Shirai, Y., Tsunoda, H., Huang, H. and Kambara, H. (2015) A highly sensitive and accurate gene expression analysis by sequencing (“bead-seq”) for a single cell. *Anal. Biochem.*, 471, 9–16.

32. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. and Lee, W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, 46, 1160–1165.

33. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L. and Garraway, L.A. (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339, 957–959.

34. Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M. et al. (2016) Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.*, 48, 500–509.

35. Cherian, M.G., Jayasurya, A. and Bay, B.H. (2003) Metallothioneins in human tumors and potential roles in carcinogenesis. *Mutat. Res.*, 533, 201–209.

36. Kmieciek, A.M., Pula, B., Suchanski, J., Olbromski, M., Gomulkiewicz, A., Owzarek, T., Kruzkak, A., Ambicka, A., Rys, J., Ugorski, M. et al. (2015) Metallothionein-3 Increases Triple-Negative Breast Cancer Cell Invasiveness via Induction of Metalloproteinase Expression. *PLoS One*, 10, e0124865.

37. Kim, H.G., Kim, J.Y., Han, E.H., Hwang, Y.P., Choi, J.H., Park, B.H. and Jeong, H.G. (2011) Metallothionein-2A overexpression increases the expression of matrix metalloproteinase-9 and invasion of breast cancer cells. *FEBS Lett.*, 585, 421–428.