refTSS: A Reference Data Set for Human and Mouse Transcription Start Sites

Imad Abugessaisa[1,†], Shuhei Noguchi[1,†], Akira Hasegawa[1], Atsushi Kondo[1], Hideya Kawaji[1,2], Piero Carninci[1] and Takeya Kasukawa[1]

1 - RIKEN Center for Integrative Medical Sciences, 1-7-22, Suehiro-Cho, Tsurumi-Ku, Yokohama, Kanagawa 230-0045, Japan
2 - RIKEN Preventive Medicine and Diagnosis Innovation Program, 2-1, Hirosawa, Wako, Saitama 351-0198, Japan

Correspondence to Takeya Kasukawa: *takeya.kasukawa@riken.jp*
https://doi.org/10.1016/j.jmb.2019.04.045

Abstract

Transcription starts at genomic positions called transcription start sites (TSSs), producing RNAs, and is mainly regulated by genomic elements and transcription factors binding around these TSSs. This indicates that TSSs may be a better unit to integrate various data sources related to transcriptional events, including regulation and production of RNAs. However, although several TSS datasets and promoter atlases are available, a comprehensive reference set that integrates all known TSSs is lacking. Thus, we constructed a reference dataset of TSSs (refTSS) for the human and mouse genomes by collecting publicly available TSS annotations and promoter resources, such as FANTOM5, DBTSS, EPDnew, and ENCODE. The data set consists of genomic coordinates of TSS peaks, their gene annotations, quality check results, and conservation between human and mouse. We also developed a web interface to browse the refTSS (http://reftss.clst.riken.jp/). Users can access the resource for collecting and integrating data and information about transcriptional regulation and transcription products.

Introduction

Recent improvements in genomic technologies for DNA sequencing and new experimental protocols have enabled us to obtain large-scale heterogeneous genomic data. To compare or combine these various genomic data sets, we usually utilize "reference data sets" (Fig. 1a). The Genome Reference Consortium has been providing reference genome assemblies for several organisms [1] (https://www.ncbi.nlm.nih.gov/grc). Several groups have been providing reference gene annotation data sets, such as NCBI Genes [2], GENCODE [3], and ENSEMBL genes [4]. We can gain further biological insights through integrating these data sets together. By mapping DNA methylated regions to a reference genome in various experimental conditions, we can analyze in which tissues and stages a specific promoter is methylated. By quantifying expression levels based on gene annotations in a reference gene set, we can easily obtain additional information (e.g., related diseases) assigned to these reference genes.

One of the research fields that utilize large-scale data sets obtained by sequencing is the field of transcriptional regulation. The RNA-seq can measure expression levels of transcripts and genes. The CAGE (Cap Analysis of Gene Expression) can identify which promoter and enhancer regions are active, and can define exact genomic positions from where transcription is starting [5,6]. The ChIP-seq can identify binding sites of transcription factors and histone modifications [7]. The ATAC-seq can find regions of open chromatin throughout the genome [8]. The whole-genome bisulfite sequencing can identify methylated nucleotides in genomic DNA [9]. By combining and analyzing these resources and experimental results, we can identify which elements are contributing to specific transcription event from genes.

For this purpose, the transcription start site (TSS) can be a good reference to integrate the above types of data (Fig. 1b). RNAs transcription is initiated at TSSs, and thus, genes and transcripts can easily associate with TSSs in the reference. Since transcription is largely regulated by events occurring

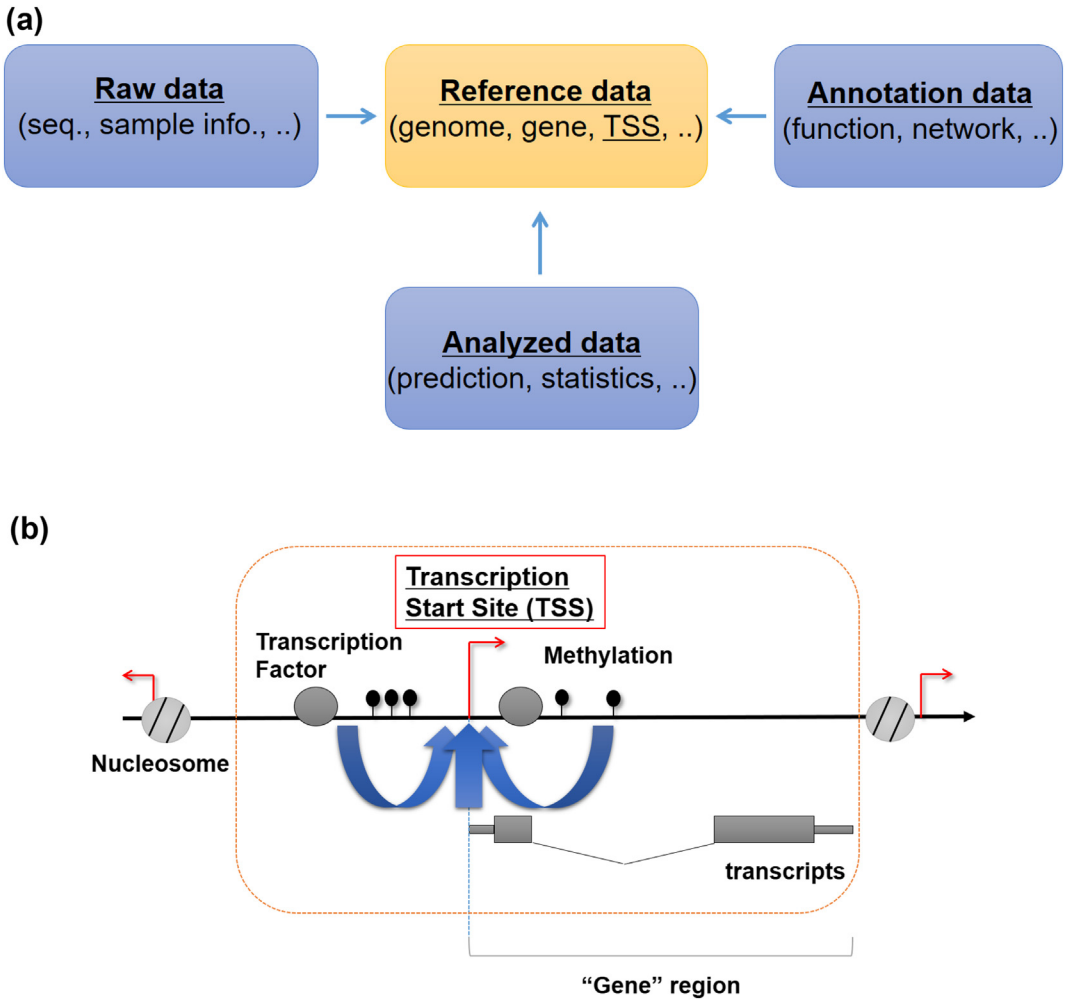


Fig. 1. Concept of refTSS. (a) Data integration based on reference data sets. Raw, annotation and analyzed data can be integrated based on units in the reference data sets. (b) Schematic view of the genomic elements around each TSS. Genomic components related to transcriptional regulation and transcriptional products are located around TSSs.

around TSSs, including binding of transcription factors, change of chromatin structure, modification of histones, and DNA methylation, we can infer relationships between these genomic elements and TSSs based on their distance. By consolidating this information for every TSSs, we can predict and analyze genome-wide transcriptional regulation networks.

However, few reference sets of TSSs with integrated transcriptional regulatory information are publicly available. In addition, the 5′-ends of annotated gene regions in these reference gene sets are not always complete (Supplemental Fig. 1). The largest TSS set currently available is the one published as part of the FANTOM5 project [10]. However, although FANTOM5 used many healthy and disease samples to obtain the promoter atlas, several cell types have yet to be analyzed [11]. Thus, we may still be missing cell type-specific TSSs, as well as cases of TSS switching [12] in which the usages of TSSs are changed among different samples. Such missing samples can be complemented by combining the FANTOM5 TSS set with other data sources. Moreover, if TSSs in the FANTOM5 promoter atlas are also detected in the different 5′-end sequencing protocols, we can enforce supporting evidences of the TSSs.

The incompleteness of available TSS data sets motivated us to build a comprehensive TSS reference with annotations about functions and transcriptional regulations. We collected publicly available data sets in order to identify 5′-end sequences of transcripts obtained by CAGE, TSS-Seq [13], and RAMPAGE [14], as well as available TSS/promoter data sets from FANTOM5 and the Eukaryote Promoter Database (EPDnew) [15]. For the integrated TSS set, we performed a quality assessment for each TSS and assigned various annotations in order to make the data set more useful for the end user. The reference set of TSS data (refTSS), along with the QC and gene annotation information and a web interface

for searching the data set, is publicly available at http://reftss.clst.riken.jp/.

Results

Reprocessing the source data

To build a comprehensive reference TSS set for the human and mouse genomes (Fig. 2a), we first collected publicly available 5′-end data sets from the FANTOM5 promoter atlas [16], ENCODE CAGE [17], ENCODE RAMPAGE [14], DBTSS [13], EPDnew [15], and CAGE profiling of human and mouse stem cells (DDBJ DRA accession number: DRA000914) [18] (Table 1 and Supplemental Tables 1–4). Each of these data sources was originally processed by different computational methods, using different genome assemblies, and are provided in different formats. Since our goal is to build the human and mouse refTSS data sets based on the latest genome assemblies (hg38 and mm10 for human and mouse, respectively), we developed a standardized set of 5′-end data reprocessing procedures (Fig. 2b and Materials and Methods). As the basic strategy for our reprocessing, we reused the genomic coordinates of TSSs or the genomic alignment of 5′-end sequences if these were provided by the data sources, as the original data generators would have likely employed the most suitable data processing for their own data sets. If we could not obtain specific processed data, we produced them ourselves by reprocessing the available files. For this purpose, we implemented a variety of reprocessing steps according to the source data set, including alignment of the raw sequence data to the hg38 or mm10 genome assemblies, TSS peaks calling, and the conversion of TSS peak coordinates from hg19/mm9 to hg38/mm10. Table 1 summarizes which steps are applied in the reprocessing of each data set.

We reprocessed the TSS regions in EPDnew by converting the original genomic coordinates to be compliant with hg38 and mm10. After the conversion, we obtained 25,500 human TSSs in hg38 and 21,236 mouse TSSs in mm10 (Table 2). The ENCODE consortium provides RAMPAGE data as 457 BAM files aligned to the hg38 genome assembly (Supplemental Table 1). From these BAM files, we identified 1,419,318,819 CAGE tag start sites (CTSSs), which indicate the 5′-ends of transcripts at 1-bp resolution. We then performed TSS peak calling using PARACLU [19] and obtained 3935 human TSS peaks (Table 2). We performed the genome mapping of 139 ENCODE CAGE raw sequence files (Supplemental Table 2). We identified 224,100,634 CTSSs from the mapping result and called 8232 TSS peaks (Table 2). We similarly reprocessed the 157 human TSS-seq data sets from DBTSS (Supplemental Table 3), identifying 59,720,912 CTSSs, and obtained 4,753 TSS peaks (Table 2). Finally, we reprocessed the 16 human and 18 mouse sequence files from CAGE profiling of human and mouse stem cells (Supplemental Table 4). The human and mouse raw sequence reads were mapped to the hg38 and mm10 assemblies, respectively. We identified 12,214,203 human and 9,171,879 mouse CTSSs, and obtained 11,082 human TSS and 16,533 mouse TSS peaks (Table 2). We used a stringent cutoff score of PARACLU to decrease false positive TSS peaks. This causes the smaller numbers of TSS peaks in the data sets using PARACLU.

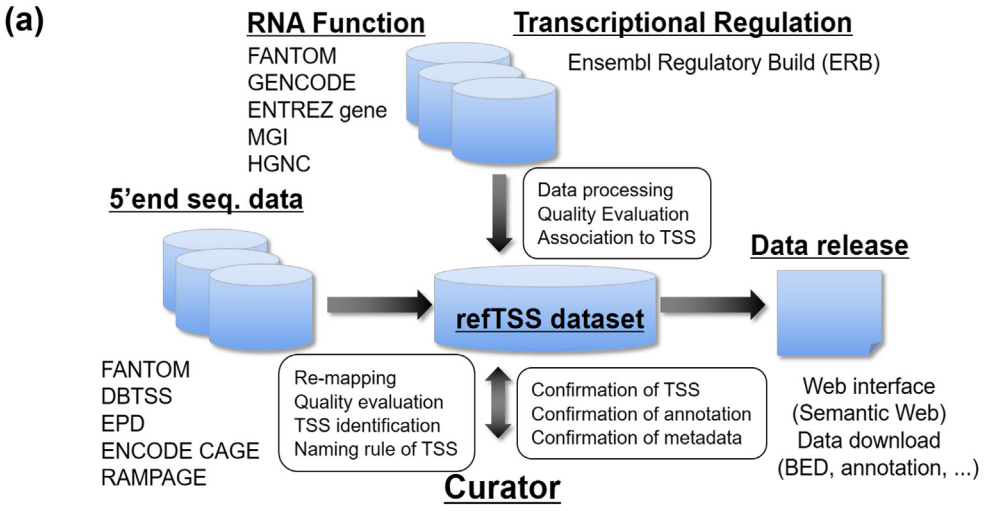
Integration of publicly available TSS sets

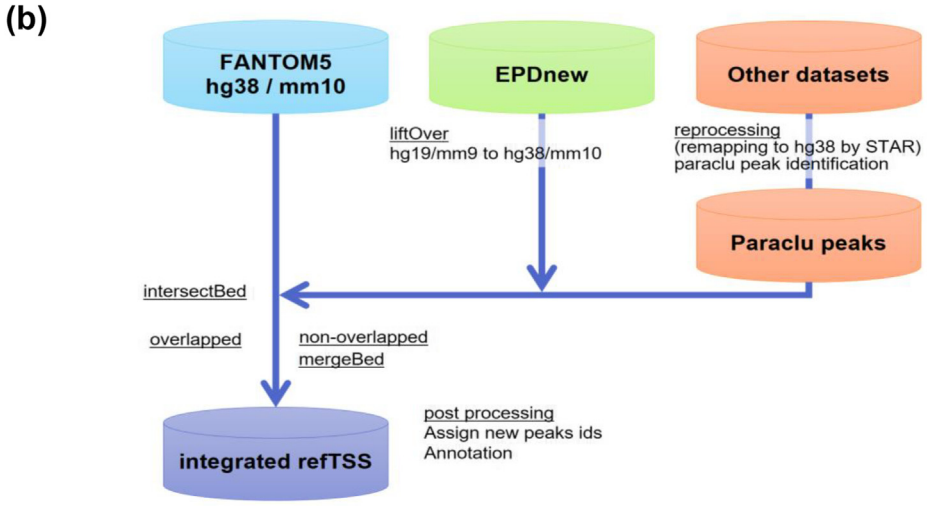
For constructing a single reference set of TSSs with the above publicly available 5′-end data sources, we developed a computational method to integrate multiple TSS sets (Fig. 2c). Since the largest TSS set is the FANTOM5 promoter atlas, we merged all other data sources with the FANTOM5 set (see details in Materials and Methods).

The numbers of integrated TSSs peaks for human and mouse are 224,694 and 173,204, respectively (Table 2). Most of these were derived from the FANTOM5 data as expected. There were 61.7% of human and 72.1% of mouse TSS peaks that are derived from non-FANTOM5 overlapped with FANTOM5 TSSs. Only 6.6% of human and 4.9% of mouse TSS peaks in refTSS are obtained exclusively from non-FANTOM5 data sets. A further 11.3% of human and 13.4% of mouse FANTOM5 peaks are also supported by other data sets and protocols. These numbers are not large, most likely because we opted for a conservative parameter for peak calling with PARACLU. Nevertheless, we can assign more supporting evidence of many TSS peaks using our integration method.

Quality check in refTSS

To evaluate the quality of each TSS peak in the refTSS, we obtained several QC metrics based on the genomic DNA sequence properties around the TSSs [20–23]. The first matric is the existence of TATA-boxes upstream of TSSs. We identified TATA-box motifs around each TSS, and the result is available in both the refTSS release file and the web interface. The summary plots of showing the distribution of TATA-boxes in human and mouse are shown in Fig. 3a and b, respectively. We found a significant enrichment of the TATA-box motif around 25 to 30 bp upstream of the TSSs. Nevertheless, the number of TSS peaks that have TATA-box motifs around them was small [26,079 (11.6%) in human and 20,588 (11.9%) in mouse], which is also reported in the previous paper [20]. Second, we calculated GC content around each TSS for an





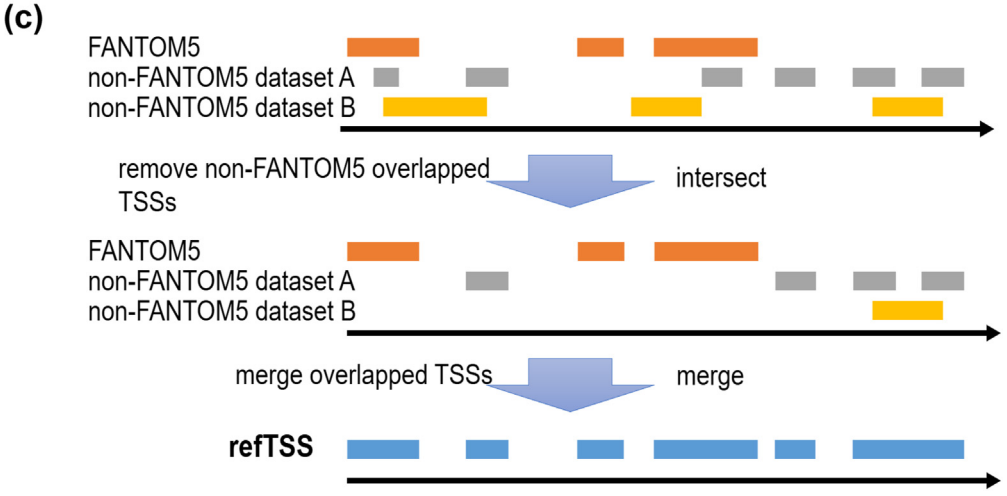


Fig. 2 (*legend on next page*)

additional evidence of TSS-ness, and the result is similarly available in the refTSS release file and the interface. The summary of the GC dinucleotide composition in flanking regions around TSSs is shown in Fig. 3c (human) and d (mouse). As expected, GC content is enriched around TSSs compared to other regions and random ones, as reported in the previous paper [20]. Finally, we performed computational classification of TSS peaks using the TSS classifier program in the TomeTools package (http://tometools.sourceforge.net/), which is a supervised classification method defining TSS peaks as either TSS-like or not based on a random decision tree model [11]. We found a TSS classification rate for the combined TSS set of about 46.6%, which is comparable to that of the FANTOM5 CAGE peaks, while the TSS classification rate of non-FANTOM5 TSSs only is higher at 68.1%. This indicates that our integration can expand the single promoter data sets to add more reliable TSS peaks.

Gene annotation

We assigned gene annotations to the consolidated TSS peaks based on a previously reported procedure [16] (Fig. 4a). We also chose the representative TSS peak for each gene as the one expressed in the highest number of samples. How many TSSs can be associated with any annotated transcripts and genes in human and mouse are summarized in Fig. 4b and c, respectively. Only about 45%–55% of TSSs can be associated to any annotated genes. As for TSS peaks that have not been associated with any transcripts, around 12% (human and mouse) are classified as TSSs using the TomeTools TSS classifier, around 51% (human) are located in regulatory regions (see the next section for more details), and about 7% human and 5% mouse TSS peaks without any gene annotations overlap with FANTOM5 active enhancer regions [24]. This may indicate that the current gene annotation is not adequate for the analysis of TSSs and promoters, and our reference set of TSSs is essential for studies of transcriptional regulation. Next, how many genes can be covered by any TSS peaks in the human and mouse refTSS are summarized in Fig. 4d and e, respectively. In Entrez Gene, the refTSS can cover about 70% of all annotated genes. Breakdowns of the gene coverages by gene types are shown in Fig. 4f and g. About 80% of protein-coding genes can be covered by refTSS, while less than half of all non-coding genes can be covered.

Regulatory annotation

We classified TSS peaks in the refTSS with the Ensembl Regulatory Build (ERB) [25] (Fig. 4h and i for human and mouse, respectively). ERB classified genomic regions related to gene regulation into promoters, promoter flanking regions, enhancers, CTCF binding sites, transcription factor binding sites, and open chromatin regions. About 47.9% of human and 42.8% of mouse TSS peaks overlapped the promoter regions in ERB, and the second most overlapped regions were the promoter flanking regions (8.4% in human and 11.5% in mouse). Overall, 56.3% human and 54.3% mouse TSS peaks in the refTSS were located around promoter regions, as expected. Two percent human and 2.4 percent mouse TSSs are located in enhancer regions, which may indicate TSSs of enhancer RNAs [26] or is due to the ambiguity in the distinction of promoters and enhancers [27].

Conservation between human and mouse

Several studies have investigated and identified conservation (similarities and differences) between human and mouse genomes at the level of regulatory networks (e.g., TF, TFBS) [28]. In the detailed analysis of the FANTOM3 promoter data set[20], they evaluated the conservation between human and mouse promoters, and they showed that evolutional speeds of promoters were varied depending on the classes of promoters. The Mouse ENCODE Consortium reported that about 50% of TF occupied sequences do not align between mouse and human genomes [28]. These kinds of studies enable us to understand the evolutionary aspect of gene regulatory mechanisms. Thus, we investigated conservation of TSS peaks between human and mouse genomes in refTSS (Fig. 5a). We identified about 45% of TSS peaks are conserved between the two genomes (the sum of “conserved” in Fig. 5b). About 15% of TSSs are not conserved, although the genomic regions themselves are conserved (“not.conserved” in Fig. 5b). About 39% of TSS peak regions are not conserved even in the genomic regions (“unmapped” in Fig. 5b). In the “conserved” TSS peaks, around half are associated with orthologous genes, and 60%–70% had matching

Fig. 2. Pipeline for refTSS construction. (a) Overview of the refTSS data set construction pipeline. The TSS data from various data sources were integrated into a unique reference TSS peak set. Functional and regulatory annotations were then assigned to the TSS peaks. The data were manually curated and published in several ways. (b) Reprocessing procedure of publicly available 5′-end data sources. All data sources were first converted to TSS peaks with hg38 or mm10 coordinates, and then integrated into a unique set of TSS peaks. (c) TSS peak integration strategy. First, non-FANTOM5 TSSs overlapping any FANTOM5 TSSs were removed. Second, all overlapping TSS peaks were merged.

Table 1. Data sources for transcriptional start sites

5′ end seq data source	Version	Seq method	Availability	Organisms	Data formats	Preprocessing
FANTOM5	hg38_v5 mm10_v5	CAGE	http://fantom.gsc.riken.jp/5/datafiles/reprocessed/	Human/ Mouse	Genomic coordiantes in BED format (hg38 and mm10)	No
EPDnew	ver. 004 (human) ver. 002 (mouse)	Manual curation	ftp://ccg.vital-it.ch/epdnew/human/004/Hs_EPDnew_004_hg19.bed ftp://ccg.vital-it.ch/epdnew/M_musculus/002/Mm_EPDnew_002_mm9.bed	Human/ Mouse	Genomic coordiantes in BED format (hg19 and mm9)	liftOver
RAMPAGE		CAGE	https://www.encodeproject.org/	Human	BAM files (hg38)	Peak calling
ENCODE CAGE		CAGE	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34448	Human	Raw sequence in Fastq	Mapping, peak calling
DBTSS	release 9.0	TSS-Seq	ftp://ftp.hgc.jp/pub/hgc/db/dbtss/dbtss_ver9	Human	Raw sequence in Fastq	Mapping, peak calling
stem cell CAGE		CAGE	https://ddbj.nig.ac.jp/DRASearch/submission?acc=DRA000914	Human/ Mouse	Raw sequence in Fastq	Mapping, peak calling

Table 2. Summary of refTSS peak integration

hg38	
Data set	The number of regions
FANTOM5	209,911
EPDnew	25,500
RAMPAGE	3935
ENCODE CAGE	8232
DBTSS	4753
Stem cell CAGE	11,082
refTSS	224,694
mm10	
Sata set	The number of regions
FANTOM5	164,672
EPDnew	21,236
Stem cell CAGE	16,533
refTSS	173,204

ERB annotations. However, about 10,000 conserved TSS peaks were not associated with any downstream genes, which could indicate that these are conserved enhancer RNAs, promoter RNAs, miRNA precursors, or other unknown RNAs. These annotations could be useful for further analysis of conservation of TSSs and promoters.

Web interface and data availability

The human and mouse refTSS files are available at our web site http://reftss.clst.riken.jp/. Users can download the files containing the genomic coordinates of TSS peaks, gene-based annotation information of TSS peaks, QC metrics, and conservation between human and mouse TSS peaks. For easy access to the refTSS data set, we developed a web-based user interface to search for TSS peaks and genes and to browse annotations and QC metrics for each TSS (Fig. 6). We also setup a TrackHub for refTSS. User can add the refTSS TrackHub to MyHub in the UCSC Genome Browser with the URL http://reftss.clst.riken.jp/trackhub/hub.txt.

Discussion

The refTSS provides a comprehensive set of consolidated human and mouse TSSs together with their quality assessment results and annotations. To build this TSS set, we collected data from multiple public resources including the FANTOM5 promoter atlas with CAGE analysis results for approximately 3000 human and mouse samples, DBTSS with various TSS-Seq results, the EPDnew resource with curated promoter regions based on the available 5′-end and gene annotation resources, and several other experimental results using a variety of protocols in public repositories. The collected data were integrated to merge overlapped TSSs, upon which was built a unique and non-redundant set of TSSs. We then performed several

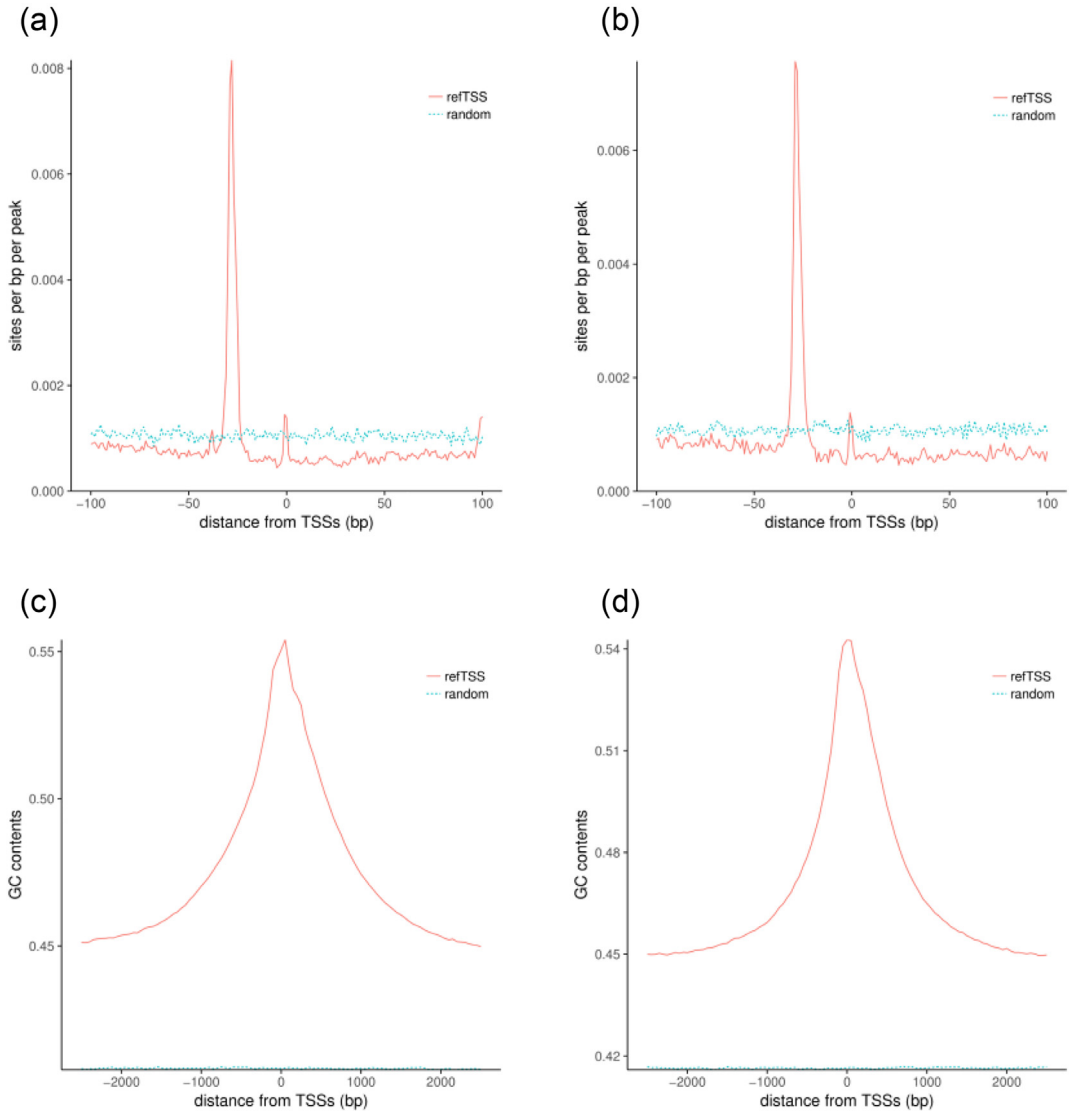
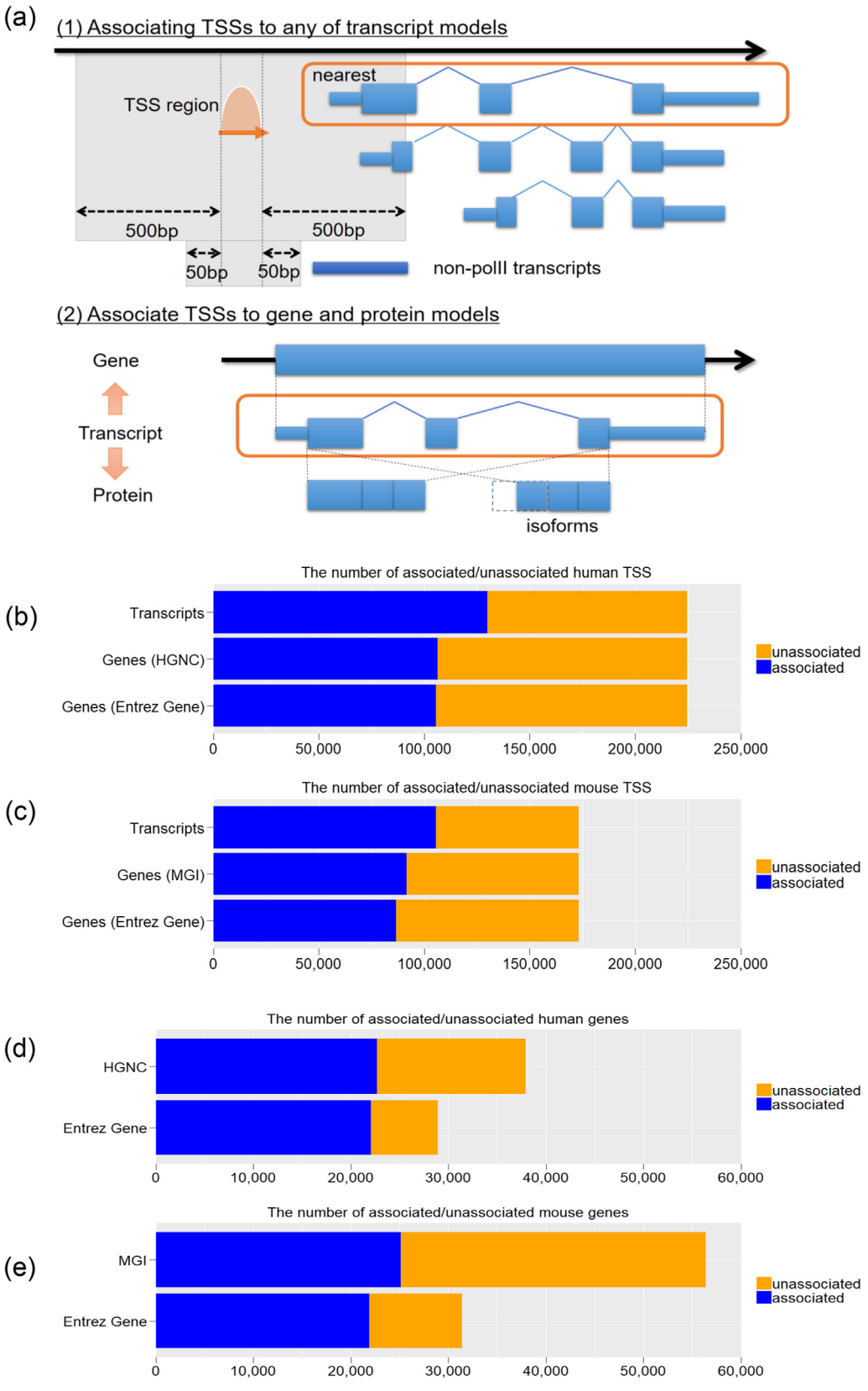


Fig. 3. QC results of refTSS. (a–b) Enrichment of TATA-boxes in refTSS of (a) human and (b) mouse. (c–d) Density of GC content in flanking regions around TSS peaks in (c) human and (d) mouse.

quality assessments in order to determine whether each TSS was likely to be a true TSS. We also assigned gene and regulatory annotations to the TSS, to provide the user with further details and characteristics for TSSs of interest. In addition, we identified conserved TSS pairs between human and mouse genomes. By adding our evaluation of quality assessments and annotations, we believe that the refTSS is a useful and reliable resource for analyzing promoters and transcriptional regulations.

The refTSS data set has various applications. One such application is the quantification of promoter activities using CAGE data. For data processing of RNA-seq results, we can use a reference gene set (e.g., GENCODE or refSeq) to quantify expression levels of genes or transcripts [29–31]. The reference genes are usually associated with rich annotations, such as gene names and Gene Ontology terms [32], and we can utilize this information without additional processing for annotation. However, in the data processing of CAGE results, additional data processing was necessary, to call TSS regions and assign gene annotations, in addition to the quantification of TSS expression and promoter activities. By using the refTSS data set similarly to reference gene sets used with RNA-seq, we can skip the annotation processes and easily obtain an expression table of annotated TSSs. The refTSS can also be used with other protocols that sequence 5′-ends of transcripts, TSS-seq, and RAMPAGE.

The refTSS covers 80%–90% of the known protein-coding genes in the public databases.


(a)
(1) Associating TSSs to any of transcript models
nearest
TSS region
500bp
500bp
50bp
50bp
non-polII transcripts
(2) Associate TSSs to gene and protein models
Gene
Transcript
Protein
isoforms
(b)
The number of associated/unassociated human TSS
Transcripts
Genes (HGNC)
Genes (Entrez Gene)
unassociated
associated
0
50,000
100,000
150,000
200,000
250,000
(c)
The number of associated/unassociated mouse TSS
Transcripts
Genes (MGI)
Genes (Entrez Gene)
unassociated
associated
0
50,000
100,000
150,000
200,000
250,000
(d)
The number of associated/unassociated human genes
HGNC
Entrez Gene
unassociated
associated
0
10,000
20,000
30,000
40,000
50,000
60,000
(e)
The number of associated/unassociated mouse genes
MGI
Entrez Gene
unassociated
associated
0
10,000
20,000
30,000
40,000
50,000
60,000

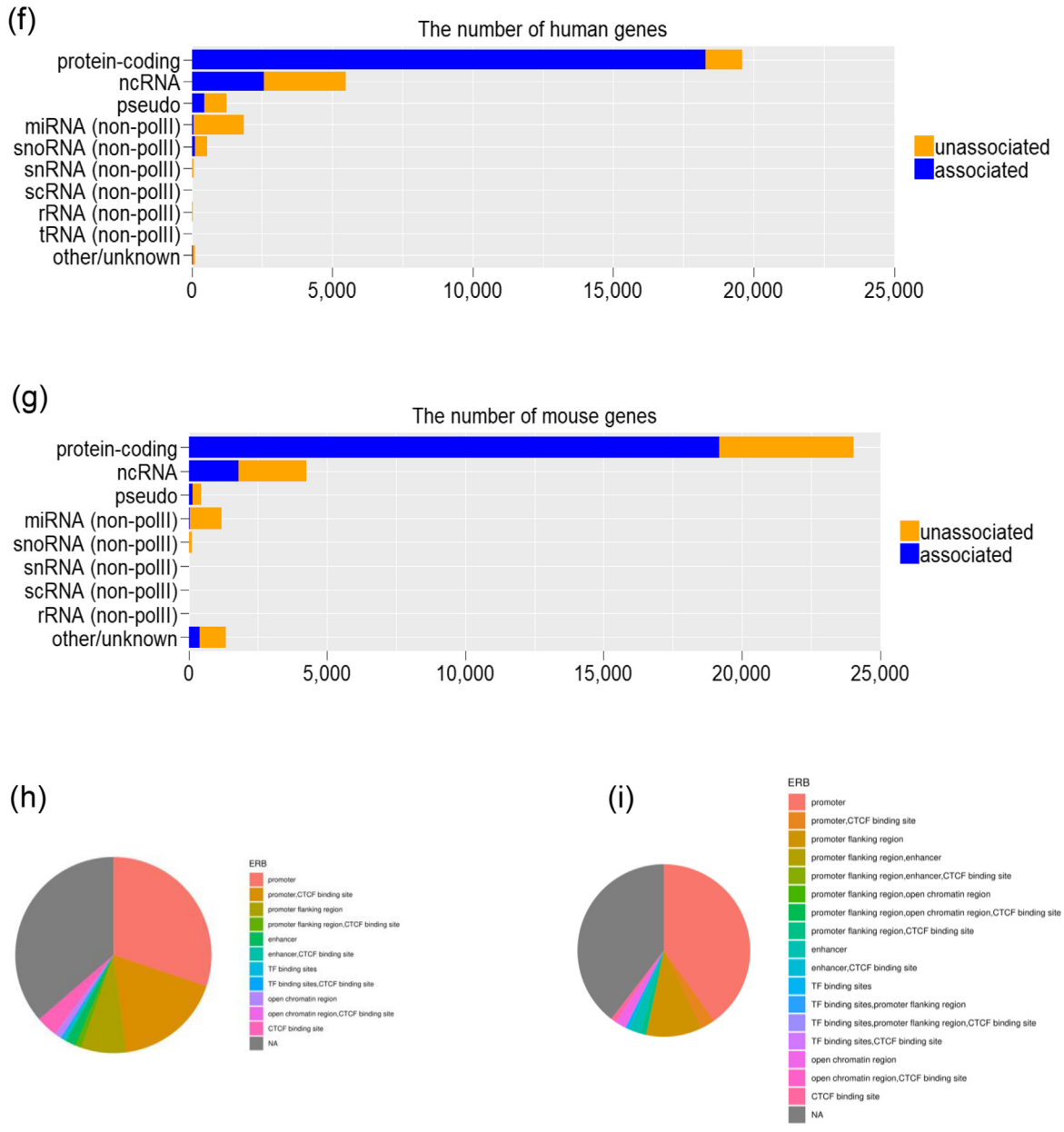


Fig. 4. Annotation of refTSS. (a) TSS annotation pipeline. First, nearest transcripts (within 500 bp for polII transcripts and 50 bp for non-polII transcripts) were chosen from public transcript sets for each TSS peak. Next, gene and protein annotations assigned to the nearest transcripts were transferred to the TSS peaks. (b–c) Human (b) and mouse (c) TSS peaks annotated to any transcripts and genes. (d–e) Human (d) and mouse (e) genes covered by any TSS peaks. (f–g) Gene categories covered by any TSS peaks. Graphs (d) and (e) are broken down into gene categories (e.g., protein-coding, ncRNA) for human (f) and mouse (g). (h,i) The number of human (h) or mouse (i) TSS peaks overlapping annotated genomic regions in ERB. Each category corresponds to the types of regulatory features in ERB, showing combinations of promoters, promoter flanking regions, enhancers, CTCF binding sites, TF binding sites, and open chromatin regions.

Although the missing genes should be recovered by future improvements to the data set and by using new RNA sequencing technologies to be able to connect TSS and genes (e.g., CAGEScan [33], Nanopore full-length RNA sequencing [34], analyses of protein-coding genes based on the refTSS can be considered almost complete. However, the refTSS covers less than half of known ncRNA genes. One reason is due to the incomplete gene models especially of ncRNAs [27], which requires the correction of TSSs in the public databases to improve the coverage. The second reason is that this may be reflecting the low abundance and the cell-type specific expression of lncRNAs [27,35],

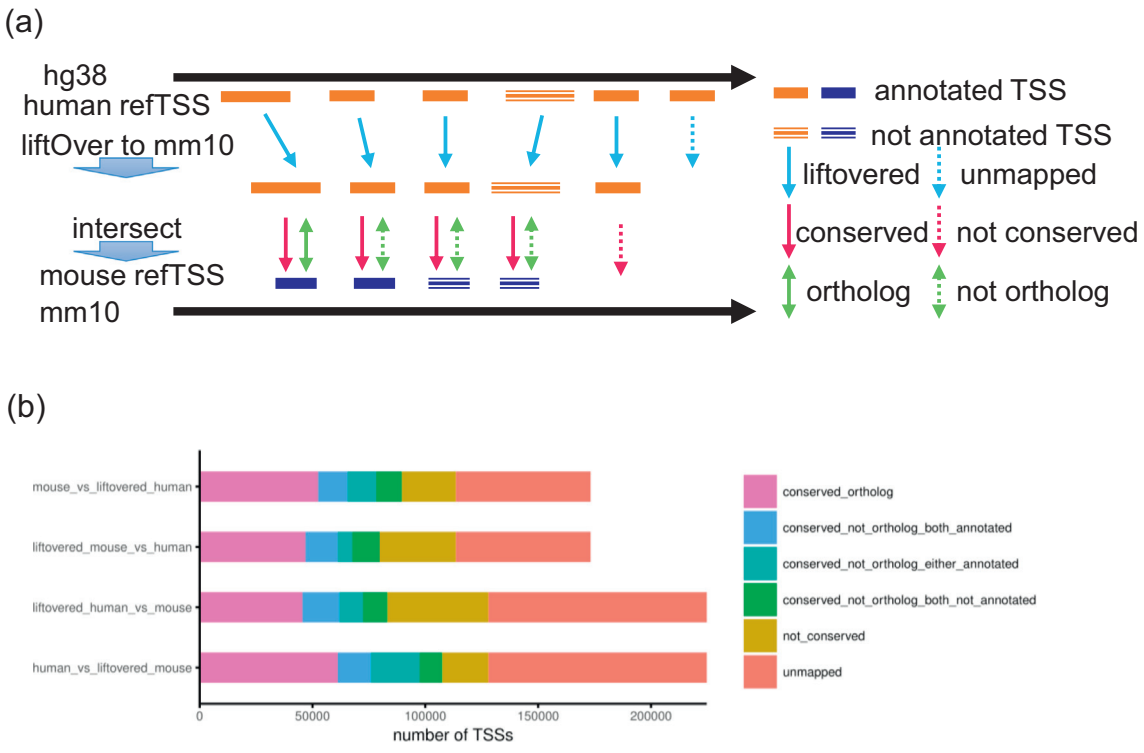


Fig. 5. Conservation between human and mouse refTSS. (a) Strategy to obtain conserved TSS peaks between human and mouse. We performed liftOver with human TSS peaks to the mouse reference genome assembly and vice versa. (b) Results of conserved TSS peaks. TSS peaks are categorized into as "conserved_ortholog" if there are conserved TSS peaks in another species and both annotated genes are orthologs; "conserved_not_ortholog_both_annotated" if there are conserved TSS peaks in another species and both human and mouse are annotated, but not orthologs; "conserved_not_ortholog_either_annotated" if there are conserved TSS peaks in one species but unannoated in the other; "conserved_not_ortholog_both_not_annotated" if there are conserved TSS peaks that are unannotated in both; "not_conserved" if the TSS region is conserved between the two genomes but there are no conserved TSS peaks in another species; and "unmapped" if genomic regions of TSS peaks are not conserved. The *y* axis is categorized into "mouse_vs_liftovered_human," the number of TSSs of mm10 overlapping TSSs liftovered from hg38 to mm10; "liftovered_mouse_vs_human," the number of TSSs liftovered from mm10 to hg38, overlapping hg38 TSSs; "liftovered_human_vs_mouse," the number of TSSs liftovered from hg38 to mm10, overlapping mm10 TSSs; "human_vs_liftovered_mouse," the number of hg38 TSSs overlapping TSSs liftovered from mm10 to hg38.

meaning that more thorough identification of 5′-ends of lncRNAs requires deeper sequencing of RNAs from more varied sample types. Another potential reason for the lower coverage is that some ncRNAs, especially those transcribed by others than RNA polymerase II, lack polyA or 5′-capping structures. To capture the 5′-ends of such RNAs, we may need to adopt other protocols [36].

In the future, we will add more sequencing results from protocols that identify 5′-ends (e.g., CAGE, TSS-Seq, RAMPAGE, STRT [37], single-cell CAGE [38]), and add more TSSs to increase the coverage of both protein-coding genes and ncRNAs and to detect new ones. For this purpose, we may use the FANTOM CAT data set, which provides a new reference gene set with 5′-ends and high numbers of non-coding genes [27]. In addition to this, we plan to add more annotation data, especially regarding transcriptional regulation and the epigenome. For example, we can use the ChIP-Atlas [39] to enrich our knowledge of transcription factors binding around TSSs, and use the information in Roadmap Epigenomics and ENCODE data sets to enrich for epigenomic elements in addition to the current ERB annotation. We also plan to develop computational tools to utilize our refTSS resources, such as a tool for the quantification of promoter activities using refTSS.

Materials and Methods

5′-End sequence data sources

We collected publicly available human and mouse 5′-end sequencing data from public repositories and databases. All the data sources used to build the

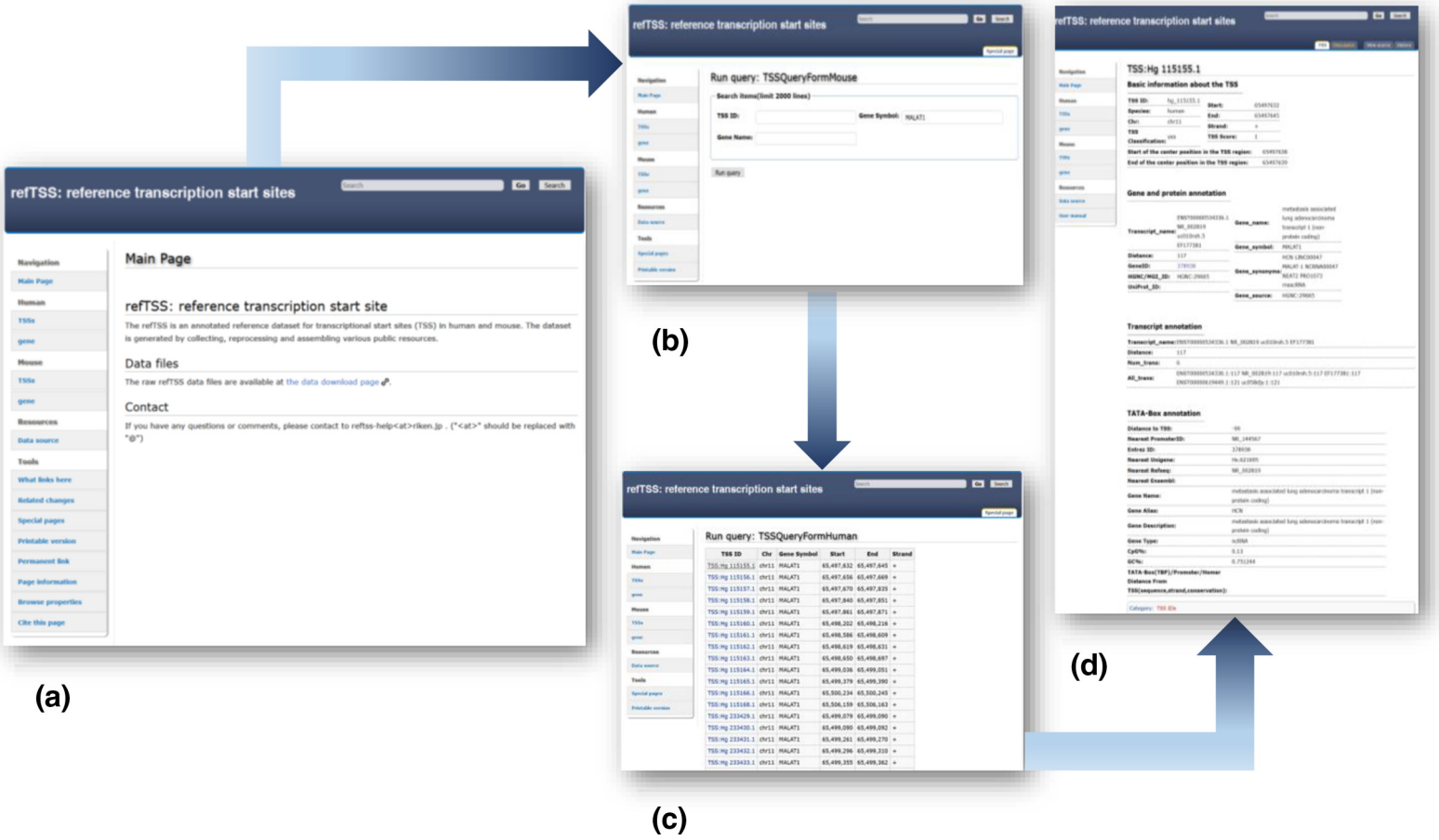
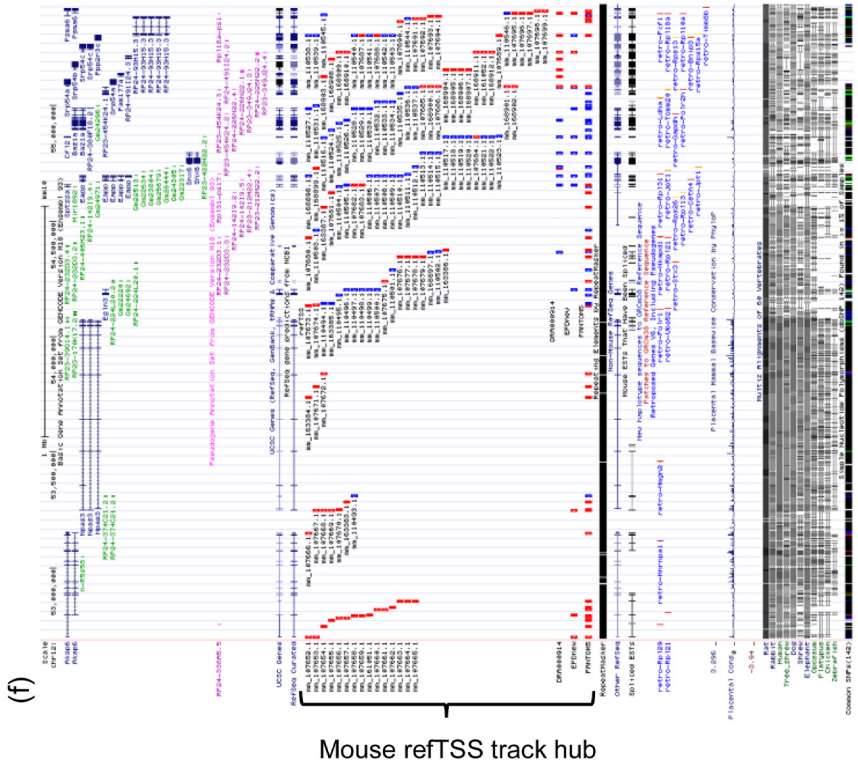


Fig. 6. Web Interface for refTSS. (a–d) Transitions among pages in the refTSS web interface. The interface includes the front (a), search (b), list of search results (c), and details of TSS peak or gene (d). (e–f) TrackHub interface for the UCSC Genome Browser. Example genomic views around GAPDH TSS (e) or MALAT1 (f) with TSS peaks in refTSS and their data sources.



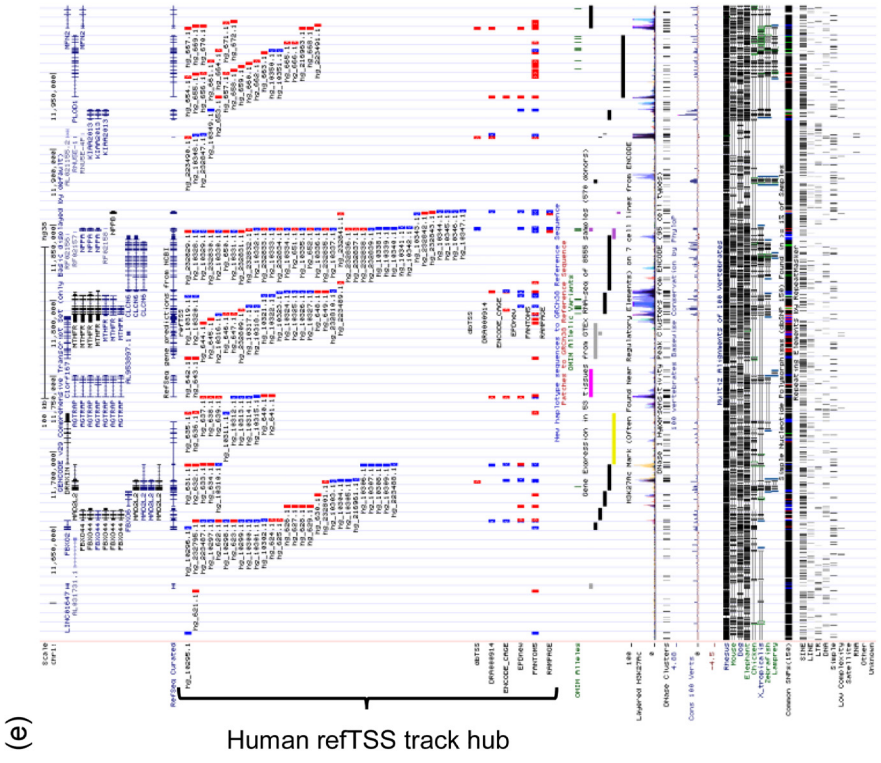


Fig. 6. (*continued*).

refTSS is in Table 1[13–16,18]. In summary, we obtained the reprocessed human and mouse CAGE peak coordinate files from the FANTOM5 promoter atlas (hg38_v5 for human and mm10_v5 for mouse). From the Eukaryotic Promoter Database (EPDnew) web site, we obtained the BED files of human (ver. 004) and mouse (ver. 002) promoter regions. For the RAMPAGE data, we downloaded the BAM files from the ENCODE web site (see metadata in Supplemental Table 1). For ENCODE CAGE (NCBI GEO accession number GSE34448), we downloaded files from NCBI Sequence Read Archive (see metadata in Supplemental Table 2). For DBTSS, we obtained raw sequence reads in the FASTQ format from DDBJ Sequence Read Archive (DRA) (see metadata in Supplemental Table 3) based on the DBTSS Release 8.0. For the CAGE profiling of human and mouse stem cells (DDBJ DRA accession number DRA000914), we downloaded the raw sequence reads from DDBJ DRA (see metadata in Supplemental Table 4). Although this data set provides TSSs generated by two different protocols (CAGEscan and CAGE), we used the human and mouse CAGE data set only.

Reprocessing of 5′-end sequence data

In order to build the refTSS in the latest genome assemblies (hg38 and mm10), we reprocessed the obtained data based on the version of the genome assembly in which the 5′-end data sets were provided, and the types of data were available. The procedures are summarized in Table 1.

Conversion of the genomic coordinates to the latest genome assembly

If the data source provided genomic coordinates in hg19 and mm9 genome assemblies as in the case of EPDnew, we extracted a 1-bp-long TSS position in the promoter regions defined in the data set and converted the genomic coordinates from hg19/mm9 to hg38/mm10 genome assemblies. For the coordinate conversion, we used the UCSC lifOver utility with the option –minMatch = 1 and the chain files downloaded from the UCSC Genome Browser site (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz and http://hgdownload.cse.ucsc.edu/goldenPath/mm9/liftOver/mm9ToMm10.over.chain.gz).

Reprocessing of raw sequence reads

For the data sets provided as raw sequence reads in FASTQ format (ENCODE CAGE, DBTSS, etc.), we carried out genome mapping and TSS peak calling. For DBTSS, divided FASTQ files from the same sequencing libraries were concatenated. Firstly, we mapped the raw sequences to the latest reference genomes (hg38 or mm10) using the STAR mapper [40] version (020201) with the default options. The resulting BAM files were further processed to extract 5′-ends of complete alignments (CTSS), and input BED files were generated by pooling all CTSS files in single data sets for the TSS peak calling. We used the latest version (ver. 9) of PARACLU [19] for peak calling with an option of minValue of 464 and maximum cluster length of 20. This cutoff was decided based on our tests of PARACLU running with different parameters in order to obtain the best peak regions for the integration process. The results of our parameter testing are summarized in Supplemental Table 5. The PARACLU program can be downloaded from http://cbrc3.cbrc.jp/~martin/paraclu/. After running PARACLU, we merged the overlapping TSS clusters with the “mergeBed” program in the BEDTools package [41]. The mapping and post-processing workflow was implemented in the MOIRAI workflow system [42].

Reprocessing of the mapped reads in BAM format

The genome mapping of the RAMPAGE data set was provided as coordinate using the hg38 genome assembly. We used the BAM files as inputs to PARACLU for TSS peak calling as described above, setting minValue of 17,233 and maximum cluster length of 20.

Integration of TSS data

All of the FANTOM5 TSS peaks were retained and intersected with the TSS peaks from non-FANTOM5 data sets using “intersectBed,” from the BEDTools package. Overlapping TSS peaks from the non-FANTOM5 data sources were removed. Non-FANTOM5 TSS peaks not overlapping the FANTOM5 peaks were merged together using “mergeBed” in BEDTools.

Quality evaluation and classification of refTSS peaks

To investigate the enrichment of TATA-box motifs and calculate GC content around the TSSs, we used the annotatePeaks.pl program in the HOMER software (v4.9.1) [43]. We used the TATA motif file embedded in the HOMER package to search for the motif within −100 to +100 bp around the center of TSS peaks. For calculating GC content, we executed the same program with a −2500 to +2500 bp range around TSS peaks. Randomized genomic positions were generated by “randomBed” in the BEDTools package. We also used the tssclass program in the TomeTools package (http://tometools.sourceforge.net/) to obtain their “TSS ness” classifications for TSS peaks in refTSS.

Table 3. Data sources for annotation of refTSS

Data source	Type	Human entries	Mouse entries	Source URL
Entrez Gene	Gene	61,190	68,527	ftp://ftp.ncbi.nih.gov/gene/DATA/
HGNC (HUGO Gene Nomenclature Committee)	Gene	41,548		ftp://ftp.ebi.ac.uk/pub/databases/genenames/
MGI (Mouse Genome Informatics)	Gene		81,303	http://www.informatics.jax.org/downloads/reports/
RefSeq	Transcript	75,893	42,574	ftp://hgdownload.soe.ucsc.edu/goldenPath/
Gencode	Transcript	203,835	136,535	ftp://ftp.ebi.ac.uk/pub/databases/gencode/
UCSC Gene	Transcript	197,782	63,814	ftp://hgdownload.soe.ucsc.edu/goldenPath/
mRNAs in UCSC Genome Browser	Transcript	421,833	399,572	ftp://hgdownload.soe.ucsc.edu/goldenPath/
UniProt	Protein	(105,088)	(77,931)	ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/

Gene annotation

We assigned gene, transcript, and protein coding annotations to refTSS peaks based on the method that we have previously reported [16]. We used the following public databases for annotation (all current versions as of October 2, 2018): GENCODE [3], Entrez Gene [2], HUGO Gene Nomenclature Committee (HGNC) database [44], the Mouse Genome Database (MGD) [45], the UCSC Genome Browser [46], UniProt [47], and FANTOM5 enhancers [24] (See Table 3).

Regulatory annotation

The hg38 refTSS was compared with the regulatory regions of the ERB[25] by overlapping the two sets by "intersectBed" in the BEDTools package. If a single refTSS region overlapped with multiple ERB regions, all overlapped ERB regions were reported.

Conservation between human and mouse peaks

To identify the conserved TSS peaks between human and mouse genomes, we used the liftOver tool [48] to convert human TSS regions to coordinates in the mouse genome assembly, and then ran intersectBed to observe overlaps with any mouse TSS peaks, and vice versa. The human or mouse TSS peaks were then classified into (1) unmapped to a counterpart genome, (2) mapped but not conserved, or (3) mapped and conserved.

Web interface development

In order to develop a web interface for the refTSS data set, we utilized Semantic MediaWiki (SMW) (Version 1.27.1) based on our previous report [49]. MediaWiki is a wiki engine developed for Wikipedia, and Semantic MediaWiki is an extension to MediaWiki that allows for the storage of semantic information (termed semantic properties) alongside wiki content. The data model of the refTSS interface consists of two classes (TSS and Gene). Each TSS is associated with one gene, while each gene is associated with one or more TSSs.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2019.04.045.

Acknowledgments

We thank Nobuyuki Takeda and Teruaki Kitakura for their support to the infrastructure of the web interface. We are also grateful for helpful comments and proofreading of the report by Dr. Marina Lizio, Dr. Andrew Kwon, and Dr. Matthew Valentine. This work was supported by research grants to the RIKEN Center for Life Science Technologies and RIKEN Center for Integrative Medical Sciences from MEXT, Japan.

Declaration of Competing Interest: None.

Received 30 December 2018;
Received in revised form 25 April 2019;
Available online 8 May 2019

Keywords:
transcription start sites;
transcriptional regulation;
data integration;
annotation;
reference data

†These authors have equally contributed to this work.

Abbreviations used:
CAGECap Analysis of Gene Expression; TSStranscription start site; refTSSreference set of TSS data; CTSSsCAGE tag start sites; ERBEnsembl Regulatory Build.

References

[1] D.M. Church, V.A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, et al., Modernizing reference genome assemblies, PLoS Biol. 9 (2011), e1001091.

[2] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez gene: gene-centered information at NCBI, Nucleic Acids Res. 39 (2011) D52–D57.

[3] J. Harrow, A. Frankish, J.M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, et al., GENCODE: the reference human genome annotation for The ENCODE Project, Genome Res. 22 (2012) 1760–1774.
[4] D.R. Zerbino, P. Achuthan, W. Akanni, M.R. Amode, D. Barrell, J. Bhai, et al., Ensembl 2018, Nucleic Acids Res. 46 (2018) D754-D61.
[5] M. Kanamori-Katayama, M. Itoh, H. Kawaji, T. Lassmann, S. Katayama, M. Kojima, et al., Unamplified cap analysis of gene expression on a single-molecule sequencer, Genome Res. 21 (2011) 1150–1159.
[6] M. Itoh, M. Kojima, S. Nagao-Sato, E. Saijo, T. Lassmann, M. Kanamori-Katayama, et al., Automated workflow for preparation of cDNA for cap analysis of gene expression on a single molecule sequencer, PLoS One 7 (2012), e30809.
[7] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions, Science 316 (2007) 1497–1502.
[8] J.D. Buenrostro, P.G. Giresi, L.C. Zaba, H.Y. Chang, W.J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position, Nat. Methods 10 (2013) 1213–1218.
[9] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315-22.
[10] S. Noguchi, T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, et al., FANTOM5 CAGE profiles of human and mouse samples, Sci Data. 4 (2017), 170112.
[11] A.R. Forrest, H. Kawaji, M. Rehli, J.K. Baillie, M.J. de Hoon, V. Haberle, et al., A promoter-level mammalian expression atlas, Nature. 507 (2014) 462–470.
[12] P. Zhang, E. Dimont, T. Ha, D.J. Swanson, W. Hide, D. Goldowitz, et al., Relatively frequent switching of transcription start sites during cerebellar development, BMC Genomics 18 (2017) 461.
[13] R. Yamashita, H. Wakaguri, S. Sugano, Y. Suzuki, K. Nakai, DBTSS provides a tissue specific dynamic view of transcription start sites, Nucleic Acids Res. 38 (2010) D98–104.
[14] P. Batut, A. Dobin, C. Plessy, P. Carninci, T.R. Gingeras, High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression, Genome Res. 23 (2013) 169–180.
[15] R. Dreos, G. Ambrosini, R.C. Périer, P. Bucher, The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools, Nucleic Acids Res. 43 (2015) D92–D96.
[16] I. Abugessaisa, S. Noguchi, A. Hasegawa, J. Harshbarger, A. Kondo, M. Lizio, et al., FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies, Sci Data. 4 (2017), 170107.
[17] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature. 489 (2012) 57–74.
[18] A. Fort, K. Hashimoto, D. Yamada, M. Salimullah, C.A. Keya, A. Saxena, et al., Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance, Nat. Genet. 46 (2014) 558–566.
[19] M.C. Frith, E. Valen, A. Krogh, Y. Hayashizaki, P. Carninci, A. Sandelin, A code for transcription initiation in mammalian genomes, Genome Res. 18 (2008) 1–12.
[20] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, et al., Genome-wide analysis of mammalian promoter architecture and evolution, Nat. Genet. 38 (2006) 626–635.
[21] B. Lenhard, A. Sandelin, P. Carninci, Metazoan promoters: emerging characteristics and insights into transcriptional regulation, Nat. Rev. Genet. 13 (2012) 233–245.
[22] V. Haberle, B. Lenhard, Promoter architectures and developmental gene regulation, Semin. Cell Dev. Biol. 57 (2016) 11–23.
[23] V. Haberle, A. Stark, Eukaryotic core promoters and the functional basis of transcription initiation, Nat. Rev. Mol. Cell Biol. 19 (2018) 621–637.
[24] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, et al., An atlas of active enhancers across human cell types and tissues, Nature. 507 (2014) 455–461.
[25] D.R. Zerbino, S.P. Wilder, N. Johnson, T. Juettemann, P.R. Flicek, The ensembl regulatory build, Genome Biol. 16 (2015) 56.
[26] E. Arner, C.O. Daub, K. Vitting-Seerup, R. Andersson, B. Lilje, F. Drabløs, et al., Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells, Science. 347 (2015) 1010–1014.
[27] C.C. Hon, J.A. Ramilowski, J. Harshbarger, N. Bertin, O.J. Rackham, J. Gough, et al., An atlas of human long non-coding RNAs with accurate 5′ ends, Nature. 543 (2017) 199–204.
[28] Y. Cheng, Z. Ma, B.H. Kim, W. Wu, P. Cayting, A.P. Boyle, et al., Principles of regulatory information conservation between mouse and human, Nature. 515 (2014) 371–375.
[29] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics. 12 (2011) 323.
[30] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics. 30 (2014) 923–930.
[31] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (2016) 525–527.
[32] The Gene Ontology Consortium, Expansion of the Gene Ontology knowledgebase and resources, Nucleic Acids Res. 45 (2017) D331-D8.
[33] N. Bertin, M. Mendez, A. Hasegawa, M. Lizio, I. Abugessaisa, J. Severin, et al., Linking FANTOM5 CAGE peaks to annotations with CAGEscan, Sci Data. 4 (2017), 170147.
[34] M. Seki, E. Katsumata, A. Suzuki, S. Sereewattanawoot, Y. Sakamoto, J. Mizushima-Sugano, et al., Evaluation and application of RNA-Seq by MinION, DNA Res. 26 (2019) 55–65.
[35] M.N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, et al., Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, Genes Dev. 25 (2011) 1915–1927.
[36] D. Canella, V. Praz, J.H. Reina, P. Cousin, N. Hernandez, Defining the RNA polymerase III transcriptome: genome-wide localization of the RNA polymerase III transcription machinery in human cells, Genome Res. 20 (2010) 710–721.
[37] S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.B. Fan, P. Lönnerberg, et al., Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing, Nat. Protoc. 7 (2012) 813–828.
[38] T. Kouno, J. Moody, A.T.-J. Kwon, Y. Shibayama, S. Kato, Y. Huang, et al., C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution, Nat. Commun. 10 (2019) 360.
[39] S. Oki, T. Ohta, G. Shioi, H. Hatanaka, O. Ogasawara, Y. Okuda, et al., ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data, EMBO Rep. 19 (2018).
[40] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics. 29 (2013) 15–21.

[41] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics. 26 (2010) 841–842.
[42] A. Hasegawa, C. Daub, P. Carninci, Y. Hayashizaki, T. Lassmann, MOIRAI: a compact workflow system for CAGE analysis, BMC Bioinformatics. 15 (2014) 144.
[43] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y.C. Lin, P. Laslo, et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, Mol. Cell 38 (2010) 576–589.
[44] B. Yates, B. Braschi, K.A. Gray, R.L. Seal, S. Tweedie, E.A. Bruford, Genenames.org: the HGNC and VGNC resources in 2017, Nucleic Acids Res. 45 (2017) D619-D25.
[45] C.L. Smith, J.A. Blake, J.A. Kadin, J.E. Richardson, C.J. Bult, Group MGD, Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse, Nucleic Acids Res. 46 (2018) D836-D42.
[46] J. Casper, A.S. Zweig, C. Villarreal, C. Tyner, M.L. Speir, K.R. Rosenbloom, et al., The UCSC Genome Browser database: 2018 update, Nucleic Acids Res. 46 (2018) D762-D9.
[47] T. UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 46 (2018) 2699.
[48] R.M. Kuhn, D. Haussler, W.J. Kent, The UCSC genome browser and associated tools, Brief. Bioinform. 14 (2013) 144–161.
[49] I. Abugessaisa, H. Shimoji, S. Sahin, A. Kondo, J. Harshbarger, M. Lizio, et al., FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki, Database (Oxford). 2016 (2016).