In the FANTOM5 project, transcription initiation events across the human and mouse genomes were mapped at a single base-pair resolution and their frequencies were monitored by CAGE (Cap Analysis of Gene Expression) coupled with single-molecule sequencing. Approximately three thousands of samples, consisting of a variety of primary cells, tissues, cell lines, and time series samples during cell activation and development, were subjected to a uniform pipeline of CAGE data production. The analysis pipeline started by measuring RNA extracts to assess their quality, and continued to CAGE library production by using a robotic or a manual workflow, single molecule sequencing, and computational processing to generate frequencies of transcription initiation. Resulting data represents the consequence of transcriptional regulation in each analyzed state of mammalian cells. Non-overlapping peaks over the CAGE profiles, approximately 200,000 and 150,000 peaks for the human and mouse genomes, were identified and annotated to provide precise location of known promoters as well as novel ones, and to quantify their activities.

Design Type(s) | organism part comparison design • species comparison design • cell type comparison design • organism development design
---|---
Measurement Type(s) | DNA-templated transcription, initiation
Technology Type(s) | cap analysis of gene expression
Factor Type(s) | Species • Organism Part • life cycle stage • cell type
Sample Characteristic(s) | Mus musculus • cerebellum • visual cortex • ileum • Peyer's patch • stomach • axillary lymph node • aorta • substantia nigra • hippocampal formation • brain • heart • liver • meningeal cluster • bone marrow • spinal cord • raphe nuclei • corpus striatum • cortex • peripheral nervous system • kidney • neural system • hemolymphoid system • blood • spleen • mesoderm • hematopoietic system • ventral wall of dorsal aorta • placenta • ganglion • spiral organ of cochlea • small intestine • intestine • adrenal gland • eyeball of camera-type eye • pituitary gland • thymus • lung • female gonad • testis • bone tissue • diencephalon • muscle organ • medulla oblongata • forelimb • pancreas • gonad • corpora quadrigemina • skin of body • tongue • colon • caecum • vesicular gland • epididymis • amnion • mammary gland • uterus • submandibular gland • prostate gland • intestinal mucosa • urinary bladder • vagina • oviduct • Homo sapiens

Correspondence and requests for materials should be addressed to H.K. (email: kawaji@gsc.riken.jp).
#A full list of authors and their affiliations appears at the end of the paper.
Background & Summary

Since the completion of the human genome sequencing, role of individual bases has been a central question. An international collaborative effort, FANTOM (Functional ANnoTation Of Mammalian Genome)\(^1\), delineated a complex landscape of transcribed RNAs (transcriptome) and their regulations. The initial key technology driving the project was to make full-length cDNA clones, representing complete primary structure of transcribed RNA molecules. Sequencing of the full-length cDNA clones uncovered unexpected number of long non-coding RNAs as well as protein coding genes\(^2\)–\(^6\). The CAGE (Cap Analysis Gene Expression)\(^7\)\(^8\) protocol, combination with high-throughput sequencing, was developed to monitor frequencies of transcription initiation by determining 5′-end of capped RNAs. The technology was devised to uncover complexity of the transcriptome\(^4\)–\(^6\) and elucidate transcriptional regulatory networks by focusing on promoter elements\(^9\)–\(^12\). By taking advantage of single molecule sequencer, HeliScopeCAGE was recently developed to provide more sensitive and accurate monitoring of transcription initiation activities\(^7\)\(^8\).

In the fifth round of the FANTOM projects, FANTOM5, the challenge was to capture the transcriptome of many varieties of cell states as possible, to understand the implication of each genomic bases in different contexts. In the first phase of the FANTOM5 project, we targeted cells in steady state, called ‘snapshot’ samples\(^13\). Our central focus was on human primary cells, while cell lines, tissues and mouse samples were chosen to cover cells inaccessible as isolated human primary samples. The resulting data provided an atlas of promoter and enhancer activities in wide range of cell states\(^14\), which is a baseline of understanding complex transcriptional regulation. In the second phase, we focused on transitions of cell states by monitoring ‘time course’ samples, such as activations, differentiations, and developments at sequential time points\(^15\). The monitored activities of promoters and enhancers demonstrated that enhancer activities is the earliest event during dynamic changes of transcriptome. These data sets are being utilized in many other studies inside and outside of the FANTOM5 consortium.

The data production scheme was implemented based on the FANTOM5 collaboration. Sample collection was performed at individual institutes, since specific types of samples require dedicated systems with special expertise or settings, as well as through purchase from commercial sources. RNA quality was firstly examined at the place where the samples were obtained (the first RNA quality check). The CAGE assay pipeline established in RIKEN GeNAS (Genome Network Analysis Support Facility) employed two workflows of HeliScopeCAGE, a manual workflow for samples with small amount of total RNAs\(^8\) and a robotic workflow for samples with standard requirements\(^7\). The assay pipeline started with checking RNA quality (the second RNA quality check), which provides a uniform quality assessment of the profiled RNA extracts. The resulting CAGE libraries were sequenced by HeliScope in RIKEN and also in Helicos Biosciences, and the obtained data were processed by the MOIRAI system\(^16\). Quality of the resulting CAGE profiles was checked with several statistics as well as manual inspection by using the ZENBU browser\(^17\). Finally CAGE profiles were shared among the consortium for further analysis.

In the course of the two phases focused on ‘snapshot’ and ‘time course’ samples, we profiled 1,816 human and 1,016 mouse samples in total, and obtained approximately four millions of single-molecule reads successfully aligned to the genome per sample on average. Based on frequencies of the observed 5′-ends of individual capped RNA molecules at a single base-pair resolution, we identified 201,802 and 158,966 peaks for human and mouse respectively, where promoters are defined as the sequence immediately upstream of the peaks and frequencies of observed CAGE reads reflect activities of the promoters. All data generated during the course of the project were deposited to a public repository (DDBJ Read Archive, DRA) and/or provided at the FANTOM5 web resource (http://fantom.gsc.riken.jp/5/)\(^18\). Here we describe the data with the processing details and quality metrics.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human</td>
<td>Mouse</td>
<td>Human</td>
</tr>
<tr>
<td>Cell lines</td>
<td>259</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Fractionations</td>
<td>12</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Primary cells</td>
<td>537</td>
<td>109</td>
<td>24</td>
</tr>
<tr>
<td>Timecourse samples</td>
<td>35</td>
<td>19</td>
<td>748</td>
</tr>
<tr>
<td>Tissues</td>
<td>150</td>
<td>237</td>
<td>33</td>
</tr>
<tr>
<td>Quality control samples</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>993</td>
<td>367</td>
<td>823</td>
</tr>
</tbody>
</table>

Table 1. Summary of FANTOM5 phase 1 and phase 2 samples.
Methods
Sample collection
Sample collection was performed as described previously.13,15 Briefly, primary cells were purchased as purified RNAs or frozen cells, or obtained as described previously through collaboration in the consortium. Purchased cells were cultured according to the manufacturer’s instructions and miRNeasy kit (QIAGEN) was used for RNA extraction. Human post mortem tissue RNAs were purchased or obtained through the Dutch Brain bank. Tissues collected through the consortium were snap-frozen in liquid nitrogen, transferred into Lysing Matrix D tubes (MP Biomedicals, Santa Ana, CA) containing chilled Trizol (Gibco), homogenized by FastPrep Homogenizer (Thermo Savant), and centrifuged. miRNeasy kit (QIAGEN) was used for RNA extraction from cultured cell lines as well as frozen cell line stocks.

For the purchased samples, lot or catalogue numbers were recorded where available. Of the collected RNAs, those with more than 1 μg, were measured by Agilent BioAnalyzer (Agilent Technologies, Santa Clara, CA) and Nanodrop spectrophotometer (Thermo Fisher Scientific, Wilmington, DE) to check RIN (RNA integrity) score and the absorbance ratio of A260/A230 and A260/A280. The rest of the samples were directly subjected to the CAGE library production to avoid wasting material. All 2,832 profiled samples are summarized in Table 1.

Single molecule CAGE and data processing
HeliScopeCAGE libraries were prepared, sequenced, and processed as described previously.13,15 Most of the RNAs were subjected to the automated HeliScopeCAGE protocol, except for RNAs with less than 1 μg that were subjected to the manual protocol optimized for low quantity RNAs. The resulting libraries were measured by OliGreen fluorescence assay kit (Life Technologies), and sequenced by following the manufacturer’s instructions (LB-016_01, LB-017_01, and LB-001_04 (ref. 13). RNAs extracted from mouse whole body embryo E17.5 (called internal control) were systematically subjected to this workflow, with one per a sequencing run.

The produced data were processed as previously described.13,15 Briefly, reads corresponding to ribosomal RNA were removed by using the program rRNAdust (http://fantom.gsc.riken.jp/5/supp/rRNAdust/), remaining reads were aligned to the reference genome of human and mouse (hg19 or mm9) by using Delve25, and alignments with a quality of less than 20 (<99% chance of true) or a sequence identity of less than 85% were discarded. Frequencies of the CAGE read 5′ ends were counted to give a unit of CAGE tag start site (CTSS), a single base-pair on the reference genome. The entire flow of the data is illustrated in Fig. 1, and the number of CAGE profiles (equivalent to CTSS files) is summarized in Table 2.

Identification of peaks and their annotations
Non-overlapping peaks based on the all CAGE profiles were identified by using DPI (decomposition-based peak identification, https://github.com/hkawaji/dpi1/) method and annotated as previously described.13,15 A ‘robust’ threshold, for which a peak must include a CTSS with more than 10 read counts and 1 TPM (tags per million) at least one sample, was employed to define a stringent subset of the CAGE peaks. The robust peaks were associated with known transcripts, such as RefSeq, UCSC known gene, GENCODE, Ensembl, and mRNAs (full-length cDNA clones), based on their 5′-end proximity to the peaks. Official gene symbols, Entrez Gene IDs, and protein (UniProt) IDs associated with the transcripts were retrieved and assigned as part of annotation. In addition to these associations, human readable names and descriptions were assigned to each of the CAGE peaks. Peaks were given a name in the form pN@GENE, where GENE indicates gene symbol or transcript name and N indicates the rank in the ranked list of promoter activities for that gene. For example, p1@SPI1 represent the peak with the highest number of observation (that is, read counts) in all of the FANTOM5 CAGE profiles, among the peaks associated with SPI1 gene.

Peak identification with the same method and the same threshold was performed two times; the first was for ‘snapshot’ samples (phase 1), and the second was for the entire samples from both the ‘snapshot’
and 'time course' studies (phase 2). We integrated these two peak sets into a hybrid set consisting of all the phase 1 peaks over the robust threshold and a subset of phase 2 peaks that did not overlap with the phase 1 peaks. Annotation of phase 1 peaks was used in the hybrid set, called phase 1+2 peaks, which provide a consistent reference in the definition of promoters.

Quantification of promoter activities
All the obtained CAGE profiles were subjected to the peak identification, even if they have some issues in quality, since all of them still represent independent observations of RNA 5′-ends. However promoter activities (that is, expression levels of CAGE peaks) were quantified only in the samples satisfying the following criteria: RIN score greater than 6, more than 500,000 successfully aligned reads to the genome, and more than 50% of the successful alignments are close to 5′-end of RefSeq gene model, for expression analysis requiring reliable quantification. After discarding a few CAGE profiles of low quality, read counts for individual CTSSs belonging to the same peak were summed up, normalization (or scaling) factors were calculated with RLE (Relative Log Expression)\textsuperscript{30} method by edgeR\textsuperscript{31}, and tags per million (that is, counts per million) was computed as expression levels.

The RLE normalization was first performed within the phase 1 samples. The naïve application of this to the entire data sets, consisting of phase 1 and phase 2 samples, might cause inconsistencies in expression levels between the two normalizations. To avoid this, we took the geometric mean of CAGE peak read counts across the phase 1 samples and used it as the reference expression for a normalization factor calculation in the same manner as RLE method. This enabled us to keep the expression levels of phase 1 as they were, and to adjust the expression levels of the phase 2 samples to be comparable\textsuperscript{15}.

Code availability
All software used in this study are publicly available. rRNAdust, for removing ribosomal RNA, is available at http://fantom.gsc.riken.jp/5/suppl/rRNAdust/. Mapping software Delve is available at http://fantom.gsc.riken.jp/5/suppl/delve/. The program to perform DPI, decomposition-based peak identification, method is available at https://github.com/hkawaji/dpi1/.

Data Records

Data record 1: Metadata
Two types of metadata are available at figshare and LSDB Archive (Data Citation 1, 10). One is for the samples, including their origins and extracted RNA. The other is for the CAGE assay, including the result of RNA quality check, library production, and post-processing of the CAGE tag sequences. Both of them are described in SDRF (Sample and Data Relationship Format)\textsuperscript{32}. Sample metadata for human and mouse are ‘HumanSamples2.0.sdrf.xlsx’ and ‘MouseSamples2.0.sdrf.xlsx’, respectively. The metadata for the CAGE assay are available as ‘sdrf.txt’.

Data record 2: CAGE profiles
All of the CAGE sequences, their alignment to the genomes, and CTSS frequencies are available at DDBJ DRA (DDBJ Sequence Read Archive) (Data Citations 2–9). The accession number of each file is summarized in ‘DRA*.txt’ at figshare (Data Citation 1).

Data record 3: CAGE peaks
Genomic coordinates, annotations and expressions of the CAGE peaks are available as “phase1and2-combined_coord.bed.gz”, “phase1and2combined_ann.txt.gz”, and “phase1and2combined_tpm.osc.txt.gz”
Technical Validation
RNA quality
Measured RNA qualities at the second check (that is, immediately before the CAGE library production) are shown in Fig. 2a–c. RNA Integrity Number (RIN) score, measured using an Agilent Bioanalyzer, was 8.96 on average (standard deviation 1.19), absorbance ratio of 260/230 nm (A260/A230) and 260/280 nm (A260/A280) were on average 2.01 (standard deviation 0.53) and 2.13 (standard deviation 0.14) respectively. These figures indicate that the majority of the RNAs were processed in good quality.

Mapped reads
The number of CAGE reads successfully aligned with the genome and the ratio of CAGE reads hitting conventional promoters are shown in Fig. 2d,e. The average number of mapped reads is 4,208,291 per CAGE profile. Of the 2,522 profiles, 98.3% (2,478) consists of at least 500,000 successfully aligned reads, which was a criterion of profiles used for expression analysis13. The average ratio of promoter-hitting reads is 76.5, and 98.6% of all profiles (2,437/2,472) have more than 50% promoter-hitting rate, which was another criterion of profiles used for expression analysis13.

Sample identity
Hierarchical clustering of the 126 mouse primary cells13 within the phase 1 was shown in Fig. 3, and the same clustering of the 571 human primary cells13 was in Supplementary Fig. 1. The average linkage method was applied to log-scale expression (TPM) profiles at promoter-level, and sample identities were assessed by expression of marker genes and also by manual inspection of the hierarchical clustering. The figures show that majority of biological replicates belonged to the same branch of the tree, that is, the same cluster, except for samples with a low number of mapped read counts.

Figure 2. RNA and mapping quality control. Distribution of RIN score (a), A260/A230 (b), A260/A280 (c), mapped reads (d), and promoter rate (e) for samples used for FANTOM5 expression analysis.
Figure 3. Hierarchical clustering of primary cells. Hierarchical clustering of primary cell samples of mouse based on logarithm of expression (TPM). Color shows anatomical categories of samples.
Usage Notes
As well as providing access to individual data files, we also set up a series of interfaces as described in the FANTOM web resource18,33. TET (Table Extraction Tool) provides an interface to obtain a subset of data by specifying the desired columns and rows. The BioMart interface33, and FANTOM5 SSTAR (Semantic catalog of Samples, Transcription initiation And Regulators) provides the metadata of the profiled samples33. The CAGE profile on the genomic axis is visible in ZENBU17 with its interactive interface and also in the UCSC genome browser36 via track data hub37.

References

Data Citations
2. DDBJ Sequence Read Archive DRA000991 (2013).
3. DDBJ Sequence Read Archive DRA001026 (2013).
4. DDBJ Sequence Read Archive DRA001027 (2013).
5. DDBJ Sequence Read Archive DRA001028 (2013).
6. DDBJ Sequence Read Archive DRA002216 (2014).
7. DDBJ Sequence Read Archive DRA002711 (2014).
8. DDBJ Sequence Read Archive DRA002747 (2014).
9. DDBJ Sequence Read Archive DRA002748 (2014).

Acknowledgements
FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Y.H. and a grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to Y.H. It was also supported by Research Grants for RIKEN Preventive Medicine and Diagnosis Innovation Program to Y.H. and RIKEN Centre for Life Science Technologies, Division of Genomic Technologies (from the MEXT, Japan).

Author Contributions

Additional Information
Supplementary Information accompanies this paper at http://www.nature.com/sdata

Competing interests: The authors declare no competing financial interests.


Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2017