

# SCIENTIFIC DATA

## OPEN Data Descriptor: Transcription start site profiling of 15 anatomical regions of the *Macaca mulatta* central nervous system

Received: 6 December 2016

Accepted: 29 August 2017

Published: 31 October 2017

Updated: 11 December 2018

Margherita Francescato<sup>1,\*</sup>, Marina Lizio<sup>2,3,\*</sup>, Ingrid Philippens<sup>4</sup>, Luba M. Pardo<sup>5</sup>, Ronald Bontrop<sup>4</sup>, Mizuho Sakai<sup>2,3</sup>, Shoko Watanabe<sup>2,3</sup>, Masayoshi Itoh<sup>2,3,6</sup>, Akira Hasegawa<sup>2,3</sup>, Timo Lassmann<sup>3,7</sup>, Jessica Severin<sup>2,3</sup>, Jayson Harshbarger<sup>2,3</sup>, Imad Abugessaisa<sup>2</sup>, Takeya Kasukawa<sup>2</sup>, Piero Carninci<sup>2,3</sup>, Yoshihide Hayashizaki<sup>3,6</sup>, Alistair R. R. Forrest<sup>3,8</sup>, Hideya Kawaji<sup>2,3,6,9</sup>, Patrizia Rizzu<sup>10</sup> & Peter Heutink<sup>10</sup>

Rhesus macaque was the second non-human primate whose genome has been fully sequenced and is one of the most used model organisms to study human biology and disease, thanks to the close evolutionary relationship between the two species. But compared to human, where several previously unknown RNAs have been uncovered, the macaque transcriptome is less studied. Publicly available RNA expression resources for macaque are limited, even for brain, which is highly relevant to study human cognitive abilities. In an effort to complement those resources, FANTOM5 profiled 15 distinct anatomical regions of the aged macaque central nervous system using Cap Analysis of Gene Expression, a high-resolution, annotation-independent technology that allows monitoring of transcription initiation events with high accuracy. We identified 25,869 CAGE peaks, representing bona fide promoters. For each peak we provide detailed annotation, expanding the landscape of 'known' macaque genes, and we show concrete examples on how to use the resulting data. We believe this data represents a useful resource to understand the central nervous system in macaque.

<sup>1</sup>Italian Institute of Technology, Department of Neuroscience and Brain Technologies, Via Morego 30, Genova 16163, Italy. <sup>2</sup>RIKEN Center for Life Science Technologies, Division of Genomic Technologies, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. <sup>3</sup>RIKEN Yokohama Institute, Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. <sup>4</sup>Biomedical Primate Research Centre, Postbox 3306, Rijswijk 2280 GH, The Netherlands. <sup>5</sup>Department of Dermatology, Erasmus MC Cancer Institute, Burg. s' Jacobplein 51 3015 CA Rotterdam, The Netherlands. <sup>6</sup>RIKEN Preventive Medicine and Diagnosis Innovation Program, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. <sup>7</sup>Telethon Kids Institute, The University of Western Australia, 100 Roberts Road, Subiaco, Western Australia 6008, Australia. <sup>8</sup>Harry Perkins Institute of Medical Research, 6 Verdun St, Nedlands, Western Australia 6009, Australia. <sup>9</sup>RIKEN Advanced Center for Computing and Communication, Preventive Medicine and Applied Genomics Unit, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan. <sup>10</sup>German Center for Neurodegenerative Diseases, Otfried-Müller Straße 23, Tübingen 72076, Germany. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to H.K.(email: kawaji@gsc.riken.jp).

<b>Design Type(s)</b>	species comparison design • organism part comparison design • transcription start site identification objective
<b>Measurement Type(s)</b>	DNA-templated transcription, initiation
<b>Technology Type(s)</b>	cap analysis of gene expression
<b>Factor Type(s)</b>	animal body part
<b>Sample Characteristic(s)</b>	<i>Macaca mulatta</i> • amygdala • caudate nucleus • cerebellum • globus pallidus • hippocampal formation • locus ceruleus • middle frontal gyrus • middle temporal gyrus • medulla oblongata • occipital gyrus • parietal lobe • putamen • spinal cord • substantia nigra • dorsal thalamus

## Background and Summary

*Macaca mulatta*, also known as rhesus macaque and named hereafter macaque for simplicity, is a small non-human primate whose natural habitat includes a large portion of India, Pakistan and China<sup>1</sup>. Because of its small size and remarkable adaptability, and its high evolutionary relatedness to humans, the macaque is as an important model organism in biomedical sciences<sup>2</sup> as mouse. After chimpanzee, it was the second non-human primate whose genome has been fully sequenced. A first macaque assembly was released in 2007, and comparisons with human and chimpanzee genomes gave initial insight into primates evolutionary relationships<sup>3</sup>. In 2015 the Baylor College of Medicine submitted a novel assembly<sup>4</sup>, doubling the contigs size thanks to the assembling of longer fragments, and fixing several assembly and annotation errors of the previous version.

Despite an improved resolution at the genome level, the macaque transcriptome has not benefited of the same level of analyses performed in human, where noncoding RNAs (including enhancer-derived RNAs), alternative transcription initiation, and RNA isoforms were uncovered<sup>5–7</sup>. While it is particularly relevant to study all these newly discovered transcript types in the context of evolution, the current annotation in macaque does not allow us to examine this aspect in depth. For example, for the latest rheMac8 genome build there are currently 6,460 Reference Sequence (RefSeq) annotated genes, against the 66,466 available for the recently released human hg38 build available from UCSC Genome Browser<sup>8</sup>. Similarly, the latest Ensembl gene build for macaque, although more comprehensive, consists of 56,748 transcript models in contrast to the 199,234 ones for human. Even though predictions from a variety of algorithms can provide an initial assessment of the genes present in a given organism, observation of RNA molecules is necessary to validate the predictions. Moreover, the expression of genes is often specific to cells or tissues<sup>9</sup>, and alternative promoter usage associated to distinct transcript isoforms has an impact on the overall expression and its regulation<sup>10,11</sup>, which can be relevant in diseases, including those of the brain.

The FANTOM (functional annotation of the mammalian genome) consortium is an international collaboration established in Japan in the year 2000 aiming to unveil all the functional elements in mammalian genomes paired with reliable annotations<sup>12–14</sup>. FANTOM5, the fifth iteration of the collaboration, revealed the most comprehensive promoter and enhancer atlases for human and mouse by using Cap Analysis of Gene Expression (CAGE) combined with single molecule sequencing<sup>5,7,15,16</sup>. The CAGE technique is a well-established expression profiling approach that allows to identify novel or previously unannotated/mis-annotated gene starts by sequencing 5'-end of RNAs transcribed by RNA polymerase II, as well as to quantify their expression levels<sup>17</sup>. Thanks to the high precision of CAGE signal<sup>18</sup> it is therefore possible to characterize transcription start sites (TSS) and promoter regions of genes.

Here we present CAGE-based expression profiles for 15 regions of the macaque central nervous system (CNS), obtained within the context of FANTOM5 (Table 1). A flow of the analysis is visualized in Fig. 1. Overall, we identify 25,869 CAGE peaks, representing bona fide transcription start sites (TSS) of genes expressed in the macaque CNS. Considering the gene models available for the rheMac8 assembly only (Table 2), 71.2% of the peaks is located within 500 bp of an annotated TSS. Overall, 80.4% of the peaks are associated to a gene when including gene models from human lifted over to macaque (also listed in Table 2). We also investigate sequence features around macaque TSSs. We observe that 14,877 CAGE peaks (57.5%) overlap a CpG island while 1,219 of them (4.7%) harbor a TATA-box site within a 70 bp region around their 5'-end. We also find that 3,471 peaks (13.4%) overlap a repetitive element, with simple repeats and LINE/L1 being the most represented classes. We perform several analyses to confirm that the expression profiles we obtained are of high quality. First and foremost, we observe a generally high correlation between samples, despite the fact that the libraries represent distinct anatomical brain regions. In addition, our expression profiles show agreement with results from Peng *et al.*<sup>19</sup> and Bakken *et al.*<sup>20</sup>.

This is, to date, the most detailed CNS dataset on macaque obtained using an annotation-independent transcription profiling technique consisting of experimentally defined promoters with detailed annotation

Library ID	Species	Tissue	Abbreviation	RIN	Mapped reads	Mapping rate (%)
CNhs14182	Rhesus macaque	Amygdala	Amy	9	4,565,587	50.9
CNhs14174	Rhesus macaque	Caudate	Cau	8.4	3,348,210	44.68
CNhs14176	Rhesus macaque	Cerebellum	Cer	9.2	5,682,996	65.99
CNhs14171	Rhesus macaque	Globus Pallidus	GIP	8.1	4,486,692	50.57
CNhs14180	Rhesus macaque	Hippocampus	Hip	8.7	3,017,562	41.6
CNhs14172	Rhesus macaque	Locus Coeruleus	LocC	8.4	4,140,956	49.5
CNhs14177	Rhesus macaque	Medial Frontal Gyrus	MFG	8.1	5,576,066	65.76
CNhs14179	Rhesus macaque	Medial Temporal Gyrus	MTG	8.2	3,968,759	47.6
CNhs14175	Rhesus macaque	Medulla Oblongata	MedO	8.2	5,407,111	61
CNhs14169	Rhesus macaque	Occipital Gyrus	Occ	8.6	5,269,914	61.3
CNhs14181	Rhesus macaque	Parietal Gyrus	Par	8.6	3,883,216	48.8
CNhs14178	Rhesus macaque	Putamen	Put	8.1	4,493,463	52.28
CNhs14168	Rhesus macaque	Spinal Cord	SpC	7.7	4,282,060	51.33
CNhs14173	Rhesus macaque	Substantia Nigra	SubN	7.7	3,167,510	40.84
CNhs14170	Rhesus macaque	Thalamus	Thal	8.4	6,211,376	71.69
CNhs12311	Homo sapiens	Amygdala	Amy	6.9	14,101,506	41.5
CNhs12321	Homo sapiens	Caudate	Cau	7.2	15,501,176	44.06
CNhs12323	Homo sapiens	Cerebellum	Cer	6.9	17,383,592	44.66
CNhs12319	Homo sapiens	Globus Pallidus	GIP	6.8	14,100,137	40.64
CNhs12312	Homo sapiens	Hippocampus	Hip	6.7	15,910,618	41.21
CNhs12322	Homo sapiens	Locus Coeruleus	LocC	7.2	14,226,392	40.88
CNhs12310	Homo sapiens	Medial Frontal Gyrus	MFG	7.2	16,201,206	43.34
CNhs12316	Homo sapiens	Medial Temporal Gyrus	MTG	7	16,329,604	40.89
CNhs12315	Homo sapiens	Medulla Oblongata	MedO	7.5	15,199,808	43.08
CNhs12320	Homo sapiens	Occipital Gyrus	Occ	8.4	15,966,482	39.88
CNhs12317	Homo sapiens	Parietal Gyrus	Par	7.1	10,980,255	33.08
CNhs13912	Homo sapiens	Putamen	Put	8.3	2,262,558	27.58
CNhs12227	Homo sapiens	Spinal Cord	SpC	7.3	5,957,420	39.95
CNhs12318	Homo sapiens	Substantia Nigra	SubN	8.1	12,393,047	41.45
CNhs12314	Homo sapiens	Thalamus	Thal	7.3	18,287,224	43.51

**Table 1.** List of the macaque and matching human samples (15 for each species) included in this study, with corresponding library ID, tissue name and basic QC values (RIN, total number of mapped reads and mapping rate). We additionally report in this table the abbreviations used in the figures for each CNS anatomical region.

and expression values. Given the availability of ‘matching’ CAGE data from human (*Homo sapiens*) in FANTOM5 (ref. 7), these data can be useful to understand CNS functions and complexity from an evolutionary perspective, rather than representing a general mapping of transcriptome in macaque. Both the macaque and the human CAGE profiles were obtained from samples extracted from aged donors and could be used to study processes in the aged primate brain or late onset brain disorders, including neurodegenerative diseases. In addition, we provide concrete examples on how to use the data presented.

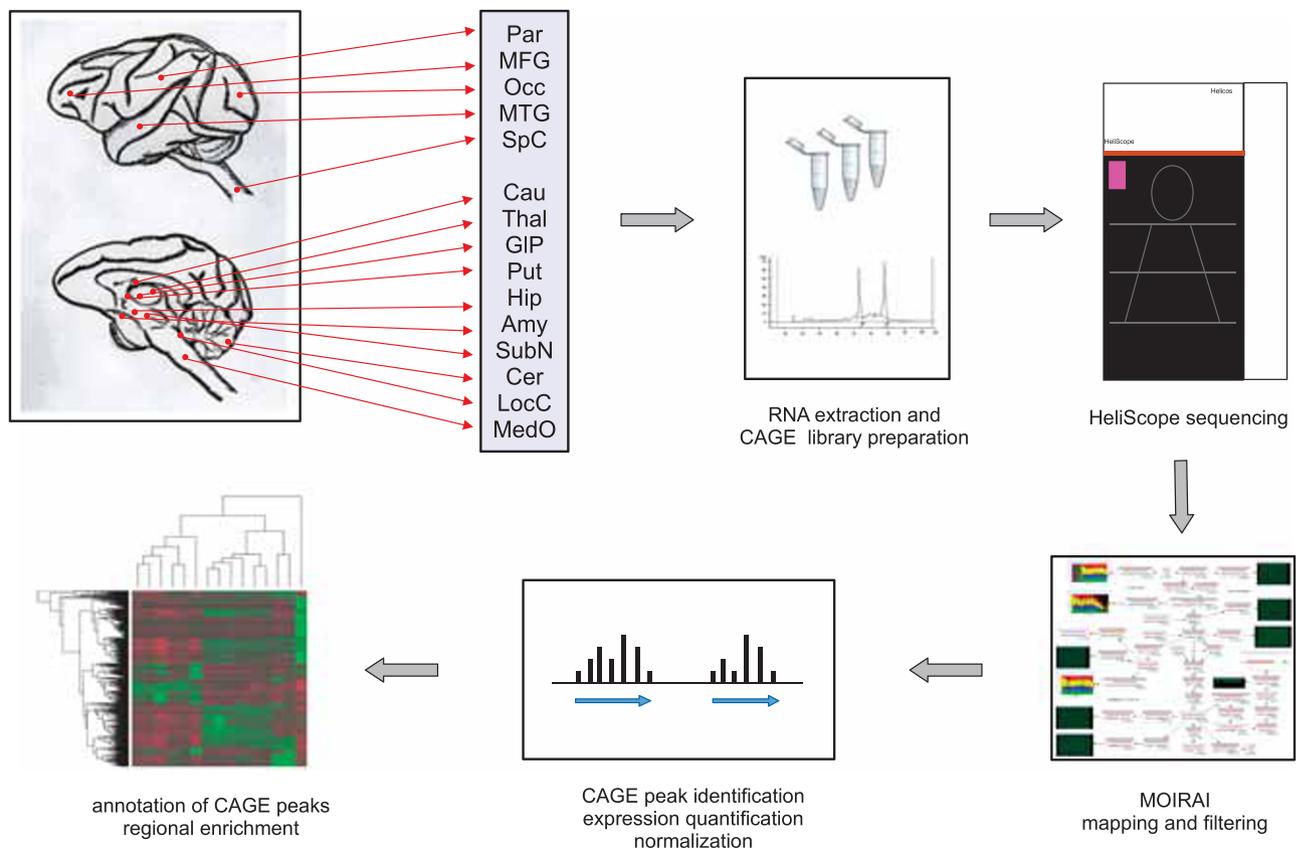
## Methods

### Sample collection and RNA extraction

The dissection of all regions was performed from fresh brain sections. Tissue was stored at  $-80^{\circ}\text{C}$  until further processing. Total RNA was extracted from frozen tissues using the Trizol tissue kit according to instructions provided by the manufacturer (Invitrogen) and further purified with RNeasy mini kit (Qiagen). RNA quality was assessed using the RNA Integrity Number (RIN) with the Agilent RNA Nano kit using the Agilent Bioanalyzer. All the samples used, their RIN values and library IDs are listed in Table 1.

### Library preparation and sequencing

For each sample one CAGE library was prepared using a protocol adapted for single molecule sequencing as previously described<sup>15</sup>. The libraries were subsequently sequenced on HeliScope sequencer<sup>16</sup> following manufacturer’s instructions.



**Figure 1. Experimental design of this study.** The figure shows a schematic workflow from sampling of CNS anatomical regions to subsequent sample and data processing. The full sample names corresponding to the abbreviations used in the figure are reported in Table 1.

### Mapping and data processing

Sequenced libraries were first filtered for reads aligning to ribosomal DNA (with up to two mismatches) by rRNA dust<sup>7</sup>. The remaining reads were aligned to rheMac8 reference genome using the probabilistic mapping tool delve (see Code Availability).

Data mapping workflow was implemented in Moirai, a system that allows creation and customization of sequencing data processing pipelines<sup>21</sup>.

Expression quantification was performed as previously described<sup>7</sup>. In short, 1-base pair TSSs were first obtained from genome alignments and then the number of aligned reads to the same 5' position was used to quantify their expression levels. We further normalized the values, as done previously<sup>7</sup>, in order to account for sample specific variations and library sizes. We kept only those reads with 99% alignment accuracy (mapping quality  $q > 20$ ) and sequence identity above 85% for downstream analyses. Mapped reads and single base frequencies of transcription initiation activities (TSS) are available via DDBJ (see Data Records, Data Citation 1).

### Identification of CAGE peaks

CAGE peaks, which represent promoters of active genes, were defined using the Decomposition Peak Identification (DPI) method, previously described<sup>7</sup> (see Code Availability). Briefly, the method performs decomposition of the CAGE signal across a set of samples and subsequently clusters the signal at TSSs, by considering a minimum distance and expression level thresholds. For each peak, normalized tags per million (TPM) were computed using the relative log expression method implemented in edgeR<sup>22</sup>.

### Comparison with publicly available annotation datasets

Peaks were annotated using all gene models available for download from UCSC genome Browser<sup>8</sup> for the rheMac8 genome build, such as Augustus<sup>23</sup> and Genescan<sup>24</sup> predictions, Ensembl genes<sup>25</sup> and experimentally validated RefSeq<sup>26</sup> and Expressed Sequence Tags (ESTs). To increase the coverage of the annotated peaks, we additionally considered human RefSeq and GENCODEv24 (ref. 27) (comprehensive annotation) models, also downloaded from UCSC genome browser. Human RefSeq and GENCODE models were lifted from hg38 to rheMac8 using UCSC liftOver tool (available from <http://hgdownload.soe.ucsc.edu/admin/exe/>), with default parameters. We considered a CAGE peak to be

rheMac8 gene model	Number of peaks	Percentage of peaks
Augustus	11,435	44.2
EST	6,514	25.2
Genscan	5,953	23
refGene	6,248	24.1
Ensembl Gene	14,376	55.6
hg38 refGene liftOver	15,562	60.2
hg38 GENCODE liftOver	18,635	72

**Table 2.** Number of CAGE peaks within 500 bp of at least one annotated gene model.

associated to a given gene model if it was located within a (−500, +500) region centered at the feature start, as in previous studies<sup>7</sup>. Annotation files were created using a combination of ‘intersectBed’ and ‘groupBy’ commands from the bedTools suite<sup>28</sup> (version 2.17.0). The total numbers of CAGE peaks annotated for each model are provided in Table 2 while complete annotation information for each peak is reported in Data Citation 2.

### Identification of region specific CAGE peaks

To identify CAGE peaks enriched in any given region we calculated relative expressions for all peaks. Given a CAGE peak P, its relative expression in a region was calculated as  $P_{rel} = \log_{10}((1+P_{region})/(1+P_{median}))$ , with  $P_{region}$  = normalized expression of peak P in the current region and  $P_{median}$  = median normalized expression of P across all regions. For each peak a relative expression above 1 indicates that it is preferentially expressed in that region. The full list of regionally enriched macaque CAGE peaks is provided in Data Citation 3, along with corresponding gene symbols (based on macaque RefSeq, macaque Ensembl and human RefSeq genes lifted over to macaque).

### Comparison with published human CAGE expression profiles

The FANTOM5 human promoter set includes normalized expression data for all 15 corresponding macaque brain regions (Data Citations 4 and 5). We extracted CAGE peaks coordinates and their expression for the matching samples from the human set, all derived from the same donor, and used them as a reference to compare sequence composition and overlap with repetitive elements, general expression profiles and regionally enriched peaks.

The dataset consists of 74,649 robust CAGE peaks, after filtering out very lowly expressed peaks, for expression comparison. Specifically, we required a CAGE peak to be expressed above 1 TPM in at least one sample. The samples and their relative quality metrics are provided in Table 1. We point out that, compared to the CAGE peaks set in macaque, the corresponding human set is larger, as the identification of peaks in that case was performed on ~1,000 samples across multiple cell and tissue types. At a 10 TPM threshold, the number of CAGE peaks in both species is comparable (~1.4 fold difference), with 14,437 peaks for macaque and 20,762 for human.

### Additional processing

Overlaps with repeat elements and CpG islands downloaded from UCSC for both macaque (rheMac8) and human (hg19), were computed using bedtools<sup>28</sup>. Enrichment of TATA binding protein (TBP) motifs across the CAGE peaks was obtained using HOMER motif discovery software<sup>29</sup> (see Code availability). In particular, the function ‘annotatePeaks.pl’ was employed to identify the CAGE peaks harboring a TATA-box motif in a region (−500, +200) around the peak. The PWM used for the TATA-box enrichment analysis corresponds to the POL012.1 motif from Jaspar<sup>30</sup>, the default HOMER motif database. The annotation function calculates the distance of a motif from the center of the target sequences. This needs to be taken into account when interpreting the resulting enrichments.

Correlations, normalization of expression values, enriched expression calculation and plots were performed using R version 3.1.2 (<http://www.R-project.org/>).

### Ethics statement

The study has been approved by the ethical committee (DEC) of the Biomedical Primate Research Centre, Lange Kleiweg 139, 2288 GJ Rijswijk, The Netherlands. Study DEC No. 606.

### Code availability

Mapping was performed using delve, a probabilistic aligner specifically developed for handling data derived from HeliScope single molecule sequencing. The software is available at [fantom.gsc.riken.jp/5/suppl/delve/](http://fantom.gsc.riken.jp/5/suppl/delve/).

The DPI peak calling was performed using our own software. It can be freely downloaded at this URL (<https://github.com/hkawaji/dpi1/>).

HOMER motif analysis suite is available at <http://homer.salk.edu/homer/>.

## Data Records

The data resulting from sequencing (fastq format), mapping (bam format) and TSS profiling (bed format) of macaque samples is deposited in the DNA Data Bank of Japan (Data Citation 1). The data is also available for direct download from FANTOM5 data repository (<http://fantom.gsc.riken.jp/5/datafiles/phase2.2/>). The data resulting from sequencing (fastq format), mapping (bam format) and TSS profiling (bed format) of human samples is deposited in the DNA Data Bank of Japan (Data Citations 4 and 5). Macrodissection microarray normalized expression profiles from multiple regions of the macaque CNS were downloaded from the NIH Blueprint Non-Human Primate (NHP) Atlas (<http://www.blueprintnhpatlas.org/static/download>). Normalized RNAseq expression profiles from multiple tissues and multiple non-human primate species were downloaded from Nonhuman Primate Reference Transcriptome Resource (<http://nhprtr.org/phase2.html>).

## Technical Validation

### RNA quality and mapped reads

To assess the quality of the starting material, a RIN (RNA Integrity Number) score was determined for each of the samples (Table 1). The RIN is a standard measurement evaluating the integrity of the RNA quality in a sample. Low RIN values indicate higher levels of degradation, suggesting that the sample should not be used to prepare CAGE libraries. The RIN numbers of our samples are all above 7.7, satisfying the minimum threshold recommended for library preparation in case of CAGE libraries (RIN = 7)<sup>31</sup>. We report, in addition, the number of mapped tags per library, which varies from 3.0 to 6.2 millions, with an average of 4.5. The total number of mapped tags can be influenced by many factors, such as RNA degradation, or low initial RNA amount, other than completeness of a genome assembly. As done previously<sup>7</sup>, we required a minimum of 0.5 M mapped tags per sample. All 15 samples were well above this threshold, with a median of 4.5 million mapped reads and median mapping rate of 50.9% (Table 1). To verify the coverage of the currently available genome by CAGE, we measured the mapping rate, defined as the ratio between all single mapped reads and all filtered extracted reads (that is, sequences remaining after artifacts and ribosomal DNA removal). We observe that the average mapping rate for macaque libraries is 53.6%, higher than the average obtained for the matching human libraries (40.4%, Table 1).

### Correlation between expression profiles across libraries

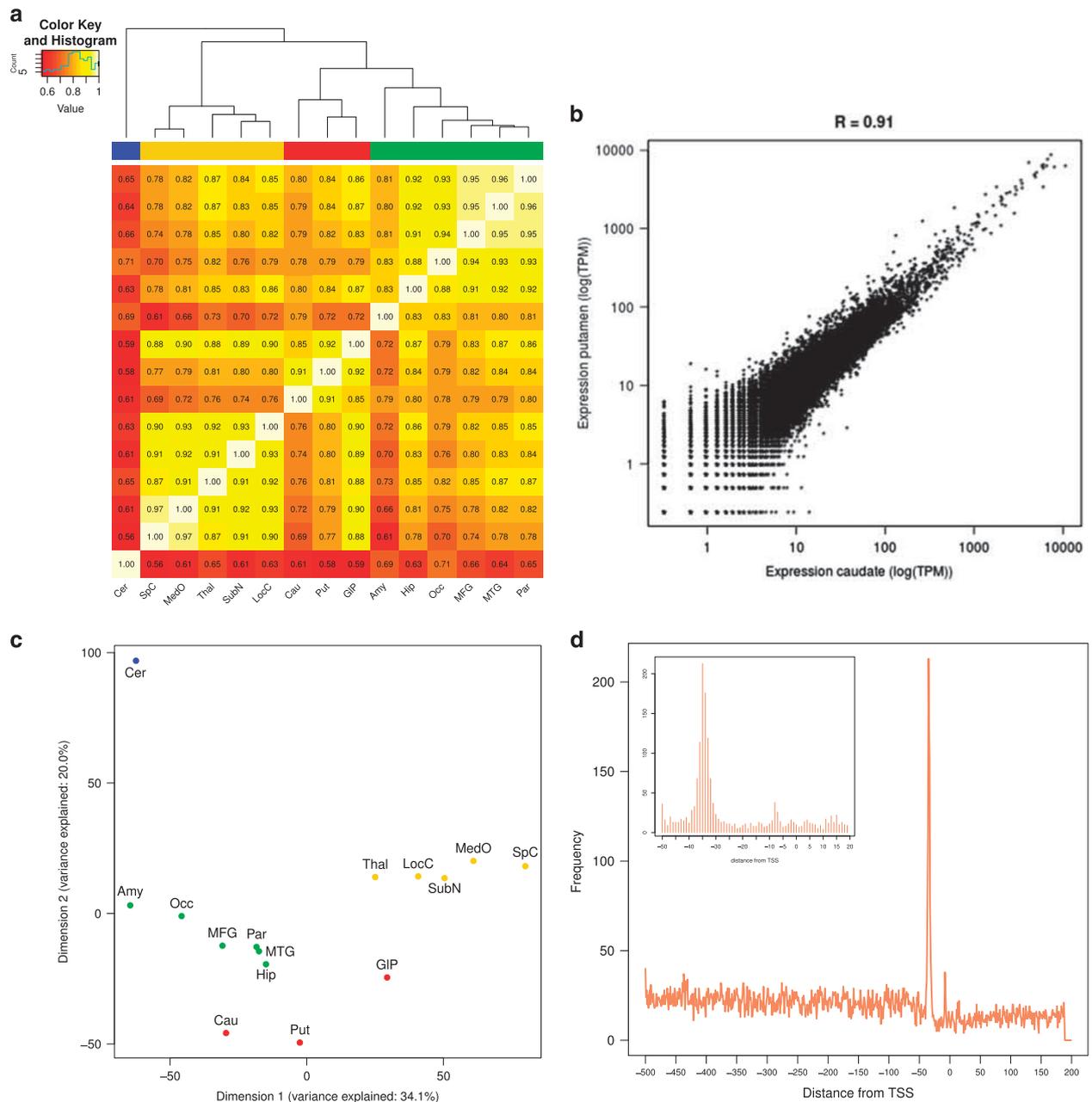
We calculated the Spearman correlation between the normalized expression profiles identified (represented as a heatmap in Fig. 2a). As reflected by the clustering, indicated by the column dendrogram, values are particularly high between structurally and functionally related regions of the CNS. For instance, the correlation between caudate and putamen (Fig. 2b), which share a similar neuronal composition together with the dorsal striatum, is 0.91. Correlation is 0.93 between substantia nigra and locus coeruleus and 0.94 between cortical samples (frontal, temporal, parietal and occipital gyri). Similarly, high correlation values between structurally and functionally related regions are observed across human expression profiles (Table 3).

### Multidimensional scaling of expression profiles reflects functionally related regions

In Fig. 2c we show a multi-dimensional scaling (MDS) representation of the samples described in this study. A clear separation of the cerebellum from all the other samples can be observed, as already suggested by correlation analysis, and previously reported in human as well<sup>32</sup>. This likely reflects both the particular cellular composition, which is distinctive with respect to other CNS districts<sup>33</sup>, and the rapid evolution of cerebellum in great apes and humans<sup>34</sup>. It is possible to distinguish three additional clusters of data points (color-coded in Fig. 2c). A first one (depicted in green) includes the cortex samples plus amygdala and hippocampus. The cortical regions make the external layer of the brain, and share functions related to the individual's awareness (cognition and sensory), so their clustering together may reflect that. The presence of hippocampus and amygdala in this group can be also explained by functional and developmental relationships (all evolved from the primitive brain). A second cluster, depicted in red, consists of globus pallidus, caudate and putamen. The latter two are both part of the striatum and are involved in motor and learning activities. The remaining cluster, in gold, is more heterogeneous and includes samples anatomically belonging to the brain stem and basal ganglia. We also note the position of the globus pallidus, midway between the red and the gold clusters, perhaps reflecting direct contact (physically and functionally) with putamen and at the same time direct targeting of substantia nigra<sup>35</sup>.

### Promoter features and sequence composition

We investigated sequence composition around the TSSs by calculating the number of CAGE peaks identified in macaque that are located within a CpG island. Since CAGE peaks precisely locate TSSs of actively transcribed genes, we treat the terms CAGE peaks and TSSs as synonyms. CpG islands are computationally predicted regions of high GC di-nucleotides frequency that tend to be found at promoters<sup>36</sup>. For comparison, we performed the same analysis on CAGE peaks defined for a matching set of CNS regions profiled in human with the same technique (see Methods). We observe an overall higher proportion of CAGE peaks overlapping CpG islands in macaque with respect to human (57.5 versus 48.8%), even though the CpG island density in the two species is very close<sup>37</sup> (Table 4). Next, we screened



**Figure 2. Promoter features and quality assessment of CAGE data.** (a) Graphical representation of the Spearman correlation values between all 15 macaque expression profiles presented in this study. The top dendrogram shows that the samples cluster in four major groups, as highlighted by the colored bar above the heatmap. (b) Scatterplot and Spearman correlation value for a pair of regions showing a high degree of similarity (caudate and putamen). Axes represent log-transformed TPM expression. (c) Multi-dimensional scaling representation of the samples included in the study, color-coded to show the clustering of samples (colors corresponding to those in a). (d) Frequency profile showing TATA-box enrichment upstream of the promoter. Insert shows a zoom-in view of the region (-50, +20), with a clear enrichment around 35 bases upstream of the TSS.

the CAGE peaks in macaque and in the corresponding human CNS samples for the presence of a TBP motif<sup>38</sup>. We observed clear enrichment for the TATA-box motif at around 35 bases upstream of the macaque TSS, in agreement with previous reports on human promoters<sup>13</sup> (Fig. 2d). We note that the distances reported by HOMER (see Methods) are calculated between the center of the TSS and the center of the motif; we also note that the median length of the peaks with a TATA motif is 13 bp, therefore our findings are consistent with what is known to be the preferred TATA-box position (between -28 and -31

	Amy	Cau	Cer	GIP	Hip	LocC	MFG	MTG	MedO	Occ	Par	Put	SpC	SubN	Thal
Amy	1	0.87	0.72	0.81	0.92	0.83	0.9	0.9	0.79	0.88	0.88	0.65	0.74	0.73	0.83
Cau		1	0.69	0.79	0.85	0.8	0.84	0.83	0.76	0.81	0.82	0.69	0.7	0.68	0.82
Cer			1	0.67	0.72	0.7	0.73	0.73	0.73	0.73	0.72	0.52	0.64	0.62	0.68
GIP				1	0.84	0.89	0.8	0.81	0.86	0.83	0.77	0.65	0.85	0.9	0.93
Hip					1	0.86	0.89	0.9	0.83	0.89	0.88	0.64	0.77	0.76	0.86
LocC						1	0.82	0.82	0.89	0.83	0.8	0.63	0.84	0.86	0.89
MFG							1	0.94	0.78	0.92	0.93	0.64	0.72	0.71	0.82
MTG								1	0.78	0.93	0.93	0.64	0.72	0.71	0.83
MedO									1	0.79	0.76	0.59	0.85	0.85	0.87
Occ										1	0.91	0.64	0.75	0.75	0.84
Par											1	0.63	0.7	0.68	0.8
Put												1	0.58	0.58	0.66
SpC													1	0.86	0.84
SubN														1	0.89
Thal															1

**Table 3.** Spearman correlation across human expression profiles derived from a matching set of CNS regions.

	CpG+,TATA+	CpG+,TATA-	CpG-,TATA+	CpG-,TATA-
macaque	2.1	55.4	2.7	39.8
human	1.5	47.3	2.7	48.6

**Table 4.** Proportions (as %) of CAGE peaks associated to TATA-box and CpG islands in macaque and human. For TATA, a region 70 bp around the TSS was used.

bp)<sup>13</sup>. The percentage of promoter regions containing a TATA motif within a (-50, +20) window is very similar between the human and macaque species (Table 4).

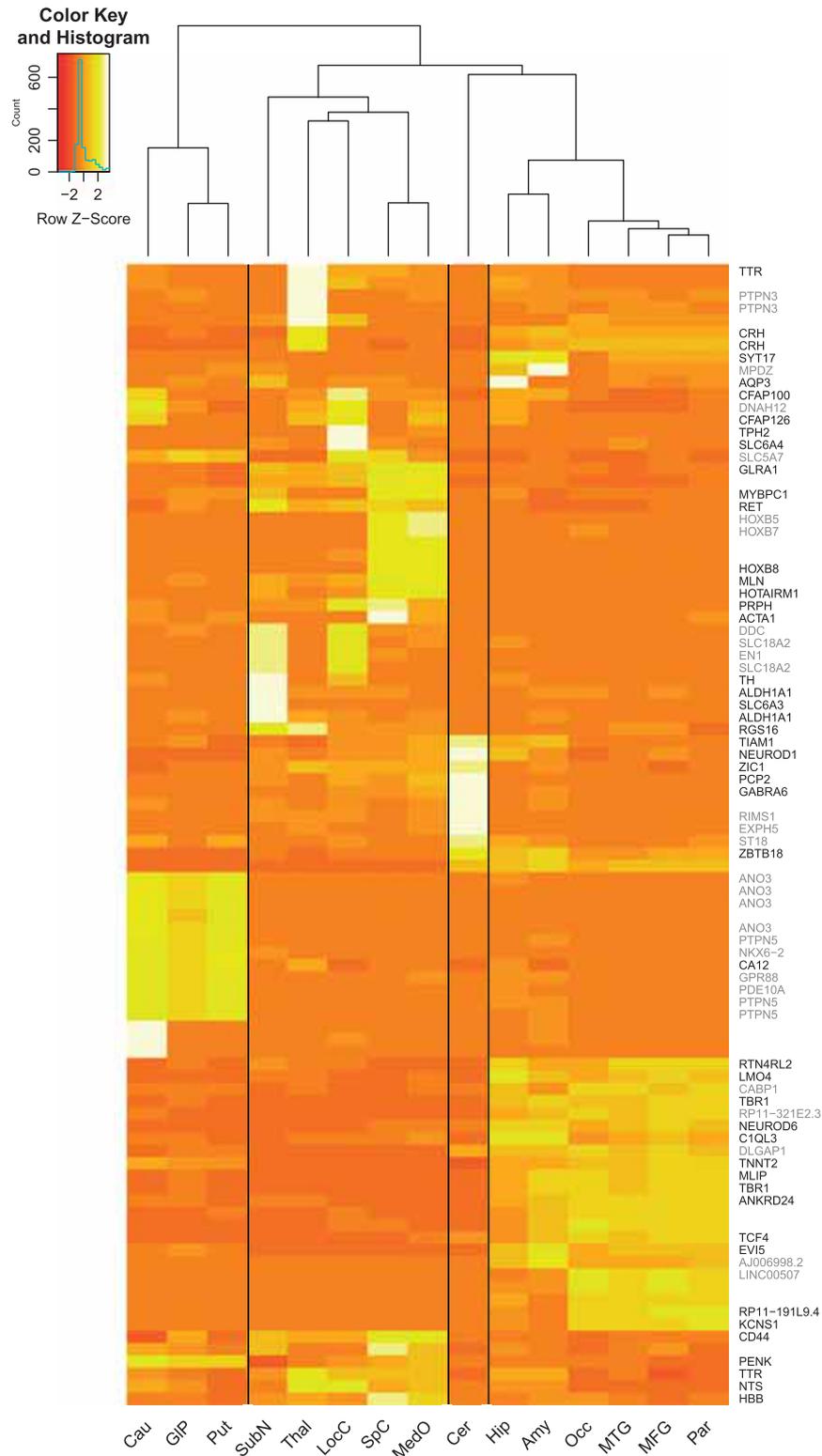
Since it was described in both human and mouse that an important portion of CAGE-defined TSS are located within repetitive elements<sup>39</sup>, we also investigated this property. We observe that 3,571 CAGE peaks (13.8%) overlap a repetitive element, with simple repeats and LINE/L1 elements being the most represented classes (peak annotation with respect to repeats, CpG and TATA-box is provided in Data Citation 6). The proportion of CAGE peaks overlapping repeat elements is overall lower than what we observe in our reference human dataset (19.6%).

### Association with annotated genes

Another way to assess the quality of our promoter dataset is comparing it with established gene models, as it is expected that most of the measured signal reflects the expression of known genes. This is particularly true for CAGE data, as the protocol is designed to specifically target regions of transcription initiation<sup>17</sup>. We downloaded 5 distinct macaque gene models from UCSC (Ensembl, RefSeq, Augustus, ESTs, Genscan) and used them to annotate the CAGE peaks identified (Data Citation 2). With the manually curated genes from RefSeq, only 24% of the CAGE peaks we identified could be annotated. By using the gene build provided by Ensembl, 55.6% of the peaks were annotated, which is higher than the rate of gene-associated CAGE peaks in human (82,150/184,827; 44% (ref. 7)). We assume that because the number of macaque CAGE libraries is much smaller than those profiled in human, the resulting peaks, which are calculated over the entire sets of samples for each species, capture mainly highly expressed promoters. By missing the rare promoters in macaque, the gene-peak association rate increases. When considering human hg38 RefSeq gene or GENCODEv24 genes projected to macaque (see Methods), 60.2 and 72% respectively of the CAGE peaks were annotated. Overall, a large proportion of the CAGE peaks identified (80.4%) results to be annotated, with differences depending on the reference chosen. A summary of the number of peaks annotated with respect to each of the gene models is provided in Table 2. In Data Citation 2 we provide the full annotation information for each CAGE peak.

### Regionally enriched CAGE peaks

To assess the resolution of our data in terms of region-specific signal, we extracted a list of enriched CAGE peaks for each region (see Methods). This resulted in 877 enriched peaks (corresponding to 642 unique peaks), which we provide in Data Citation 3 with the corresponding gene symbols for rhesus Mac8 Ensembl and RefSeq, and projected hg38 RefSeq models. We then created a heatmap of region-specific



**Figure 3. Expression patterns and annotation of regionally enriched peaks.** The heatmap shows the TPM normalized expression profiles of the top 10 enriched CAGE peaks in each region. The black vertical lines highlight the separation of clustered regions as induced by the expression profiles (top dendrogram). Labels on the right side report the gene symbols (lifted over to rheMac8) corresponding to each peak, when such association is available. Black labels indicate symbols available in the native macaque annotations (either RefSeq or Ensembl), gray labels indicate symbols from human annotations (either RefSeq or GENCODE) lifted over to macaque.

expression patterns using the TPM normalized expression profiles (scaled in order to render the differences in expression easily identifiable) of the corresponding top 10 peaks enriched in each region (Fig. 3). Gene names are displayed for all those peaks that could be annotated, in black are all those available from macaque annotations and in gray those assigned thanks to the lift over of human annotations. As some of the samples are biologically and/or functionally extremely close, most of the peaks are specific for a subset of regions rather than for a single region. For example frontal, temporal, occipital and parietal gyri share the majority of the enriched CAGE peaks, as shown in bright yellow in the heatmap, with the Potassium Voltage-Gated Channel Modifier Subfamily S Member 1 (*KCNS1*) showing the highest enrichment in all four regions (score > 1.5, Data Citation 3). The most notable exception is given by cerebellum, for which most of the enriched genes are truly specific (97.0%). These include Purkinje Cell Protein 2 (*PCP2*), Cerebellin 1 and Cerebellin 3 precursors (*CBLN1*, *CBLN3*), all reported to be highly expressed in this region<sup>40–42</sup>.

It is important to notice that the choice of gene models can have a significant impact on the results: considering experimentally validated models only, such as rheMac8 RefSeq gene annotations, would lead us to conclude that the vast majority of the region-enriched CAGE peaks is un-annotated (only 11.1% of the peaks would be assigned a gene symbol). However, a larger number of annotations can be rescued when considering the rheMac8 Ensembl gene set (35.8% annotated), or the more comprehensive hg38 RefSeq genes lifted over to macaque (43.3% annotated). For example, the cerebellum-enriched genes Gamma-Aminobutyric Acid Type A Receptor Alpha6 Subunit (*GABRA6*), Regulating Synaptic Membrane Exocytosis 1 (*RIMS1*) and Neuronal Differentiation 1 (*NEUROD1*) are not found in the macaque RefSeq set, but are present in both the Ensembl and human RefSeq sets. So, in general, leveraging the human annotations lifted over to macaque allows us to assign a gene to a larger number of peaks, an example of which is the possibility to include more gene symbols as done in Fig. 3, giving us a tool to improve gene accuracy.

## Usage Notes

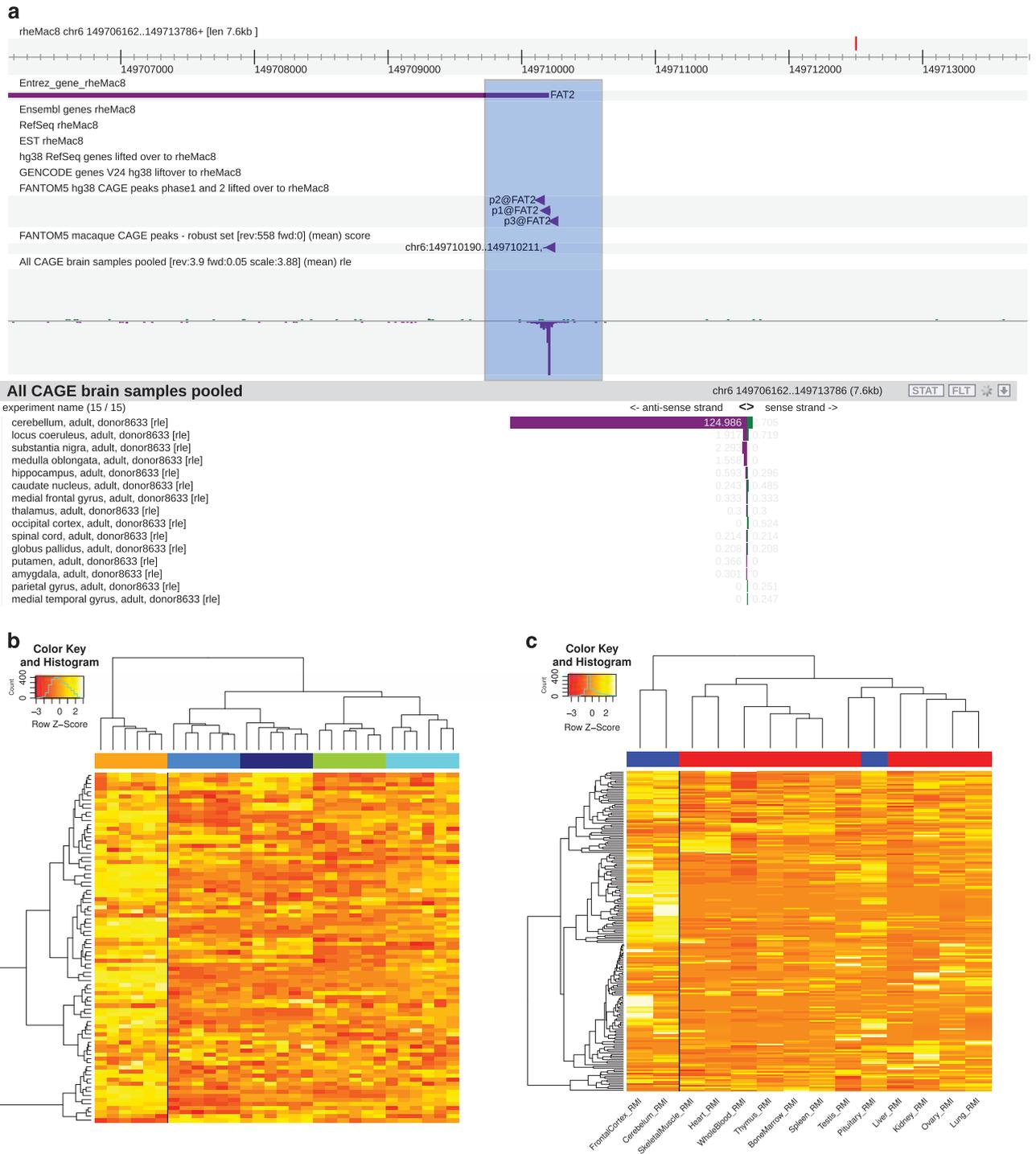
### Data exploration using the FANTOM5 portal

An online website is available (<http://fantom.gsc.riken.jp>) where all data generated within the FANTOM5 project are collected, and visualization and browsing tools are publicly accessible. Users interested in downloading the whole genomic coordinates of CAGE TSSs, peaks, expression values and gene associations for macaque as well as other organisms, such as human, can access our ftp site [fantom.gsc.riken.jp/5/datafiles/latest/](http://fantom.gsc.riken.jp/5/datafiles/latest/).

Moreover, the expression profile for a sample, or even the expression levels for a single gene across samples, can be visualized, together with dynamic statistics and data pooling operations, via our original ZENBU genome browser<sup>43</sup> (<http://fantom.gsc.riken.jp/zenbu/>). A tabular version of the expression values for a subset of genes and/or for a subset of samples from the FANTOM5 collection can be obtained from the Table Extraction Tool TET (<http://fantom.gsc.riken.jp/5/tet/>).

An example demonstration of how to navigate our resource, and how to find out more about the data, is given below.

The semantic browser SSTAR<sup>44</sup> stores information about samples and/or CAGE peaks. A human cerebellum sample page (<http://fantom.gsc.riken.jp/5/sstar/FF:10166-103B4>) lists all information and analysis results associated to that sample. Among the most enriched genes in cerebellum is *NEUROD1* (p1@NEUROD1), a known neuronal transcription factor involved in nervous system development and insulin regulation. We can compare the human *NEUROD1* CAGE peak (or better the peak associated to *NEUROD1*) to the corresponding macaque CAGE peak by determining, for example, whether this gene is also enriched in macaque. Indeed, among the top 10 enriched peaks in macaque cerebellum we find one that according to Ensembl gene annotation is associated to *NEUROD1* (CAGE peak ID 'chr12:68735984..68735999,-' in Data Citation 3). Conversely, the precision of CAGE in finding TSSs can help to resolve cases of ambiguous or unreported genes; for instance, the macaque cerebellum-enriched CAGE peak 'chr6:149710190..149710211,-' seems to have no association to known genes although (a) it is located at the 5'-end of the *FAT2* Entrez gene locus, (b) there is a cluster of lifted over human promoters to the *FAT2* gene (Fig. 4a), and (c) *FAT2* gene search in SSTAR browser shows it is highly enriched in human cerebellum (<http://fantom.gsc.riken.jp/5/sstar/EntrezGene:2196>), all points suggesting the possibility of this peak being a promoter for a *FAT2* gene in macaque. Similarly, another top enriched CAGE peak in macaque (chr4:70244748..70244762,+) is reported to have no associated RefSeq or Ensembl genes. This peak is located in an intron of the annotated macaque *RIMS1*, a gene involved in enhanced cognitive abilities in humans<sup>45</sup>, shows nearly exclusive expression in cerebellum, and matches the liftOver of human p5@RIMS1 promoter, almost certainly representing a promoter for a shorter *RIMS1* gene isoform in macaque (<http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=TJD69jEXM5yIAJTRKij45D;loc=rheMac8:chr4:70244705..70244804+>). A search in SSTAR semantic browser for *RIMS1* gene in human confirms expression enrichment in cerebellum as well (<http://fantom.gsc.riken.jp/5/sstar/EntrezGene:22999>), specifically for p5@RIMS1 promoter ([http://fantom.gsc.riken.jp/5/sstar/FFCP\\_PHASE2:Hg19::chr6:72922590..72922605,%2B](http://fantom.gsc.riken.jp/5/sstar/FFCP_PHASE2:Hg19::chr6:72922590..72922605,%2B)). Proof of the fact that several un-annotated peaks represent the promoter of a gene can also be ascertained by exploring the dedicated configuration for macaque in ZENBU (<http://fantom.gsc.riken.jp/zenbu/gLyphs/#config=TJD69jEXM5yIAJTRKij45D>).



**Figure 4. Visual exploration of the dataset presented and comparison with external data.** (a) Screenshot from ZENBU, one of the resources available to explore the expression data presented in this study, showing one example of missing gene annotation. (b) Example of a comparison between CAGE and NIH Blueprint NPH Atlas data. The colors on the top bar label distinct macroscopical anatomical regions (orange = basal ganglia, dodger blue = occipital cortex, dark blue = medial frontal cortex, yellow green = hippocampal cortex, cyan = amygdaloid complex). (c) Example of a comparison between CAGE and NHPRT datasets. The top bar distinguishes CNS (blue) and non-CNS (red) samples.

These examples also expose the dangers of using currently available, incomplete gene models, and show that gene models annotation can benefit from a technology such as CAGE.

### Comparison with other publicly available primate CNS expression datasets

The thorough annotation we provide for the expression profiles presented in this study makes it easy to perform comparisons with external datasets. As an example, we considered two recently published expression atlases including macaque samples corresponding to several regions of the CNS (see Data Records). We extracted normalized expression data of basal ganglia published from NIH Blueprint NHP Atlas<sup>20</sup> (Data Records) for donors of 12 and 48 months of age and compared the expression signature between this dataset and ours. We selected the CAGE peaks enriched (see Methods) in the corresponding regions profiled (caudate, putamen, globus pallidus and substantia nigra). By matching gene symbols from our annotation with those reported in the study, we extracted the corresponding expression profiles. Figure 4b shows a heatmap for these expression profiles in which the subgroup of basal ganglia samples is well discriminated (samples labeled in orange). In a similar fashion, we obtained normalized expression data, relatively to the subset corresponding to macaque, from the non-human primate reference transcriptome resource<sup>19</sup> (Data Records). We again selected all CAGE peaks regionally enriched and we extracted the corresponding NHPRTTR expression profiles based on matched gene symbols. Figure 4c shows a heatmap of such profiles, with CNS samples labeled in blue and non-CNS samples in red. We note that while frontal cortex and cerebellum profiles are well separated from non-CNS samples, the pituitary gland, although being part of the CNS, falls within non-CNS tissues, as a hormone-secreting gland.

In conclusion, we describe a new and important resource to help understand central nervous system in macaque that can be compared with similar resources in humans and other species such as mouse.

### References

- Behie, A. M. & Groves, C. P. China's primates: preserve wild species. *Nature* **534**, 179 (2016).
- Vallender, E. J. & Miller, G. M. Nonhuman primate models in the genomic era: a paradigm shift. *ILAR J* **54**, 154–165 (2013).
- Gibbs, R. A. *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
- Zimin, A. V. *et al.* A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* **9**, 20 (2014).
- Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
- Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
- Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44**, D717–D725 (2016).
- Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Ayoubi, T. A. & Van De Ven, W. J. Regulation of gene expression by alternative promoters. *FASEB J* **10**, 453–460 (1996).
- Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T. H. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**, 167–177 (2008).
- Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
- Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626–635 (2006).
- Consortium, F. *et al.* The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet* **41**, 553–562 (2009).
- Kanamori-Katayama, M. *et al.* Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res* **21**, 1150–1159 (2011).
- Thompson, J. F. & Steinmann, K. E. Single molecule sequencing with a HeliScope genetic analysis system. *Curr Protoc Mol Biol* **Chapter 7**, Unit7 10 (2010).
- Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786**, 181–200 (2012).
- Kawaji, H. *et al.* Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**, 708–717 (2014).
- Peng, X. *et al.* Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTTR). *Nucleic Acids Res* **43**, D737–D742 (2015).
- Bakken, T. E. *et al.* A comprehensive transcriptional map of primate brain development. *Nature* **535**, 367–375 (2016).
- Hasegawa, A., Daub, C., Carninci, P., Hayashizaki, Y. & Lassmann, T. MOIRAI: a compact workflow system for CAGE analysis. *BMC Bioinformatics* **15**, 144 (2014).
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
- Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**(Suppl 2), ii215–ii225 (2003).
- Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
- Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res* **30**, 38–41 (2002).
- Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61–D65 (2007).
- Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–589 (2010).
- Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110–D115 (2016).
- Takahashi, H., Lassmann, T., Murata, M. & Carninci, P. 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc* **7**, 542–561 (2012).

32. Roth, R. B. *et al.* Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**, 67–80 (2006).
33. Marzban, H. *et al.* Cellular commitment in the developing cerebellum. *Front Cell Neurosci* **8**, 450 (2014).
34. Barton, R. A. & Venditti, C. Rapid evolution of the cerebellum in humans and other great apes. *Curr Biol* **24**, 2440–2444 (2014).
35. Purvez, D., Augustine, G. & Fitzpatrick, D. *et al.* Neuroscience. Circuits within the Basal Ganglia System. *Sunderland (MA): Sinauer Associates*, (2001).
36. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J Mol Biol* **196**, 261–282 (1987).
37. Han, L., Su, B., Li, W. H. & Zhao, Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol* **9**, R79 (2008).
38. Cormack, B. P. & Struhl, K. The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells. *Cell* **69**, 685–696 (1992).
39. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**, 563–571 (2009).
40. Guan, J., Luo, Y. & Denker, B. M. Purkinje cell protein-2 (Pcp2) stimulates differentiation in PC12 cells by Gbetagamma-mediated activation of Ras and p38 MAPK. *Biochem J* **392**, 389–397 (2005).
41. Twine, N. A., Janitz, C., Wilkins, M. R. & Janitz, M. Sequencing of hippocampal and cerebellar transcriptomes provides new insights into the complexity of gene regulation in the human brain. *Neurosci Lett* **541**, 263–268 (2013).
42. Urade, Y., Oberdick, J., Molinar-Rode, R. & Morgan, J. I. Precerebellin is a cerebellum-specific protein with similarity to the globular domain of complement C1q B chain. *Proc Natl Acad Sci USA* **88**, 1069–1073 (1991).
43. Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat Biotechnol* **32**, 217–219 (2014).
44. Abugessaisa, I. *et al.* FANTOM5 transcriptome catalog of cellular states based on Semantic MediaWiki. *Database (Oxford)* **2016**, baw105 (2016).
45. Sisodiya, S. M. *et al.* Genetic enhancement of cognition in a kindred with cone-rod dystrophy due to RIMS1 mutation. *J Med Genet* **44**, 373–380 (2007).

## Data Citations

1. DNA Data Bank of Japan DRA005281 (2016).
2. Francescato, M., Lizio, M. & Kawaji, H. *Figshare* <https://doi.org/10.6084/m9.figshare.5414035.v1> (2017).
3. Francescato, M., Lizio, M. & Kawaji, H. *Figshare* <https://doi.org/10.6084/m9.figshare.5414041.v1> (2017).
4. DNA Data Bank of Japan DRA000991 (2013).
5. DNA Data Bank of Japan DRA002747 (2015).
6. Francescato, M., Lizio, M. & Kawaji, H. *Figshare* <https://doi.org/10.6084/m9.figshare.5414047.v1> (2017).

## Acknowledgements

FANTOM5 project was funded by: Research Grant for RIKEN Omics Science Center from MEXT to Y.H.; Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from MEXT to Y.H.; Research Grant from MEXT to the RIKEN Center for Life Science Technologies; Research Grant to RIKEN Preventive Medicine and Diagnosis Innovation Program from MEXT to Y.H. We thank the BPRC for providing the material and the GeNAS facility for data production. We also thank Dr Anemieke Rozemuller for helping sample collection and Dr Cornelis Blauwendraat for critically reading the manuscript.

## Author Contributions

A.R.R.F. and P.H. designed the experiments. I.P., L.P. and R.B. provided the samples and performed the dissections. P.R. performed RNA extractions. M.O., S.W., M.I. generated the CAGE data. M.F., M.L., A. H., T.L., J.S., J.H., I.A., T.K., H.C.C. and H.K. processed the data. P.C., Y.H., A.R.R.F. and H.K. were responsible for FANTOM5 concept and management. M.F. and M.L. wrote the manuscript with input from all authors. All authors read and approved the final version of the manuscript.

## Additional Information

**Competing interests:** The authors declare no competing financial interests.

**How to cite this article:** Francescato, M. *et al.* Transcription start site profiling of 15 anatomical regions of the *Macaca mulatta* central nervous system. *Sci. Data* **4**:170163 doi: 10.1038/sdata.2017.163 (2017).

**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

OPEN

# Author Correction: Transcription start site profiling of 15 anatomical regions of the *Macaca mulatta* central nervous system

Margherita Francescato<sup>1</sup>, Marina Lizio<sup>2,3</sup>, Ingrid Philippens<sup>4</sup>, Luba M. Pardo<sup>5</sup>, Ronald Bontrop<sup>4</sup>, Mizuho Sakai<sup>2,3</sup>, Shoko Watanabe<sup>2,3</sup>, Masayoshi Itoh<sup>2,3,6</sup>, Akira Hasegawa<sup>2,3</sup>, Timo Lassmann<sup>3,7</sup>, Jessica Severin<sup>2,3</sup>, Jayson Harshbarger<sup>2,3</sup>, Imad Abugessaia<sup>2</sup>, Takeya Kasukawa<sup>2</sup>, Piero Carninci<sup>2,3</sup>, Yoshihide Hayashizaki<sup>3,6</sup>, Alistair R. R. Forrest<sup>3,8</sup>, Hideya Kawaji<sup>2,3,6,9</sup>, Patrizia Rizzu<sup>10</sup> & Peter Heutink<sup>10</sup>

Correction to: *Scientific Data*, <https://doi.org/10.1038/sdata.2017.163>, Published online 31 October 2017

The authors regret that Luba M. Pardo was omitted in error from the author list of the original version of this Data Descriptor. This omission has now been corrected in the HTML and PDF versions.

The authors also regret that Anemieke Rozemuller was omitted in error from the Acknowledgements of the original version of this Data Descriptor. This omission has now been corrected in the HTML and PDF versions.



**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 Author(s) 2018

<sup>1</sup>Italian Institute of Technology, Department of Neuroscience and Brain Technologies, Via Morego 30, Genova, 16163, Italy. <sup>2</sup>RIKEN Center for Life Science Technologies, Division of Genomic Technologies, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan. <sup>3</sup>RIKEN Yokohama Institute, Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan. <sup>4</sup>Biomedical Primate Research Centre, Postbox 3306, Rijswijk, 2280 GH, The Netherlands. <sup>5</sup>Department of Dermatology, Erasmus MC Cancer Institute, Burg. s' Jacobplein 51, 3015, CA Rotterdam, The Netherlands. <sup>6</sup>RIKEN Preventive Medicine and Diagnosis Innovation Program, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan. <sup>7</sup>Telethon Kids Institute, The University of Western Australia, 100 Roberts Road, Subiaco, Western Australia, 6008, Australia. <sup>8</sup>Harry Perkins Institute of Medical Research, 6 Verdun St, Nedlands, Western Australia, 6009, Australia. <sup>9</sup>RIKEN Advanced Center for Computing and Communication, Preventive Medicine and Applied Genomics Unit, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan. <sup>10</sup>German Center for Neurodegenerative Diseases, Otfried-Müller Straße 23, Tübingen, 72076, Germany. Margherita Francescato and Marina Lizio contributed equally. Correspondence and requests for materials should be addressed to H.K. (email: [kawaji@gsc.riken.jp](mailto:kawaji@gsc.riken.jp))