ENCODEprojectCAGE: an R data package with CAGE data from ENCODE and modENCODE projects

Vanja Haberle [*]

November 28, 2014

Contents

1 Introduction

This document briefly describes the content of the *ENCODEprojectCAGE* data package. *ENCODEprojectCAGE* is a Bioconductor-compliant R package that contains Cap Analysis of Gene Expression (CAGE) sequencing data produced by ENCODE consortium (http://genome.ucsc.edu/ENCODE/). CAGE (Kodzius et al. (2006)) is a high-throughput method for transcriptome analysis that utilizes "cap-trapping" (Carninci et al. (1996)), a technique based on the biotinylation of the 7-methylguanosine cap of Pol II transcripts, to pulldown the 5'-complete cDNAs reversely transcribed from the captured transcripts. This enables the sequencing of short fragments from 5' ends, which can be mapped back to the referent genome to infer the exact position of the transcription start sites (TSSs) used for transcription of captured RNAs. Number of CAGE tags supporting each TSS gives the information on relative frequency of its usage and can

[*] Department of Biology, University of Bergen, Bergen, Norway

be used as a measure of expression from that specific TSS. Thus, CAGE provides information on two aspects of capped transcriptome: genome-wide 1bp-resolution map of transcription start sites and transcript expression levels. This information can be used for various analyses, from 5' centered expression profiling (Takahashi et al. (2012)) to studying promoter architecture (Carninci et al. (2006)).

This data package contains genomic coordinates of TSSs and number of CAGE tags supporting each TSS in various human cell line samples analysed by CAGE within ENCODE project. The data was originally published in the main ENCODE publication (Djebali et al. (2012)). Human CAGE data mapped to hg19 assembly of the genome was downloaded from the ENCODE web resource at UCSC (Consortium (2011), http://genome.ucsc.edu/). Obtained mapped CAGE tags were processed with the CAGEr Bioconductor package (http://www.bioconductor.org/packages/release/bioc/html/CAGEr.ht to correct for the G nucleotide addition bias and to obtain positions of individual TSSs and number of CAGE tags supporting each TSS. The data is organized into datasets by cell line and cellular compartments.

In addition, this package contains CAGE data for fruit fly (*Drosophila melanogaster*) embryos analysed within modENCODE project that was originally published in the modENCODE publication (Hoskins et al. (2011)). Fruit fly CAGE data mapped to the dm3 assembly of the genome was downloaded from the modENCODE web resource (http://data.modencode.org/).

Figure 1 schematically describes the organization and the structure of the data in the *ENCODEprojectCAGE* package. The datasets that can be loaded via call to `data()` function are shaded in blue.

2 Getting started

To load the *ENCODEprojectCAGE* package into your R envirnoment type:

```
> library(ENCODEprojectCAGE)
```

2.1 Listing available human CAGE samples

The `ENCODEhumanCellLinesSamples` dataset is a `data.frame` that lists all available CAGE samples. To load the list of human cell line samples type:

```
> data(ENCODEhumanCellLinesSamples)
> head(ENCODEhumanCellLinesSamples, 10)
```

```
   dataset   group            sample
1     A549    cell    A549_cell_rep1
2     A549    cell    A549_cell_rep2
3     A549 cytosol A549_cytosol_rep1
4     A549 cytosol A549_cytosol_rep2
```

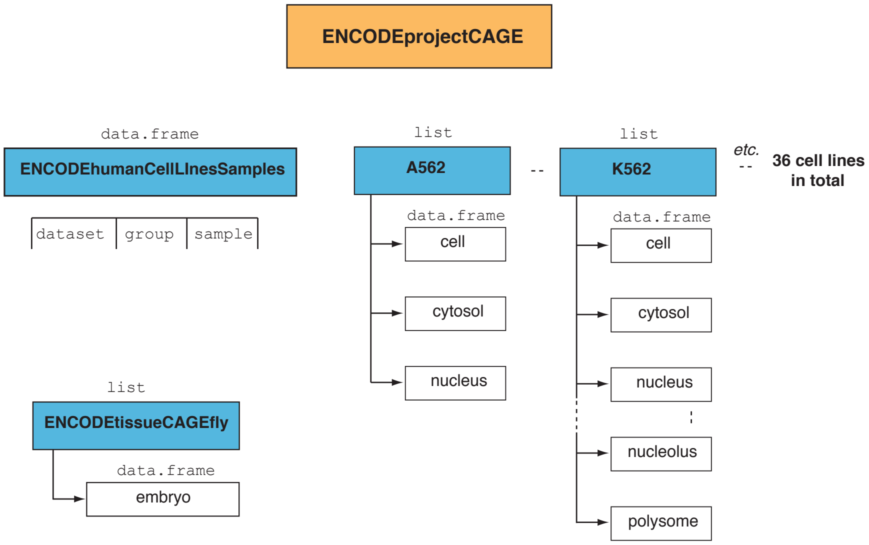

Figure 1: Content and structure of data in *ENCODEprojectCAGE* data package

```
5      A549 nucleus A549_nucleus_rep1
6      A549 nucleus A549_nucleus_rep2
7   AG04450    cell AG04450_cell_rep1
8   AG04450    cell AG04450_cell_rep2
9        BJ    cell      BJ_cell_rep1
10       BJ    cell      BJ_cell_rep2
```

The information is organized into three columns:

- `dataset`: the name of the dataset that can be loaded using `data()` function. There is one dataset per cell line named according to that cell line (*e.g.* A549)
- `group`: the name of the group of samples within one cell line that originate from the same cellular compartment (*e.g.* cytosol)
- `sample`: the name of the specific sample

All available datasets (cell lines) can be listed by typing:

```
> unique(ENCODEhumanCellLinesSamples[,"dataset"])

 [1] "A549"              "AG04450"           "BJ"
 [4] "B_CDC20+"          "CD34+_Mobilized"   "GM12878"
 [7] "H1-hESC"           "HAoAF"             "Haoec"
```

```
[10] "HCH"             "HeLa-S3"         "HepG2"
[13] "HFDPC"           "HMEpC"           "hMSC-AT"
[16] "hMSC-BM"         "hMSC-UC"         "HOB"
[19] "HPC-PL"          "HPIEpC"          "HSaVEC"
[22] "HUVEC"           "HVMF"            "HWP"
[25] "IMR90"           "K562"            "MCF-7"
[28] "Monocytes_CD14+" "NHDF"            "NHEK"
[31] "NHEM.f_M2"       "NHEM_M2"         "Prostate"
[34] "SkMC"            "SK-N-SH"         "SK-N-SHra"
```

2.2 Datasets for inidvidual human cell lines

Each dataset listed in the ENCODEhumanCellLinesSamples data frame can be loaded via call to data() function. For example, data for A549 cell line can be loaded by typing:

```
> data("A549")
> cellLineData <- get("A549")
> names(cellLineData)
```

```
[1] "cell"    "cytosol" "nucleus"
```

The dataset for each cell line is a named list, where names correspond to entries in the group column (in the ENCODEhumanCellLinesSamples data frame listing all the samples) and indicate the cellular compartment. Each element of the list is a data.frame with genomic coordinates of TSSs detected in that group of samples followed by columns with numbers of CAGE tags supporting each TSS in every individual sample. The names of columns correspond to entries in the sample column (in the ENCODEhumanCellLinesSamples data frame listing all the samples) and describe individual samples.

```
> cellLineDataNucleus <- cellLineData[["nucleus"]]
> head(cellLineDataNucleus)
```

```
   chr   pos strand A549_nucleus_rep1 A549_nucleus_rep2
1 chr1 10643      -                 0                 1
2 chr1 10648      -                 0                 1
3 chr1 14946      -                 2                 0
4 chr1 14956      -                 0                 1
5 chr1 16222      +                 0                 3
6 chr1 16470      -                 0                 1
```

2.3 Fruit fly embryos CAGE

In addition to CAGE data for various human cell lines, this package contains CAGE data for fruit fly embyos. To load this dataset type:

```
> data(ENCODEtissueCAGEfly)
> names(ENCODEtissueCAGEfly)

[1] "embryo"

> head(ENCODEtissueCAGEfly[["embryo"]])

    chr  pos strand mixed_embryos_0-24hr
1 chr2L 5238      -                    1
2 chr2L 6162      -                    2
3 chr2L 6188      -                    1
4 chr2L 6211      -                    4
5 chr2L 6581      -                    1
6 chr2L 6794      -                    2
```

This dataset is a list with only one element named ”embryo”. This element is a `data.frame` with genomic coordinates of TSSs and number of supporting CAGE tags in a mixture of fruit fly embryos (0-24 hours past fertilization).

3 Importing data to *CAGEr* package

The data provided in this package can be further processed and analyzed with *CAGEr* package and can be directly imported using the `importPublicData()` function from *CAGEr*. Here is an example of how to import whole cell CAGE data for three different cell lines.

```
> library(CAGEr)
> myCAGEset <- importPublicData(source="ENCODE",
+ dataset=c("A549", "H1-hESC", "IMR90"), group = c("cell", "cell", "cell"),
+ sample=c("A549_cell_rep1", "H1-hESC_cell_rep1", "IMR90_cell_rep1"))
```

For further details please refer to the vignette of the *CAGEr* package.

4 Session Info

```
> sessionInfo()

R version 3.1.1 (2014-07-10)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=C                 LC_NUMERIC=C
 [3] LC_TIME=C                  LC_COLLATE=C
 [5] LC_MONETARY=C              LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
[1] ENCODEprojectCAGE_1.0.0

loaded via a namespace (and not attached):
[1] tools_3.1.1
```

References

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., and Schneider, C. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3):327–336.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635.

Consortium, T. E. P. (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, 9:e1001046.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 488(7414):101–108.

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., Yu, C., Booth, B. W., Zhang, D., Wan, K. H., Yang, L., Boley, N., Andrews, J., Kaufman, T. C., Graveley, B. R., Bickel, P. J., Carninci, P., Carlson, J. W., and Celniker, S. E. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research*, 21(2):182–192.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nature Methods*, 3(3):211–222.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nature Protocols*, 7(3):542–561.