FANTOM3and4CAGE: an R data package with CAGE data from FANTOM3 and FANTOM4 projects

Vanja Haberle *

February 6, 2013

Contents

1 Introduction

This document briefly describes the content of the *FANTOM3and4CAGE* data package. *FANTOM3and4CAGE* is a Bioconductor-compliant R package that contains Cap Analysis of Gene Expression (CAGE) sequencing data produced by FANTOM consortium (http://fantom.gsc.riken.jp/). 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 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

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

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 mouse and human samples analysed by CAGE in FANTOM3 and FANTOM4 projects. The data was originally published in main FANTOM publications (Carninci et al. (2005), Carninci et al. (2006), Suzuki et al. (2009), Faulkner et al. (2009)), and represents a valuable resource of genome-wide TSSs for mouse and human. All of the data was downloaded from the FANTOM web resource (Kawaji et al. (2010), http://fantom.gsc.riken.jp/4/download/) and was organized into datasets by organism and tissue of origin. All human data is mapped to hg18 assembly, and mouse data to mm9 assembly of the genome. Figure 1 schematically describes the organization and the structure of the data in the *FANTOM3and4CAGE* package.

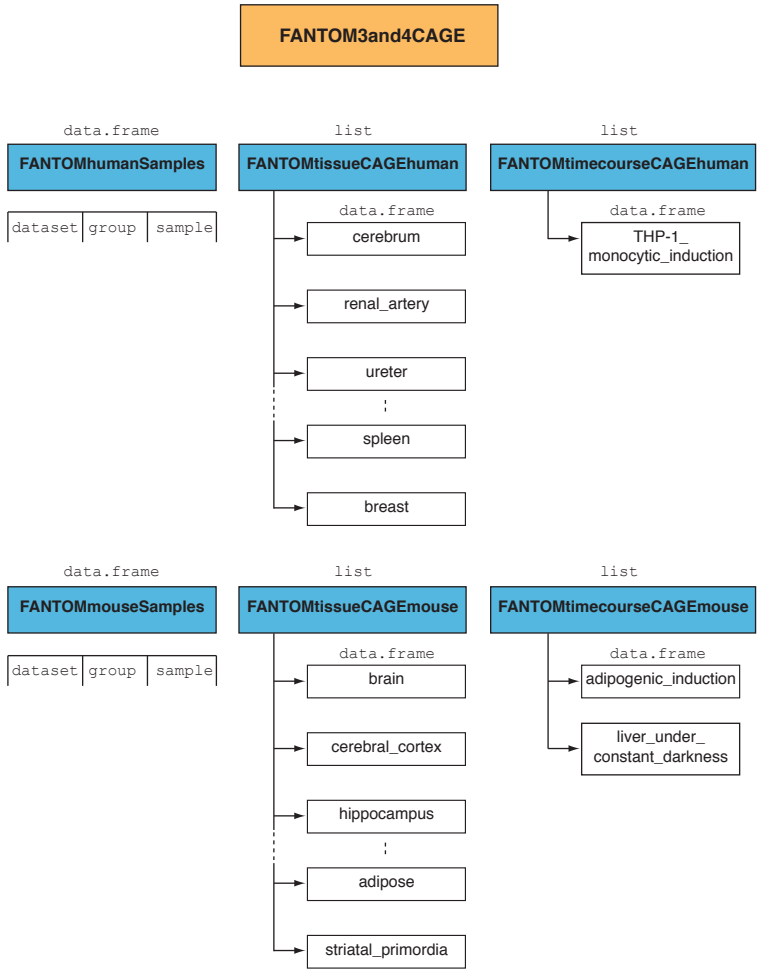

Figure 1: Content and structure of data in *FANTOM3and4CAGE* data package

2 Getting started

To load the *FANTOM3and4CAGE* package into your R envirnoment type:

```
> library(FANTOM3and4CAGE)
```

2.1 Listing available CAGE samples

As shown in Figure 1, there are six datasets (shaded in blue) that can be loaded via call to `data()` function. Two of them are `data.frames` that describe the content of the remaining four datasets. These are `FANTOMhumanSamples` and `FANTOMmouseSamples`, for human and mouse, respectively.
To load the list of human samples type:

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

```
                  dataset           group          sample
1  FANTOMtissueCAGEhuman        cerebrum        cerebrum
2  FANTOMtissueCAGEhuman    renal_artery    renal_artery
3  FANTOMtissueCAGEhuman          ureter          ureter
4  FANTOMtissueCAGEhuman urinary_bladder urinary_bladder
5  FANTOMtissueCAGEhuman          kidney      malignancy
6  FANTOMtissueCAGEhuman          kidney          kidney
7  FANTOMtissueCAGEhuman small_intestine small_intestine
8  FANTOMtissueCAGEhuman          rectum      malignancy
9  FANTOMtissueCAGEhuman          rectum          rectum
10 FANTOMtissueCAGEhuman           cecum      malignancy
```

The information is organized into three columns:

- `dataset`: the name of the dataset that can be loaded using `data()` function
- `group`: the name of the group of samples that originate from the same tissue (*e.g.* blood)
- `sample`: the name of the specific sample

2.2 CAGE datasets for various tissues

The `FANTOMtissueCAGEhuman` and `FANTOMtissueCAGEmouse` datasets contain CAGE data organized by tissue of origin:

```
> data(FANTOMtissueCAGEhuman)
> names(FANTOMtissueCAGEhuman)
```

```
 [1] "cerebrum"        "renal_artery"    "ureter"
 [4] "urinary_bladder" "kidney"          "small_intestine"
 [7] "rectum"          "cecum"           "liver"
[10] "large_intestine" "prostate_gland"  "mammary_gland"
[13] "epididymis"      "skin"            "adipose"
[16] "pancreas"        "thymus"          "undefined"
[19] "blood"           "lung"            "adrenal_gland"
[22] "colon"           "brain"           "cerebellum"
[25] "testis"          "embryo"          "bone_marrow"
[28] "heart"           "muscle"          "frontal_lobe"
[31] "occipital_lobe"  "parietal_lobe"   "spleen"
[34] "breast"
```

It is a named list, where names correspond to entries in the `group` column (in the `data.frame` listing all the samples) and indicate tissue of origin. 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 `data.frame` listing all the samples) and describe individual samples.

```
> lung_group <- FANTOMtissueCAGEhuman[["lung"]]
> head(lung_group)
```

```
   chr    pos strand RCB-0702_WI-38 RCB-0098_A549
1 chr1 558799      +              0             1
2 chr1 559777      +              0             1
3 chr1 703878      -              0             1
4 chr1 703904      -              0             0
5 chr1 752741      -              0             0
6 chr1 752754      -              0             0
  RCB-0465_Lu-130 lung
1               0    0
2               0    0
3               0    0
4               1    0
5               1    0
6               1    0
```

2.3 CAGE timecourse datasets

In addition to CAGE data for various tissue types, there are timecourse datasets available in *FANTOM3and4CAGE* package. These are `FANTOMtimecourseCAGEhuman` and `FANTOMtimecourseCAGEmouse`, for human and mouse, respectively.

```
> data(FANTOMtimecourseCAGEmouse)
> names(FANTOMtimecourseCAGEmouse)

[1] "adipogenic_induction"
[2] "liver_under_constant_darkness"

> head(FANTOMtimecourseCAGEmouse[["adipogenic_induction"]])

   chr     pos strand DFAT-D1_preadipocytes_0days
1 chr1 3091394      +                           0
2 chr1 3641339      -                           1
3 chr1 3661787      -                           0
4 chr1 3936146      +                           0
5 chr1 3968879      -                           0
6 chr1 4569858      +                           0
  DFAT-D1_preadipocytes_2days DFAT-D1_preadipocytes_4days
1                           0                           1
2                           0                           0
3                           0                           0
4                           0                           1
5                           0                           0
6                           0                           0
  DFAT-D1_preadipocytes_6days DFAT-D1_preadipocytes_8days
1                           0                           0
2                           0                           0
3                           0                           1
4                           0                           0
5                           0                           1
6                           0                           1
```

They are organized in the same way as tissue datasets described above, *i.e.* each element of the list is a `data.frame` with CAGE detected TSSs for one timecourse.

3 Session Info

```
> sessionInfo()

R version 3.1.1 Patched (2014-09-25 r66681)
Platform: x86_64-apple-darwin10.8.0 (64-bit)

locale:
[1] C/C/C/C/C/en_US.UTF-8
```

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

other attached packages:
[1] FANTOM3and4CAGE_1.1.1

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

References

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Gatta, G. D., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Babu, M. M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., SchÃűnbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Group, R. G. E. R., Group), G. S. G. G. N. P. C., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima, M., Kondo,

S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., and Hayashizaki, Y. (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563.

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.

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., Waki, K., Hornig, N., Arakawa, T., Takahashi, H., Kawai, J., Forrest, A. R. R., Suzuki, H., Hayashizaki, Y., Hume, D. A., Orlando, V., Grimmond, S. M., and Carninci, P. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics*, 41(5):563–571.

Kawaji, H., Severin, J., Lizio, M., Forrest, A. R. R., van Nimwegen, E., Rehli, M., Schroder, K., Irvine, K., Suzuki, H., Carninci, P., Hayashizaki, Y., and Daub, C. O. (2010). Update of the FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Research*, 39(Database):D856–D860.

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.

Suzuki, H., Forrest, A. R. R., van Nimwegen, E., Daub, C. O., Balwierz, P. J., Irvine, K. M., Lassmann, T., Ravasi, T., Hasegawa, Y., de Hoon, M. J. L., Katayama, S., Schroder, K., Carninci, P., Tomaru, Y., Kanamori-Katayama, M., Kubosaki, A., Akalin, A., Ando, Y., Arner, E., Asada, M., Asahara, H., Bailey, T., Bajic, V. B., Bauer, D., Beckhouse, A. G., Bertin, N., Björkegren, J., Brombacher, F., Bulger, E., Chalk, A. M., Chiba, J., Cloonan, N., Dawe, A., Dostie, J., Engström, P. G., Essack, M., Faulkner, G. J., Fink, J. L., Fredman, D., Fujimori, K., Furuno, M., Gojobori, T., Gough, J., Grimmond, S. M., Gustafsson, M., Hashimoto, M., Hashimoto, T.,

Hatakeyama, M., Heinzel, S., Hide, W., Hofmann, O., Hörnquist, M., Huminiecki, L., Ikeo, K., Imamoto, N., Inoue, S., Inoue, Y., Ishihara, R., Iwayanagi, T., Jacobsen, A., Kaur, M., Kawaji, H., Kerr, M. C., Kimura, R., Kimura, S., Kimura, Y., Kitano, H., Koga, H., Kojima, T., Kondo, S., Konno, T., Krogh, A., Kruger, A., Kumar, A., Lenhard, B., Lennartsson, A., Lindow, M., Lizio, M., MacPherson, C., Maeda, N., Maher, C. A., Maqungo, M., Mar, J., Matigian, N. A., Matsuda, H., Mattick, J. S., Meier, S., Miyamoto, S., Miyamoto-Sato, E., Nakabayashi, K., Nakachi, Y., Nakano, M., Nygaard, S., Okayama, T., Okazaki, Y., Okuda-Yabukami, H., Orlando, V., Otomo, J., Pachkov, M., Petrovsky, N., Plessy, C., Quackenbush, J., Radovanovic, A., Rehli, M., Saito, R., Sandelin, A., Schmeier, S., Schönbach, C., Schwartz, A. S., Semple, C. A., Sera, M., Severin, J., Shirahige, K., Simons, C., St Laurent, G., Suzuki, M., Suzuki, T., Sweet, M. J., Taft, R. J., Takeda, S., Takenaka, Y., Tan, K., Taylor, M. S., Teasdale, R. D., Tegnér, J., Teichmann, S., Valen, E., Wahlestedt, C., Waki, K., Waterhouse, A., Wells, C. A., Winther, O., Wu, L., Yamaguchi, K., Yanagawa, H., Yasuda, J., Zavolan, M., Hume, D. A., Arakawa, T., Fukuda, S., Imamura, K., Kai, C., Kaiho, A., Kawashima, T., Kawazu, C., Kitazume, Y., Kojima, M., Miura, H., Murakami, K., Murata, M., Ninomiya, N., Nishiyori, H., Noma, S., Ogawa, C., Sano, T., Simon, C., Tagami, M., Takahashi, Y., Kawai, J., and Hayashizaki, Y. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, 41(5):553–562.

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.