ZebrafishDevelopmentalCAGE: an R data package with CAGE data for zebrafish developmental time course

Vanja Haberle *

May 27, 2014

Contents

1 Introduction

This document briefly describes the content of the *ZebrafishDevelopmentalCAGE* data package. *ZebrafishDevelopmentalCAGE* is a Bioconductor-compliant R package that contains Cap Analysis of Gene Expression (CAGE) sequencing data produced by ZEPROME consortium and originally published by Nepal *et al.* (**?**). CAGE (**?**) is a high-throughput method for transcriptome analysis that utilizes "cap-trapping" (**?**), 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 transcript expression levels. This information can be used for various analyses, from 5' centered expression profiling (**?**) to studying promoter architecture (**?**).

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

This data package contains genomic coordinates of individual TSSs and number of CAGE tags supporting each TSS in 12 developmental stages throughout zebrafish (*Danio rerio*) early embryonic development:

- unfertilized egg
- fertilized egg
- 64 cells
- 512 cells
- high
- oblong
- sphere / dome
- dome / 30% epiboly
- shield
- 14 somites
- prim 6
- prim 20

The CAGE data was produced by the ZEPROME consortium and originally published in the resource paper by Nepal *et al.* (**?**) and subsequently used in the publication by Haberle *et al.* (**?**). The data is mapped to the Zv9 (danRer7) zebrafish genome assembly. The *ZebrafishDevelopmentalCAGE* package contains only one dataset named `ZebrafishCAGE` that can be loaded via call to `data()` function. The dataset is a `list` with only one element named development:

2 Getting started

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

```
> library(ZebrafishDevelopmentalCAGE)
```

To list all CAGE samples contained within this package type:

```
> data(ZebrafishSamples)
> ZebrafishSamples
```

```
         dataset       group              sample
1  ZebrafishCAGE development zf_unfertilized_egg
2  ZebrafishCAGE development   zf_fertilized_egg
3  ZebrafishCAGE development          zf_64cells
4  ZebrafishCAGE development         zf_512cells
5  ZebrafishCAGE development             zf_high
6  ZebrafishCAGE development           zf_oblong
7  ZebrafishCAGE development      zf_sphere_dome
8  ZebrafishCAGE development      zf_30perc_dome
9  ZebrafishCAGE development           zf_shield
10 ZebrafishCAGE development        zf_14somites
11 ZebrafishCAGE development            zf_prim6
12 ZebrafishCAGE development           zf_prim20
```

This data.frame lists the names of all CAGE samples alongside with the name of the dataset (ZebrafishCAGE) and the group (development) within which they are contained. Note that the correct `dataset`, `group` and `sample` labels are required for importing and analyzing this CAGE data with *CAGEr* package, as explained further below.

To load the dataset type:

```
> data(ZebrafishCAGE)
> names(ZebrafishCAGE)

[1] "development"

> head(ZebrafishCAGE[["development"]])

   chr  pos strand zf_unfertilized_egg zf_fertilized_egg
1 chr1 3716      +                   0                 2
2 chr1 3718      +                   0                 0
3 chr1 3724      +                   0                 0
4 chr1 3760      +                   0                 0
5 chr1 3783      +                   0                 0
6 chr1 3788      +                   0                 0
  zf_64cells zf_512cells zf_high zf_oblong zf_sphere_dome
1          0           0       0         0              0
2          1           0       0         0              0
3          0           0       0         0              0
4          0           0       0         0              0
5          0           0       0         0              0
6          0           0       0         0              0
  zf_30perc_dome zf_shield zf_14somites zf_prim6 zf_prim20
```

```
1              0         0              0         0         0
2              0         0              0         0         0
3              0         1              0         0         0
4              0         0              0         0         1
5              0         0              0         1         2
6              0         0              0         1         0
```

The data is contained within a `data.frame` with 15 columns. First 3 columns give the coordinates of individual TSS positions (chromosome, 1-based coordinate of the TSS, strand) and subsequent columns provide counts of CAGE tags supporting given TSS in the 12 zebrafish developmental samples.

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 data for single developmental stage.

```
> library(CAGEr)
> myCAGEset <- importPublicData(source="ZebrafishDevelopment",
+ dataset="ZebrafishCAGE", group="development", sample="zf_prim6")
```

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

4 Session Info

```
> sessionInfo()

R version 3.1.0 (2014-04-10)
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_US.UTF-8
 [7] LC_PAPER=no_NO.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=no_NO.UTF-8 LC_IDENTIFICATION=C

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

```
other attached packages:
[1] ZebrafishDevelopmentalCAGE_0.99.0

loaded via a namespace (and not attached):
[1] tools_3.1.0
```