diggitdata, a data package required for the examples and vignette of the diggit package

Mariano J. Alvarez, James Chen, Andrea Califano
Department of Systems Biology, Columbia University, New York, USA

April 28, 2022

1 Overview of diggitdata data package

The *diggitdata* data package provides some example datasets, including mRNA expression and copy number variation (CNV) profiles for human glioblastoma, CNV for normal blood samples, and two human glioma-context specific regulatory networks, including a transcriptional regulatory network assembled by the ARACNe algorithm[2] and a post-translational regulatory network reverse engineered by the MINDy algorithm[3].

Human glioblastoma mRNA expression dataset The human glioblastoma dataset consists of 250 human glioblastoma samples profiled by The Cancer Genome Atlas (TCGA) on Affymetrix HT-HGU133A arrays. The raw data was was pre-processed by the cleaner algorithm [1] and then MAS5 normalized. The dataset is contained in an ExpressionSet object with 6,215 features (genes) x 250 samples. We can access this dataset with the following code:

```
> library(diggitdata)
> data(gbm.expression)
> print(gbmExprs)
```

```
ExpressionSet (storageMode: lockedEnvironment)
assayData: 9215 features, 245 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: TCGA-02-0071-01 TCGA-02-0086-01 ... TCGA-06-0747-01 (245
    total)
  varLabels: subtype
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation:
```

Human glioblastoma Copy Number Variation (CNV) dataset The human glioblastoma CNV dataset contains 230 human glioblastoma samples profiled by TCGA on Agilent HG-CGH-244A arrays. The arrays data was summarized at the gene level and stored in a numerical matrix format, with genes in rows and samples in columns. To access this dataset we can use the code:

```
> data(gbm.cnv)
> print(gbmCNV[1:3, 1:3])
```

```
         TCGA-02-0001-01 TCGA-02-0002-01 TCGA-02-0003-01
B4GALNT1     -0.01903801    -0.008458616        1.251183
DTX3         -0.01391792    -0.007144781        1.099719
SEC61G        0.20503025    -0.157275603        1.440273
```

Human blood CNV dataset The human blood CNV dataset contains 33 normal human blood samples profiled by TCGA on Agilent HG-CGH-244A arrays. The arrays data was summarized at the gene level and stored in a numerical matrix format, with genes in rows and samples in columns. To access this dataset we can use the code:

```
> data(gbm.cnv.normal)
> print(gbmCNVnormal[1:3, 1:3])
```

```
          TCGA-08-0344-11A TCGA-08-0345-11A TCGA-08-0349-11A
LOC440900       0.02076955       0.05375572      0.002933042
FAM83D          0.03465642       0.01119457      0.049300981
SLK             0.02287745      -0.02443086     -0.021355070
```

Human glioma context-specific transcriptional network The human glioma transcriptional regulatory network (transcriptional interactome) represents 183,774 inferred regulatory interactions between 835 transcription factors and 8,365 target genes. It is contained in a *regulon* class S3 object, and methods to access it are included in the *viper* package, which is available from Bioconductor and it is imported by the diggitdata package.

```
> data(gbm.aracne)
> print(gbmTFregulon)
```

```
Object of class regulon with 835 regulators, 8365 targets and 183774 interactions
```

Human glioma context-specific post-translational network for CEBPB, CEBPD and STAT3 The human glioma post-translational regulatory network (post-translational interactome) represents 43 inferred modulatory interactions between 38 signaling genes and the 3 considered transcription factors. It is contained in a *regulon* class S3 object, and methods to access it are included in the *viper* package, which is available from Bioconductor and it is imported by the diggitdata package.

```
> data(gbm.mindy)
> print(gbmMindy)
```

```
Object of class regulon with 157 regulators, 3 targets and 178 interactions
```

References

[1] Alvarez,M.J. et al. (2009) Correlating measurements across samples improves accuracy of large-scale expression profile experiments. Genome Biol., 10, R143.

[2] Margolin,A.A. et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics, 7 Suppl 1, S7.

[3] Wang,K. et al. (2009) Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. Nat. Biotechnol., 27, 829-39.