Intro to the *CellMapperData* Package

Brad Nelms

May 2, 2024

Contents

1 Introduction

CellMapperData contains microarray data from several large expression compendia that have been pre-processed for use with the *CellMapper* package. These pre-processed datasets are recommended for routine searches using *CellMapper*. This introduction contains a brief overview of the datasets included in the package. For more examples on how to use *CellMapper* and *CellMapperData*, please refer to the *CellMapper* Vignette and reference [1].

2 Dataset overview

First, access the *CellMapperData* from ExperimentHub:

```
> library(ExperimentHub)
> hub <- ExperimentHub()
> x <- query(hub, "CellMapperData")
> x
```

```
ExperimentHub with 6 records
# snapshotDate(): 2024-04-29
# $dataprovider: GEO, ArrayExpress, Allen Brain Atlas
# $species: Homo sapiens, Mus musculus
# $rdataclass: CellMapperList
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH170"]]'

          title
  EH170 | Pre-processed microarray data from the Allen Brain Atlas
  EH171 | Pre-processed microarray data from the Affymetrix HG-U133PlusV2 ...
  EH172 | Pre-processed microarray data from the Affymetrix HG-U133A platform
```

```
  EH173 | Pre-processed microarray data from the Affymetrix MG-U74Av2 plat...
  EH174 | Pre-processed microarray data from the human small and large int...
  EH175 | Pre-processed microarray data from the human kidney
```

The *CellMapperData* package contains 6 microarray datasets that have been pre-processed for use with *CellMapper*. More information about each dataset can be found in the CellMapperData manual page:

```
> ?CellMapperData
```

These datasets can be extracted using their ExperimentHub accession numbers. For instance, to download and load the Allen Brain Atlas dataset (accession 'EH170'), run the following code:

```
> BrainAtlas <- x[["EH170"]]
> BrainAtlas

An object of class "CellMapperList"
#  Provide as input to the 'CMsearch' function of the 'CellMapper' package
#  Derived from an expression dataset with 20787 genes and 3702 samples
#     Dataset source:  'Allan Brain Atlas'
#  The type of gene ID used is: 'Human Entrez IDs'
#     Example gene IDs:  '733', '735', '740', '741', '744', '745', ...
```

Each of these pre-processed datasets can be provided directly to the *CellMapper* CMsearch function to predict genes selectively expressed in a specific cell type. See the *CellMapper* vignette for more details.

3 Technical background of dataset processing

Each dataset is a CellMapperList object that has been pre-processed using the CMprep function. The CMprep function transforms the data using singular value decomposition (SVD), resulting in a matrix, 'B', with the left-singular vectors of original data matrix and a vector, 'd', with the singular values. Singular vectors that account for less variance than an individual sample in the original dataset have been trimed (Kaiser's criterion), thereby removing singular vectors that mainly account for noise:

```
> names(BrainAtlas)

[1] "B" "d"

> dim(BrainAtlas$B)

[1]  1010 20787

> length(BrainAtlas$d)

[1] 1010
```

The advantage of this transformation is that it reduces dataset size, and avoids the need to perform a time-consuming SVD transformation before running *CellMapper*.

References

[1] Bradlee D. Nelms, Levi Waldron, Luis A. Barrera, Andrew W. Weflen, Jeremy A. Goettel, Guoji Guo, Robert K. Montgomery, Marian R. Neutra, David T. Breault, Scott B. Snapper, Stuart H. Orkin, Martha L. Bulyk, Curtis Huttenhower, and Wayne I. Lencer. CellMapper: rapid and accurate inference of gene expression in difficult-to-isolate cell types. *Genome Biology*, 17(1), sep 2016. URL: http://dx.doi.org/10.1186/s13059-016-1062-5, doi:10.1186/s13059-016-1062-5.