Using the TENET.ExperimentHub datasets

Rhie Lab at the University of Southern California

2024-11-14

Abstract

This vignette describes the basic usage of the TENET.ExperimentHub package, which contains datasets for use in the TENET package's vignette and function examples. These include a variety of different objects to illustrate different datasets used in TENET functions. See our GitHub repository (https://github.com/rhielab/TENET.ExperimentHub) for more information. Where applicable, all datasets are aligned to the hg38 human genome.

Contents

Introduction

The TENET.ExperimentHub package contains 6 datasets for use in the TENET package's examples and vignettes. These datasets include an example MultiAssayExperiment object with matched gene expression and DNA methylation data from a subset of both tumor (case) and adjacent normal (control) samples in The Cancer Genome Atlas (TCGA)'s breast adenocarcinoma (BRCA) cohort with essential information used in all TENET functions, an example GRanges object produced by the TENET `step1MakeExternalDatasets` function, a SummarizedExperiment object with example purity data to pass to the TENET `step2GetDifferentiallyMethylatedSites` function, a data frame with example patient clinical data (matching the data in the example MultiAssayExperiment object), and two additional GRanges objects containing example peak and topologically associating domain (TAD) data, respectively. Where applicable, all datasets are aligned to the hg38 human genome.

Acquiring and installing TENET.ExperimentHub

R 4.4 or a newer version is required.

On Ubuntu 22.04, installation was successful in a fresh R environment after adding the official R Ubuntu repository using the instructions at https://cran.r-project.org/bin/linux/ubuntu/ and running the following command in a terminal:

```
sudo apt-get install r-base-core r-base-dev libcurl4-openssl-dev libfreetype6-dev
libfribidi-dev libfontconfig1-dev libharfbuzz-dev libtiff5-dev libxml2-dev
```

No dependencies other than R are required on macOS or Windows.

Two versions of this package are available.

To install the stable version from Bioconductor, start R and run:

```
## Install BiocManager, which is required to install packages from Bioconductor
if (!requireNamespace("BiocManager", quietly = TRUE)) {
    install.packages("BiocManager")
}

BiocManager::install(version = "devel")
BiocManager::install("TENET.ExperimentHub")
```

The development version containing the most recent updates is available from our GitHub repository (https://github.com/rhielab/TENET.ExperimentHub).

To install the development version from GitHub, start R and run:

```
## Install prerequisite packages to install the development version from GitHub
if (!requireNamespace("BiocManager", quietly = TRUE)) {
    install.packages("BiocManager")
}
if (!requireNamespace("remotes", quietly = TRUE)) {
    install.packages("remotes")
}

BiocManager::install(version = "devel")
BiocManager::install("rhielab/TENET.ExperimentHub")
```

Loading TENET.ExperimentHub

To load the TENET.ExperimentHub package, start R and run:

```
library(TENET.ExperimentHub)
```

Using the included datasets

Wrapper functions are provided to allow easy access to all included datasets. Usage of each wrapper function is demonstrated below.

Included datasets

exampleTENETMultiAssayExperiment

A MultiAssayExperiment dataset created using a modified version of the `TCGADownloader` function from the TENET package utilizing TCGAbiolinks package functionality. This object contains two SummarizedExperiment objects, `expression` and `methylation`, with expression data for 11,637 genes annotated to the GENCODE v36 dataset, including all 1,637 identified human TF genes, and DNA methylation data for 20,000 probes from the Illumina HM450 methylation array. The data are aligned to the human hg38 genome. Expression and methylation values were matched from 200 tumor and 42 adjacent normal tissue samples subset from the TCGA BRCA dataset. Additionally, results from running the TENET step 1-6 functions on these samples are included in the metadata of this MultiAssayExperiment object. Clinical data for these samples are included in the colData of the MultiAssayExperiment object. (A separate data frame object containing a subset of the clinical data for these samples is available as `exampleTENETClinicalDataFrame`.) This dataset is included to demonstrate TENET functions. Note: Because this dataset is a small subset of the overall BRCA dataset, results generated by TENET from this dataset differ from those presented for the BRCA dataset at large in TENET publications.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETMultiAssayExperiment(metadata = TRUE)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9587
#> # package(): TENET.ExperimentHub
#> # $dataprovider: TCGA
#> # $species: Homo sapiens
#> # $rdataclass: MultiAssayExperiment
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETMultiAssayExperiment
#> # $description: A MultiAssayExperiment dataset created using a modified vers...
#> # $taxonomyid: 9606
#> # $genome: hg38
#> # $sourcetype: Multiple
#> # $sourceurl: https://bioconductor.org/packages/release/bioc/html/TCGAbiolin...
#> # $sourcesize: NA
#> # $tags: c("CancerData", "clinical", "CopyNumberVariationData",
#> #   "DNAMethylationData", "ExpressionData", "Homo_sapiens_Data",
#> #   "Survival", "TCGA", "TENET")
#> # retrieve record with 'object[["EH9587"]]'

## Retrieve the object itself
exampleTENETMultiAssayExperiment()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#> require("MultiAssayExperiment")
#> A MultiAssayExperiment object of 2 listed
#>  experiments with user-defined names and respective classes.
#>  Containing an ExperimentList class object of length 2:
#>  [1] expression: RangedSummarizedExperiment with 11637 rows and 242 columns
#>  [2] methylation: RangedSummarizedExperiment with 20000 rows and 242 columns
#> Functionality:
#>  experiments() - obtain the ExperimentList instance
#>  colData() - the primary/phenotype DataFrame
```

```
#>  sampleMap() - the sample coordination DataFrame
#>  `$`, `[`, `[[` - extract colData columns, subset, or experiment
#>  *Format() - convert into a long or wide DataFrame
#>  assays() - convert ExperimentList to a SimpleList of matrices
#>  exportClass() - save data to flat files
```

exampleTENETClinicalDataFrame

A data frame containing example and simulated clinical information corresponding to the samples in the `exampleTENETMultiAssayExperiment` object, used to demonstrate how TENET functions can import clinical data from a specified data frame. Clinical data are utilized by the `step2GetDifferentiallyMethylatedSites`, `step7TopGenesSurvival`, and `step7ExpressionVsDNAMethylationScatterplots` functions. The data frame consists of vital status and time variables for use by the `step7TopGenesSurvival` function, simulated purity data for each sample, and simulated copy number variation (CNV) and somatic mutation (SM) data for the top 10 genes by number of linked hypermethylated and hypomethylated probes derived from analyses done using the `exampleTENETMultiAssayExperiment` object. These data are a subset of the clinical data contained in the colData of the `exampleTENETMultiAssayExperiment` object.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETClinicalDataFrame(metadata = TRUE)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9588
#> # package(): TENET.ExperimentHub
#> # $dataprovider: Multiple
#> # $species: Homo sapiens
#> # $rdataclass: data.frame
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETClinicalDataFrame
#> # $description: A data frame containing example and simulated clinical infor...
#> # $taxonomyid: 9606
#> # $genome: NA
#> # $sourcetype: Multiple
#> # $sourceurl: https://bioconductor.org/packages/release/bioc/html/TCGAbiolin...
#> # $sourcesize: NA
#> # $tags: c("CancerData", "clinical", "CopyNumberVariationData",
#> #   "ExpressionData", "Homo_sapiens_Data", "Survival", "TCGA", "TENET")
#> # retrieve record with 'object[["EH9588"]]'

## Retrieve the object itself
exampleTENETClinicalDataFrame()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#>      vital_status time purity ENSG00000165821_CNV
#>  [ reached 'max' / getOption("max.print") -- omitted 39 columns ]
#>  [ reached 'max' / getOption("max.print") -- omitted 231 rows ]
```

exampleTENETStep1MakeExternalDatasetsGRanges

A GenomicRanges dataset representing putative enhancer regions relevant to BRCA, created using the `step1MakeExternalDatasets` function in the TENET package with the `consensusEnhancer`,

consensusNDR, publicEnhancer, publicNDR, and ENCODEdELS arguments all set to TRUE, and the cancerType argument set to "BRCA". The data are aligned to the human hg38 genome. This dataset is included to demonstrate TENET's step2GetDifferentiallyMethylatedSites function.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETStep1MakeExternalDatasetsGRanges(metadata = TRUE)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9589
#> # package(): TENET.ExperimentHub
#> # $dataprovider: Multiple
#> # $species: Homo sapiens
#> # $rdataclass: GRanges
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETStep1MakeExternalDatasetsGRanges
#> # $description: A GenomicRanges dataset representing putative enhancer regio...
#> # $taxonomyid: 9606
#> # $genome: hg38
#> # $sourcetype: Multiple
#> # $sourceurl: https://github.com/rhielab/TENET.AnnotationHub_files
#> # $sourcesize: NA
#> # $tags: c("CancerData", "ChipSeq", "CopyNumberVariationData",
#> #   "DnaseSeq", "ENCODE", "EpigenomeRoadMap", "ExpressionData",
#> #   "FANTOM5", "GEO", "H3K27ac", "Homo_sapiens_Data", "peaks", "TCGA",
#> #   "TENET")
#> # retrieve record with 'object[["EH9589"]]'

## Retrieve the object itself
exampleTENETStep1MakeExternalDatasetsGRanges()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#> GRanges object with 1971031 ranges and 0 metadata columns:
#>             seqnames        ranges strand
#>                <Rle>     <IRanges>  <Rle>
#>         [1]     chr1   10121-10270      *
#>         [2]     chr1   10389-10400      *
#>         [3]     chr1   16141-16290      *
#>         [4]     chr1   20061-20210      *
#>         [5]     chr1 135126-135275      *
#>         ...      ...           ...    ...
#>   [1971027]     chrM     8917-9607      *
#>   [1971028]     chrM     9665-9974      *
#>   [1971029]     chrM   10079-10766      *
#>   [1971030]     chrM   11143-12241      *
#>   [1971031]     chrM   12302-16539      *
#>   -------
#>   seqinfo: 25 sequences from an unspecified genome; no seqlengths
```

exampleTENETStep2GetDifferentiallyMethylatedSitesPuritySummarizedExperiment

SummarizedExperiment object

A SummarizedExperiment object with three DNA methylation datasets each composed of 10 adjacent

normal colorectal adenocarcinoma (COAD) samples from The Cancer Genome Atlas (TCGA), retrieved using the TCGAbiolinks package. Each dataset has data for 20,000 probes from the Illumina HM450 methylation array, to match the number of probes in the `exampleTENETMultiAssayExperiment` object. The data are aligned to the human hg38 genome. This object is representative of a `purity` dataset, which would contain DNA methylation data from potentially confounding sources, used with TENET's `step2GetDifferentiallyMethylatedSites` function.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETStep2GetDifferentiallyMethylatedSitesPuritySummarizedExperiment(
    metadata = TRUE
)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9590
#> # package(): TENET.ExperimentHub
#> # $dataprovider: TCGA
#> # $species: Homo sapiens
#> # $rdataclass: SummarizedExperiment
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETStep2GetDifferentiallyMethylatedSitesPuritySummarizedE...
#> # $description: A SummarizedExperiment object with three DNA methylation dat...
#> # $taxonomyid: 9606
#> # $genome: hg38
#> # $sourcetype: Multiple
#> # $sourceurl: https://bioconductor.org/packages/release/bioc/html/TCGAbiolin...
#> # $sourcesize: NA
#> # $tags: c("CancerData", "CopyNumberVariationData",
#> #   "DNAMethylationData", "ExpressionData", "Homo_sapiens_Data", "TCGA",
#> #   "TENET")
#> # retrieve record with 'object[["EH9590"]]'

## Retrieve the object itself
exampleTENETStep2GetDifferentiallyMethylatedSitesPuritySummarizedExperiment()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#> class: RangedSummarizedExperiment
#> dim: 20000 10
#> metadata(0):
#> assays(2): purityMethylationExampleA purityMethylationExampleB
#> rownames(20000): cg00002190 cg00002809 ... rs4331560 rs6982811
#> rowData names(52): address_A address_B ... MASK_extBase MASK_general
#> colnames: NULL
#> colData names(0):
```

`exampleTENETPeakRegions`

A GenomicRanges dataset with example genomic regions (peaks) of interest, used to demonstrate TENET's `step7TopGenesUserPeakOverlap` function. The peaks are derived from a ChIP-seq experiment on FOXA1 in MCF-7 cells and aligned to the human hg38 genome. They were downloaded from the ENCODE portal (file ENCFF112JVK in experiment ENCSR126YEB). **Citation:** ENCODE Project Consortium; Moore JE, Purcaro MJ, Pratt HE, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul;583(7818):699-710. doi: 10.1038/s41586-020-2493-4. Epub

2020 Jul 29. Erratum in: Nature. 2022 May;605(7909):E3. PMID: 32728249; PMCID: PMC7410828.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETPeakRegions(metadata = TRUE)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9591
#> # package(): TENET.ExperimentHub
#> # $dataprovider: ENCODE
#> # $species: Homo sapiens
#> # $rdataclass: GRanges
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETPeakRegions
#> # $description: A GenomicRanges dataset with example genomic regions (peaks)...
#> # $taxonomyid: 9606
#> # $genome: hg38
#> # $sourcetype: BED
#> # $sourceurl: https://www.encodeproject.org/files/ENCFF112JVK/@@download/ENC...
#> # $sourcesize: NA
#> # $tags: c("CancerData", "ChIPSeqData", "CopyNumberVariationData",
#> #   "ENCODE", "ExpressionData", "FOXA1", "Homo_sapiens_Data", "peaks",
#> #   "TENET")
#> # retrieve record with 'object[["EH9591"]]'

## Retrieve the object itself
exampleTENETPeakRegions()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#> GRanges object with 37386 ranges and 0 metadata columns:
#>           seqnames              ranges strand
#>              <Rle>           <IRanges>  <Rle>
#>       [1]    chr20   41650340-41650989      *
#>       [2]     chr1 147612278-147612917      *
#>       [3]    chr20   48812030-48812609      *
#>       [4]    chr15   69594337-69595180      *
#>       [5]     chr8 101607580-101608404      *
#>       ...      ...                 ...    ...
#>   [37382]    chr14 101762956-101763345      *
#>   [37383]    chr16   15875210-15875599      *
#>   [37384]     chr5 179821648-179822037      *
#>   [37385]    chr12   84236865-84237254      *
#>   [37386]    chr17   64985135-64985570      *
#>   -------
#>   seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

exampleTENETTADRegions

A GenomicRanges dataset with example topologically associating domains (TADs), used to demonstrate TENET's `step7TopGenesTADTables` function. The TADs are derived from T47D cells (mistakenly labeled as 'T470'), and aligned to the human hg38 genome. They were downloaded from the 3D Genome Browser at http://3dgenome.fsm.northwestern.edu. **Citation:** Wang Y, Song F, Zhang B, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interac-

tions. Genome Biol. 2018 Oct 4;19(1):151. doi: 10.1186/s13059-018-1519-9. PMID: 30286773; PMCID: PMC6172833.

```
## Retrieve the ExperimentHub metadata for the object
exampleTENETTADRegions(metadata = TRUE)
#> ExperimentHub with 1 record
#> # snapshotDate(): 2024-11-13
#> # names(): EH9592
#> # package(): TENET.ExperimentHub
#> # $dataprovider: 3D Genome Browser
#> # $species: Homo sapiens
#> # $rdataclass: GRanges
#> # $rdatadateadded: 2024-09-13
#> # $title: exampleTENETTADRegions
#> # $description: A GenomicRanges dataset with example topologically associati...
#> # $taxonomyid: 9606
#> # $genome: hg38
#> # $sourcetype: BED
#> # $sourceurl: http://3dgenome.fsm.northwestern.edu/downloads/hg38.TADs.zip
#> # $sourcesize: NA
#> # $tags: c("CancerData", "CopyNumberVariationData", "ExpressionData",
#> #   "Homo_sapiens_Data", "TAD", "TENET")
#> # retrieve record with 'object[["EH9592"]]'

## Retrieve the object itself
exampleTENETTADRegions()
#> see ?TENET.ExperimentHub and browseVignettes('TENET.ExperimentHub') for documentation
#> downloading 1 resources
#> retrieving 1 resource
#> loading from cache
#> GRanges object with 1889 ranges and 0 metadata columns:
#>        seqnames              ranges strand
#>           <Rle>           <IRanges>  <Rle>
#>      1     chr1      800001-3680000      *
#>      2     chr1     3800001-6000000      *
#>      3     chr1     6520001-7640000      *
#>      4     chr1     7960001-8920000      *
#>      5     chr1     9240001-9600000      *
#>    ...      ...                 ...    ...
#>   1886     chrX 148000001-149520000      *
#>   1887     chrX 149840001-150720000      *
#>   1888     chrX 151080001-152920000      *
#>   1889     chrX 152960001-153520000      *
#>   1890     chrX 154560001-156040895      *
#>   -------
#>   seqinfo: 23 sequences from an unspecified genome; no seqlengths
```

Session info

```
sessionInfo()
#> R Under development (unstable) (2024-10-21 r87258)
#> Platform: x86_64-pc-linux-gnu
#> Running under: Ubuntu 24.04.1 LTS
```

```
#>
#> Matrix products: default
#> BLAS:   /home/biocbuild/bbs-3.21-bioc/R/lib/libRblas.so
#> LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=en_GB              LC_COLLATE=C
#>  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
#>  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
#>  [9] LC_ADDRESS=C               LC_TELEPHONE=C
#> [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
#>
#> time zone: America/New_York
#> tzcode source: system (glibc)
#>
#> attached base packages:
#> [1] stats4    stats     graphics  grDevices utils     datasets  methods
#> [8] base
#>
#> other attached packages:
#>  [1] MultiAssayExperiment_1.33.0 SummarizedExperiment_1.37.0
#>  [3] Biobase_2.67.0              GenomicRanges_1.59.0
#>  [5] GenomeInfoDb_1.43.0         IRanges_2.41.0
#>  [7] S4Vectors_0.45.1            BiocGenerics_0.53.2
#>  [9] generics_0.1.3              MatrixGenerics_1.19.0
#> [11] matrixStats_1.4.1           TENET.ExperimentHub_0.99.0
#>
#> loaded via a namespace (and not attached):
#>  [1] KEGGREST_1.47.0         xfun_0.49               lattice_0.22-6
#>  [4] vctrs_0.6.5             tools_4.5.0             curl_6.0.1
#>  [7] tibble_3.2.1            fansi_1.0.6             AnnotationDbi_1.69.0
#> [10] RSQLite_2.3.7           blob_1.2.4              BiocBaseUtils_1.9.0
#> [13] pkgconfig_2.0.3         Matrix_1.7-1            dbplyr_2.5.0
#> [16] lifecycle_1.0.4         GenomeInfoDbData_1.2.13 compiler_4.5.0
#> [19] Biostrings_2.75.1       htmltools_0.5.8.1       yaml_2.3.10
#> [22] pillar_1.9.0            crayon_1.5.3            DelayedArray_0.33.1
#> [25] cachem_1.1.0            abind_1.4-8             mime_0.12
#> [28] ExperimentHub_2.15.0    AnnotationHub_3.15.0    tidyselect_1.2.1
#> [31] digest_0.6.37           purrr_1.0.2             dplyr_1.1.4
#> [34] BiocVersion_3.21.1      fastmap_1.2.0           grid_4.5.0
#> [37] cli_3.6.3               SparseArray_1.7.1       magrittr_2.0.3
#> [40] S4Arrays_1.7.1          utf8_1.2.4              withr_3.0.2
#> [43] filelock_1.0.3          UCSC.utils_1.3.0        rappdirs_0.3.3
#> [46] bit64_4.5.2             rmarkdown_2.29          XVector_0.47.0
#> [49] httr_1.4.7              bit_4.5.0               png_0.1-8
#> [52] memoise_2.0.1           evaluate_1.0.1          knitr_1.49
#> [55] BiocFileCache_2.15.0    rlang_1.1.4             glue_1.8.0
#> [58] DBI_1.2.3               BiocManager_1.30.25     jsonlite_1.8.9
#> [61] R6_2.5.1                zlibbioc_1.53.0
```