Other types of packages

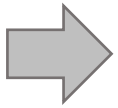
Physalia course 2023

Instructor: Jacques Serizay

Incentivize other people to use your package

Your computer:

```
myPackage/
    R/
       functions.R
       utils.R
    man/
       myfunction.Rd
    tests/
       testthat.R
       testthat/
          test-myfun.R
    inst/
       extdata/
          <raw-data-file>
    data/
       <data>.Rda
    vignettes/
       myPackage.Rmd
    DESCRIPTION
    NAMESPACE
    README.md
    NEWS
    LICENSE
```



Many computers:

```
myPackage/
    R/
       functions.R
       utils.R
    man/
       myfunction.Rd
    tests/
       testthat.R
       testthat/
          test-myfun.R
    inst/
       extdata/
          <raw-data-file>
    data/
       <data>.Rda
    vignettes/
       myPackage.Rmd
    DESCRIPTION
    NAMESPACE
    README.md
    NEWS
    LICENSE
```

Incentivize other people to use your package

- By providing high-quality public resources

- By serving a polished documentation website for your package

Different types of public resources

http://bioconductor.org/help/course-materials/2014/summerx/Annotation-slides.pdf

AnnotationHub: retrieving release-specific files

The AnnotationHub package:

- Provides a client interface to resources stored at the AnnotationHub web service.
- Offers large-scale genome resources, lightly curated for easy access from R.
- Supports tab-completion, metadata discovery, selection and filtering.

AnnotationHub: retrieving release-specific files

There are currently around 55K “records”, or files, available from the AnnotationHub.

```
> ah <- AnnotationHub::AnnotationHub()
snapshotDate(): 2020-10-27
> ah
AnnotationHub with 54989 records
# snapshotDate(): 2020-10-27
# $dataprovider: Ensembl, BroadInstitute, UCSC, ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, Haemcode, FungiDB, Inparanoid8, TriTrypDB, Plasmo...
# $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus, Pan troglodytes, Rattus norvegicus, Danio rerio, Gallus gal...
# $rdataclass: GRanges, TwoBitFile, BigWigFile, EnsDb, Rle, OrgDb, ChainFile, TxDb, Inparanoid8Db, data.frame
# additional mcols(): taxonomyid, genome, description, coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH5012"]]'

            title
  AH5012  | Chromosome Band
  AH5013  | STS Markers
  AH5014  | FISH Clones
  AH5015  | Recomb Rate
  AH5016  | ENCODE Pilot
  ...       ...
  AH89321 | Ensembl 102 EnsDb for Xiphophorus couchianus
  AH89322 | Ensembl 102 EnsDb for Xiphophorus maculatus
  AH89323 | Ensembl 102 EnsDb for Xenopus tropicalis
  AH89324 | Ensembl 102 EnsDb for Zonotrichia albicollis
  AH89325 | Ensembl 102 EnsDb for Zalophus californianus
```

AnnotationHub: retrieving release-specific files

Queries are done using query(ah, “keyword”).

```
> query(ah, c('sacCer3', 'TwoBitFile'))
AnnotationHub with 1 record
# snapshotDate(): 2020-10-27
# names(): AH14104
# $dataprovider: UCSC
# $species: Saccharomyces cerevisiae
# $rdataclass: TwoBitFile
# $rdatadateadded: 2014-12-15
# $title: sacCer3.2bit
# $description: UCSC 2 bit file for sacCer3
# $taxonomyid: 4932
# $genome: sacCer3
# $sourcetype: TwoBit
# $sourceurl: http://hgdownload.cse.ucsc.edu/goldenpath/sacCer3/bigZips/sacCer3.2bit
# $sourcesize: NA
# $tags: c("2bit", "UCSC", "genome")
# retrieve record with 'object[["AH14104"]]'
```

AnnotationHub: retrieving release-specific files

mcols(ah) can also parse the AnnotationHub as a data frame, for convenient filtering.

```
 > AnnotationHub::mcols(ah) |> tibble::as_tibble() |> dplyr::filter(genome == 'hg19', rdataclass == 'BigWigFile')
# A tibble: 9,932 × 15
   title  datap…¹ species taxon…² genome descr…³ coord…⁴ maint…⁵ rdata…⁶ prepa…⁷
   <chr>  <chr>   <chr>     <int> <chr>  <chr>     <int> <chr>   <chr>   <chr>
 1 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 2 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 3 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 4 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 5 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 6 E001-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 7 E002-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 8 E002-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
 9 E002-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
10 E002-… BroadI… Homo s…    9606 hg19   Bigwig…       0 Biocon… 2015-0… Epigen…
# … with 9,922 more rows, 5 more variables: tags <I<list>>, rdataclass <chr>,
#   rdatapath <chr>, sourceurl <chr>, sourcetype <chr>, and abbreviated
#   variable names ¹dataprovider, ²taxonomyid, ³description,
#   ⁴coordinate_1_based, ⁵maintainer, ⁶rdatadateadded, ⁷preparerclass
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

AnnotationHub: retrieving release-specific files

Objects (i.e. **files**) are retrieved using `[[`.

```
 > AnnotationHub::query(ah, 'UCSD.H1.H3K4me3.LL227.fc.signal.bigwig')
AnnotationHub with 1 record
# snapshotDate(): 2022-10-26
# names(): AH34904
# $dataprovider: BroadInstitute
# $species: Homo sapiens
# $rdataclass: BigWigFile
# $rdatadateadded: 2015-05-07
# $title: UCSD.H1.H3K4me3.LL227.fc.signal.bigwig
# $description: Bigwig File containing fold enrichment signal tracks from Ep...
# $taxonomyid: 9606
# $genome: hg19
# $sourcetype: BigWig
# $sourceurl: http://egg2.wustl.edu/roadmap/data/byFileType/signal/unconsoli...
# $sourcesize: 97131347
# $tags: c("EpigenomeRoadMap", "signal", "unconsolidated",
#   "foldChange", "NA")
# retrieve record with 'object[["AH34904"]]'
 > ah[['AH34904']]
loading from cache
require("rtracklayer")
BigWigFile object
resource: /Users/jacques/Library/Caches/org.R-project.R/R/AnnotationHub/39047900466_40344
```

AnnotationHub: retrieving release-specific files

The files retrieved as **stored on your computer!**

Their location can be identified using `resource` function from the BiocIO package.

```
 > BiocIO::resource(ah[['AH34904']])
loading from cache
                                                                  AH34904
"/Users/jacques/Library/Caches/org.R-project.R/R/AnnotationHub/39047900466_40344"
```

ExperimentHub: retrieving release-specific files

Same as AnnotationHub, but for actual experimental datasets.

```
 > AnnotationHub::query(eh, 'HiContacts')
ExperimentHub with 7 records
# snapshotDate(): 2022-10-24
# $dataprovider: Jacques Serizay
# $species: Saccharomyces cerevisiae, Mus musculus
# $rdataclass: character
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH7701"]]'

           title
  EH7701 | WT yeast Hi-C cool
  EH7702 | WT yeast Hi-C mcool
  EH7703 | WT yeast Hi-C pairs
  EH7704 | Eco1-AID yeast Hi-C mcool
  EH7705 | Eco1-AID yeast Hi-C pairs
  EH7706 | mESC Hi-C mcool
  EH7707 | mESC Hi-C pairs
```

ExperimentHub: retrieving release-specific files

Same as AnnotationHub, but for actual experimental datasets.

Generally, the datasets are provided to help demonstrating the use of a package.

```
 > eh['EH7701']
ExperimentHub with 1 record
# snapshotDate(): 2022-10-24
# names(): EH7701
# package(): HiContactsData
# $dataprovider: Jacques Serizay
# $species: Saccharomyces cerevisiae
# $rdataclass: character
# $rdatadateadded: 2022-08-23
# $title: WT yeast Hi-C cool
# $description: Hi-C performed on wild-type S288C yeast strain processed wit...
# $taxonomyid: 4932
# $genome: S288C
# $sourcetype: HDF5
# $sourceurl: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM5918358
# $sourcesize: NA
# $tags: c("HiCDdata", "SequencingData")
# retrieve record with 'object[["EH7701"]]'
 > eh['EH7701']$description
[1] "Hi-C performed on wild-type S288C yeast strain processed with tinyMapper and represented as a single-resolution cool file. Data representation derived from SRA run results SRR13994279."
 > eh[['EH7701']]
see ?HiContactsData and browseVignettes('HiContactsData') for documentation
loading from cache
                                                                    EH7701
"/Users/jacques/Library/Caches/org.R-project.R/R/ExperimentHub/680a1a15142_7751"
```

ExperimentData packages

Experiment data packages enable access to data sets that are used, often by software packages, to illustrate particular analyses.

These packages provide a gateway to curated data from an experiment, teaching course or publication.

The actual data will be managed by the ExperimentHub.

ExperimentData packages

- Similar requirements as software packages.
- Most importantly, proper documentation for the data included within the package is required.
- Traditional Annotation and Experiment packages are not ideal.
- AnnotationHub and ExperimentHub interfaces and packages are desirable.

Developing an experimentData package

The package should minimally contain:

1. The resource metadata
2. Man pages describing the resources
3. A vignette

It may also contain supporting R functions the author wants to provide

Resource metadata

Stored in inst/extdata/metadata.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Title	Description	BiocVe	Genome	SourceType	SourceUrl	SourceVersion	Species	TaxonomyId	Coordinate_	DataProvider	Maintainer	RDataClass	DispatchCla	RDataPath	Tags
2	WT yeast Hi-C cool	Hi-C performed	3.16	S288C	HDF5	https://wwv	Aug 17 2022	Saccharom	4932	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/S288C.cool	HiCDdata
3	WT yeast Hi-C mcool	Hi-C performed	3.16	S288C	HDF5	https://wwv	Aug 17 2022	Saccharom	4932	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/S288C.mcool	HiCDdata
4	WT yeast Hi-C pairs	Hi-C performed	3.16	S288C	TXT	https://wwv	Aug 17 2022	Saccharom	4932	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/S288C.pairs.gz	HiCDdata
5	Eco1-AID yeast Hi-C mcool	Hi-C performed	3.16	S288C	HDF5	https://wwv	Aug 17 2022	Saccharom	4932	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/S288C_Eco1-AID.mcool	HiCDdata
6	Eco1-AID yeast Hi-C pairs	Hi-C performed	3.16	S288C	Zip	https://wwv	Aug 17 2022	Saccharom	4932	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/S288C_Eco1-AID.pairs.g	HiCDdata
7	mESC Hi-C mcool	Hi-C performed	3.16	mm10	HDF5	https://wwv	Aug 17 2022	Mus muscu	10090	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/mESCs.mcool	HiCDdata
8	mESC Hi-C pairs	Hi-C performed	3.16	mm10	TXT	https://wwv	Aug 17 2022	Mus muscu	10090	FALSE	Jacques Serizay	Jacques Seri	character	FilePath	HiContactsData/mESCs.pairs.gz	HiCDdata
9																
10																

0.1 Introduction to HiContactsData

HiContactsData is a companion data package giving programmatic access to several processed Hi-C files for demonstration, such as cool, mcool and pairs files. It is meant to be used with `HiContacts`.

```
library(HiContactsData)
```

The only function provided by HiContactsData package is `HiContactsData()`. Several files are avaible using this function, namely:

- S288C.cool (`sample` : `yeast_wt` , `format` = `cool`)
- S288C.mcool (`sample` : `yeast_wt` , `format` = `mcool`)
- S288C.pairs.gz for chrII only (`sample` : `yeast_wt` , `format` = `pairs`)
- S288C_Eco1-AID.mcool (`sample` : `yeast_Eco1` , `format` = `mcool`)
- S288C_Eco1-AID.pairs.gz for chrII only (`sample` : `yeast_Eco1` , `format` = `pairs`)
- mESCs.mcool (`sample` : `mESCs` , `format` = `mcool`)
- mESCs.pairs.gz for chr13 only (`sample` : `mESCs` , `format` = `pairs`)

Yeast data comes from Bastie, Chapard et al., Nature Structural & Molecular Biology 2022 and mouse ESC data comes from Bonev et al., Cell 2017.

To download one of these files, one can specify a `sample` and a file `format` :

```
cool_file <- HiContactsData()
#> Available files:
#>       sample   format genome    condition                                notes
#> 1   yeast_wt     cool  S288C    wild-type      cool file @ resolution of 1kb
#> 2   yeast_wt    mcool  S288C    wild-type               multi-res mcool file
#> 3   yeast_wt pairs.gz  S288C    wild-type only pairs from chrII are provided
#> 4 yeast_eco1    mcool  S288C Eco1-AID+IAA               multi-res mcool file
#> 5 yeast_eco1 pairs.gz  S288C Eco1-AID+IAA only pairs from chrII are provided
#> 6      mESCs    mcool   mm10        mESCs               multi-res mcool file
#> 7      mESCs pairs.gz   mm10        mESCs only pairs from chr13 are provided
#>     EHID
#> 1 EH7701
#> 2 EH7702
#> 3 EH7703
#> 4 EH7704
#> 5 EH7705
#> 6 EH7706
#> 7 EH7707
#>
cool_file <- HiContactsData(sample = 'yeast_wt', format = 'cool')
#> snapshotDate(): 2022-10-24
#> see ?HiContactsData and browseVignettes('HiContactsData') for documentation
#> loading from cache
cool_file
#>                                                    EH7701
#> "/home/biocbuild/.cache/R/ExperimentHub/3508b72098fa84_7751"
```

Accessory function

```r
#' HiContactsData
#'
#' @description Downloads different types of Hi-C processed files
#'    (cool, mcool, pairs.gz) and returns the path of the cached file.
#' @param sample sample
#' @param format format
#'
#' @return Local path of the queried file cached with BiocFileCache.
#' @import ExperimentHub
#' @import AnnotationHub
#' @import BiocFileCache
#' @export
#'
#' @examples
#' HiContactsData(sample = 'yeast_wt', format = 'cool')

HiContactsData <- function(sample = NULL, format = NULL) {
    ehub_entry <- HiContactsDataFiles[
        which(HiContactsDataFiles$sample == sample &
            HiContactsDataFiles$format == format),
        "EHID"
    ]
    if (length(ehub_entry) == 0) {
        message('Available files: \n')
        print(HiContactsDataFiles)
        message('')
        if (!is.null(sample) | !is.null(format)) {
            stop('Unknown combination of `sample` and `format`.\n  ',
                'Please check which files are available from ',
                'the data frame printed above.'
            )
        }
        else {
            return()
        }
    }
    ehub <- ExperimentHub::ExperimentHub()
    res <- ehub[[ehub_entry]]
    return(res)
}
```

Developing an experimentData package

1. Open a submission issue for your ExperimentData package

2. Wait until a reviewer is assigned

3. Add your analysis package to the issue by commenting in the issue:

`AdditionalPackage: https://github.com/<yourGH>/<yourPackage>`

This ensures that the ExperimentData package is first available to Bioconductor submission building machines, so that the software package has access to it when submitted next to the same virtual machines.