The Bioconductor “Hubs”

AnnotationHub/ExperimentHub

Lori Shepherd-Kern
Bioc2022
lori.shepherd@roswellpark.org

Overview

- What is the AnnotationHub and ExperimentHub
- How you find resources with AnnotationHub/ExperimentHub
- Common Misconceptions regarding the “Hubs”
- How you contribute/add data to the appropriate Hub
- Common pitfalls of submitting data to the Hubs

What is the AnnotationHub and ExperimentHub?

Databases

that store metadata about resources,

Including download location

AnnotationHub( ) / ExperimentHub( )

- create a hub object
- sqlite database backend
  - Metadata about the resources including download location
- Files are stored remotely and downloaded as needed/requested
  - Bioconductor Microsoft Azure Data Lake
    - May be stored elsewhere
  - After downloaded, file cached for quick access for future runs

How do you find resources of interest?

- query ( )

  general search (recommended)

- subset ( )

  exact search

Against the metadata of the database

What is metadata?

- Provided by contributor at the time of data inclusion in the hub database
- Metadata columns:
  - Title
  - Data provider
  - Data Class
  - Species
  - Taxonomy Id
  - Genome
  - Description
  - Source Type
  - Data Date Added
  - Preparer Class (package name/recipe)
  - Tags
  - Maintainer
  - Data Path
  - Source URL
  - Coordinate 1-based

Example:

```
> library(AnnotationHub)
> hub = AnnotationHub()
snapshotDate(): 2022-06-28
```

```
> hub
AnnotationHub with 65055 records
# snapshotDate(): 2022-06-28
# $dataprovider: Ensembl, BroadInstitute, UCSC, ftp://ftp.ncbi.nlm.nih.gov/g...
# $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus,...
# $rdataclass: GRanges, TwoBitFile, BigWigFile, EnsDb, Rle, OrgDb, ChainFile...
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH5012"]]'
```

As of 7-6-2022

Example:

```
> names(mcols(hub))
 [1] "title"              "dataprovider"       "species"
 [4] "taxonomyid"         "genome"             "description"
 [7] "coordinate_1_based" "maintainer"         "rdatadateadded"
[10] "preparerclass"      "tags"               "rdataclass"
[13] "rdatapath"          "sourceurl"          "sourcetype"
```

Example:

```
> length(unique(tolower(hub$species)))
[1] 2557
```

```
> head(unique(tolower(hub$species)))
[1] "homo sapiens"        "vicugna pacos"       "dasypus novemcinctus"
[4] "otolemur garnettii" "papio hamadryas"     "papio anubis"
```

```
> length(unique(hub$rdataclass))
[1] 28
```

```
> unique(hub$rdataclass)
 [1] "GRanges"          "data.frame"                          "Inparanoid8Db"    “TwoBitFile”
 [5] "ChainFile"        "SQLiteConnection"                    "biopax"           “BigWigFile”
 [9] "AAStringSet"      "MSnSet"                              "mzRident"         “list”
[13] "TxDb"             "Rle"                                 "EnsDb"            "VcfFile"
[17] "igraph"           "data.frame, DNAStringSet, GRanges" "sqlite"           "data.table"
[21] "character"        "SQLite"                              "SQLiteFile"       "Tibble"
[25] "Rda"              "FaFile"                              "String"           "OrgDb"
```

Example:

```
> query(hub, "Canis familiaris")
AnnotationHub with 223 records
# snapshotDate(): 2022-06-28
# $dataprovider: Ensembl, UCSC, NCBI,DBCLS, FANTOM5,DLRP,IUPHAR,HPRD,STRING,...
# $species: Canis familiaris, Canis familiaris_dingo
# $rdataclass: GRanges, TwoBitFile, TxDb, EnsDb, SQLiteFile, OrgDb, Tibble, ...
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH5816"]]'

            title
  AH5816   | Assembly
  AH5817   | Gap
  AH5818   | RefSeq Genes
  AH5819   | Other RefSeq
  AH5820   | Ensembl Genes
  ...        ...
  AH97851  | MeSHDb for Canis familiaris (Dog, v002)
  AH100316 | MeSHDb for Canis familiaris (Dog, v003)
  AH100402 | org.Cf.eg.db.sqlite
  AH100439 | LRBaseDb for Canis familiaris (Dog, v003)
  AH100927 | org.Canis_familiaris_dingo.eg.sqlite
```

Example:

```
> query(hub, c("Canis familiaris", “GRanges”))
AnnotationHub with 125 records
# snapshotDate(): 2022-06-28
# $dataprovider: UCSC, Ensembl
# $species: Canis familiaris
# $rdataclass: GRanges
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH5816"]]'

           title
  AH5816  | Assembly
  AH5817  | Gap
  AH5818  | RefSeq Genes
  AH5819  | Other RefSeq
  AH5820  | Ensembl Genes
  ...     | ...
  AH75259 | Canis_familiaris.CanFam3.1.98.chr.gtf
  AH75260 | Canis_familiaris.CanFam3.1.98.gtf
  AH79001 | Canis_familiaris.CanFam3.1.99.abinitio.gtf
  AH79002 | Canis_familiaris.CanFam3.1.99.chr.gtf
  AH79003 | Canis_familiaris.CanFam3.1.99.gtf
```

Example:

```
> query(hub, c("Canis familiaris", “GRanges”, “ensembl”, “release-99”))
AnnotationHub with 3 records
# snapshotDate(): 2022-06-28
# $dataprovider: Ensembl
# $species: Canis familiaris
# $rdataclass: GRanges
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH79001"]]'

            title
  AH79001 | Canis_familiaris.CanFam3.1.99.abinitio.gtf
  AH79002 | Canis_familiaris.CanFam3.1.99.chr.gtf
  AH79003 | Canis_familiaris.CanFam3.1.99.gtf
```

The more specific search terms, the more specialized results

Example:

```
           title
  AH79001 | Canis_familiaris.CanFam3.1.99.abinitio.gtf
  AH79002 | Canis_familiaris.CanFam3.1.99.chr.gtf
  AH79003 | Canis_familiaris.CanFam3.1.99.gtf

> hub["AH79003"]
AnnotationHub with 1 record
# snapshotDate(): 2022-06-28
# names(): AH79003
# $dataprovider: Ensembl
# $species: Canis familiaris
# $rdataclass: GRanges
# $rdatadateadded: 2019-10-29
# $title: Canis_familiaris.CanFam3.1.99.gtf
# $description: Gene Annotation for Canis familiaris
# $taxonomyid: 9615
# $genome: CanFam3.1
# $sourcetype: GTF
# $sourceurl: ftp://ftp.ensembl.org/pub/release-99/gtf/canis_familiaris/Cani...
# $sourcesize: 17961052
# $tags: c("GTF", "ensembl", "Gene", "Transcript", "Annotation")
# retrieve record with 'object[["AH79003"]]'
```

A single bracket
[
Will let you investigate the metadata of a resources WITHOUT downloading

Example:

A double bracket
[[
Will download resource
and cache it locally

```
> grangesObj = hub[["AH79003"]]

downloading 1 resources
retrieving 1 resource
  |======================================================================| 100%

loading from cache
Importing File into R ..
require("rtracklayer")


> summary(grangesObj)
[1] "GRanges object with 1397974 ranges and 21 metadata columns"


> grangesObj = hub[["AH79003"]]  # skips download and loads from cache = faster access

loading from cache
```

Example:

```
> library(GenomicFeatures)

> TxDb = makeTxDbFromGRanges(grangesObj)
Warning message:
In .get_cds_IDX(mcols0$type, mcols0$phase) :
  The "phase" metadata column contains non-NA values for features of type
  stop_codon. This information was ignored.

> TxDb
TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Genome: CanFam3.1
# Nb of transcripts: 60994
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2022-07-06 12:45:10 -0400 (Wed, 06 Jul 2022)
# GenomicFeatures version at creation time: 1.49.5
# RSQLite version at creation time: 2.2.14
# DBSCHEMAVERSION: 1.2
```

Example:

```
> query(hub, c("Canis familiaris", “TxDb”))
AnnotationHub with 9 records
# snapshotDate(): 2022-06-28
# $dataprovider: UCSC
# $species: Canis familiaris
# $rdataclass: TxDb
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH52251"]]'

            title
  AH52251 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH57985 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH61791 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH66168 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH70585 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH75754 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH79589 | TxDb.Cfamiliaris.UCSC.canFam3.refGene.sqlite
  AH95969 | TxDb.Cfamiliaris.UCSC.canFam4.refGene.sqlite
  AH95970 | TxDb.Cfamiliaris.UCSC.canFam5.refGene.sqlite
```

Example:

```
> TxDb2 = hub[["AH95970"]]

> TxDb2

TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: canFam5
# Organism: Canis familiaris
# Taxonomy ID: 9615
# UCSC Table: refGene
# UCSC Track: RefSeq Genes
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Entrez Gene ID
# Full dataset: yes
# miRBase build ID: NA
# Nb of transcripts: 2322
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2021-09-16 17:16:23 +0000 (Thu, 16 Sep 2021)
# GenomicFeatures version at creation time: 1.45.2
# RSQLite version at creation time: 2.2.8
# DBSCHEMAVERSION: 1.2
```

Example:

ExperimentHub works the exact same way

```
> library(ExperimentHub)
> hub = ExperimentHub()
snapshotDate(): 2022-06-29


> hub
ExperimentHub with 6332 records
# snapshotDate(): 2022-06-29
# $dataprovider: Eli and Edythe L. Broad Institute of Harvard and MIT, NCBI,...
# $species: Homo sapiens, Mus musculus, Saccharomyces cerevisiae, human gut ...
# $rdataclass: SummarizedExperiment, ExpressionSet, matrix, character, list,...
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH1"]]'
```

As of 7-6-2022

Example:

```
> query(hub, c("SingleCellExperiment", "Mus musculus", "GEO"))
ExperimentHub with 12 records
# snapshotDate(): 2022-06-29
# $dataprovider: GEO (GSE60749), Kumar et al. (2014), GEO, Wellcome Trust Sa...
# $species: Mus musculus
# $rdataclass: SingleCellExperiment
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH1433"]]'

            title
  EH1433 | GEO accession data GSE71585 as a SingleCellExperiment
  EH1508 | sce_full_Kumar
  EH1509 | sce_filteredExpr10_Kumar
   ...      ...
  EH3297 | Crowell19_4vs4
  EH5433 | GTseq_transcriptomic
  EH6747 | zeisel
```

Example:

```
> unique(mcols(query(hub, c("SingleCellExperiment", "Mus musculus", "GEO")))$preparerclass)
[1] "allenpvc"                "DuoClustering2018"
[3] "muscData"                "SingleCellMultiModal"
[5] "benchmark.data.scRNAseq"

> mcols(query(hub, c("SingleCellExperiment", "Mus musculus", "GEO")))$genome
 [1] "mm10"   "GRCm38" "GRCm38" "GRCm38" "GRCm38" "GRCm38" "GRCm38" "GRCm38"
 [9] "GRCm38" NA        NA    "mm10"
```

Example:

Example of subset. Subset requires a priori knowledge

```
> subset(hub, preparerclass=="GSE62944" & rdataclass=="SummarizedExperiment")
ExperimentHub with 2 records
# snapshotDate(): 2022-06-29
# $dataprovider: GEO
# $species: Homo sapiens
# $rdataclass: SummarizedExperiment
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH1043"]]'

            title
  EH1043 | RNA-Sequencing and clinical data for 9246 tumor samples from The...
  EH1044 | RNA-Sequencing and clinical data for 741 normal samples from The...
```

Common Misconceptions

- Bioconductor Core Team provides all the data in the Hubs
- The data is updated by Bioconductor Core Team
- All the data is hosted on Bioconductor Microsoft Azure Data Lakes / When submitting data I have to use the Bioconductor provided location to host data

How to contribute

All data must have an accompanied package!

Helpful:

Bioconductor Package HubPub

Vignette “Create A Hub Package”

https://bioconductor.org/packages/devel/bioc/vignettes/HubPub/inst/doc/CreateAHubPackage.html

What makes a hub package different

Description BiocViews terms:

ExperimentHub, AnnotationHub, ExperimentHubSoftware, AnnotationHubSoftware

inst/

extdata/ - **metadata.csv** file with metadata on resources being added to hub database

scripts/ - files on how the data was generated. May be code, sudo-code, or text

Everything else is the same as a normal package submission/requirements

http://contributions.bioconductor.org/

Metadata file

- Title
  - Name of the resource
  - If taking advantage of creating ExperimentHub access functions should avoid spaces and punctuation
- Description
  - Brief description of the resource. Try to avoid special characters
- BiocVersion
  - First Bioconductor version the **resource** will be available. Generally this is the current **devel** version of Bioconductor

Metadata file

- Genome
  - Can be NA if not appropriate
- Species
  - For valid species see `getSpeciesList`, `validSpecies` or `suggestSpecies` functions.
  - Can be NA if not appropriate
- TaxonomyId
  - There are checks for valid taxonomy id given a Species
  - Full validation table is from `GenomeInfoDb::loadTaxonomyDb()`
  - Can be NA if not appropriate
- Coordinate_1_based
  - Logical indicating if data are 1-based
  - Can be NA if not appropriate

Metadata file

- SourceType
  - Format of original data
  - See `AnnotationHubData::getValidSourceTypes()` for currently acceptable values
  - Simulated is acceptable type
- SourceUrl
  - Location of original data files.
  - Multiple urls should be provided as comma separated string “url1, url2, url3”
  - If data is simulated we recommend putting the Lab url or future bioc package url
- SourceVersion
  - Version of original data
- DataProvider
  - Name of company or institution that provided the original data

Metadata file

- RDataClass
  - R/Bioconductor class structure the data is stored as e.g. GRanges, SummarizedExperiment, etc
  - When the file or object is loaded or read into R what is the class of the object

- DispatchClass
  - Determines how the data is loaded into R using the AnnotationHub interface
  - See `AnnotationHub::DispatchClassList()` for available implementations
  - Example:
    - Data created with save() should use `Rda`
    - Data created with saveRDS should use `Rds`
  - If unsure, we recommend **FilePath** . FilePath instead of trying to load or read an object into R will simply return the path to the locally downloaded file.

Metadata file

- Location_Prefix
  - **DO NOT** include this column if you are using the Bioconductor provided storage location
  - This should be the base url to the data resource
  - Has trailing slash

- RDataPath
  - Remainder of the path to the resource that will be concatenated to location_prefix
  - If using Bioconductor default storage location, this is generally the paths of resources uploaded, including subdirectories, and often starts with the package name
  - Does not have leading slash
  - Includes the resource name with extension

Location_Prefix + RDataPath == full url to resource

Metadata file

Location_Prefix and RDataPath example 1:

Let’s say you upload a directory for your package to the Bioconductor storage location. The uploaded directory contains subdirectory sub1 and sub2. Each of those directories have two files file1 a csv and file 2 a rda. You would NOT include a Location_Prefix column and your RDataPath column would look something like the following:

RDataPath
YourPackageName/sub1/file1.csv
YourPackageName/sub1/file2.rda
YourPackageName/sub2/file1.csv
YourPackageName/sub2/file2.rda

Metadata file

Location_Prefix and RDataPath example 2:

Let's say you are hosting the data at an institutional server. The path to the resources you would like to include is something like the following:
https://myinstitutionwebsite.org/dataserver/mylab/file
https://myinstitutionwebsite.org/dataserver/mylab/file2

Your columns might look like the following:

dataserver could be included in either location_prefix or rdatapath which ever is more appropriate

Location_Prefix	RDataPath
https://myinstitutionwebsite.org/dataserver/	mylab/file
https://myinstitutionwebsite.org/dataserver/	mylab/file2

Metadata file

- Maintainer
  - Maintainer name and email
  - Who is responsible for the data in the hub
  - “Maintainer_Name username@address”

- Tags
  - Optional set of keywords to be associated with the resource
  - Users will be able to query against these values to find resources
  - If multiple tags are used, it should be a single character vector with tags separated by a colon. E.g. “tag1:tag2:tag3”
  - biocViews of description file are automatically added as additional tags.

Other columns may be included in the metadata file but will be ignored and not added to the hub database

Metadata file location and name

- Should be in inst/exdata
- Should have an accompanied inst/script file that describes how the resources uploaded were generated (code, sudo-code, text). Should be a how-to if a user wanted to replicate creating a similar resource structure. Minimally contains source information and appropriate licensing if applicable.
- CSV file with the required columns previously discussed
- We have been referring to “metadata.csv” but it can be named anything “resources.csv”, “dataset1.csv”, etc.
- You may have more than one csv file. This may be useful to have separate files for different datasets or versions.

Submission Process

- Create your template package with biocViews and inst/extdata/metadata.csv completed
- Email hubs@bioconductor.org and maintainer@bioconductor.org to add data to database and provide link to the package
- If using Bioconductor storage, request access to upload data when reaching out above, and upload data accordingly
- A team member will reach out with any issues or when the data is available through the current hub interface
- Update package to use the hub interface
- Submit the package to the new package submission tracker

HubPub

- Package to assist with the creation of a hub package and helper functions to populate necessary files

- Documentation:
  - Creating a Hub Package Vignette:
    - https://bioconductor.org/packages/release/bioc/vignettes/HubPub/inst/doc/CreateAHubPackage.html
  - HubPub Package Vignette:
    - https://bioconductor.org/packages/release/bioc/vignettes/HubPub/inst/doc/HubPub.html

Common pitfalls when submitting

- R is case sensitive! Metadata Location_Prefix/RDataPath must match case of destination (example: lori.rda is not the same as LoRi.rda)
- Most important metadata is location_prefix/rdatapath: Be mindful of typos between metadata and destination (example: lori.rda becomes oops loir.rda)
- Double check number of uploads vs rows of metadata. Almost always they should match (exception: data types with two files like bam/bai)
- Include tags or sufficient biocViews. Must have 2 valid
- Try to avoid special characters in metadata
- Please run the validation function: AnnotationHubData::makeAnnotationHubMetadata or ExperimentHubData::makeExperimentHubMetadata

Acknowledgments

Research reported in this presentation is supported by the National Human Genome Research Institute and the National Cancer Institute of the National Institutes of Health under award numbers U24HG004059 (Bioconductor) and U24CA180996 (ITCR).

Core Team contributions throughout the years to create the Hub packages and database infrastructure

Everyone who contributes to the Hubs from the community

Questions?

Contact Information for further questions:

Lori Shepherd

lori.shepherd@roswellpark.org

Slack: @lshepherd