AnnotationHub

Liepa

3/27/2020

```
library(AnnotationHub)
```

```
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min

## Loading required package: BiocFileCache

## Loading required package: dbplyr
```

Create AnnotationHub object

```
ah <- AnnotationHub()
```

```
## snapshotDate(): 2019-10-29
```

```
ah
```

```
## AnnotationHub with 48092 records
## # snapshotDate(): 2019-10-29
## # $dataprovider: BroadInstitute, Ensembl, UCSC, ftp://ftp.ncbi.nlm.nih.gov/g...
## # $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus,...
## # $rdataclass: GRanges, BigWigFile, TwoBitFile, Rle, OrgDb, EnsDb, ChainFile...
## # 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
##   ...       ...
##   AH79558 | Xiphophorus_maculatus.X_maculatus-5.0-male.99.abinitio.gtf
##   AH79559 | Xiphophorus_maculatus.X_maculatus-5.0-male.99.chr.gtf
##   AH79560 | Xiphophorus_maculatus.X_maculatus-5.0-male.99.gtf
##   AH79561 | Zonotrichia_albicollis.Zonotrichia_albicollis-1.0.1.99.abinitio.gtf
##   AH79562 | Zonotrichia_albicollis.Zonotrichia_albicollis-1.0.1.99.gtf
```

data providers:

```
unique(ah$dataprovider)
```

```
##  [1] "UCSC"
##  [2] "Ensembl"
##  [3] "RefNet"
##  [4] "Inparanoid8"
##  [5] "NHLBI"
##  [6] "ChEA"
##  [7] "Pazar"
##  [8] "NIH Pathway Interaction Database"
##  [9] "Haemcode"
## [10] "BroadInstitute"
## [11] "PRIDE"
## [12] "Gencode"
## [13] "CRIBI"
## [14] "Genoscope"
## [15] "MISO, VAST-TOOLS, UCSC"
## [16] "UWashington"
## [17] "Stanford"
## [18] "dbSNP"
## [19] "BioMart"
## [20] "GeneOntology"
## [21] "KEGG"
## [22] "URGI"
## [23] "EMBL-EBI"
## [24] "MicrosporidiaDB"
## [25] "FungiDB"
## [26] "TriTrypDB"
## [27] "ToxoDB"
## [28] "AmoebaDB"
## [29] "PlasmoDB"
## [30] "PiroplasmaDB"
## [31] "CryptoDB"
## [32] "TrichDB"
## [33] "GiardiaDB"
## [34] "The Gene Ontology Consortium"
## [35] "ENCODE Project"
## [36] "SchistoDB"
## [37] "NCBI/UniProt"
## [38] "GENCODE"
## [39] "http://www.pantherdb.org"
## [40] "ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/"
```

```
head(unique(ah$species))
```

```
## [1] "Homo sapiens"         "Vicugna pacos"        "Dasypus novemcinctus"
## [4] "Otolemur garnettii"   "Papio hamadryas"      "Papio anubis"
```

```
length(unique(ah$species))
```

```
## [1] 2280
```

```
unique(ah$rdataclass)
```

```
##  [1] "GRanges"                          "data.frame"
##  [3] "Inparanoid8Db"                    "TwoBitFile"
##  [5] "ChainFile"                        "SQLiteConnection"
##  [7] "biopax"                           "BigWigFile"
##  [9] "AAStringSet"                      "MSnSet"
## [11] "mzRpwiz"                          "mzRident"
## [13] "list"                             "TxDb"
## [15] "Rle"                              "EnsDb"
## [17] "VcfFile"                          "igraph"
## [19] "data.frame, DNAStringSet, GRanges" "sqlite"
## [21] "data.table"                       "character"
## [23] "SQLite"                           "OrgDb"
```

```
dm <- query(ah, c("ChainFile", "UCSC", "Drosophila melanogaster"))
dm
```

```
## AnnotationHub with 45 records
## # snapshotDate(): 2019-10-29
## # $dataprovider: UCSC
## # $species: Drosophila melanogaster
## # $rdataclass: ChainFile
## # additional mcols(): taxonomyid, genome, description,
## #   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
## #   rdatapath, sourceurl, sourcetype
## # retrieve records with, e.g., 'object[["AH15102"]]'
##
##             title
##   AH15102 | dm3ToAnoGam1.over.chain.gz
##   AH15103 | dm3ToApiMel3.over.chain.gz
##   AH15104 | dm3ToDm2.over.chain.gz
##   AH15105 | dm3ToDm6.over.chain.gz
##   AH15106 | dm3ToDp3.over.chain.gz
##   ...       ...
##   AH15142 | dm2ToDroVir3.over.chain.gz
##   AH15143 | dm2ToDroWil1.over.chain.gz
##   AH15144 | dm2ToDroYak1.over.chain.gz
##   AH15145 | dm2ToDroYak2.over.chain.gz
##   AH15146 | dm1ToDm2.over.chain.gz
```

```
df <- mcols(dm)
df
```

```
## DataFrame with 45 rows and 15 columns
##                               title dataprovider                 species
##                         <character>  <character>             <character>
## AH15102 dm3ToAnoGam1.over.chain.gz         UCSC Drosophila melanogaster
## AH15103 dm3ToApiMel3.over.chain.gz         UCSC Drosophila melanogaster
## AH15104     dm3ToDm2.over.chain.gz         UCSC Drosophila melanogaster
## AH15105     dm3ToDm6.over.chain.gz         UCSC Drosophila melanogaster
## AH15106     dm3ToDp3.over.chain.gz         UCSC Drosophila melanogaster
## ...                            ...          ...                     ...
## AH15142 dm2ToDroVir3.over.chain.gz         UCSC Drosophila melanogaster
## AH15143 dm2ToDroWil1.over.chain.gz         UCSC Drosophila melanogaster
## AH15144 dm2ToDroYak1.over.chain.gz         UCSC Drosophila melanogaster
## AH15145 dm2ToDroYak2.over.chain.gz         UCSC Drosophila melanogaster
## AH15146     dm1ToDm2.over.chain.gz         UCSC Drosophila melanogaster
##         taxonomyid      genome                                  description
##          <integer> <character>                                  <character>
## AH15102       7227         dm3 UCSC liftOver chain file from dm3 to anoGam1
## AH15103       7227         dm3 UCSC liftOver chain file from dm3 to apiMel3
## AH15104       7227         dm3     UCSC liftOver chain file from dm3 to dm2
## AH15105       7227         dm3     UCSC liftOver chain file from dm3 to dm6
## AH15106       7227         dm3     UCSC liftOver chain file from dm3 to dp3
## ...            ...         ...                                          ...
## AH15142       7227         dm2 UCSC liftOver chain file from dm2 to droVir3
## AH15143       7227         dm2 UCSC liftOver chain file from dm2 to droWil1
## AH15144       7227         dm2 UCSC liftOver chain file from dm2 to droYak1
## AH15145       7227         dm2 UCSC liftOver chain file from dm2 to droYak2
## AH15146       7227         dm1     UCSC liftOver chain file from dm1 to dm2
##         coordinate_1_based
##                  <integer>
## AH15102                  0
## AH15103                  0
## AH15104                  0
## AH15105                  0
## AH15106                  0
## ...                    ...
## AH15142                  0
## AH15143                  0
## AH15144                  0
## AH15145                  0
## AH15146                  0
##                                                     maintainer rdatadateadded
##                                                    <character>    <character>
## AH15102 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15103 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15104 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15105 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15106 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## ...                                                       ...            ...
## AH15142 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15143 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15144 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15145 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
## AH15146 Bioconductor Maintainer <maintainer@bioconductor.org>     2014-12-15
##             preparerclass                    tags  rdataclass
##               <character>                  <list> <character>
## AH15102 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15103 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15104 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15105 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15106 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## ...                   ...                     ...         ...
## AH15142 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15143 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15144 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15145 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
## AH15146 UCSCChainPreparer liftOver,chain,UCSC,...   ChainFile
##                                                   rdatapath
##                                                 <character>
## AH15102 goldenPath/dm3/liftOver/dm3ToAnoGam1.over.chain.gz
## AH15103 goldenPath/dm3/liftOver/dm3ToApiMel3.over.chain.gz
## AH15104     goldenPath/dm3/liftOver/dm3ToDm2.over.chain.gz
## AH15105     goldenPath/dm3/liftOver/dm3ToDm6.over.chain.gz
## AH15106     goldenPath/dm3/liftOver/dm3ToDp3.over.chain.gz
## ...                                                    ...
## AH15142 goldenPath/dm2/liftOver/dm2ToDroVir3.over.chain.gz
## AH15143 goldenPath/dm2/liftOver/dm2ToDroWil1.over.chain.gz
## AH15144 goldenPath/dm2/liftOver/dm2ToDroYak1.over.chain.gz
## AH15145 goldenPath/dm2/liftOver/dm2ToDroYak2.over.chain.gz
## AH15146     goldenPath/dm1/liftOver/dm1ToDm2.over.chain.gz
##                                                                                    sourceurl
##                                                                                  <character>
## AH15102 http://hgdownload.cse.ucsc.edu/goldenpath/dm3/liftOver/dm3ToAnoGam1.over.chain.gz
## AH15103 http://hgdownload.cse.ucsc.edu/goldenpath/dm3/liftOver/dm3ToApiMel3.over.chain.gz
## AH15104     http://hgdownload.cse.ucsc.edu/goldenpath/dm3/liftOver/dm3ToDm2.over.chain.gz
## AH15105     http://hgdownload.cse.ucsc.edu/goldenpath/dm3/liftOver/dm3ToDm6.over.chain.gz
## AH15106     http://hgdownload.cse.ucsc.edu/goldenpath/dm3/liftOver/dm3ToDp3.over.chain.gz
## ...                                                                                     ...
## AH15142 http://hgdownload.cse.ucsc.edu/goldenpath/dm2/liftOver/dm2ToDroVir3.over.chain.gz
## AH15143 http://hgdownload.cse.ucsc.edu/goldenpath/dm2/liftOver/dm2ToDroWil1.over.chain.gz
## AH15144 http://hgdownload.cse.ucsc.edu/goldenpath/dm2/liftOver/dm2ToDroYak1.over.chain.gz
## AH15145 http://hgdownload.cse.ucsc.edu/goldenpath/dm2/liftOver/dm2ToDroYak2.over.chain.gz
## AH15146     http://hgdownload.cse.ucsc.edu/goldenpath/dm1/liftOver/dm1ToDm2.over.chain.gz
##          sourcetype
##         <character>
## AH15102       Chain
## AH15103       Chain
## AH15104       Chain
## AH15105       Chain
## AH15106       Chain
## ...             ...
## AH15142       Chain
## AH15143       Chain
## AH15144       Chain
## AH15145       Chain
## AH15146       Chain
```

```r
ahs <- query(ah, c('inparanoid8', 'ailuropoda'))
ahs
```

```
## AnnotationHub with 1 record
## # snapshotDate(): 2019-10-29
## # names(): AH10451
## # $dataprovider: Inparanoid8
## # $species: Ailuropoda melanoleuca
## # $rdataclass: Inparanoid8Db
## # $rdatadateadded: 2014-03-31
## # $title: hom.Ailuropoda_melanoleuca.inp8.sqlite
## # $description: Inparanoid 8 annotations about Ailuropoda melanoleuca
## # $taxonomyid: 9646
## # $genome: inparanoid8 genomes
## # $sourcetype: Inparanoid
## # $sourceurl: http://inparanoid.sbc.su.se/download/current/Orthologs/A.melan...
## # $sourcesize: NA
## # $tags: c("Inparanoid", "Gene", "Homology", "Annotation")
## # retrieve record with 'object[["AH10451"]]'
```

```
dm["AH15146"]
```

```
## AnnotationHub with 1 record
## # snapshotDate(): 2019-10-29
## # names(): AH15146
## # $dataprovider: UCSC
## # $species: Drosophila melanogaster
## # $rdataclass: ChainFile
## # $rdatadateadded: 2014-12-15
## # $title: dm1ToDm2.over.chain.gz
## # $description: UCSC liftOver chain file from dm1 to dm2
## # $taxonomyid: 7227
## # $genome: dm1
## # $sourcetype: Chain
## # $sourceurl: http://hgdownload.cse.ucsc.edu/goldenpath/dm1/liftOver/dm1ToDm...
## # $sourcesize: NA
## # $tags: c("liftOver", "chain", "UCSC", "genome", "homology")
## # retrieve record with 'object[["AH15146"]]'
```

(BiocManager::install("rtracklayer"))

```
dm[["AH15106"]]
```

```
## loading from cache
```

```
## require("rtracklayer")
```

```
## Chain of length 15
## names(15): chr2L chr2R chr3L chr3R chr4 ... chr3RHet chrXHet chrYHet chrUextra
```

```
ah_Hsapiens <- subset(ah, species == "Homo sapiens")
ah_Hsapiens
```

```
## AnnotationHub with 26104 records
## # snapshotDate(): 2019-10-29
## # $dataprovider: BroadInstitute, UCSC, Ensembl, GENCODE, UWashington, Stanfo...
## # $species: Homo sapiens
## # $rdataclass: GRanges, BigWigFile, Rle, ChainFile, TwoBitFile, list, data.f...
## # 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
##   ...       ...
##   AH78783 | Ensembl 99 EnsDb for Homo sapiens
##   AH79158 | Homo_sapiens.GRCh38.99.abinitio.gtf
##   AH79159 | Homo_sapiens.GRCh38.99.chr.gtf
##   AH79160 | Homo_sapiens.GRCh38.99.chr_patch_hapl_scaff.gtf
##   AH79161 | Homo_sapiens.GRCh38.99.gtf
```

```r
orgdb <- query(ah, c("OrgDb", "maintainer@bioconductor.org"))
length(orgdb$species)
```

```
## [1] 1708
```

```r
orgdb_mouse <- query(ah, c("OrgDb", "maintainer@bioconductor.org", "Mus musculus"))
orgdb_mouse
```

```
## AnnotationHub with 1 record
## # snapshotDate(): 2019-10-29
## # names(): AH75743
## # $dataprovider: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
## # $species: Mus musculus
## # $rdataclass: OrgDb
## # $rdatadateadded: 2019-10-29
## # $title: org.Mm.eg.db.sqlite
## # $description: NCBI gene ID based annotations about Mus musculus
## # $taxonomyid: 10090
## # $genome: NCBI genomes
## # $sourcetype: NCBI/ensembl
## # $sourceurl: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, ftp://ftp.ensembl.org/p...
## # $sourcesize: NA
## # $tags: c("NCBI", "Gene", "Annotation")
## # retrieve record with 'object[["AH75743"]]'
```

retrieve mouse data

```r
mouse <- orgdb_mouse[[1]]
```

```
## loading from cache
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
## The following object is masked from 'package:AnnotationHub':
##
##     cache
```

```
mouse
```

```
## OrgDb object:
## | DBSCHEMAVERSION: 2.1
## | Db type: OrgDb
## | Supporting package: AnnotationDbi
## | DBSCHEMA: MOUSE_DB
## | ORGANISM: Mus musculus
## | SPECIES: Mouse
## | EGSOURCEDATE: 2019-Jul10
## | EGSOURCENAME: Entrez Gene
## | EGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
## | CENTRALID: EG
## | TAXID: 10090
## | GOSOURCENAME: Gene Ontology
## | GOSOURCEURL: ftp://ftp.geneontology.org/pub/go/godatabase/archive/latest-lite/
## | GOSOURCEDATE: 2019-Jul10
## | GOEGSOURCEDATE: 2019-Jul10
## | GOEGSOURCENAME: Entrez Gene
## | GOEGSOURCEURL: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA
## | KEGGSOURCENAME: KEGG GENOME
## | KEGGSOURCEURL: ftp://ftp.genome.jp/pub/kegg/genomes
## | KEGGSOURCEDATE: 2011-Mar15
## | GPSOURCENAME: UCSC Genome Bioinformatics (Mus musculus)
## | GPSOURCEURL:
## | GPSOURCEDATE: 2019-Sep3
## | ENSOURCEDATE: 2019-Jun24
## | ENSOURCENAME: Ensembl
## | ENSOURCEURL: ftp://ftp.ensembl.org/pub/current_fasta
## | UPSOURCENAME: Uniprot
## | UPSOURCEURL: http://www.UniProt.org/
## | UPSOURCEDATE: Mon Oct 21 14:37:15 2019
```

```
##
## Please see: help('select') for usage information
```

```
keytypes <- keytypes(mouse)
keytypes
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
##  [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
## [11] "GO"           "GOALL"        "IPI"          "MGI"          "ONTOLOGY"
## [16] "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"         "PROSITE"
## [21] "REFSEQ"       "SYMBOL"       "UNIGENE"      "UNIPROT"
```

```
egid <- keys(mouse, "ENTREZID")
ids_paths_GO <- select(mouse, egid, c("SYMBOL", "UNIPROT", "GO", "PATH"), "ENTREZID")
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
head(ids_paths_GO, 30)
```

```
##    ENTREZID SYMBOL UNIPROT         GO EVIDENCE ONTOLOGY  PATH
## 1     11287    Pzp    <NA> GO:0002020      IBA       MF  <NA>
## 2     11287    Pzp    <NA> GO:0004866      IBA       MF  <NA>
## 3     11287    Pzp    <NA> GO:0004866      IDA       MF  <NA>
## 4     11287    Pzp    <NA> GO:0004867      IEA       MF  <NA>
## 5     11287    Pzp    <NA> GO:0005576      TAS       CC  <NA>
## 6     11287    Pzp    <NA> GO:0005615      IEA       CC  <NA>
## 7     11287    Pzp    <NA> GO:0007566      IGI       BP  <NA>
## 8     11287    Pzp    <NA> GO:0010466      IEA       BP  <NA>
## 9     11287    Pzp    <NA> GO:0030414      IEA       MF  <NA>
## 10    11287    Pzp    <NA> GO:0044877      ISO       MF  <NA>
## 11    11287    Pzp    <NA> GO:0048403      ISO       MF  <NA>
## 12    11287    Pzp    <NA> GO:0048406      ISO       MF  <NA>
## 13    11287    Pzp    <NA> GO:0062023      HDA       CC  <NA>
## 14    11298  Aanat  O88816 GO:0004059      IBA       MF 00380
## 15    11298  Aanat  O88816 GO:0004059      IBA       MF 01100
## 16    11298  Aanat  O88816 GO:0004059      ISO       MF 00380
## 17    11298  Aanat  O88816 GO:0004059      ISO       MF 01100
## 18    11298  Aanat  O88816 GO:0004060      IDA       MF 00380
## 19    11298  Aanat  O88816 GO:0004060      IDA       MF 01100
## 20    11298  Aanat  O88816 GO:0005737      IBA       CC 00380
## 21    11298  Aanat  O88816 GO:0005737      IBA       CC 01100
## 22    11298  Aanat  O88816 GO:0005737      ISO       CC 00380
## 23    11298  Aanat  O88816 GO:0005737      ISO       CC 01100
## 24    11298  Aanat  O88816 GO:0005829      ISO       CC 00380
## 25    11298  Aanat  O88816 GO:0005829      ISO       CC 01100
## 26    11298  Aanat  O88816 GO:0006474      ISO       BP 00380
## 27    11298  Aanat  O88816 GO:0006474      ISO       BP 01100
## 28    11298  Aanat  O88816 GO:0007623      IBA       BP 00380
## 29    11298  Aanat  O88816 GO:0007623      IBA       BP 01100
## 30    11298  Aanat  O88816 GO:0007623      ISO       BP 00380
```