

Bioconductor Packages For Cached File Management

BiocFileCache, AnnotationHub, ExperimentHub

BiocFileCache

Local File Management

Motivation:

It can be time consuming to download remote resource from the web. Let's design a way to check a local resource to see if it needs to be updated or not.

Let's also have a way to better organize local files

BiocFileCache()

- creates a cache object
- sqlite database backend
- add 'resources' (files) to the cache object to track

Cache Info:

- `bfcache ()`
- `length ()`
- `show ()`
- `bfcinfo ()`

Adding Resources:

- `bfcadd()`
- `bfcnew ()`

Removing Resources:

- `bfcremove ()`
- `bfcsync ()`

Investigating Resources:

- `bfcquerycols ()`
- `bfcquery ()`
- `bfccount ()`
- `bfcrid ()`
- `bfcpath ()`
- `bfcrcpath ()`
- `[`

Web Resources:

- `bfcneedsupdate ()`
- `bfcdownload ()`

Updating Resources:

- `bfcupdate ()`
- `[[`

MetaData:

- `bfcmetalists ()`
- `bfcmeta ()`
- `bfcmeta () <-`
- `bfcmetaremove ()`

Export/Import Cache:

- `importbfc ()`
- `exportbfc ()`
- `makeBiocFileCacheFromDataFrame()`

Clean/Remove Cache:

- `cleanbfc ()`
- `removebfc ()`

Example:

```
> BiocFileCache()  
class: BiocFileCache  
bfccache: /home/lori/.cache/BiocFileCache  
bfccount: 0  
For more information see: bfcinfo() or bfcquery()
```

```
> bfcinfo()  
# A tibble: 0 x 10  
# ... with 10 variables: rid <chr>, rname <chr>, create_time <dbl>,  
#   access_time <dbl>, rpath <chr>, rtype <chr>, fpath <chr>,  
#   last_modified_time <dbl>, etag <chr>, expires <dbl>
```

Example:

```
> bfcadd(rname="Wiki", fpath="https://en.wikipedia.org/wiki/Bioconductor")
|=====| 100%
                                     BFC1
"/home/lori/.cache/BiocFileCache/282e8be47f6_Bioconductor"

> bfcinfo()
# A tibble: 1 x 10
  rid    rname create_time  access_time rpath rtype fpath last_modified_t... etag
  <chr> <chr> <chr>          <chr>      <chr> <chr> <chr> <chr>          <chr>
1 BFC1 Wiki  2018-07-12 ... 2018-07-12... /hom... web    http... 2018-07-07 07:1... NA
# ... with 1 more variable: expires <chr>

> library(dplyr)
> bfcinfo() %>% select(last_modified_time, rpath)
# A tibble: 1 x 2
  last_modified_time  rpath
  <chr>              <chr>
1 2018-07-07 07:13:52 /home/lori/.cache/BiocFileCache/282e8be47f6_Bioconductor
```

Example:

```
> pathToSave = bfcnew(rname="My RDS File", ext=".rds")
```

```
> pathToSave
```

BFC2

```
"/home/lori/.cache/BiocFileCache/2feb30a96058_2feb30a96058.rds"
```

```
> bfcinfo()
```

```
# A tibble: 2 x 10
```

	rid	rname	create_time	access_time	rpath	rtype	fpath	last_modified_t...	etag
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	BFC1	Wiki	2018-07-12...	2018-07-12...	/hom...	web	http...	2018-07-07 07:1...	NA
2	BFC2	My RD...	2018-07-12...	2018-07-12...	/hom...	rela...	388d...	NA	NA

```
# ... with 1 more variable: expires <chr>
```

```
> saveRDS(myObj, file=pathToSave)
```

Example:

```
> bfcneedsupdate()  
BFC1  
TRUE
```


Utilizes functions from httr to capture Expires, Last-modified time, and Etag

1. HEAD()
2. cache_info()

```
> library(httr)
```

```
> cache_info(HEAD("https://en.wikipedia.org/wiki/Bioconductor"))
```

```
<cache_info> https://en.wikipedia.org/wiki/Bioconductor
Cacheable:    TRUE
Expires:      Thu, 12 Jul 2018 13:37:06 GMT <expired>
Last-Modified: Sat, 07 Jul 2018 07:13:52 GMT
Etag:
```

```
> cache_info(HEAD("https://bioconductor.org/packages/3.8/data/annotation/src/contrib/PANTHER.db_1.0.4.tar.gz"))
```

```
<cache_info> https://bioconductor.org/packages/3.8/data/annotation/src/contrib/PANTHER.db_1.0.4.tar.gz
Cacheable:    TRUE
Last-Modified: Wed, 27 Sep 2017 17:09:56 GMT
Etag:         "608b685-55a2edc70632a"
```

Example:

```
> bfcquery(query="RDS")
# A tibble: 1 x 10
  rid    rname  create_time access_time rpath rtype fpath last_modified_t... etag
  <chr> <chr>   <chr>         <chr>    <chr> <chr> <chr>          <dbl> <chr>
1 BFC2  My RD... 2018-07-12... 2018-07-12... /hom... rela... 388d...          NA NA
# ... with 1 more variable: expires <dbl>
```

```
> bfcrid(bfcquery(query="RDS"))
[1] "BFC2"
```

```
> bfcrpath(rids="BFC2")
```

```
                                BFC2
"/home/lori/.cache/BiocFileCache/2feb30a96058_2feb30a96058.rds"
```

```
> readRDS(bfcrpath(rids="BFC2"))
```

Example:

```
# data.frame or tibble

> meta = data.frame(rid="BFC2", info="pipeLine project X", numSamples=2000)

> bfc = BiocFileCache()

> bfcmeta(bfc, name="pipeLineXmeta") <- meta
> bfcmetalist()
[1] "pipeLineXmeta"

> library(dplyr)
> bfcinfo(bfc) %>% select(rid, rname, info, numSamples)
# A tibble: 2 x 4
   rid          rname          info numSamples
<chr>      <chr>      <chr>      <dbl>
1  BFC1         Wiki      <NA>         NA
2  BFC2 My RData File pipeLine project X    2000
```

Example:

```
> bfcquery(query="project X", field="info")
# A tibble: 1 x 12
  rid    rname  create_time access_time rpath  rtype fpath last_modified_t... etag
  <chr> <chr>   <chr>         <chr>         <chr> <chr> <chr>         <dbl> <chr>
1 BFC2  My RD... 2018-07-12... 2018-07-12... /hom... rela... 388d...         NA NA
# ... with 3 more variables: expires <dbl>, info <chr>, numSamples <dbl>
```

```
> bfcquerycols()
[1] "rid"           "rname"         "create_time"
[4] "access_time"  "rpath"         "rtype"
[7] "fpath"        "last_modified_time" "etag"
[10] "expires"      "info"          "numSamples"
```

AnnotationHub/ExperimentHub

AnnotationHub

AnnotationHub()

- creates a hub object
- sqlite database backend
- Files are stored remotely and downloaded as needed
 - Bioconductor AWS S3 Buckets
 - After downloaded, cached for quick access for future runs

Example:

```
> hub = AnnotationHub()  
snapshotDate(): 2018-06-27
```

```
> hub  
AnnotationHub with 44923 records  
# snapshotDate(): 2018-06-27  
# $dataprovder: BroadInstitute, Ensembl, UCSC, ftp://ftp.ncbi.nlm.nih.gov/g...  
# $species: Homo sapiens, Mus musculus, Drosophila melanogaster, Bos taurus,...  
# $rdataclass: GRanges, BigWigFile, FaFile, TwoBitFile, Rle, ChainFile, OrgD...  
# additional mcols(): taxonomyid, genome, description,  
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,  
#   rdatapath, sourceurl, sourcetype  
# retrieve records with, e.g., 'object[["AH2"]]'
```


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] 1879
```

```
> head(unique(tolower(hub$species)))  
[1] "ailuropoda melanoleuca" "anolis carolinensis" "bos taurus"  
[4] "caenorhabditis elegans" "callithrix jacchus" "canis familiaris"
```

```
> length(unique(hub$rdataclass))  
[1] 20
```

```
> unique(hub$rdataclass)  
[1] "FaFile" "GRanges" "data.frame" "Inparanoid8Db"  
[5] "TwoBitFile" "ChainFile" "SQLiteConnection" "biopax"  
[9] "BigWigFile" "AAStringSet" "MSnSet" "mzRpwiz"  
[13] "mzRident" "list" "TxDb" "Rle"  
[17] "EnsDb" "VcfFile" "OrgDb"
```

Example:

```
> query(hub, c("Homo sapien", "UCSC", "GRanges"))
AnnotationHub with 5788 records
# snapshotDate(): 2018-06-27
# $dataprovder: UCSC, Gencode
# $species: Homo sapiens
# $rdataclass: GRanges
# 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
...	...
AH27622	wgEncodeUwTfbsWi38CtcfStdPkRep2.narrowPeak.gz
AH49554	gencode.v23.2wayconspseudos.gff3.gz
AH53176	UCSC cytoBand track for hg18
AH53177	UCSC cytoBand track for hg19
AH53178	UCSC cytoBand track for hg38

Example:

```
> hub["AH53178"]
AnnotationHub with 1 record
# snapshotDate(): 2018-06-27
# names(): AH53178
# $dataprovder: UCSC
# $species: Homo sapiens
# $rdataclass: GRanges
# $rdatadateadded: 2017-01-05
# $title: UCSC cytoBand track for hg38
# $description: Approximate location of bands seen on Giemsa-stained chromos...
# $taxonomyid: 9606
# $genome: hg38
# $sourcetype: UCSC track
# $sourceurl: http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/cytoBa...
# $sourcesize: NA
# $tags: c("cytoBand", "AHCytoBands")
# retrieve record with 'object[["AH53178"]]'
```

Example:

```
> gr = hub[["AH53178"]]
```

```
downloading 1 resources
```

```
retrieving 1 resource
```

```
|=====| 100%
```

```
loading from cache
```

```
  '/home/lori//.AnnotationHub/59916'
```

```
> summary(gr)
```

```
[1] "GRanges object with 1293 ranges and 2 metadata columns"
```

```
> head(gr)
```

```
GRanges object with 6 ranges and 2 metadata columns:
```

	seqnames	ranges	strand	name	gieStain
	<Rle>	<IRanges>	<Rle>	<factor>	<factor>
[1]	chr1	[1, 2300000]	*	p36.33	gneg
[2]	chr1	[2300001, 5300000]	*	p36.32	gpos25
[3]	chr1	[5300001, 7100000]	*	p36.31	gneg
[4]	chr1	[7100001, 9100000]	*	p36.23	gpos25
[5]	chr1	[9100001, 12500000]	*	p36.22	gneg
[6]	chr1	[12500001, 15900000]	*	p36.21	gpos50

```
-----
```

```
seqinfo: 455 sequences from an unspecified genome; no seqlengths
```

Example:

```
> gr = hub[["AH53178"]]
downloading 1 resources
retrieving 1 resource
|=====| 100%
loading from cache
  '/home/lori//.AnnotationHub/59916'
```

```
> gr = hub[["AH53178"]]
downloading 0 resources
loading from cache
  '/home/lori//.AnnotationHub/59916'
```

Example:

```
> recordStatus(hub, "AH53178")
```

```
record status
```

```
1 AH53178 Public
```

```
> subset(hub, species == "Homo sapiens" & genome=="GRCh38" & rdataclass=="VcfFile")
```

```
AnnotationHub with 4 records
```

```
# snapshotDate(): 2018-06-27
```

```
# $dataprovder: dbSNP
```

```
# $species: Homo sapiens
```

```
# $rdataclass: VcfFile
```

```
# additional mcols(): taxonomyid, genome, description,
```

```
# coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
```

```
# rdatapath, sourceurl, sourcetype
```

```
# retrieve records with, e.g., 'object[["AH57960"]]'
```

```
title
```

```
AH57960 | clinvar_20160203.vcf.gz
```

```
AH57961 | clinvar_20160203_papu.vcf.gz
```

```
AH57962 | common_and_clinical_20160203.vcf.gz
```

```
AH57963 | common_no_known_medical_impact_20160203.vcf.gz
```

ExperimentHub

ExperimentHub()

- creates a hub object
- sqlite database backend
- Files are stored remotely and downloaded as needed
 - Bioconductor AWS S3 Buckets
 - After downloaded, cached for quick access for future runs

ExperimentHub data is associated with a Bioconductor package!

Example:

```
> eh = ExperimentHub()  
snapshotDate(): 2018-06-29
```

```
> length(eh)  
[1] 1233
```

```
> eh  
ExperimentHub with 1233 records  
# snapshotDate(): 2018-06-29  
# $dataprovder: Eli and Edythe L. Broad Institute of Harvard and MIT, NA, D...  
# $species: Homo Sapiens, Homo sapien, Homo sapiens, Mus musculus, Mus Muscu...  
# $rdataclass: ExpressionSet, SummarizedExperiment, RaggedExperiment, DataFr...  
# additional mcols(): taxonomyid, genome, description,  
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,  
#   rdatapath, sourceurl, sourcetype  
# retrieve records with, e.g., 'object[["EH1"]]'
```

Example:

```
> names(mcols(eh))
```

[1]	"title"	"dataprovder"	"species"
[4]	"taxonomyid"	"genome"	"description"
[7]	"coordinate_1_based"	"maintainer"	"rdatadateadded"
[10]	"preparerclass"	"tags"	"rdataclass"
[13]	"rdatapath"	"sourceurl"	"sourcetype"

```
> unique(eh$preparerclass)
```

[1]	"GSE62944"	"alpineData"	"CellMapperData"
[4]	"HumanAffyData"	"curatedMetagenomicData"	"SeqSQC"
[7]	"restfulSEData"	"curatedTCGAData"	"HarmonizedTCGAData"
[10]	"HMP16SData"	"TENxBrainData"	"MetaGxOvarian"
[13]	"CLLmethylation"	"tissueTreg"	"MetaGxBreast"
[16]	"HDCytoData"	"MetaGxPancreas"	"FlowSorted.Blood.EPIC"

Example:

```
> query(eh, "TENxBrainData")
ExperimentHub with 4 records
# snapshotDate(): 2018-06-29
# $dataprovder: 10X Genomics
# $species: 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[["EH1039"]]'
```

	title
EH1039	Brain scRNA-seq data, 'RLE-compressed'
EH1040	Brain scRNA-seq data, 'rectangular'
EH1041	Brain scRNA-seq data, sample (column) annotation
EH1042	Brain scRNA-seq data, gene (row) annotation

Example:

```
> query(eh, c("Mus musculus", "rna-seq"))
ExperimentHub with 7 records
# snapshotDate(): 2018-06-29
# $dataprovder: 10X Genomics, DKFZ
# $species: Mus musculus
# $rdaclass: character, SummarizedExperiment
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["EH1039"]]'
```

	title
EH1039	Brain scRNA-seq data, 'RLE-compressed'
EH1040	Brain scRNA-seq data, 'rectangular'
EH1041	Brain scRNA-seq data, sample (column) annotation
EH1042	Brain scRNA-seq data, gene (row) annotation
EH1074	RNA-seq data from tissue Tregs (RPKM values)
EH1075	RNA-seq data from tissue Tregs (htseq values)
EH1433	GEO accession data GSE71585 as a SingleCellExperiment

Example:

```
> # query(eh, c("Mus musculus", "rna-seq"))
```

```
> package(query(eh, c("Mus musculus", "rna-seq")))
```

EH1039	EH1040	EH1041	EH1042	EH1074
"TENxBrainData"	"TENxBrainData"	"TENxBrainData"	"TENxBrainData"	"tissueTreg"
EH1075	EH1433			
"tissueTreg"	"allenpvc"			

Example:

```
> unique(eh$rdataclass)
 [1] "ExpressionSet"          "GAlignmentPairs"
 [3] "CellMapperList"        "gds.class"
 [5] "RangedSummarizedExperiment" "GRanges"
 [7] "DataFrame"             "RaggedExperiment"
 [9] "SummarizedExperiment"  "list"
[11] "List"                  "Character"
[13] "data.frame"            "character"
[15] "bsseq"                 "flowSet"
[17] "RGChannelSet"
```


Selected rows:

Return rows to R session

Show entriesSearch:

idx	dataprovider	species	genome	description	coordinate_1_based	maintainer	rdatadateadded	preparerclass	tags	rdataclass	rdatapath
idx	dataprovider	species	genome	description	coordinate_1_based	maintainer	rdatadateadded	preparerclass	tags	rdataclass	rdatapath
EH1	GEO	Homo sapiens	hg19	TCGA RNA-seq Rsubread-summarized raw count data for 7706 tumor samples, represented as an ExpressionSet. R data representation derived from GEO accession GSE62944.	1	Bioconductor Package Maintainer <maintainer@bioconductor.org>	2016-02-23	GSE62944	TCGA, RNA-seq, Expression, Count	ExpressionSet	GSE62944/GSE629
EH166	GEUVADIS	Homo sapiens	GRCh38	Subset of aligned reads from sample ERR188297	1	Michael Love <michaelsaiahlove@gmail.com>	2016-07-21	alpineData	Sequencing, RNASeq, GeneExpression, Transcription	GAlignmentPairs	alpineData/ERR188
EH167	GEUVADIS	Homo sapiens	GRCh38	Subset of aligned reads from sample ERR188088	1	Michael Love <michaelsaiahlove@gmail.com>	2016-07-21	alpineData	Sequencing, RNASeq, GeneExpression, Transcription	GAlignmentPairs	alpineData/ERR188
EH168	GEUVADIS	Homo sapiens	GRCh38	Subset of aligned reads from sample ERR188204	1	Michael Love <michaelsaiahlove@gmail.com>	2016-07-21	alpineData	Sequencing, RNASeq, GeneExpression, Transcription	GAlignmentPairs	alpineData/ERR188
EH169	GEUVADIS	Homo sapiens	GRCh38	Subset of aligned reads from sample ERR188317	1	Michael Love <michaelsaiahlove@gmail.com>	2016-07-21	alpineData	Sequencing, RNASeq, GeneExpression, Transcription	GAlignmentPairs	alpineData/ERR188
EH170	Allen Brain Atlas	Homo sapiens	hg19	Large collection of microarrays on microdissected human brain samples from the Allen Brain Atlas, pre-processed for use with the CellMapper R package.	1	Brad Nelms <bnelms.research@gmail.com>	2016-08-08	CellMapperData	ExperimentData, MicroarrayData, ExpressionData	CellMapperList	CellMapperData/Bra

What's the advantage?

From a user perspective:

Public Accessible data!

Easy access to either more data or a second set of validation data

What's the advantage?

From a developer perspective:

Keeps the Package Lightweight!

Only download data as needed

Make large files accessible as simple objects

Resources are documented through package documentation

ExperimentHub Associated Package

<http://bioconductor.org/packages/devel/bioc/vignettes/ExperimentHubData/inst/doc/CreateAnExperimentHubPackage.html>

Requires inst/scripts/metadata.csv

- Gives the information provided when you view a resources (the mcols)

Requires inst/scripts/make-data.R

- Shows the preprocessing of raw files to R objects for reproducibility

R files/functions

- Potentially can construct complex structures from simple objects, behind the scene
- Helper functions to download data directly

User Application Demo by Levi...

Lori Shepherd

Bioconductor Core Team

lori.shepherd@roswellpark.org