

# restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

*Vincent J. Carey, stvjc at channing.harvard.edu, Shweta Gopaulakrishnan, reshg at channing.harvard.edu, Samuela Pollack, spollack at jimmy.harvard.edu*

April 30, 2018

## Contents

1	restfulSE . . . . .	2
1.1	HDF5 server backed SummarizedExperiment . . . . .	2
2	Some details . . . . .	5
2.1	Motivation . . . . .	5
2.2	Executive summary . . . . .	5
2.3	10xGenomics examples . . . . .	7
2.4	Background . . . . .	9
2.5	Hierarchy of server resources . . . . .	10

# 1 restfulSE

---

This R package includes proof-of-concept code illustrating several approaches to Summarized-Experiment design with assays stored out-of-memory.

## 1.1 HDF5 server backed SummarizedExperiment

**HDF Server** “extends the HDF5 data model to efficiently store large data objects (e.g. up to multi-TB data arrays) and access them over the web using a RESTful API.” In this `restfulSE` package, several data structures are introduced

- to model the server data architecture and
- to perform targeted extraction of numerical data from HDF5 arrays stored on the server.

We maintain, thanks to a grant from the National Cancer Institute, the server <http://h5s.channingremotedata.org:5000/>. Visit this URL to get a flavor of the server structure: datasets, groups, and datatypes are high-level elements to be manipulated to work with data values from the server.

### 1.1.1 Illustration with 10x genomics 1.3 million neurons

We used Martin Morgan’s `TENxGenomics` package to transform the sparse-formatted HDF5 supplied by 10x into a dense HDF5 matrix to support natural slicing. Thanks to native compression in HDF5, the data volume expansion is modest.

A helper function in the `restfulSE` package creates a `RESTfulSummarizedExperiment` instance that points to the full numerical dataset.

```
library(restfulSE)
my10x = se1.3M()
## analyzing groups for their links...
## done
## snapshotDate(): 2018-04-27
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation
## downloading 0 resources
## loading from cache
##      '/home/biocbuild//.ExperimentHub/554'
my10x
## class: RangedSummarizedExperiment
## dim: 27998 1306127
## metadata(0):
## assays(1): counts
## rownames(27998): ENSMUSG00000051951 ENSMUSG00000089699 ...
##      ENSMUSG00000096730 ENSMUSG00000095742
## rowData names(2): geneid symbol
## colnames(1306127): AACCTGAGATAGGAG-1 AACCTGAGCGGCTTC-1 ...
##      TTTGTCAGTTAAAGTG-133 TTTGTCATCTGAAAGA-133
## colData names(4): Barcode Sequence Library Mouse
```

As an exercise, we acquire the ENSEMBL identifiers for mouse genes annotated to hippocampus development, which has GO ID [GO:0021766](#), and check counts for 10 genes on 6 samples:

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
library(org.Mm.eg.db)
##
hippdev = select(org.Mm.eg.db,
  keys="G0:0021766", keytype="G0", column="ENSEMBL")$ENSEMBL
## 'select()' returned 1:many mapping between keys and columns
hippdev = intersect(hippdev, rownames(my10x))
unname(assay(my10x[ hippdev[1:10], 10001:10006]))
## <10 x 6> DelayedMatrix object of type "double":
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]  0  0  0  0  0  0
## [2,]  0  0  0  0  0  0
## [3,]  0  0  0  1  0  0
## [4,]  0  1  2  6  5  0
## [5,]  0  0  0  0  0  0
## [6,]  1  2  4  8  7  3
## [7,]  0  0  0  0  0  0
## [8,]  0  0  0  0  0  2
## [9,]  0  0  0  0  0  0
## [10,] 3  0  3  0  1  9
```

The result:

```
      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  0  0  0  0  0  0
[2,]  0  0  0  0  0  0
[3,]  0  0  0  1  0  0
[4,]  0  1  2  6  5  0
[5,]  0  0  0  0  0  0
[6,]  1  2  4  8  7  3
[7,]  0  0  0  0  0  0
[8,]  0  0  0  0  0  2
[9,]  0  0  0  0  0  0
[10,] 3  0  3  0  1  9
```

### 1.1.2 Illustration with GTEx tissue expression

We exported the content of the [recount2 GTEx gene-level quantifications](#) to our HDF5 server. A convenience function is available:

```
tiss = gtexTiss()
## analyzing groups for their links...
## done
## snapshotDate(): 2018-04-27
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation
## downloading 0 resources
## loading from cache
##      '/home/biocbuild//.ExperimentHub/556'
tiss
## class: RangedSummarizedExperiment
## dim: 58037 9662
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
## metadata(0):  
## assays(1): recount  
## rownames(58037): ENSG00000000003.14 ENSG00000000005.5 ...  
## ENSG00000283698.1 ENSG00000283699.1  
## rowData names(3): gene_id bp_length symbol  
## colnames(9662): SRR660824 SRR2166176 ... SRR612239 SRR615898  
## colData names(82): project sample ... title characteristics
```

We'll use this remote data as a tool for investigating transcriptional patterns in brain anatomy. We can identify the samples from brain using the 'smtsd' colData element:

```
binds = grep("Brain", tiss$smtsd)  
table(tiss$smtsd[binds][1:100]) # check diversity in 100 samples  
##  
## Brain - Amygdala  
## 4  
## Brain - Anterior cingulate cortex (BA24)  
## 5  
## Brain - Caudate (basal ganglia)  
## 10  
## Brain - Cerebellar Hemisphere  
## 9  
## Brain - Cerebellum  
## 13  
## Brain - Cortex  
## 13  
## Brain - Frontal Cortex (BA9)  
## 10  
## Brain - Hippocampus  
## 8  
## Brain - Hypothalamus  
## 5  
## Brain - Nucleus accumbens (basal ganglia)  
## 7  
## Brain - Putamen (basal ganglia)  
## 4  
## Brain - Spinal cord (cervical c-1)  
## 3  
## Brain - Substantia nigra  
## 9
```

We'll identify genes annotated to neurotrophic functions using another convenience function in this package:

```
ntgenes = goPatt(termPattern="neurotroph")  
## 'select()' returned 1:1 mapping between keys and columns  
## 'select()' returned 1:many mapping between keys and columns  
head(ntgenes)  
## GO EVIDENCE ONTOLOGY ENSEMBL SYMBOL  
## 1 GO:0004897 IDA MF ENSG00000122756 CNTFR  
## 2 GO:0004897 IMP MF ENSG00000160712 IL6R  
## 3 GO:0004897 IDA MF ENSG00000134352 IL6ST
```

## 4	G0:0004897	IDA	MF	ENSG00000113594	LIFR
## 5	G0:0043121	ISS	MF	ENSG00000064300	NGFR
## 6	G0:0043121	IBA	MF	ENSG00000198400	NTRK1

## 2 Some details

### 2.1 Motivation

Extensive human and computational effort is expended on downloading and managing large genomic data at site of analysis. Interoperable formats that are accessible via generic operations like those in RESTful APIs may help to improve cost-effectiveness of genome-scale analyses.

In this report we examine the use of HDF5 server as a back end for assay data, mediated through the RangedSummarizedExperiment API for interactive use.

A modest server configured to deliver HDF5 content via a RESTful API has been prepared and is used in this vignette.

### 2.2 Executive summary

We want to provide rapid access to array-like data. We'll work with the Banovich 450k data as there is a simple check against an in-memory representation.

```
library(restfulSE)
bigec2 = H5S_source("http://h5s.channingremotedata.org:5000")
## analyzing groups for their links...
## done
bigec2
## H5serv server url : http://h5s.channingremotedata.org:5000
## There are 10 groups.
## Use groups(), links(), ..., to probe and access metadata.
## Use dsmeta() to get information on datasets within groups.
## Use [[ [dsname] ]] to get a reference suitable for [i, j] subsetting.
dsmeta(bigec2)[1:2,] # two groups
## DataFrame with 2 rows and 3 columns
##   groupnum
##   <integer>
## 1         1
## 2         2
##
##                                     dsnames
##                                     <CharacterList>
## 1                                     tenx_400k_sorted,mike,darmgcls,...
## 2 newlinkedtype.datatypetest.test,compound.datatypetest.test,newdtype.datatypetest.test,...
##                                     grp.uuid
##                                     <character>
## 1 ea87fd6e-ffba-11e7-985f-0aa4f45de806
## 2 ea8901c8-ffba-11e7-985f-0aa4f45de806
dsmeta(bigec2)[1,2][[1]] # all dataset candidates in group 1
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
## [1] "tenx_400k_sorted" "mike" "darmgcls"  
## [4] "neurons400k" "tenx_full" "tissues"  
## [7] "assays" "patelGBMSC" "neurons100k"  
## [10] "tenx_100k_sorted"
```

We use double-bracket subscribing to grab a reference to a dataset from an H5S source.

```
banref = bigec2[["assays"]] # arbitrary name assigned long ago  
banref  
## H5S_dataset instance:  
## dsname intl.dim1 intl.dim2 created type.base  
## 1 assays 64 329469 2017-04-05T18:02:37Z H5T_IEEE_F64LE
```

We build a SummarizedExperiment by combining an assay-free RangedSummarizedExperiment with this reference.

```
ehub = ExperimentHub::ExperimentHub()  
## snapshotDate(): 2018-04-27  
myfiles <- AnnotationHub::query(ehub, "restfulSEData")  
banoSE = myfiles[["EH551"]] # no assay data  
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation  
## downloading 0 resources  
## loading from cache  
## '/home/biocbuild//.ExperimentHub/551'  
ds = H5S_Array("http://h5s.channingremotedata.org:5000", "assays")  
## analyzing groups for their links...  
## done  
assays(banoSE) = SimpleList(betas=ds)  
banoSE  
## class: RangedSummarizedExperiment  
## dim: 329469 64  
## metadata(0):  
## assays(1): betas  
## rownames(329469): cg000000029 cg000000165 ... ch.9.98989607R  
## ch.9.991104F  
## rowData names(10): addressA addressB ... probeEnd probeTarget  
## colnames(64): NA18498 NA18499 ... NA18489 NA18909  
## colData names(35): title geo_accession ... data_row_count naid
```

We can update the SummarizedExperiment metadata through subsetting operations, and then extract the relevant assay data. The data are retrieved from the remote server with the `assay` method.

```
rbanoSub = banoSE[5:8, c(3:9, 40:50)]  
## Loading required package: Biostrings  
## Loading required package: XVector  
##  
## Attaching package: 'Biostrings'  
## The following object is masked from 'package:DelayedArray':  
##  
## type  
## The following object is masked from 'package:base':
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
##
##   strsplit
## assay(rbanoSub)
## <4 x 18> DelayedMatrix object of type "double":
##           NA18501      NA18502      NA18516 ...      NA19138
## cg00000363  0.325433263  1.377820005  0.596699897 . 0.966695669
## cg00000622  0.003436888 -0.668499289 -1.210634762 . 0.076062477
## cg00000714 -1.184443665 -1.654047967 -0.174729357 . 0.325742947
## cg00000734  0.153831565 -1.299289359  1.903976827 . 1.185320424
##           NA19140
## cg00000363  1.203765271
## cg00000622  0.958031578
## cg00000714 -0.008202908
## cg00000734  0.319937329
```

## 2.3 10xGenomics examples

### 2.3.1 t-SNE for a set of genes annotated to hippocampus

We have used Martin Morgan's `TENxGenomics` package to create a dense HDF5 representation of the assay data, and placed it on the `bigec2` server. The metadata are available as `se100k` in this package; we have used `EnsDb.Mmusculus.v79` to supply gene ranges where available; genes reported but without addresses are addressed at `chr1:2` with width 0. The rows are sorted by genomic address within chromosomes.

```
tenx100k = se100k()
## analyzing groups for their links...
## done
## snapshotDate(): 2018-04-27
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation
## downloading 0 resources
## loading from cache
##   '/home/biocbuild//.ExperimentHub/552'
tenx100k
## class: RangedSummarizedExperiment
## dim: 27998 100000
## metadata(1): source
## assays(1): counts
## rownames(27998): ENSMUSG000000109048 ENSMUSG000000109510 ...
##   ENSMUSG000000096768 ENSMUSG000000096850
## rowData names(6): gene_id gene_name ... seq_coord_system symbol
## colnames(100000): AAACCTGAGATAGGAG-1 AAACCTGAGCGGCTTC-1 ...
##   GACGTTAGTCATACTG-11 GACGTTAGTCCGTGAC-11
## colData names(4): Barcode Sequence Library Mouse
```

We will subset genes annotated to hippocampus development. Here are some related categories:

```
12092 G0:0021766      hippocampus development
12096 G0:0021770      parahippocampal gyrus development
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

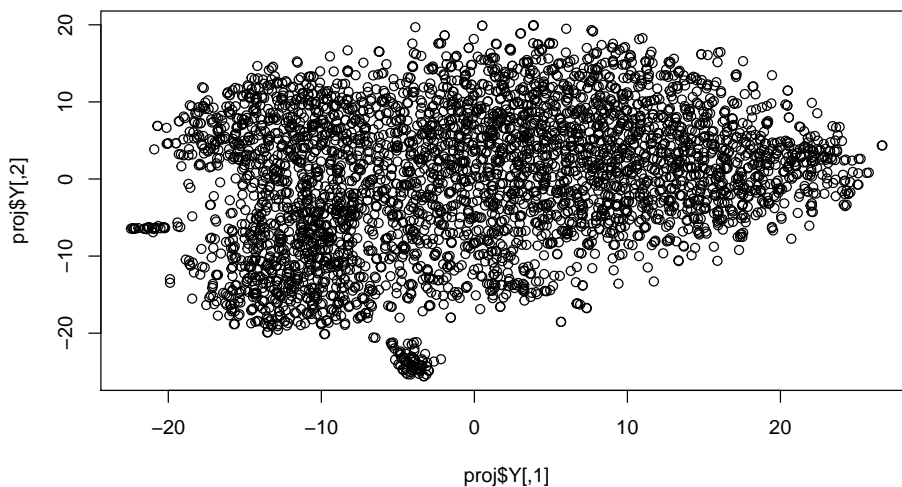
```
34609 G0:0097410 hippocampal interneuron differentiation
34631 G0:0097432 hippocampal pyramidal neuron differentiation
34656 G0:0097457 hippocampal mossy fiber
35169 G0:0098686 hippocampal mossy fiber to CA3 synapse
42398 G0:1990026 hippocampal mossy fiber expansion
```

```
library(org.Mm.eg.db)
atab = select(org.Mm.eg.db, keys="G0:0021766", keytype="G0", columns="ENSEMBL")
## 'select()' returned 1:many mapping between keys and columns
hg = atab[, "ENSEMBL"]
length(hgok <- intersect(hg, rownames(tenx100k)))
## [1] 50
```

This is a very scattered collection of rows in the matrix. We acquire expression measures for genes annotated to hippocampus on 4000 samples. t-SNE is then used to project the log-transformed measures to the plane.

```
hipn = assay(tenx100k[hgok, 1:4000]) # slow
d = dist(t(log(1+hipn)), method="manhattan")
proj = Rtsne(d)
```

```
plot(proj$Y)
```



### 2.3.2 A set of genes related to the visual cortex

Tasic et al. (Nature neuro 2016, DOI 10.1038/nn.4216) describe single cell analysis of the adult murine brain, identify clusters of cells with distinct transcriptional profiles and anatomic location, and enumerate lists of genes that discriminate these clusters. The tasicST6 DataFrame provides details.

```
#data("tasicST6", package = "restfulSEData")
ehub = ExperimentHub::ExperimentHub()
## snapshotDate(): 2018-04-27
myfiles <- AnnotationHub::query(ehub, "restfulSEData")
```



## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
myfiles[["EH557"]] -> tasicST6
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation
## downloading 0 resources
## loading from cache
##   '/home/biocbuild//.ExperimentHub/557'
tasicST6
## DataFrame with 49 rows and 4 columns
##      clid      txtype1      txtype2
##   <character> <character> <character>
## 1          f01         Vip         Mybpc1
## 2          f02         Vip          Parm1
## 3          f03         Vip          Sncg
## 4          f04         Vip          Chat
## 5          f05         Vip          Gpc3
## ...      ...          ...          ...
## 45         f45         Oligo 9630013A20Rik
## 46         f46         Oligo          Opalin
## 47         f47         Micro          Ctss
## 48         f48         Endo          Xdh
## 49         f49         SMC          Myl9
##
##
## 1
## 2
## 3
## 4
## 5
## ...
## 45                                     c("Brca1", "Rnf122", "Mbp", "Zcchc12", "Er
## 46
## 47                                     c("Cx3cr1", "Clqb", "Cd53", "Csf1r", "Itgam", "Abi3", "Clqa",
## 48                                     c("Tbc1d4", "AI467606", "Exosc7", "Eltd1", "Fas", "Hmgcs2", "Nos
## 49 c("Bgn", "Nupr1", "Casq2", "Mylk", "Gprc5c", "Slc38a11", "Slc6a20a", "Pcolce", "Vtn", "Cnn2", "Nid1",
```

Key high-level discrimination concerns cells regarded as GABAergic vs. glutamatergic (inhibitory vs excitatory neurotransmission).

## 2.4 Background

Banovich et al. published a subset of DNA methylation measures assembled on 64 samples of immortalized B-cells from the YRI HapMap cohort.

```
library(restfulSE)
#data("banoSEMeta", package = "restfulSEData")
ehub = ExperimentHub::ExperimentHub()
## snapshotDate(): 2018-04-27
myfiles <- AnnotationHub::query(ehub, "restfulSEData")
myfiles[["EH551"]] -> banoSEMeta
## see ?restfulSEData and browseVignettes('restfulSEData') for documentation
## downloading 0 resources
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
## loading from cache
##   '/home/biocbuild//.ExperimentHub/551'
banoSEMeta
## class: RangedSummarizedExperiment
## dim: 329469 64
## metadata(0):
## assays(0):
## rownames(329469): cg00000029 cg00000165 ... ch.9.98989607R
##   ch.9.991104F
## rowData names(10): addressA addressB ... probeEnd probeTarget
## colnames(64): NA18498 NA18499 ... NA18489 NA18909
## colData names(35): title geo_accession ... data_row_count naid
```

The numerical data have been exported using H. Pages' `saveHDF5SummarizedExperiment` applied to the `banovichSE` `SummarizedExperiment` in the `yriMulti` package. The HDF5 component is simply copied into the server data space on the remote server.

## 2.5 Hierarchy of server resources

### 2.5.1 Server

Given the URL of a server running HDF5 server, we create an instance of `H5S_source`:

```
mys = H5S_source(serverURL="http://h5s.channingremotedata.org:5000")
## analyzing groups for their links...
## done
mys
## H5serv server url : http://h5s.channingremotedata.org:5000
## There are 10 groups.
## Use groups(), links(), ..., to probe and access metadata.
## Use dsmeta() to get information on datasets within groups.
## Use [[ [dsname] ]] to get a reference suitable for [i, j] subsetting.
```

### 2.5.2 Groups

The server identifies a collection of 'groups'. For the server we are working with, only one group, at the root, is of interest.

```
groups(mys)
## DataFrame with 10 rows and 2 columns
##           groups      nlinks
##           <character> <integer>
## 1 ea87fd6e-ffba-11e7-985f-0aa4f45de806      12
## 2 ea8901c8-ffba-11e7-985f-0aa4f45de806       4
## 3 ea88d144-ffba-11e7-985f-0aa4f45de806      28
## 4 ea88a142-ffba-11e7-985f-0aa4f45de806       3
## 5 ea885872-ffba-11e7-985f-0aa4f45de806      87
## 6 ea88b8b2-ffba-11e7-985f-0aa4f45de806       1
## 7 ea887096-ffba-11e7-985f-0aa4f45de806       1
```

## restfulSE – experiments with HDF5 server content wrapped in SummarizedExperiment

```
## 8 ea88ea1c-ffba-11e7-985f-0aa4f45de806 1
## 9 ea882fdc-ffba-11e7-985f-0aa4f45de806 1
## 10 ea888914-ffba-11e7-985f-0aa4f45de806 1
```

### 2.5.3 Links for a group

There is a class to hold the link set for any group:

```
lin1 = rhdf5client::links(mys,1)
lin1
## HDF5 server link set for group ea87fd6e-ffba-11e7-985f-0aa4f45de806
## There are 12 links.
## Use targets([linkset]) to extract target URLs.
```