Recent Cloud-scale Bioconductor Innovations

S Gopaulakrishnan, S Pollack, A Culhane, BJ Stubbs, S Davis, V Carey

2018-05-23

Overview

The potential effectiveness of cloud computing in genome biology is well-established[1]. This document reviews relatively new approaches to using Bioconductor's data structures to work with remote cloud-scale resources. **All demonstrations are based on the "devel branch" of Bioconductor in R 3.5.**

[1] Langmead, Ben, and Abhinav Nellore. 2018. *Cloud computing for genomic data analysis and collaboration.* Nature Reviews Genetics. Nature Publishing Group. doi:10.1038/nrg.2017.113.

HDF server for SummarizedExperiment assays

We have created an HDF5 server at `h5s.channingremotedata.org:5000` whose use is described in the `rhdf5client` package vignette (devel branch until October 2018).

The following code can be used to work with the Recount2[2] version of the GTEx RNA-seq tissue expression data.

[2] Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E. Ellis, Margaret A. Taub, Kasper D. Hansen, Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2017. *Reproducible RNA-seq analysis using recount2.* Nature Biotechnology 35 (4): 319–21. doi:10.1038/nbt.3838.

```
library(restfulSE)
gt = gtexTiss()
gt
```

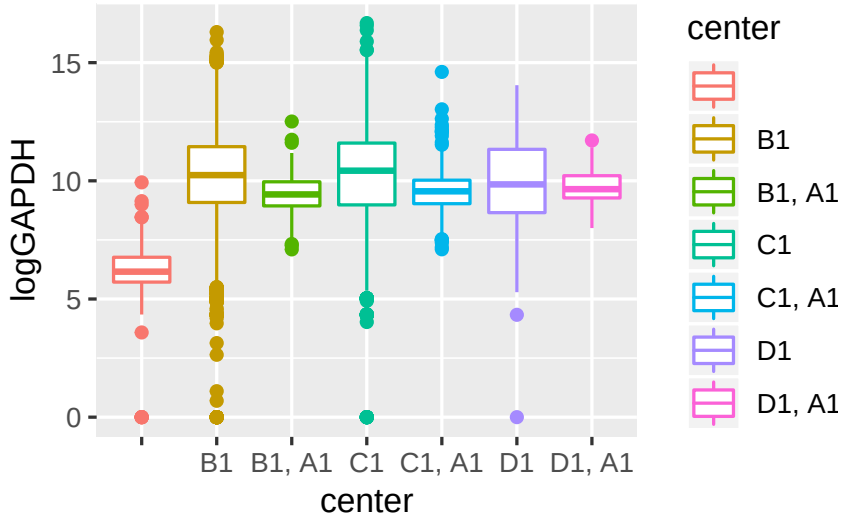
```
## class: RangedSummarizedExperiment
## dim: 58037 9662
## metadata(0):
## assays(1): recount
## rownames(58037): ENSG00000000003.14 ENSG00000000005.5 ...
##   ENSG00000283698.1 ENSG00000283699.1
## rowData names(3): gene_id bp_length symbol
## colnames(9662): SRR660824 SRR2166176 ... SRR612239 SRR615898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

```
table(gt$smts)
```

```
##
##                  Adipose Tissue   Adrenal Gland         Bladder           Blood
##               5             620             159              11             595
##    Blood Vessel     Bone Marrow           Brain          Breast    Cervix Uteri
##             750             102            1409             218              11
## ...
```

A quick visual check of center labeling and normalization success:

```
gind = which(unlist(rowData(gt)$symbol) == "GAPDH")
mydf = data.frame(logGAPDH=log(as.numeric(assay(gt[gind,]))+1),
  center=gt$smcenter)
library(ggplot2)
ggplot(mydf, aes(x=center, y=logGAPDH, colour=center)) + geom_boxplot()
```



Other resources conveyed via this HDF5 Server include the 10x Genomics 1.3 million neuron dataset (use `restfulSE::se1.3M()`), CONQUER-based quantifications[3] of single-cell RNA-seq studies by Patel et al.[4] and Darmanis et al.[5] and the "tabula muris"[6] quantifications distributed at the Human Cell Atlas. Check with pamphlet authors for more details.

HDF Cloud for DelayedArrays

HDF Cloud is newer and more scalable than HDF5 Server, but is not yet open source. We are experimenting with HDF Cloud thanks to support from John Readey of the HDF Group. You can start working with the 1.3 million neuron data through the DelayedArray interface as follows. We illustrate summation of selected columns.

```
dm10x = HSDS_Matrix("http://52.4.181.237:5101",
   "/home/stvjc/tenx_full.h5")
dim(dm10x)

## [1]   27998 1306127

colSums(dm10x[,1:6])

## [1]  4046  2087  4654  3193  8444 11178
```

[3] Soneson, Charlotte, and Mark D. Robinson. 2018. *Bias, robustness and scalability in single-cell differential expression analysis.* Nature Methods 15 (4). Nature Publishing Group: 255–61. doi:10.1038/nmeth.4612.

[4] Patel, Anoop P, Itay Tirosh, John J Trombetta, Alex K Shalek, M Shawn, Hiroaki Wakimoto, Daniel P Cahill, et al. 2014. *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science 344 (6190): 1396–1401. doi:10.1126/science.1254257.

[5] Darmanis, Spyros, Steven A. Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, et al. 2017. *Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma.* Cell Reports 21 (5). ElsevierCompany.: 1399–1410. doi:10.1016/j.celrep.2017.10.030.

[6] Quake, Stephen R., Tony Wyss-Coray, and Spyros Darmanis. 2017. *Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris.* bioRxiv, 237446. doi:10.1101/237446.

The result of `HSDS_Matrix` can be used as the assay component of a SummarizedExperiment.

To *write* data to this instance of HDF Cloud, you will need privileges. Given any HDF5 dataset, the utilities for porting to the cloud are part of the h5pyd python package.

Google BigQuery for MultiAssayExperiments derived from the Pan-Cancer Atlas

The restfulSE and BiocOntoTK packages help authenticated users to derive SummarizedExperiment instances for TCGA and PanCancerAtlas with Google BigQuery back ends, as managed by the Institute for Systems Biology.

```
library(BiocOncoTK)
bq = pancan_BQ() # authenticate
brca_mir = pancan_SE(bq) # defaults are BRCA micro-RNA assays
brca_mir
```

```
## class: SummarizedExperiment
## dim: 743 1068
## metadata(0):
## assays(1): assay
## rownames(743): hsa-miR-9-3p hsa-miR-191-3p ... hsa-miR-516a-5p
##   hsa-miR-892b
## rowData names(0):
## colnames(1068): TCGA-3C-AAAU TCGA-3C-AALJ ... TCGA-Z7-A8R6 TCGA-Z7-A8R5
## colData names(746): bcr_patient_uuid bcr_patient_barcode ...
##   bilirubin_upper_limit days_to_last_known_alive
```

The BiocOncoTK vignette indicates how to construct MultiAssayExperiment[7] instances on the basis of SummarizedExperiment instances like this.

[7] Ramos, Marcel, Lucas Schiffer, Angela Re, Rimsha Azhar, Azfar Basunia, Rodriguez Cabrera, Tiffany Chan, Philip Chapman, Sean Davis, and David Gomez-cabrero. 2017. *Software for the integration of multi-omics experiments in Bioconductor.* Cancer Research 77: e39–e42. doi:10.1158/0008-5472.CAN-17-0344.

The omicdx/BigRNA project

As the number of high throughput datasets available in public repositories grows, flexible, performant search engines that expose full metadata records become increasingly important to researchers. In addition, access to computable metadata allows researchers to enhance the value of data by augmenting the metadata with other data resources, machine learning approaches, and data "mashups".

The Sequence Read Archive (SRA) is the largest repository of publicly available sequencing data. We have mined the entire metadata set from the SRA using the open source Big Data platform, Apache

Spark. After extracting and transforming these metadata from NCBI files, we have loaded them into the high-performance search and analytics Elasticsearch engine. These metadata are exposed via a RESTful application programming interface (API), conforming to the OpenAPI specification[8], facilitating client template development in any of dozens of supported languages. Finally, we have developed an R-based client that provides a bridge between the SRA resource and the Bioconductor developer and user community.

[8] https://swagger.io/docs/specification/about/

The metadata are a key component for organizing results of uniform reprocessing of all RNA-seq experiments. An initial demonstration of this process generates a SummarizedExperiment with assay data in HDF Cloud for 1222 RNA-seq runs on glioblastoma samples.

```
## class: RangedSummarizedExperiment
## dim: 200401 1222
## metadata(3): origin sessionInfo limitation
## assays(5): abundance count_lstpm length bsmedian bsMAD
## rownames(200401): ENST00000456328.2 ENST00000450305.2 ...
##   ENST00000387460.2 ENST00000387461.2
## rowData names(6): source gene_id ... transcript_support_level
##   protein_id
## colnames(1222): SRX1558818 SRX1558809 ... SRX683527 SRX683523
## colData names(53): experiment_Insdc experiment_LastMetaUpdate ...
##   study_study_type study_title
```

With this object users can decompose the data by study or combine data across studies using familiar R idioms. The omicidx resource provides the metadata; below we print titles and sample counts for the eight largest studies, filtering the `colData` of the SummarizedExperiment.

```
##   accession                                                 title n.expt
## 1  SRP057500 RNA-seq of tumor-educated platelets enables blood-...    285
## 2  ERP013498                      CRISPR_Screening_in_Glioblastoma    109
## 3  SRP044668 MRI-localized biopsies reveal subtype-specific dif...     90
## 4  SRP109992 Sox5/6/21 prevent oncogene driven transformation o...     72
## 5  SRP092795 Ion Channel Expression Patterns in Glioblastoma St...     69
## 6  SRP072494 Transcriptional changes induced by bevacizumab com...     36
## 7  SRP069235       Gene expression in human glioblastoma specimens     32
## 8  SRP087439 Genome-wide mRNA expression in human glioblastoma ...     29
```

Funding support

This work was supported by NCI U01 CA214846, V. Carey, PI, NCI U24 CA180996, M. Morgan PI. Work of Sean Davis was supported by the NCI Center for Cancer Research.