SOUND
Statistical Multi-Omics Understanding

Project Deliverable

Project acronym: **SOUND**	GA number: **633974**
Project title: **Statistical Multi-Omics Understanding of Patient Data**	
Funding Scheme: Collaborative Project (H2020-PHC-2014-2015/H2020-PHC-2014-two-stage) Health, novel medical developments	
Project start date: **01 September 2015**	Duration: **36 months**
Project's coordinator: Dr Wolfgang Huber (European Molecular Biology Laboratory, Heidelberg)	

D8.2 SOUNDData and SOUNDHub public data resources

Due date of deliverable: Month 18 – 28.02.2017
Actual submission date: 23.08.2018

Organization name of lead contractor for this deliverable: Roswell Park Cancer Institute (RPCI)
Organization name of other involved partners:
Personnel involved: Shepherd, Lori; Obenchain, Valerie; Morgan, Martin.

Project co-funded by the European Commission within the H2020 Program (2015-2018)		
Dissemination Level		
PU	Public	x
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Deliverable description and summary

The deliverable includes 'annotation' and 'experiment' data results to be distributed through 'SOUNDHub' data resources, as described in Deliverable 8.1. To maximize impact of this software development, the functionality proposed for SOUNDHub have been developed and incorporated into *Bioconductor* 'AnnotationHub' and 'ExperimentHub' resources. The development consists of server-based and *R / Biocondutor* client facilities. The deliverable also includes the *BiocFileCache* package for managing file-based resources like those generated by research projects in *SOUND*.

The server is implemented in Ruby / sinatra and deployed through a robust MySQL database and Apache-based server stack; the servers are hosted in the Amazon cloud as a very modest instance, with data resources stored in S3 buckets or via redirects managed by the server software to the original data source.

Client software is implemented in the *R / Bioconductor* software packages *AnnotationHub* and *ExperimentHub*. The software provides a straight-forward interface to retrieve the current database of resources, to query for particular resources, to retrieve the resource to the local disk, and to input the resource into *R* in standard formats for further processing.

Our activities to develop *AnnotationHub* and *ExperimentHub* have also contributed to the development of *BiocFileCache*, a facility for managing file-based resources. *BiocFileCache* will be used to implement the caching mechanism of the *AnnotationHub* and *ExperimentHub* clients, and has proved popular for tasks common in analysis related to SOUND, e.g., managing intermediate files during complicated and distributed work flows. *BiocFileCache* also provides functionality for internal data sharing and standardization of file-based resources prior to public availability on *AnnotationHub* or *ExperimentHub*, as well as opportunities for federation across users within a geographically disparate group.

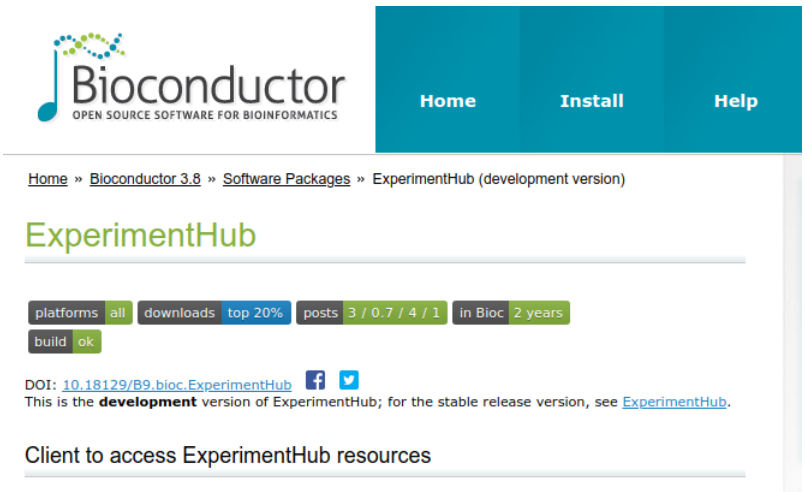
Software availability

Server software is available via the *Bioconductor* github account

- https://github.com/Bioconductor/BioconductorHubServer

The software repository includes the Ruby / sinatra application, as well as detailed instructions (see the README.md file) for deploying stand-alone instances for use within groups or institutions.

Client software is available via the *Bioconductor* github account and *Bioconductor* git repository



- https://github.com/Bioconductor/AnnotationHub
- https://github.com/Bioconductor/ExperimentHub
- git clone https://git.bioconductor.org:packages/AnnotationHub
- git clone https://git.bioconductor.org:packages/ExperimentHub

Client software is also available via standard *Bioconductor* installation procedures

- BiocManager::install("AnnotationHub")
- BiocManager::install("ExeperimentHub")

Client software 'landing pages'

- https://bioconductor.org/packages/AnnotationHub
- https://bioconductor.org/packages/ExperimentHub

include extensive documentation of individual functions (help manual), overall use cases (vignettes), and documentation for contributing new resources.

The *BiocFileCache* software is available via *Bioconductor* GitHub and git accounts, *Bioconductor* standard installation procedures, and through the 'landing page' and associated help manual and vignettes.

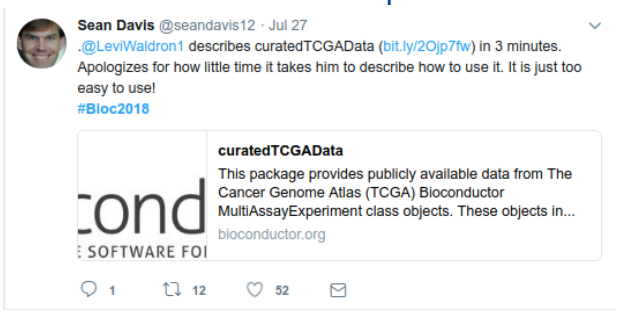
- https://github.com/Bioconductor/BiocFileCache
- git clone git@git.bioconductor.org:packages/BiocFileCache
- BiocManager::install("BiocFileCache")
- https://bioconductor.org/packages/BiocFileCache

Status

The AnnotationHub and ExperimentHub services are used extensively in *Bioconductor*. The *AnnotationHub* client is downloaded to approximately 4000 unique IP addresses per month; the newer *ExperimentHub* receives 300 – 400 unique IP downloads. More than 37000 resources have been retrieved from AnnotationHub, and more than 13000 from ExperimentHub.

A further accomplishment that we have achieved with SOUND funding has been to simplify and standardize the addition of new resources. The AnnotationHub has grown from 36000 records at the start of the funding period to 45676 records currently. The ExperimentHub has 1331 resources currently; it did not exist at the start of the funding period.

The protocol for submitting ExperimentHub resources has been standardized and simplified. ExperimentHub resources are typically heavily curated data derived from consortium or other large scale project activities like SOUND. The heavy curation requires documentation, and it is not unusual for the data to be presented in several versions. For these reasons, ExperimentHub resources are now typically contributed as *Bioconductor* 'experiment data' packages that contain a vignette describing data processing as well as version information. The packages contain simple functions to facilitate navigation and discovery of particular resources, so that is possible for users to very easily retrieve very rich data sets; an example of this is in the ironic tweet from Dr. Sean Davis lamenting the ease with which TCGA data can be accessed, making a 'tutorial' almost unnecessary. A second example of the activities ExperimentHub enables is the 'TENxBrainData' package, which retrieves and encapsulates the iconic 10x Genomics 'million neuron' single cell RNA sequence data in a standard *R /Bioconductor* object with a single line of code




```
> TENxBrainData::TENxBrainData()
snapshotDate(): 2018-08-03
see ?TENxBrainData and browseVignettes('TENxBrainData') for
documentation
downloading 0 resources
loading from cache
    '/home/mtmorgan//.ExperimentHub/1042'
class: SingleCellExperiment
dim: 27998 1306127
metadata(0):
assays(1): counts
rownames: NULL
rowData names(2): Ensembl Symbol
colnames(1306127): AAACCTGAGATAGGAG-1 AAACCTGAGCGGCTTC-1 ...
  TTTGTCAGTTAAAGTG-133 TTTGTCATCTGAAAGA-133
colData names(4): Barcode Sequence Library Mouse
reducedDimNames(0):
spikeNames(0):
```

Since it's introduction in March, 2017, *BiocFileCache* has been receiving increasing usage. It currently has more than 500 unique IP downloads per month. There are several innovative uses of *BiocFileCache* in the community, including as an 'on-the-fly' repository for expensive-to-create but frequently re-used files in analysis of RNA-seq differential expression in the *txmeta Bioconductor* package.