R and Bioconductor

A (very) short introduction and a few examples

Fabio Macciardi, MD PhD
Lab of Molecular Psychiatry, UCI
How to access R

5.4. R
1. Single core/CPU example R-single-cpu.sub:

```
#!/bin/bash

#SBATCH -p standard  ## run on the standard partition
#SBATCH -N 1        ## run on a single node
#SBATCH -n 1        ## request 1 task (one CPU)
#SBATCH -t 1-       ## 1-day run time limit

module load R/3.6.2
R CMD BATCH --no-save mycode.R
```

The above will submit the R code `mycode.R` with specified requested resources. Note: because the default is one CPU per task, `-n 1` can be thought of as requesting just one CPU.

The equivalent command-line method:

```
[user@login-x:-]$ module load R/3.6.2
[user@login-x:-]$ sbatch -p standard -N 1 -n 1 -t 1- --wrap="R CMD BATCH --no-save mycode.R"
```
2. Multiple core/CPU example **R-multi-cpu.sub**:

**Nearly all R jobs will only use a single core.** Please make sure your job is multi-threaded or is explicitly using R parallel libraries.

```
#!/bin/bash

#SBATCH -p standard  ## run on the standard partition
#SBATCH -N 1          ## run on a single node
#SBATCH -n 12         ## request 12 tasks (12 CPUs)
#SBATCH -t 00:20:00    ## 20 min run time limit

module load R/3.6.2
R CMD BATCH --no-save mycode.R
```

The above will submit the R code **mycode.R** with specified requested resources. Note: because the default is one CPU per task, `-n 12` can be thought of as requesting 12 CPUs.

The equivalent command-line method:

```
[user@login-x:-]$ module load R/3.6.2
[user@login-x:-]$ sbatch -p standard -N 1 -n 12 -t 00:20:00 \
   --wrap="R CMD BATCH --no-save mycode.R"
```
5.5. Rstudio

There are a couple of ways to run Rstudio.

1. **Windows users**: this method usually works for users who connect to the cluster using **MobaXterm**.

   Once logged in, claim an interactive session and start Rstudio after loading Rstudio module and R of your choice module via:

   ```bash
   [user@login-x-:~]$ srun -p free --pty --x11 /bin/bash -i  # claim an interactive session
   [user@login-x-:~]$ module load rstudio/1.4.1106         # load rstudio module
   [user@login-x-:~]$ module load R/4.0.2                # load R module
   [user@login-x-:~]$ rstudio                            # start Rstudio
   ```

2. **Mac users**: Your local Mac needs to have **XQuartz** installed. This is a standard application that provides X Window system for Mac OS. Follow Mac installation guide for installing applications if you don't have XQuartz installed.

   - login on the cluster using X forwarding. This means using `-X` or `-X -Y` option in the ssh command. For example:

     ```bash
     ssh -X panteater@hpc3.rcic.uci.edu
     ```

   - Once logged in, claim an interactive session, load Rstudio and R modules. Enforce software rendering engine in the `rstudio` command:

     ```bash
     [user@login-x-:~]$ srun -p free --pty --x11 /bin/bash -i  # claim an interactive session
     [user@login-x-:~]$ module load rstudio/1.4.1106         # load rstudio module
     [user@login-x-:~]$ module load R/4.0.2                # load R module
     [user@login-x-:~]$ QMLSCENE_DEVICE=softwarecontext rstudio  # enforce rendering in rstudio
     ```

3. **All users**: If the above method does not work for you (common for Mac users), the alternative way is to use our Jupyterhub portal and a container with Rstudio. Please see **Jupyter portal** for step by step instructions.
5.1. Jupyterhub Portal

Sometimes applications are available via containers on our Jupyterhub portal. This includes many bioinformatics applications, Rstudio, etc. Below are the steps to start a container that will provide Rstudio along with a few other applications.

1. Using your favorite browser go to: https://hpc3.rcic.uci.edu/biojhub3/hub/login You will see the following screen where you will use your usual HPC3 credentials to sign in:

![Sign in screen](image)

*Figure 1. Sign in screen*
2. After authentication you will see a screen with server options as in the figure below:

**Server Options**

- **Select Partition/Reservation to Use**
  - Standard

- **Select Account to Charge**
  - npw

- **Specify number of CPU cores (max 8)**
  - 1

- **memory per CPU core (max 4Gb per core)**
  - 4

- **Select a Containerized Notebook Image**
  - Default Base Jupyter w/R4.0.2, STATA, Rstudio, Rshiny + Slurm Su

- **Resume last session if available**
  - 

[Start button]

*Figure 2. Jupyter Server options screen*
2. After authentication you will see a screen with server options as in the figure below:

Server Options

- Select Partition/Reservation to Use: Standard
- Select Account to Charge: npw
- Specify number of CPU cores (max 8): 1
- memory per CPU core (max 4Gb per core): 4
- Select a Containerized Notebook Image: Default Base Jupyter w/R4.0.2, STATA, Rstudio, Rshiny + Slurm Su
- Resume last session if available: checked

Start

*Figure 2. Jupyter Server options screen*

3. Modify the *Select Account to Charge* to be one of your Slurm accounts, change number of CPUs and amount of memory if needed, and press **Start**.

Once the notebook is done spawning, you will get a Launcher screen with a number GUI apps you can use. One of those buttons is RStudio.
4. Be sure to stop your Jupyterhub notebook server after you are done with RStudio. From the **File** menu choose **Hub Control Panel** and you will be forwarded to a screen similar where you can press on **Stop My Server** to shut down the server:

![Stop My Server](image)

**Named Servers**

In addition to your default server, you may have additional 3 server(s) with names. This allows you to have more than one server running at the same time.

<table>
<thead>
<tr>
<th>Server name</th>
<th>URL</th>
<th>Last activity</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name your server</td>
<td>Add New Server</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 3. Jupyter Server stop screen*
What is Bioconductor
[http://www.bioconductor.org/]

The mission of the *Bioconductor* project is to develop, support, and disseminate free open-source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays.

*Bioconductor is* dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

*Bioconductor* uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community.
About Bioconductor

The mission of the Bioconductor project is to develop, support, and disseminate free open source software that facilitates rigorous and reproducible analysis of data from current and emerging biological assays. We are dedicated to building a diverse, collaborative, and welcoming community of developers and data scientists.

Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as Docker images.

News

- Bioconductor Bioc 3.15 Released.
- Bioconductor browsable code base now available.
- See our google calendar for events, conferences, meetings, forums, etc. Add your event with email to events at bioconductor.org.
- Bioconductor P1000 Research Channel is available.
- Orchestrating single-cell analysis with Bioconductor (abstract: website) and other recent literature.
- Bioconductor 3.14 release schedule announced. Please view for important deadlines.

Bic022 Conference

This is a hybrid in-person and virtual conference.

Registration Now Open! Limited in-person tickets available.

Scholarships to attend the conference are available. To apply please see Registration Page

Become a Bioconductor Sponsor

See Bio2022 Conference Website for more details

Abstract Submissions are now closed.

Install

- Discover 2140 software packages available in Bioconductor release 3.15.
- Get started with Bioconductor
  - Install Bioconductor
  - Get support
  - Latest newsletter
  - Follow us on twitter
  - Install R

Learn

- Master Bioconductor tools
  - Courses
  - Education and Training
  - Support site
  - Package vignettes
  - Literature citations
  - Common work flows
  - FAQ
  - Community resources
  - Videos

Use

Create bioinformatic solutions with Bioconductor

- Software, Annotation, and Experiment packages
- Docker
- Latest release announcement
- Use Bioconductor in the AnVIL. See our project updates.
- Community Slack sign-up
- Support site
- Events calendar; email events at bioconductor.org to add an event.

Develop

Contribute to Bioconductor

- Developer resources
- Use Bioc `devel`
- Devel packages
- Package guidelines
- New package submission
- Git source control
- Build reports
- Troubleshooting Build Report
- Browsable code base
Start Rstudio on hPC3 – see steps from previous slides using jupyter notebook ....

Install Bioconductor .....
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

    Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> if (!require("BiocManager", quietly = TRUE))
+   install.packages("BiocManager")
Bioconductor version 3.12 (BiocManager 1.30.10), BiocManager::install for help
Bioconductor version '3.12' is out-of-date; the current release version '3.15' is available with R version '4.2'; see https://bioconductor.org/install
> BiocManager::install(version = "3.15")
Error: Bioconductor version '3.15' requires R version '4.2'; see https://bioconductor.org/install
>
Using Bioconductor

The current release of Bioconductor is version 3.15; it works with R version 4.2.0. Users of older R and Bioconductor must update their installation to take advantage of new features and to access packages that have been added to Bioconductor since the last release.

The development version of Bioconductor is version 3.16; it works with R version 4.2.0. More recent 'dev' versions of R (if available) will be supported during the next Bioconductor release cycle.

Install the latest release of R, then get the latest version of Bioconductor by starting R and entering the commands

```r
if (!require("Bioconductor", quietly = TRUE))
  install.packages("Bioconductor")
Bioconductor::install(version = "3.15")
```

It may be possible to change the Bioconductor version of an existing installation; see the 'Changing version' section of the BioManager vignette.

Details, including instructions to install additional packages and to update, find, and troubleshoot are provided below. A `dev` version of Bioconductor is available. There are good reasons for using `bioconductor::install()` for managing Bioconductor resources.

Install R

1. Download the most recent version of R. The R FAQs and the R Installation and Administration Manual contain detailed instructions for installing R on various platforms (Linux, OS X, and Windows being the main ones).
2. Start the R program; on Windows and OS X, this will usually mean double-clicking on the R application, on UNIX-like systems, type "R" at a shell prompt.
3. As a first step with R, start the R help browser by typing `help.start()` in the R command window. For help on any function, e.g. the "mean" function, type `mean`.

Install Bioconductor Packages

Previous Releases

Bioconductor (R) versions:

- 3.13 (4.1)
- 3.12 (4.0)
- 3.11 (4.0)
- 3.10 (3.6)
- 3.9 (3.6)
- 3.8 (3.5)
- 3.7 (3.5)
- 3.6 (3.4)
- 3.5 (3.4)
- 3.4 (3.3)
- 3.3 (3.3)
- 3.2 (3.2)
- 3.1 (3.2)
- 3.0 (3.1)
- 2.14 (3.1)
- 2.13 (3.0)
- 2.12 (3.0)
- 2.11 (2.15)
- 2.10 (2.15)
- 2.9 (2.14)
- 2.8 (2.13)
- 2.7 (2.12)
- 2.6 (2.11)
- 2.5 (2.10)
- 2.4 (2.9)
- 2.3 (2.8)
- 2.2 (2.7)
- 2.1 (2.6)
- 2.0 (2.5)
- 1.9 (2.4)
- 1.8 (2.3)
- 1.7 (2.2)
- 1.6 (2.1)
As of today (June 27, 2022) there is NOT the R/4.2.x available on HPC3

[fmacciar@hpc3-19-12:/dfs7/gene]$ module avail

Thus you need to choose “your” R version and “your” Bioconductor version ...

[fmacciar@hpc3-19-12:/dfs7/gene]$ module load R/4.0.2
[fmacciar@hpc3-19-12:/dfs7/gene]$ R

R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit.
R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> if (!require("Bioconductor", quietly = TRUE))
> install.packages("Bioconductor")
Bioconductor version 3.12 (Bioconductor 1.30.10), ?Bioconductor::install for help
Bioconductor version '3.12' is out-of-date; the current release version '3.15' is available with R version '4.2'; see https://bioconductor.org/install

> Bioconductor::install(version = "3.15")
Current Bioconductor version 3.12.10 seems R version '4.2' see https://bioconductor.org/install

installing package(s) 'Bioconductor'
Content type 'application/x-gzip' length 981 bytes

* installing *source* package 'Bioconductor' ...
** using staged installation
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (Bioconductor)

The downloaded source packages are in

  '/tmp/RtmpNqN4dN/downloaded_packages'

Installation path not writable, unable to update packages: BH, Bioconductor, boot, class, cli, cluster, coda, coda.tools, commonmark, crayon, digest, ellipsis, evaluate, fansi, fastmap, foreign, glue, htmltools, httpuv, IRdisplay, IRkernel, jsonlite, KernSmooth, lattice, lifecycle, magrittr, MASS, Matrix, mgcv, mime, nle, nnet, phangorn, pillar, promises, Rs, Rd, repr, rlang, rpart, shiny, shinyjs, shinythemes, spatial, survival, utf8, uuid, vctrs, withr

>
Find Bioconductor Packages

Visit the software package list to discover available packages.

To search through available packages programmatically, use the following:

```
BiocManager::available()
```

For example, using a "^org" search pattern will show all of the available organism annotation packages.

- `BiocManager::available()`
- [1] "A3" "a4" "a4Base"
- [4] "a4Classif" "a4Core" "a4Preproc"
- [7] "a4Reporting" "AATtools" "ABACUS"
- [10] "ABAData" "ABAEEnrichment" "ABArray"
- [13] "abbreviate" "abbyyR" "abc"
- ....
- [985] "basifoR" "basilisk" "basilisk.utils"
- [988] "BASiNET" "BASIX" "BaSkePro"
- [991] "basket" "BasketballAnalyzeR" "baskexact"
- [994] "BASS" "BaSTA" "bastah"
- [997] "BAT" "batata" "batch"
- [1000] "batchelor"
- [ reachedgetOption("max.print") -- omitted 20412 entries ]
Find which Bioconductor packages are already installed (on HPC3) ...
## Bioconductor version 3.12

- Developers: check this box to toggle the visibility of childless biocViews.

### Autocomplete biocViews search:

<table>
<thead>
<tr>
<th>Software</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>AssayDomain</td>
<td>790</td>
</tr>
<tr>
<td>BiologicalQuestion</td>
<td>823</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>455</td>
</tr>
<tr>
<td>ResearchField</td>
<td>903</td>
</tr>
<tr>
<td>StatisticalMethod</td>
<td>727</td>
</tr>
<tr>
<td>Technology</td>
<td>1251</td>
</tr>
<tr>
<td>WorkflowStep</td>
<td>1081</td>
</tr>
<tr>
<td>AnnotationData</td>
<td>971</td>
</tr>
<tr>
<td>ExperimentData</td>
<td>398</td>
</tr>
<tr>
<td>Workflow</td>
<td>28</td>
</tr>
</tbody>
</table>

### Packages found under Software:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show ▼ ▼ entries

<table>
<thead>
<tr>
<th>Package</th>
<th>Maintainer</th>
<th>Title</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioversion</td>
<td>Bioconductor Package Maintainer</td>
<td>Set the appropriate version of Bioconductor packages</td>
<td>1</td>
</tr>
<tr>
<td>BiocGenerics</td>
<td>Bioconductor Package Maintainer</td>
<td>S4 generic functions used in Bioconductor</td>
<td>2</td>
</tr>
<tr>
<td>S4Vectors</td>
<td>Bioconductor Package Maintainer</td>
<td>Foundation of vector-like and list-like containers in Bioconductor</td>
<td>3</td>
</tr>
<tr>
<td>IRanges</td>
<td>Bioconductor Package Maintainer</td>
<td>Foundation of integer range manipulation in Bioconductor</td>
<td>4</td>
</tr>
<tr>
<td>Biobase</td>
<td>Bioconductor Package Maintainer</td>
<td>Biobase: Base functions for Bioconductor</td>
<td>5</td>
</tr>
<tr>
<td>zlibioc</td>
<td>Bioconductor Package Maintainer</td>
<td>An R packaged zlib-1.2.5</td>
<td>6</td>
</tr>
<tr>
<td>GenomeInfoDb</td>
<td>Bioconductor Package Maintainer</td>
<td>Utilities for manipulating chromosome names, including modifying them to follow a particular naming style</td>
<td>7</td>
</tr>
</tbody>
</table>
Support for Bioconductor …..
A very simple example: installing the Bioconductor package “VennDetail”

VennDetail

DOI: 10.18129/B9.bioc.VennDetail
This package is for version 3.12 of Bioconductor; for the stable, up-to-date release version, see VennDetail.

A package for visualization and extract details

Bioconductor version: 3.12
A set of functions to generate high-resolution Venn, Vennpie plot, extract and combine details of these subsets with user datasets in dataframe is available.

Author: Kai Guo, Brett McGregor
Maintainer: Kai Guo <guokai8 at gmail.com>

Citation (from within R, enter citation("VennDetail")):


Installation
To install this package, start R (version "4.0") and enter:

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("VennDetail")
```

For older versions of R, please refer to the appropriate Bioconductor release.

Documentation
To view documentation for the version of this package installed in your system, start R and enter:

```r
browseVignettes("VennDetail")
```
... and the package is installed in YOUR directory on HPC3, NOT in /usr/bin ...
Now, let’s work with VennDetail and try some examples ....

First, make the VennDetail library and its dependencies active =

```
> library(VennDetail)
>
```

Then, upload data example (in this case, three sets of Differentially Expressed Genes (DEG) obtained by comparing diabetic mice with or without a given drug (pioglitazone) treatment.

```
> data(T2DM)

Using the VennDetail package, we want to first explore whether there are common DEG across 3 different tissues sequenced from mice, kidney Cortex, kidney glomerula, and sciatic nerve.

```
> ven <- venndetail(list(Cortex = T2DM$Cortex$Entrez, SCN = T2DM$SCN$Entrez, + Glom = T2DM$Glom$Entrez))
```
And we plot the results ...

> plot(ven)

Which generates a plot in your RStudio

Then by clicking on the “Export” button, you can choose how to create a graph to copy or download
Save Plot as PDF

PDF Size: US Letter 8.50 x 11.00 inches

Orientation: Portrait

Options: Use cairo_pdf device

Directory...

File name: VennDiagram_test

View plot after saving

Preview  Save  Cancel
And/or we can customize the graph to our like, for example a "pie"

```r
> plot(ven, type = "vennpi")
Warning messages:
1: `filter()` was deprecated in dplyr 0.7.0.
Please use `filter()` instead.
See vignette('programming') for more help
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
2: `select()` was deprecated in dplyr 0.7.0.
Please use `select()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```
Or a “bar” graph:

```r
> plot(ven, type = "upset")
```
```
> head(getFeature(ven, subset = "Shared", rlist = T2DM))
     Subset Detail Cortex_Entrez Cortex_Symbol  Cortex_Annotation Cortex
   x_log2FC   Cortex_FDR
 1 Shared            229599       Gm129              predicted gene 129
        4.851041      0.00156529
 2 Shared            243385       Gprin3    GPRIN family member 3
        2.588754      0.00156529
 3 Shared             99899       Iifi44 interferon-induced protein 44
        2.186102      0.00156529
 4 Shared            17001       Ltc4s     leukotriene C4 synthase
        3.916510      0.00156529
 5 Shared            18143       Npas2 neuronal PAS domain protein 2
        3.527904      0.00156529
 6 Shared            64136       Sdf2l1 stromal cell-derived factor 2-like 1
        2.723979      0.00156529
   SCN_Entrez   SCN_Symbol  SCN_Annotation SCN_log2FC  SCN_FDR Glom
     1  229599       Gm129              predicted gene 129  3.638130 0.000772111
    229599
    2  243385       Gprin3    GPRIN family member 3  2.942612 0.002032400
    243385
    3  99899        Iifi44 interferon-induced protein 44 -2.042164 0.012997000
       99899
    4  17001        Ltc4s     leukotriene C4 synthase  2.852832 0.000772111
     17001
    5  18143        Npas2 neuronal PAS domain protein 2 -2.219165 0.015590600
       18143
    6  64136       Sdf2l1 stromal cell-derived factor 2-like 1 -2.092271 0.000772111
     64136
    Glom_Symbol  Glom_Annotation Glom_log2FC  Glom_FDR
    1         Gm129              predicted gene 129  2.223499 0.025568700
    2        Gprin3    GPRIN family member 3 -2.186954 0.000962798
    3         Iifi44 interferon-induced protein 44 -2.146200 0.000962798
    4         Ltc4s     leukotriene C4 synthase  2.471602 0.011659400
    5        Npas2 neuronal PAS domain protein 2 -11.845227 0.000962798
    6       Sdf2l1 stromal cell-derived factor 2-like 1 -2.875391 0.000962798
```
> dplot(ven, order = TRUE, textsize = 4)
```
set.seed(123)
A <- sample(1:1000, 400, replace = FALSE)
B <- sample(1:1000, 600, replace = FALSE)
C <- sample(1:1000, 350, replace = FALSE)
D <- sample(1:1000, 550, replace = FALSE)
E <- sample(1:1000, 450, replace = FALSE)
ven <- venn(data.frame(A = A, B = B, C = C, D = D, E = E))
plot(ven)
plot(ven)
head(getFeature(ven, subset = "Shared", rlist = T2M5))
```

```
1 Shared 229599 229599 Gm129 predicted gene 129 4.851041 0.00156529
2 Shared 243385 243385 Gm129 Gm129 LEPRE1 FCRN 2.588756 0.00156529
3 Shared 99899 99899 Ifi44 Interferon-induced protein 44 -2.186182 0.00156529
4 Shared 17001 17001 Ltc4s Leukotriene C4 synthase 3.915638 0.00156529
5 Shared 18143 18143 Npas2 Neuronal PAS domain protein 2 -3.827994 0.00156529
6 Shared 64136 64136 Sdf211 stromal cell-derived factor 2-like 1 -2.723997 0.00156529
```

```
plot(ven, order = TRUE, textsize = 4)
```