

"No doubt about it, Ellington—we've mathematically expressed the purpose of the universe. God, how I love the thrill of scientific discovery!"

DCMB BioComputing BootCamp

Day 3, Session IV:

Working with Packages from CRAN and Bioconductor

Armand Bankhead

bankhead@umich.edu

8/21/2019



Overview

1. CRAN
2. Bioconductor
3. Package Installation
4. Package Documentation
5. Package Source Code
6. Tidyverse
7. Example: BioMart bioconductor package

CRAN: The Comprehensive R Archive Network

- CRAN is a website dedicated to hosting the R language and R packages
 - packages are shared collections of R code
- There are currently 14,762 available packages
 - search via google
 - table (index) of available packages
 - CRAN Task Views

Available CRAN Packages By Name

[A](#)[B](#)[C](#)[D](#)[E](#)[F](#)[G](#)[H](#)[I](#)[J](#)[K](#)[L](#)[M](#)[N](#)[O](#)[P](#)[Q](#)[R](#)[S](#)[T](#)[U](#)[V](#)[W](#)[X](#)[Y](#)[Z](#)

[A3](#)
[abbyyR](#)
[abc](#)
[abc.data](#)
[ABC.RAP](#)
[ABCanalysis](#)
[abcdeFBA](#)
[ABCOptim](#)
[ABCp2](#)
[abcrf](#)
[abctools](#)
[abd](#)
[abe](#)
[abf2](#)
[ABHgenotypeR](#)
[abind](#)
[abjutils](#)
[abn](#)
[abnormality](#)
[abodOutlier](#)
[ABPS](#)

Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
Access to Abbyy Optical Character Recognition (OCR) API
Tools for Approximate Bayesian Computation (ABC)
Data Only: Tools for Approximate Bayesian Computation (ABC)
Array Based CpG Region Analysis Pipeline
Computed ABC Analysis
ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
Implementation of Artificial Bee Colony (ABC) Optimization
Approximate Bayesian Computational Model for Estimating P2
Approximate Bayesian Computation via Random Forests
Tools for ABC Analyses
The Analysis of Biological Data
Augmented Backward Elimination
Load Gap-Free Axon ABF2 Files
Easy Visualization of ABH Genotypes
Combine Multidimensional Arrays
Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association
Modelling Multivariate Data with Additive Bayesian Networks
Measure a Subject's Abnormality with Respect to a Reference Population
Angle-Based Outlier Detection
The Abnormal Blood Profile Score to Detect Blood Doping

CRAN Package Table

What do R Packages Contain?

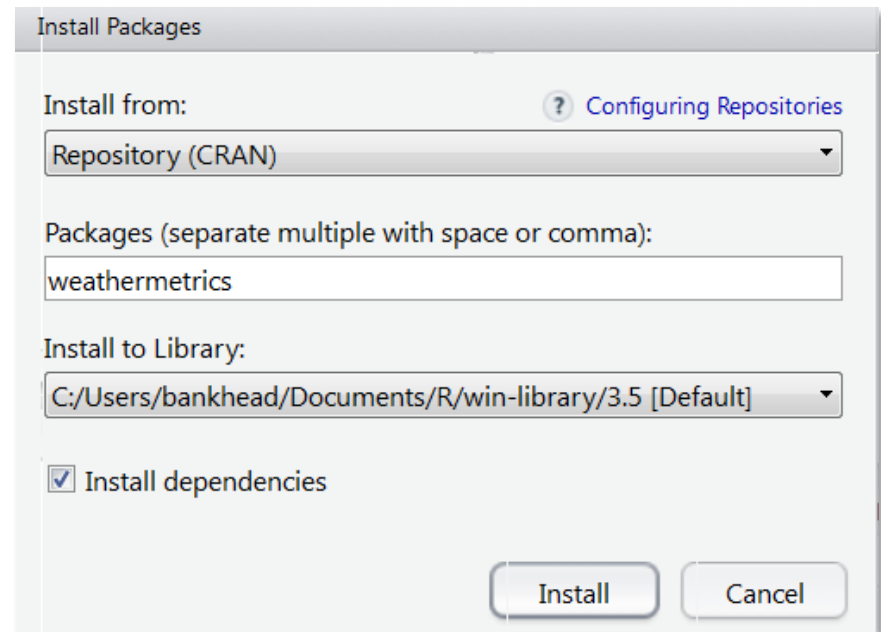
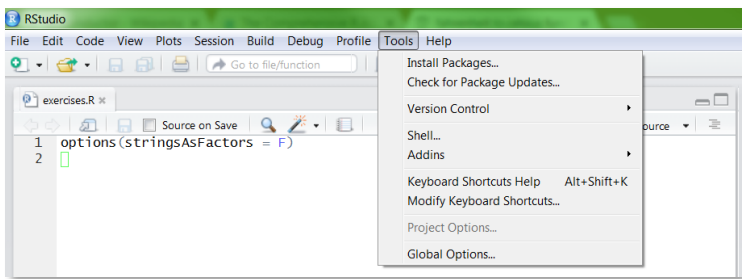
- R code
- Help pages
- Data: example data sets included for demonstration purposes
- Other documentation
 - vignettes: documents that integrate code and how to perform a specific task
 - tutorials
- Code written in other languages (e.g. C, FORTRAN)
- Directives used to help install the package

Many Ways to Install a Package

1. On the command prompt type:

```
> install.packages('weathermetrics')
```

2. In RStudio select Tools -> Install Packages from the top of the main screen



Useful Commands for Exploring R Packages

In-Class Exercise:
Try It Out!

`search()` – lists packages currently loaded

`sessionInfo()` – lists packages with version #'s

`packageDescription()` – prints brief description

`install.packages()` - install an R package from CRAN

`update.packages()` – checks for package updates and installs updates if necessary

`available.packages()` – returns packages that are available from CRAN

`installed.packages()` – returns packages that are installed on the user's system

Exercise: What packages are currently loaded? What packages are installed on your computer?

Overview

1. CRAN
- 2. Bioconductor**
3. Package Installation
4. Package Documentation
5. Package Source Code
6. Tidyverse
7. Example: BioMart bioconductor package




- R packages and datasets for working with high-throughput genomic data
 - <http://bioconductor.org>
 - Bioconductor was started in Fall 2001
 - Two releases each year
 - v3.9 has 1,741 packages
- Packages undergo formal initial review and continuous automated testing

nature methods

Perspective | Published: 29 January 2015

Orchestrating high-throughput genomic analysis with Bioconductor

Wolfgang Huber , Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron & Martin Morgan - Show fewer authors

Nature Methods **12**, 115–121 (2015) | Download Citation 

Version	Release Date	Package Count	Dependency
1.0	1 May 2002	15	R 1.5
1.1	19 Nov 2002	20	R 1.6
1.2	29 May 2003	30	R 1.7
1.3	30 Oct 2003	49	R 1.8
1.4	17 May 2004	81	R 1.9
1.5	25 Oct 2004	100	R 2.0
1.6	18 May 2005	123	R 2.1
1.7	14 Oct 2005	141	R 2.2
1.8	27 Apr 2006	172	R 2.3
1.9	4 Oct 2006	188	R 2.4
2.0	26 Apr 2007	214	R 2.5
2.1	8 Oct 2007	233	R 2.6
2.2	1 May 2008	260	R 2.7
2.3	22 Oct 2008	294	R 2.8
2.4	21 Apr 2009	320	R 2.9
2.5	28 Oct 2009	352	R 2.10
2.6	23 Apr 2010	389	R 2.11
2.7	18 Oct 2010	418	R 2.12
2.8	14 Apr 2011	466	R 2.13
2.9	1 Nov 2011	517	R 2.14
2.10	2 Apr 2012	554	R 2.15
2.11	3 Oct 2012	610	R 2.15
2.12	4 Apr 2013	671	R 3.0
2.13	15 Oct 2013	749	R 3.0
2.14	14 Apr 2014	824	R 3.1
3.0	14 Oct 2014	934	R 3.1
3.1	17 Apr 2015	1024	R 3.2
3.2	14 Oct 2015	1104	R 3.2
3.3	4 May 2016	1211	R 3.3
3.4	18 Oct 2016	1296	R 3.3
3.5	25 Apr 2017	1383	R 3.4
3.6	31 Oct 2017	1473	R 3.4
3.7	1 May 2018	1560	R 3.5

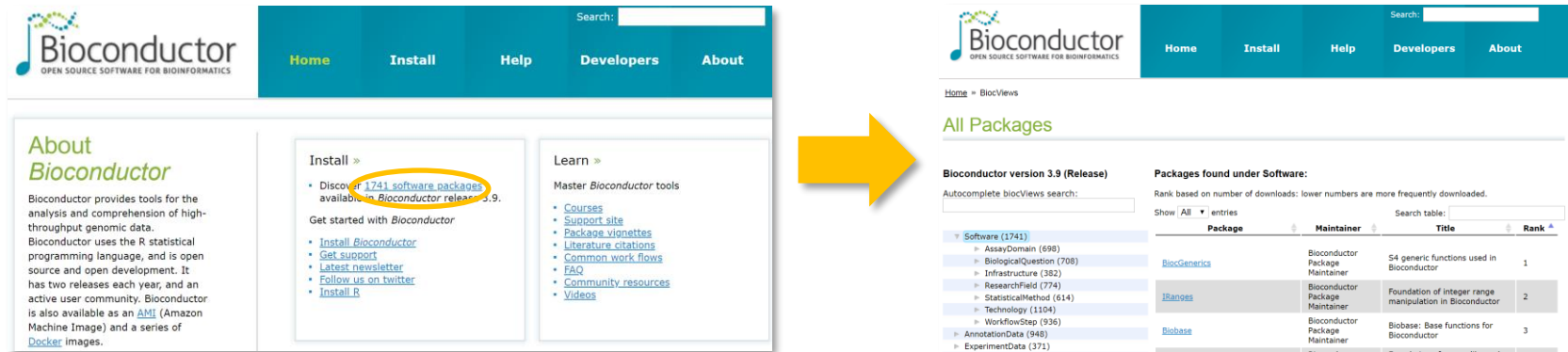


Project Goals

1. Provide **widespread access** to a broad range of powerful statistical and graphical methods for the analysis of genomic data.
2. Facilitate the inclusion of **biological metadata** in the analysis of genomic data, e.g. literature data from PubMed, annotation data from LocusLink/Entrez.
3. Provide a **common software platform** that enables the rapid development and deployment of plug-able, scalable, and interoperable software.
4. Further scientific understanding by producing **high-quality documentation** and **reproducible** research.
5. **Train researchers** on computational and statistical methods for the analysis of genomic data.

How to Install Bioconductor Packages

In-Class Exercise:
Try It Out!



The image shows two screenshots of the Bioconductor website. The left screenshot shows the 'Install' section with a link to 'Discover 1741 software packages available in Bioconductor release 3.9.' circled in yellow. A yellow arrow points from this link to the right screenshot. The right screenshot shows the 'All Packages' page, which lists various software packages. The 'Software (1741)' section is expanded, showing a list of packages including AssayDomain, BiologicalQuestion, Infrastructure, ResearchField, StatisticalMethod, WorkflowStep, AnnotationData, and ExperimentData. The 'Packages found under Software:' section shows a table with columns for Package, Maintainer, Title, and Rank. The table lists packages like BiocGenerics, Ranges, and Biobase.

Package	Maintainer	Title	Rank
BiocGenerics	Bioconductor Package Maintainer	54 generic functions used in Bioconductor	1
Ranges	Bioconductor Package Maintainer	Foundation of integer range manipulation in Bioconductor	2
Biobase	Bioconductor Package Maintainer	Biobase: Base functions for Bioconductor	3

First install core packages:

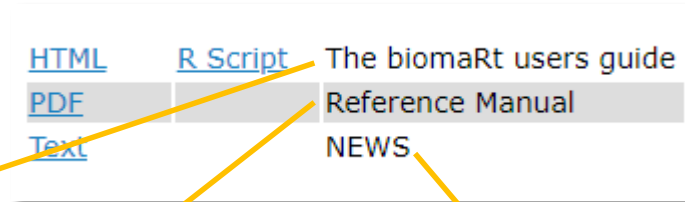
```
> install.packages("BiocManager")
```

Next install specific packages of interest:

```
> BiocManager::install("biomaRt")
```

**Exercise: Install Bioconductor. Install the biomaRt package.
This may take time...**

Package Reference Manuals, Vignettes, and News



The biomaRt users guide

Steffen Durinck, Wolfgang Huber, Mike Smith

23 May 2018

Package

biomaRt 2.36.1

Contents

- 1 Introduction
- 2 Selecting a BioMart database and dataset
- 3 How to build a biomaRt query
- 4 Examples of biomaRt queries
 - 4.1 Annotate a set of Affymetrix identifiers with HUGO symbol and chromosomal locations of corresponding genes
 - 4.2 Annotate a set of EntrezGene identifiers with GO annotation
 - 4.3 Retrieve all HUGO gene symbols of genes that are located on chromosomes 17, 20 or Y, and are associated with specific GO terms
 - 4.4 Annotate set of identifiers with INTERPRO protein domain identifiers
 - 4.5 Select all Affymetrix identifiers on the hgu133plus2 chip and Ensembl gene identifiers for genes located on chromosome 16 between basepair 1100000 and 1250000.
 - 4.6 Retrieve all entrezgene identifiers and HUGO gene symbols of genes which have a "MAP kinase activity" GO term associated with it.

Vignette

exportFASTA	Exports getSequence results to FASTA format				
<hr/>					
Description	Exports getSequence results to FASTA format				
Usage	<code>exportFASTA(sequences, file)</code>				
Arguments	<table><tr><td>sequences</td><td>A data.frame that was the output of the getSequence function</td></tr><tr><td>file</td><td>File to which you want to write the data</td></tr></table>	sequences	A data.frame that was the output of the getSequence function	file	File to which you want to write the data
sequences	A data.frame that was the output of the getSequence function				
file	File to which you want to write the data				
Author(s)	Steffen Durinck				
Examples	<pre>if(interactive()){ mart <- useMart("ensembl", dataset="hsapiens_gene_ensembl") #seq<-getSequence(chromosome=c(2,2),start=c(100000,300000),end=c(100300,30500),mart=mart) #exportFASTA(seq,file="test.fasta") martDisconnect(mart = mart) }</pre>				

User's Manual

CHANGES IN VERSION 2.36.0

BUG FIXES
<ul style="list-style-type: none">o Patched problem returning the list of available datasets, if the description of one or more datasets included an apostrophe (introduced with new primate species in Ensembl).o Caught scenario where ensemblRedirect=FALSE was still being ignored.o Changed query submission when redirection is detected to cope with apparently new behaviour of the Ensembl mirrors.
MINOR CHANGES
<ul style="list-style-type: none">o Increase query timeout limit to 5 minutes.
CHANGES IN VERSION 2.34.0

NEW FEATURES
<ul style="list-style-type: none">o Added the listEnsemblArchives() function. This returns a table of the available Ensembl archives, and replaces the archive = TRUE argument to several functions, which was no longer working.
BUG FIXES
<ul style="list-style-type: none">o The Ensembl BioMart server doesn't always respond well if queries with more than 500 filter values are submitted. If a query that exceed this is detect biomaRt will now submit the query in batches and concatenate the result when completed.
MINOR CHANGES
<ul style="list-style-type: none">o You can now provide a host with 'http:///' at the start, or a trailing '/' (typically copy/pasted from a browser) and useMarts() etc will cope.

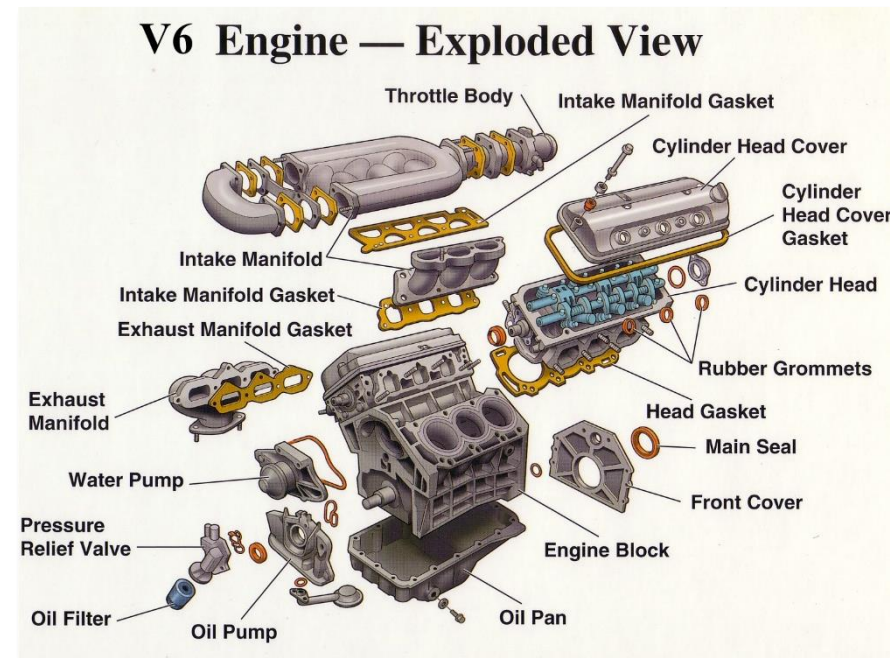
Package News

Overview

1. CRAN
2. Bioconductor
3. Package Installation
4. Package Documentation
- 5. Package Source Code**
6. Tidyverse
7. Example: BioMart bioconductor package

Package Source Code

- What if you wanted to find out more about how a package is implemented?
- By default the binary form of a package is downloaded and installed
- Download the “source” version of the package to view the code



Exercise: Download the source code package for weathermetrics and view the code for the function `fahrenheit.to.celsius()`

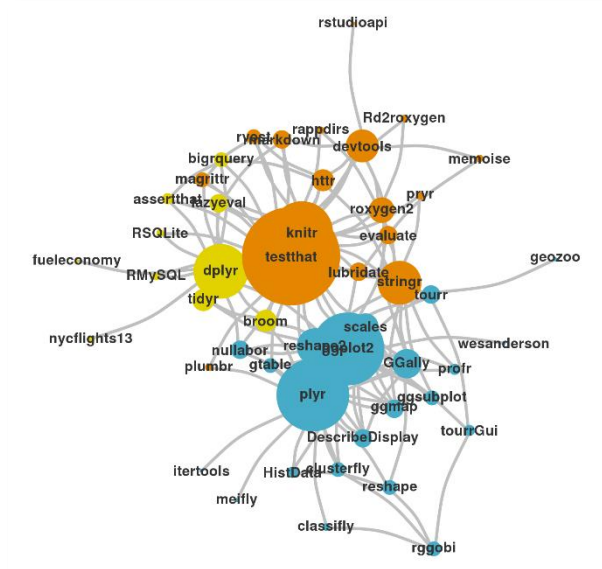
In-Class Exercise:
Try It Out!

- Here are two ways you can download the source code for the weathermetrics package:
 1. Go to <https://cran.r-project.org/web/packages/weathermetrics/index.html> and download "Package source"
 2.

```
download.packages(pkgs = "weathermetrics",  
destdir = ".", type = "source")
```
- Use gunzip or winzip to decompress the package
- Use grep to find the file with `fahrenheit.to.celsius`
 - HINT: `R/temperature_conversions.R`
- Open the file and find the function

- 
- A portrait of a man with a beard and short brown hair, wearing a blue and white plaid button-down shirt. He is smiling slightly and looking towards the camera. The background is a plain, light color.

Hadley Wicham



Tidyverse



packages

- ggplot2: data graphics and visualization
- dplyr: data frame manipulation
- tidyr: data wrangling (reshape replacement)
- readr: methods to read rectangular data formats
- purr: supports piping data to replace loops
- tibble: updated data.frames
- stringr: string manipulation methods
- forcats: updated factors

Note: R programming styles vary, tidyverse is something you should be aware of but not all R programmers use it.

Overview

1. CRAN
2. Bioconductor
3. Package Installation
4. Package Documentation
5. Package Source Code
6. Tidyverse
7. **Example: BioMart bioconductor package**

biomaRt Allows Users to Search Across Genomic Annotation Databases

- biomaRt is an R interface to the BioMart software suite (<http://biomart.org>) that is a repository of inter-connected genome annotation databases
- Challenges in genome annotation are a widespread in the field of bioinformatics for several reasons:
 1. Our understanding the genome continues to evolve
 2. Many researchers investigating the genome across the globe may use slightly different conventions
 3. Genome annotation is biased by historical nomenclature
- Example: AKT1 also has an Entrez gene ID of 207 and Refseq transcripts called NM_005163.2, NM_001014432.1, NM_001014431.1
- Example: AKT1 also has Ensembl transcripts called ENST00000555528.5, ENST00000349310.7, ENST00000407796.6

Exercise: Retrieve Entrez and Ensembl IDs for AKT1, TP53, EGFR, PIK3CA using biomaRt

In-Class Exercise:
Try It Out!

1. Load the package (biomaRt should already be installed from earlier in this session)

```
> library(biomaRt)
```

2. Select a mart and a dataset to query

```
> ensembl = useMart("ensembl", dataset =  
"hsapiens_gene_ensembl")
```

3. Use the biomaRt vignette “How to build a biomaRt query” (Section 3) to construct a query

HINT: filters = c('hgnc_symbol')

HINT: attributes = c('hgnc_symbol','entrezgene','ensembl_gene_id')

HINT: inputValues = c('AKT1','TP53','EGFR','PIK3CA')

HINT: mart = ensembl

Closing Remarks/Advice

- Comment your code
- Short programs are better
- Plan!
- Be prepared to iterate
 - Make a change
 - Run
 - Make another change
 - Run
 - ...

References

- Gentleman, Robert. R Programming for Bioinformatics. CRC Press, 2009.
- Slides sourced in part from Barry Grant