

The *Bioconductor* Project: Current Status

Martin Morgan

Roswell Park Cancer Institute
Buffalo, NY, USA
martin.morgan@roswellpark.org

5 December, 2017



Analysis and comprehension of high-throughput genomic data.

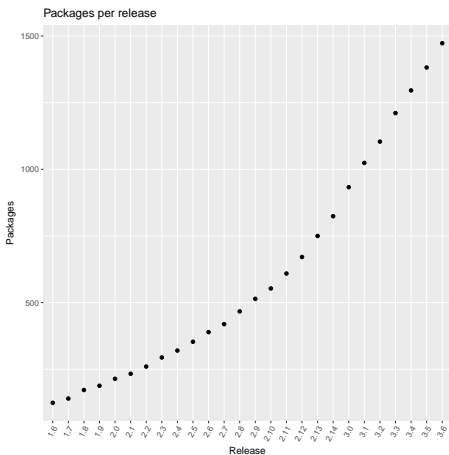
- Started 2002
- 1473 *R* packages – developed by 'us' and user-contributed.

Well-used and respected.

- 53k unique IP downloads / month.
- 21,700 PubMedCentral citations.

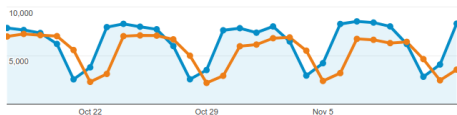
More than 1000 maintainers!

State of the project



- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project



1.	United States	58,384 (32.78%)
2.	China	20,910 (11.74%)
3.	United Kingdom	12,265 (6.89%)
4.	Germany	10,024 (5.63%)
5.	France	5,536 (3.11%)
6.	Canada	4,999 (2.81%)
7.	Spain	4,864 (2.73%)
8.	Japan	4,539 (2.55%)
9.	India	4,397 (2.47%)
10.	Australia	4,043 (2.27%)

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project

<https://bioconductor.org>

<https://support.bioconductor.org>

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project

The screenshot displays the GitHub Issues page for the Bioconductor/Contributions repository. The page shows a list of 33 open issues. The top issue is 'mcSEAdata' with a 'review in progress' label and a 'WARNING' badge. Other issues include 'pace', 'CHARGE', 'BASINET', 'Numero', and 'KeyZEnrich: An all-in-one R/Bioconductor package for gene list enrichment analysis and pathway visualization'. The interface includes navigation tabs for Code, Issues, Pull requests, Projects, Wiki, Insights, and Settings. A 'New Issue' button is visible in the top right.

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

State of the project



SOUND

- Packages
- Users
- Web & support sites
- Training, workflows, & meetings
- New package submission
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & permanence.
- Interoperability.
- Documentation.
- Support.
- Tested.



Why use or contribute to *Bioconductor*?

- Recognition.
- Access & permanence.
- Interoperability.
- Documentation.
- Support.
- Tested.

in Bioc > 12.5 years

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & permanance.
- Interoperability.
- Documentation.
- Support.
- Tested.

```
git$ grep -l SummarizedExperiment \  
*/DESCRIPTION | wc -l  
165
```

Why use or contribute to *Bioconductor*?

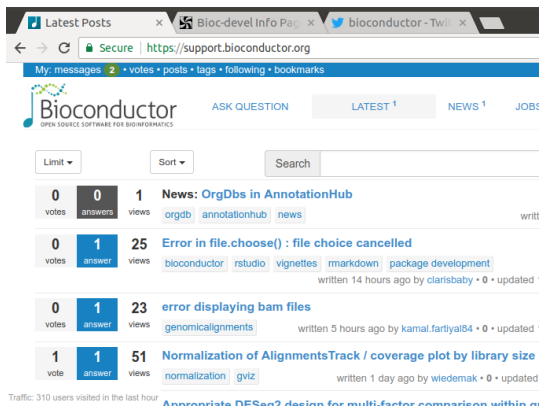
- Recognition.
- Access & permanence.
- Interoperability.
- Documentation.
- Support.
- Tested.

Documentation

HTML	R Script	Analyzing RNA-seq data with DESeq2
PDF		Reference Manual
Text		NEWS

Why use or contribute to *Bioconductor*?

- Recognition.
- Access & permanence.
- Interoperability.
- Documentation.
- Support.
- Tested.



The screenshot shows the Bioconductor support forum interface. At the top, there are navigation tabs for 'Latest Posts', 'Bioc-devel Info Page', and 'bioconductor - Twitter'. The browser address bar shows the URL 'https://support.bioconductor.org'. Below the navigation, there are links for 'My: messages', 'votes', 'posts', 'tags', 'following', and 'bookmarks'. The Bioconductor logo is prominently displayed, along with the tagline 'OPEN SOURCE SOFTWARE FOR BIOINFORMATICS'. There are buttons for 'ASK QUESTION', 'LATEST 1', 'NEWS 1', and 'JOBS'. A search bar is located below the navigation. The main content area displays a list of posts with the following details:

Votes	Answers	Views	Title	Tags	Author	Updated
0	0	1	News: OrgDbs in AnnotationHub	orgdb, annotationhub, news		written 14 hours ago by clarisbaby • 0 • updated
0	1	25	Error in file.choose(): file choice cancelled	bioconductor, rstudio, vignettes, markdown, package development		written 14 hours ago by clarisbaby • 0 • updated
0	1	23	error displaying bam files	genomicalignments		written 5 hours ago by kamal.fartiyal84 • 0 • updated
1	1	51	Normalization of AlignmentsTrack / coverage plot by library size	normalization, gviz		written 1 day ago by wiedemak • 0 • updated

Traffic: 310 users visited in the last hour

Why use or contribute to Bioconductor?

- Recognition.
- Access & permanence.
- Interoperability.
- Documentation.
- Support.
- Tested.

Multiple platform build/check report for BioC 3.6
This page was generated on 2017-11-15 15:06:34 -0500 (Wed, 15 Nov 2017).

git log
Snapshot Date: 2017-11-14 17:00:31 -0500 (Tue, 14 Nov 2017)

Hostname	OS	Arch	Platform label (P)	R session	Installed pkg
malbec7	Linux (Ubuntu 16.04.1 LTS)	x86_64	x86_64-linux-gnu	3.4.2 (2017-09-28) -- "Short Summer"	1951
tokay7	Windows Server 2012 R2 Standard	x84	mingw32/x86_64-w64-mingw32	3.4.2 Patched (2017-10-07 r73456) -- "Short Summer"	1918
veracruz7	OS X 10.11.6 El Capitan	x86_64	apple-darwin15.6.0	3.4.2 (2017-09-28) -- "Short Summer"	1930

Package status is indicated by one of the following glyphs

- TIMEOUT** INSTALL, BUILD, CHECK or BUILD BIN of package took more than 60 minutes
- ERROR** INSTALL, BUILD, or BUILD BIN of package failed, or CHECK produced errors
- WARNINGS** CHECK of package produced warnings
- OK** INSTALL, BUILD, CHECK or BUILD BIN of package was OK
- NotNeeded** INSTALL of package was not needed (click on glyph to see why)
- skipped** CHECK or BUILD BIN of package was skipped because the BUILD step failed
- NA** BUILD, CHECK or BUILD BIN result is not available because of an anomaly in the Build System

Package propagation status is indicated by one of the following LEDs

- GREEN** Package was propagated because it didn't previously exist or version was bumped
- NO** Package was not propagated because of a problem (impossible dependencies, or version lower than what is already propagated)
- UNNEEDED** Package was not propagated because it is already in the repository with this version. A version bump is required in order to propagate it.

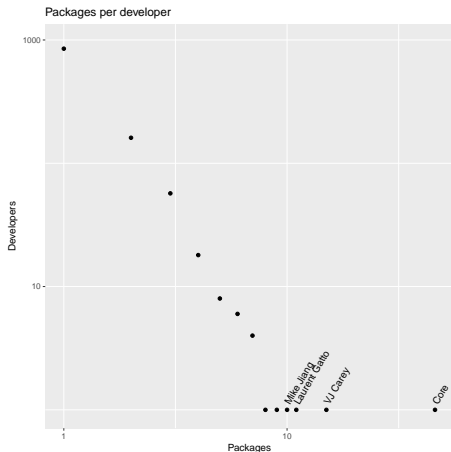
A crossed-out package name indicates the package is **deprecated**

SUMMARY	OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
malbec7	Linux (Ubuntu 16.04.1 LTS) / x86_64	0 4 513 856 2 33 1438	2 16 146 1271	0 0 1407	0 0 1426
tokay7	Windows Server 2012 R2 Standard / x84	0 3 488 330 4 29 1407	5 18 1445 1244	0 0 1407	0 0 1426
veracruz7	OS X 10.11.6 El Capitan / x86_64	0 2 509 882 3 34 1426	1 16 163 1246	0 0 1426	0 0 1426

A 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

Package	OS / Arch	INSTALL	BUILD	CHECK	BUILD BIN
base 3.4.2	malbec7 Linux (Ubuntu 16.04.1 LTS) / x86_64	NotNeeded	OK	OK	
base 3.4.2	tokay7 Windows Server 2012 R2 Standard / x84	NotNeeded	OK	OK	OK
base 3.4.2	veracruz7 OS X 10.11.6 El Capitan / x86_64	NotNeeded	OK	OK	OK
base 3.4.2	malbec7 Linux (Ubuntu 16.04.1 LTS) / x86_64	NotNeeded	OK	OK	

Our contributors



- 1060 unique maintainers.
- 791 'first time' authors.

Lessons learned from package reviews I

1 Interoperability

- ▶ Use feature \times sample `SummarizedExperiment`, not sample \times feature matrix.
- ▶ Use paradigms familiar to *Bioconductor* users.

2 Reuse

- ▶ Use `rtracklayer::import.bed()`, not custom parser.

3 Robust code

- ▶ Edge cases: `seq_len()` / `seq_along()`, not `1:n`.
- ▶ Code complexity: `vapply()`, not `sapply()`.

4 Performant code

- ▶ *Vectorize* rather than *iterate* (`for`, `lapply()`, `apply()` are all iterative).
- ▶ Reuse (e.g., `matrixStats`) before C / C++ implementation.

Lessons learned from package reviews II

- 5 Tested code
 - ▶ Essential: evaluated example and vignette code chunks.
 - ▶ Desirable: unit tests, e.g., *testthat*.
- 6 Time and space limits.
 - ▶ Excessive computation may represent inefficient code.
 - ▶ Challenging to identify rich but modest data for illustration.
 - ▶ Experiment data packages, work flows, F1000 papers as venues for more expensive / comprehensive reproducible analysis.
- 7 Ambition
 - ▶ Implement essential features well.
 - ▶ Avoid dependencies on packages for marginal value.
- 8 Pretty
 - ▶ 'Poetry' with short lines, consistent and ample spacing, standard formatting.

Recent developments

- Git!

```
git clone https://git.bioconductor.org/packages/limma  
git clone git@git.bioconductor.org:packages/DESeq2
```

- Large Single Cell

- ▶ *SingleCellExperiment*
- ▶ *HDF5Array*

Large single-cell data

```
> sce = TENxBrainData::TENxBrainData()
snapshotDate(): 2017-10-30
> sce
class: SingleCellExperiment
dim: 27998 1306127
metadata(0):
assays(1): counts
rownames: NULL
rowData names(2): Ensembl Symbol
colnames(1306127): AACCTGAGATAGGAG-1 AACCTGAGCGGCTTC-1 ...
  TTTGTCAGTTAAAGTG-133 TTTGTCATCTGAAAGA-133
colData names(4): Barcode Sequence Library Mouse
reducedDimNames(0):
spikeNames(0):
```

Large single-cell data

- Chunk-wise iteration (often transparent to the user / developer).
- Marginal summaries in `rowData`, `colData`.
- Supporting infrastructure: *ExperimentHub*, *rhdf5*, *HDF5Array*, *DelayedMatrixStats*, *beachmat*.

Cloud computing

Possible visions

- As now, but 'in the cloud' – <https://rstudio.cloud>.
- Exploit cloud services, e.g., BigQuery.
- Pay-as-you-play – use existing *Bioconductor* AMIs or docker containers.
- Integrated with 'third party' compute efforts, e.g., NCI, NIH in the United States.
- Federated data access.

Events

- CSAMA (training), Brixen / Bressanone, Italy, 8 - 13 July.
- *Bioc2018* Toronto, Canada, 25 - 27 July.

Acknowledgments

Core team: Qian Liu, Valerie Obenchain, Hervé Pagès, Marcel Ramos, Lori Shepherd, Nitesh Turaga, Daniel van Twisk.

Technical advisory board: Vincent Carey, Kasper Hansen, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Levi Waldron, Michael Lawrence, Sean Davis, Aedin Culhane

Scientific advisory board: Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Rafael Irizzary (Dana Farber), Jan Vitek (Northeastern University), Robert Gentleman (23andMe).

Research reported in this presentation was supported by the National Human Genome Research Institute and the National Cancer Institute of the National Institutes of Health under award numbers U41HG004059 and U24CA180996. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.