Bioconductor Annual Report 2022

Lori Kern
Roswell Park Comprehensive Cancer Center
Vince Carey
Channing Division of Network Medicine
Harvard Medical School

July 28, 2023

Contents

1 Project Scope

Bioconductor provides access to software for the analysis and comprehension of high throughput genomic data. Packages are written in the *R* programming language by members of the *Bioconductor* team and the international community. *Bioconductor* was started in Fall, 2001 by Dr. Robert Gentleman and others, and now consists of 2183 software packages for the analysis of data ranging from single-cell sequencing to flow cytometry. There are also specialized packages for data experiments, annotations, and workflows. Bioconductor also hosts 3 comprehensive online books.

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.

1.1 Funding

Core team is funded primarily by NHGRI 5U24HG004059-19, "Bioconductor: An Open-Source, Open-Development Computing Resource for Genomics", R. Irizarry, contact PI. The grant is primed at Dana-Farber Cancer Institute with total annual budget of $1.3 million/year ending Feb 28 2026. Subcontracts are arranged with Mass General Brigham (V. Carey, MPI), Roswell Park Comprehensive Cancer Center (M. Morgan, co-PI), Fred Hutchinson Cancer Center (O. Hyrien, co-PI), City University of New York School of Public Health (L. Waldron, co-PI). Additional projects have been funded by NCI (5U24CA180996-10, "Cancer Genomics: Integrative and Scalable Solutions in R/Bioconductor"), NHGRI (2U24HG010263-06 "Expanding the AnVIL (Analysis, Visualization, and Informatics Lab-space)"). Several investigators in the Bioconductor community have received Chan Zuckerberg Initiative Essential Open Source Software and Single Cell Biology awards.

The organization chart indicates the locations and key task areas for core developer team members.

1.2 Package and Annotation Resources

R software packages represent the primary product of the *Bioconductor* project. Packages are produced by the *Bioconductor* team and from international contributors. Table **??** summarizes growth in the number of packages hosted by *Bioconductor*, with 2183 software packages available in release 3.16. The project produces 910 'annotation' packages to help researchers place analytic results into biological context. Annotation packages are

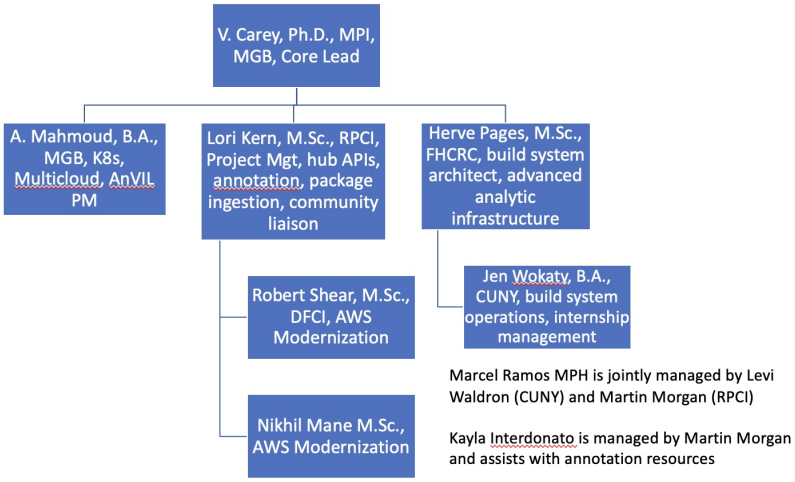

Figure 1: Core development organizational chart for 2023.

Table 1: Number of contributed packages included in each *Bioconductor* release. Releases occur twice per year.

Release		N	Release		N	Release		N	Release		N			
2002	1.0	15	2006	1.8	172	2010	2.6	389	2014	2.14	824	2018	3.7	1560
	1.1	20		1.9	188		2.7	419		3.0	936		3.8	1649
2003	1.2	30	2007	2.0	214	2011	2.8	467	2015	3.1	1024	2019	3.9	1741
	1.3	49		2.1	233		2.9	517		3.2	1104		3.10	1823
2004	1.4	81	2008	2.2	260	2012	2.10	554	2016	3.3	1211	2020	3.11	1903
	1.5	100		2.3	294		2.11	610		3.4	1294		3.12	1974
2005	1.6	123	2009	2.4	320	2013	2.12	671	2017	3.5	1381	2021	3.13	2042
	1.7	141		2.5	352		2.13	749		3.6	1473		3.14	2083
												2022	3.15	2140
													3.16	2183

curated resources derived from external data sources, and are updated at each release. The project also produces 416 'experiment data' packages to provide heavily curated results for pedagogic and comparative purposes. We have standardized reproducible, cross-package protocols into 28 'workflow' packages. There are also 3 'books' for in-depth analysis mostly focused on Single Cell Analysis.

The project has developed, over the last several years, the 'AnnotationHub' and 'ExperimentHub' resources for serving and managing genome-scale annotation data, e.g., from the TCGA, NCBI, and Ensembl. There are 69797 records in the AnnotationHub, and 6543 ExperimentHub records.

The number of distinct IP addresses downloading software continues to grow in an approximately exponential fashion (Figure **??**).

1.3 Courses and Conferences

Our annual conferences include:

- BioC2022 North American conference pivoted to a 3-day hybrid conference from July 27-29; in person components in Seattle were greatly reduced due to continued covid 2022 restrictions and protocols. Participation reached 94 in person and 328 virtual.

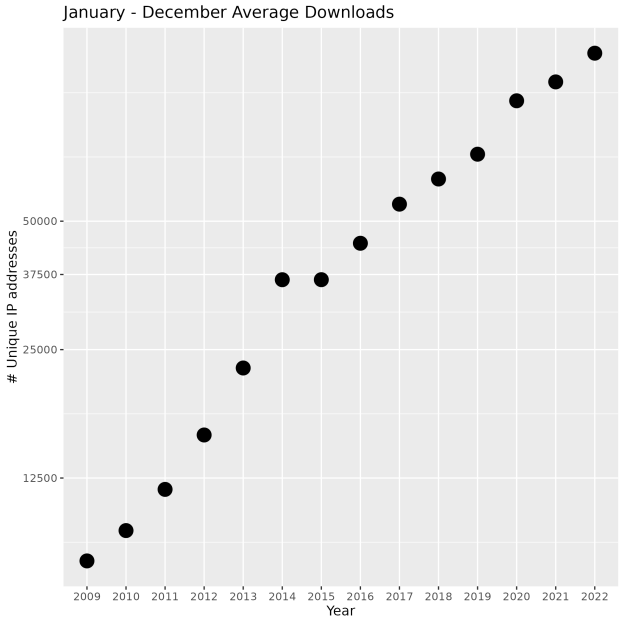

Figure 2: *Bioconductor* package download statistics, average number of unique downloads, first six months of each year.

- BioCAsia2022, held as a hybrid conference on December 2, 2022 with an in person component in Melbourne. There were estimated 96 in person and 121 virtual attendees.
- European Bioconductor Meeting, was held September 14-16, 2022 in in Heidelberg, Germany. Participation estimated at 90.

The course materials section of the web site summarizes material from these and some of the many other courses and conferences offered by *Bioconductor*and the events calendar list conferences, workshops, and other events that involve *Bioconductor*.

A summary of Bioc2022 presentations and workshops is given in Figure **??**.

1.4 Community Support

The *Bioconductor* support site is used for help, announcements, and outreach worldwide. From January 01, 2022 to December 31, 2022, there are 29064 new users, 1966 'top-level' posts, and 5127 comments (answers+comments).

The support site was upgraded and standardized to be consistent with the Biostars code base. Natay Aberra from Biostars has been instrumental in the transition; Lori Kern from the core team is working with Natay to be able to update and troubleshoot as necessary.

Another form of communication for the Bioconductor community is the *Bioconductor* community slack. As of December 31, 2022, there are 2141 members of the community slack channel with 114 different channels.

We continue to provide the bioc-devel, mailing list forum for package contributors' questions and discussion relating to the development of *Bioconductor* packages. There are 1960 subscribers on this list. Table **??** lists the number of posts and number of unique authors per month as a monthly average since 2002. Recent increase in activity is likely due to (1) enforced requirement that new package maintainers subscribe to the mailing list, and (b) using the bioc-devel mailing list as a support forum for use of git.bioconductor.org.

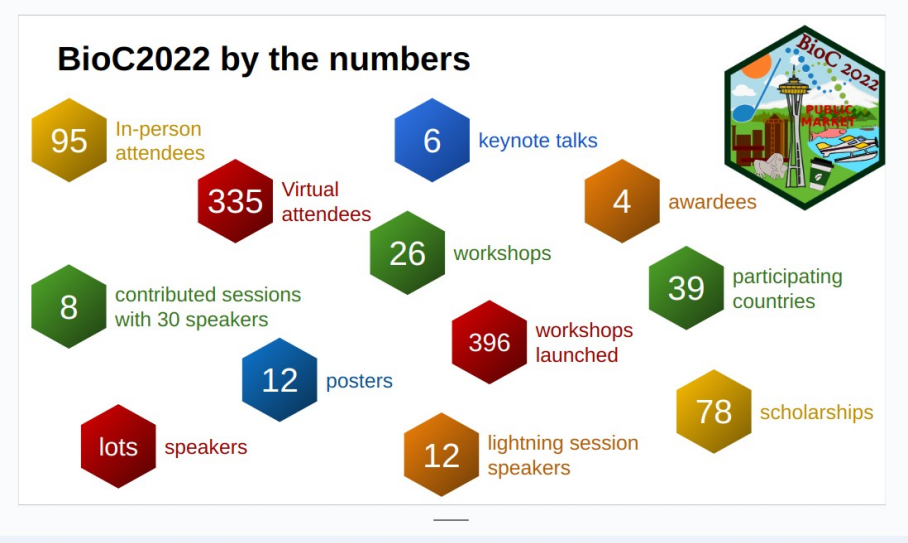

Figure 3: Review of events and participation in Bioc2022

Table 2: Monthly average number of posts and number of unique authors for the *Bioconductor* 'devel' mail list from January, 2005 to December, 2022

Year	Posts per month	Authors per month	Year	Posts per month	Authors per month	Year	Posts per month	Authors per month
2005	27	13	2011	52	24	2017	186	45
2006	39	19	2012	75	25	2018	160	48
2007	50	23	2013	97	34	2019	123	44
2008	27	18	2014	139	41	2020	134	47
2009	26	17	2015	142	43	2021	104	38
2010	30	18	2016	153	45	2022	51	23

1.5 Publication

Bioconductor has become a vital software platform for the worldwide genomic research community, with more than 60,500 PubMedCentral full-text citations for 'Bioconductor'. Table **??** summarizes PubMed author / title / abstract or PubMedCentral full-text citations since 2003.

Featured and recent publications citing *Bioconductor* are available on the *Bioconductor* web site, and are updated daily.

Table 3: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January, 2003 – July, 2020.

Year	N	Year	N	Year	N	Year	N	Year	N
2003	7	2007	44	2011	68	2015	3138	2019	5939
2004	13	2008	52	2012	1386	2016	3415	2020	4977
2005	19	2009	62	2013	2048	2017	3988	2021	5984
2006	30	2010	52	2014	2401	2018	4610	2022	5754

2 New and Ongoing Accomplishments

2.1 Leadership structure & community engagement

The Technical Advisory Board meets monthly to discuss technical issues important to establishing and maintaining project momentum. In 2022, Vincent Carey acted as chair, Levi Waldron vice-chair, and Charlotte Soneson as secretary. A view of the current TAB membership is here.

The Community Advisory Board meets monthly to discuss community driven issues including outreach, education, diversity, and inclusion. In 2022, Aedin Culhane and Susan Holmes acted as co-chairs and Lori Kern as secretary. A view of the current CAB membership is here.

The Community and Technical boards, and the overall leadership structure of the project remains a work-in-progress. There is a need for established lines of communication and coordination between boards, as well as a clear organizational plan describing the relationship between them.

Both the Technical Advisory Board and Community Advisory Board have open nominations and elections yearly with elected members serving three year terms.

A newer concept introduced in 2022, Bioconductor would like to explore collaboration through working groups and committees. Anyone from the community is welcome to start or join an existing working group to try to encourage and foster collaboration. The Working Group materials can be browser here.

2.2 Software

Bioconductor continues to have biannual releases where new packages are included into the Bioconducto ecosystem. During 2022, release 3.15 and 3.16 occurred. See full release announcements and Table **??**.

Table 4: Bioconductor Release 3.15 and 3.16

Release	Release Date	R Version	Packages	Announcement
3.15	April 27, 2022	4.2	2140	3.15
3.16	November 2, 2022	4.2	2183	3.16

2.3 Infrastructure

Version control We continue to use package-specific git repositories hosted at git.bioconductor.org for package maintenance.

Some aspects of the use of git.bioconductor.org, especially access management, represent pain points (as evidenced by frequent reports of difficulties on the bioc-devel mailing list) where further effort is needed to smooth the experience.

New package contributions use Github and a public review process. The process has been updated so that new maintainers become familiar with use of the *Bioconductor* git repository during the package ingestion process. Effort expended on reviewing packages is considerable; generally, the review process has become both more protracted and less comprehensive in response to this.

Bioconductor Build System (BBS) is an open system for orchestrating package testing and assembly into distributable source tarballs or binaries for Windows or Mac (in Intel and ARM flavors). The system was ported to do checks for the ARM Linux platform through a pro bono collaboration with cloud engineers from Huawei. See the extended build reports for platform kunpeng2.

Single package builder (SPB) is used to build packages in the review process on commit. Builds occur across Linux, macOS, and Windows environments that closely resemble the *Bioconductor* build system, providing developers with immediate feedback for iterative improvement of their packages. The SPB has been updated to use git.bioconductor.org as a repository source. This was done so that new package contributes use the *Bioconductor* git repository during the review process (see previous point). An easy technical extension

is to allow build-on-commit for existing as well as new packages; this would provide a build experience, complementary to the nightly builds, that is more comparable to the continuous integration systems many of our developers are familiar with. A necessary step before implementing this is to understand whether a build-on-commit system would be too taxing to the physical resources of the build system.

Cloud services Amazon Web Services is used to manage git version control, assembly of download statistics, distribution of software and data via CloudFront, and resource archiving and distribution via S3 buckets. Microsoft Azure is also significant for the project

Workshop authoring and deployment platform https://workshop.bioconductor.org/ is a Kubernetes-backed Galaxy deployment that serves Rstudio instances to workshop attendees in a scalable way. This originated in Sean Davis' app.orchestra.cancerdatasci.org.

2.4 User Support

Support site has established itself as an important resource. We have been engaged in an extended collaboration with Biostars author to harmonize our code base with upstream code, to enhance security, and to prepare for the release of an updated support forum.

Workflows provide cross-package training material and integrate with the F1000 *Bioconductor* channel. Workflows are now distributed as standard *R* packages built regularly, distributed through CRAN-style repositories, and organized on the web site using the same approach as other package types.

Slack channels for the core team and *Bioconductor* community are providing new avenues for communication. The community slack channel was an important catalyst in the HCA grantsmanship process, and in several significant collaborative software initiatives lead by community members.

Use of slack within the community poses several challenges. Support channels have become fragmented, with users and developers posting requests to the support site, developer mailing list, specific issues on github repositories, and slack. Even with a substantial discount, the slack channel is increasingly expensive. The large number of channels, and the opportunity for private messaging, poses challenges for ensuring community code of conduct and appropriate use.

Course Materials organize and make accessible recent course and training material.

3 Core Tasks & Capabilities

3.1 Core Tasks

1. Package Building and Testing. The *Bioconductor* project provides access to its packages through repositories hosted at `bioconductor.org`. One of the services provided to the *Bioconductor* community is nightly automated build and check of all packages. Maintaining the automated build and test suite and keeping the published package repositories updated requires a significant amount of time on the part of the Roswell *Bioconductor* team. As the project has grown, the organizational and computational resources required to sustain the package build system have also increased; see section **??**.
2. Package dissemination via https://bioconductor.org and underlying CRAN-style repository using Amazon CloudFront for global distribution.
3. Software development.
4. End-user support via https://support.bioconductor.org and the bioc-community slack channel.
5. Developer support via the bioc-devel mailing list.
6. New package submission. The *Bioconductor* project relies on technical review process of candidate packages to ensure they contain high-quality software.
7. Annotation data packages. The *Bioconductor* project synthesizes genomic and proteomic information available in public data repositories in order to annotate genomic sequences and probes of standard microarray chips. These annotation data packages are made available to the community and allow *Bioconductor* users to easily access meta data relating to their experimental platform. We maintain automated tools to parse the available information.

8. Semi-annual releases, typically in March and October.

3.2 Hardware and Infrastructure

The *Bioconductor* project provides packages for computing platforms common in the informatic community. We provide source packages that can be installed on Linux and most UNIX-like variants, as well as binary packages for Windows and macOS. To ensure that packages are consistently documented, easy to install, and functioning properly, we run a nightly build during which we test all packages in the release and development repositories.

The build system currently consists of at least two Windows machines, two Linux machines, and two macOS machines. The Windows, Linux, and one macOS machines are physical servers located at Roswell Park, the remaining macOS machine is rented via MacStatdium. The web site, support site, AnnotationHub, and additional servers are hosted on virtual machines, some of which are Amazon machine instances. The build machines are recently updated, with adequate room for growth.

New infrastructure has been introduced to produce Linux binaries for use with Docker containers. The production system is based on Kubernetes and has run in Google Kubernetes Engine and in the NSF Jetstream2 academic cloud. Binary production can also be accomplished at zero cost using GitHub actions. An example repository that currently produces over 2151 software package binaries for ARM linux containers is available. This system was conceived and deployed by Alex Mahmoud.

4 Key Personnel

The **Core Development Team** are responsible for developing software and other infrastructure and ensuring day-to-day operation of the project. Core team members in the period covered by this report include Vince Carey (Project Lead), Lori Kern (Project Manager) Hervé Pagès, Marcel Ramos, Alexandru Mahmoud, Robert Shear, Jennifer Wokaty and Kayla Interdonato. The core team is stable but in chronic need of additional members.

The **Technical Advisory Board** provides guidance through monthly telephone conference calls. Current members include: Vince Carey, Brigham & Women's, Harvard Medical School, USA. Chair; Levi Waldron CUNY School of Public Health at Hunter College, New York, NY, Vice-Chair; Charlotte Soneson, Friedrich Miescher Institute, Basel, Switzerland, Secretary; Aedin Culhane, Dana-Farber Cancer Institute, Harvard School of Public Health, USA; Sean Davis, Univesiry of Colorado Anschutz School of Medicine, USA; Laurent Gatto Institut de Duve, Belgium; Robert Gentleman, Harvard Medical School, USA; Shila Ghazanfar, Cancer Research UK, Cambridge; Kasper Daniel Hansen, Bloomberg School of Public Health, Johns Hopkins University; Stephanie Hicks Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, USA; Wolfgang Huber European Molecular Biology Laboratory, Heidelberg, Germany; Rafael Irizarry Dana-Farber Cancer Institute, USA; Lori Kern, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA; Michael Love, University of North Carolina-Chapel Hill, USA; Davide Risso, University of Padova, Italy.

The **Community Advisory Board** supports the *Bioconductor* mission by empowering user and developer communities by coordinating training and outreach activities, and enabling productive and respectful participation by Bioconductor users and developers at all levels of experience. Current members include: Aedin Culhane, Dana-Farber Cancer Institute, USA (co-chair); Susan Holmes, Stanford, USA (co-chair); Lori Kern, Roswell Park Comprehensive Cancer Center, USA (secretary); Yagoub Adam, Covenant University, Nigeria; Benilton Carvalho, University of Campinas, Brazil; Leonardo Collado-Torres, Lieber Institute for Brain Development, USA; Kozo Nishida, RIKEN Center for Biosystems Dynamics Research, Japan; Johannes Rainer, Eurac Research, Italy; Matt Ritchie, The Walter and Eliza Hall Institute of Medical Research, Australia; Daniela Cassol, University of California, Riverside, USA; Leo Lahti, University of Turku, Finland; Estefania Mancini, Centre for Genomic Regulations, Spain; Nicole Ortogero, NanoString Technologies, USA; Janai Ravi, University of Colorado Anschultz, USA; Kevin Rue-Albrecht, University of Oxford, UK; Mike Smith, EMBL, Germany; Hedia Tnani, CNAG, Spain.

The **Scientific Advisory Board** provides oversight through yearly meetings. Current members include: Rafael

Irizarry, DFCI; Benilton S Carvalho, Universidade Estadual de Campinas; Benjamin Neale, Broad Institute; Daniela Witten, U Washington; Martin Morgan, RPCCC; Robert Gentleman, HMS; Sandrine Dudoit, UC Berkeley; Susan Holmes, Stanford; Wolfgang Huber, EMBL; Mike Schatz, JHU; Chris Wellington, NHGRI.