Bioconductor 2023 Annual Report

Lori Kern (Roswell Park) Maria Doyle (Limerick) Vincent Carey (Harvard)

April 17, 2024

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 2266 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 4 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.

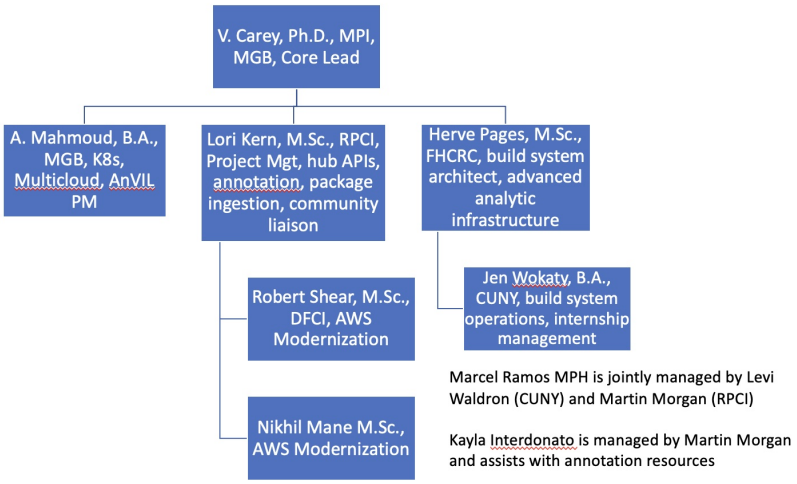

Figure 1: Core development organizational chart for 2023.

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~1 summarizes growth in the number of packages hosted by *Bioconductor*, with 2266 software packages available in release 3.18. The project produces 920

'annotation' packages to help researchers place analytic results into biological context. Annotation packages are curated resources derived from external data sources, and are updated at each release. The project also produces 429 'experiment data' packages to provide heavily curated results for pedagogic and comparative purposes. We have standardized reproducible, cross-package protocols into 30 'workflow' packages. There are also 4 '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 71082 records in the AnnotationHub, and 7282 ExperimentHub records.

The number of distinct IP addresses downloading software continues to grow in an approximately exponential fashion (Figure~2).

Bioconductor allows public and private mirrors; traffic on these mirrors is not included in the provided download statistics. The current number of public mirrors and private mirrors are 14 and 27.

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	2007	2.0	214	2012	2.10	554	2017	3.5	1381	2022	3.15	2140
	1.1	20		2.1	233		2.11	610		3.6	1473		3.16	2183
2003	1.2	30	2008	2.2	260	2013	2.12	671	2018	3.7	1560	2023	3.17	2230
	1.3	49		2.3	294		2.13	749		3.8	1649		3.18	2266
2004	1.4	81	2009	2.4	320	2014	2.14	824	2019	3.9	1741			
	1.5	100		2.5	352		3.0	936		3.10	1823			
2005	1.6	123	2010	2.6	389	2015	3.1	1024	2020	3.11	1903			
	1.7	141		2.7	419		3.2	1104		3.12	1974			
2006	1.8	172	2011	2.8	467	2016	3.3	1211	2021	3.13	2042			
	1.9	188		2.9	517		3.4	1294		3.14	2083			

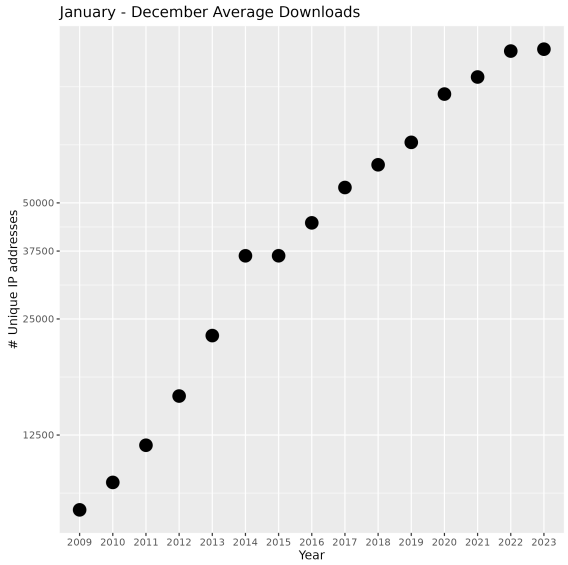

Figure 2: *Bioconductor* package download statistics, average number of unique downloads, first six months of each year.

1.3 Courses and Conferences

Our annual conferences include:

- BioC2023 North American conference pivoted to a 3-day hybrid conference from August 2-4; in person components in Boston were greatly reduced due to continued covid 2023 restrictions and protocols. Participation reached 163 in person and 433 virtual. See this blog post for recap.
- European Bioconductor Meeting, was held September 20-22, 2023 in in Gent, Belgium. Participation estimated at 120.
- BioCAsia2023, held October 15-17, 2023 in Hong Kong. There was estimated 164 in person attendees. A key outcome of this conference was the formation of a working group to establish a Hong Kong Bioinformatics Society.

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 Bioc2023 presentations and workshops is given in Figure 3.

Figure 3: Review of events and participation in Bioc2023

1.4 Community Support

The *Bioconductor* support site is used for help, announcements, and outreach worldwide. From January 01, 2023 to December 31, 2023, there are 18297 new users, 1679 'top-level' posts, and 4531 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. There were 842 new accounts created in 2023. As of December 31, 2023, there are 2980 members of the community slack channel with 139 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 2048 subscribers on this list. Table~2 lists the number of posts and number of unique authors per month as a monthly average since 2003. Recent decline is likely due to the introduction of the community slack mentioned above.

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	2012	75	25	2019	124	44
2006	39	19	2013	97	34	2020	134	47
2007	50	23	2014	139	41	2021	99	37
2008	27	18	2015	142	43	2022	51	23
2009	26	17	2016	153	45	2023	61	28
2010	30	18	2017	186	54			
2011	52	24	2018	160	48			

The Bioconductor Blog gained some momentum. The blog features community contributed artibles related to Bioconductor projects.

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~3 summarizes PubMed author / title / abstract or PubMedCentral full-text citations since 2003.

Table 3: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January 2003 – December 2023.

Year	N	Year	N	Year	N	Year	N	Year	N
2003	7	2008	52	2013	2048	2018	4610	2023	5163
2004	13	2009	62	2014	2401	2019	5939		
2005	19	2010	52	2015	3138	2020	4977		
2006	30	2011	68	2016	3415	2021	5984		
2007	44	2012	1386	2017	3988	2022	5754		

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

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 2023, 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 2023, Kevin Rue-Albrecht and Janani Ravi acted as co-chairs and Johannes Rainer 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 browsed here.

Bioconductor started particpating in Outreachy in 2022. See the intern blog post for an account of experiences of the first set of sponsored interns.

Bioconductor hired a Community Manager, Maria Doyle, in September 2022, to oversee community driven projects, such as the website redesign and the global Bioconductor Carpentries training program. See this CSCCE profile for a summary of the first year of these community management activities.

2.2 Fiscal sponsorship

After consultation with La Piana's Lara Jabukowski, we applied for fiscal sponsorship with https://numfocus.org. From their website:

> The mission of NumFOCUS is to promote open practices in research, data, and scientific computing by serving as a fiscal sponsor for open source projects and organizing community-driven educational programs.

The application was accepted and the Bioconductor Foundation of N.A., Inc., is being dissolved in 2024 with transfer of assets to NumFOCUS. This will simplify administration of industry sponsorships of conferences, management of donations, and administration of international grants and contracts.

2.3 Software

Bioconductor continues to have biannual releases where new packages are included into the Bioconductor ecosystem. During 2023, release 3.17 and 3.18 occurred. See full release announcements and Table 4.

Table 4: Bioconductor Release 3.17 and 3.18

Release	Release Date	R Version	Packages	Announcement
3.17	April 26, 2023	4.3	2230	3.17
3.18	October 25, 2023	4.3	2266	3.18

2.4 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.

Significant effort during 2023 involved transitioning the *Bioconductor* git.bioconductor.org default branch from designation 'master' to 'devel'. *Bioconductor* recognizes use of certain terminology should be avoided to help advance efforts of inclusivity and that are consistent with the project's Code Of Conduct.

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. Contributors will also now receive build reports before a review is assigned with the intention that build reports should have no ERRORs, few or no WARNINGs, and justifable NOTEs before an indepth review is necessary. This lightens the burden on reviewers to comment on issues noted in the reports.

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 and macOS 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 and database like structure for recording past and recent builds of packages.

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 hosting resources for the ExperimentHub and AnnotationHub, as well as providing virtual machines used in the BBS.

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.

2.5 User Support

Website bioconductor.org provides information for users and developers about *Bioconductor*. It provides helpful guides and lists all available *Bioconductor* packages.

Bioconductor enlisted Nearform to help redesign the main project website in 2023. The overall goal of the redesign was to simplify and organize material in a more comprehensive manner especially to newcomers. For more details, see the blog posts

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

Workflows Workflow Packages 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 led 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 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 Core *Bioconductor* team. As the project has grown, the organizational and computational resources required to sustain the package build system have also increased; see section 3.2.
2. Package dissemination via https://bioconductor.org and underlying CRAN-style repository using Amazon CloudFront for global distribution.
3. Container development and maintenance, in GitHub. Compatible binaries are produced and are made available through BiocManager.
4. Software development and maintenance.
5. End-user support via https://support.bioconductor.org and the bioc-community slack channel.
6. Developer support via the bioc-devel mailing list.
7. New package submission. The *Bioconductor* project relies on technical review process of candidate packages to ensure they contain high-quality software.
8. 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.
9. 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.

Package build and check processes are conducted for Windows, macOS, and Linux systems. The Windows builders are run in Microsoft Azure, thanks to support of Microsoft Genomics and Erdal Cosgun. The macOS builders consist of an ARM Mac Pro at Dana-Farber, and Intel Macs rented at MacStadium.com.
Linux systems with Intel hardware are managed at Dana-Farber. Martin Grigorov of Huawei runs an ARM Linux builder for testing on that platform. The Microsoft Azure Cloud is used for hosting the services underlying ExperimentHub and AnnotationHub. An NSF-ACCESS allocation (BIR190004) provides 30 TB of egress-free S3 storage, and substantial access to virtual machines and GPUs.

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.

A new project "bioc2u" produces fully linked Debian packages for almost all Bioconductor packages. See the bioc2u repository.

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, Nikhail Mane, 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; Henrik Bengtsson, University of California San Francisco, USA; Helena Crowell, National Center of Genomic Analysis, Spain; Sean Davis, University of Colorado Anschutz School of Medicine, USA; Laurent Gatto Institut de Duve, Belgium; Ludwig Geistlinger, Center for Computational Biomedicine, Harvard Medical School, USA; 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: Enis Afgan, Johns Hopkins University, USA; Umar Ahmad, Bauchi State University, Nigeria; Daniela Cassol, University of California, Riverside, USA; Aedin Culhane, University of Limerick, Ireland; Xueyi Dong, Walter and Eliza Hall Institute of Medical Research, Australia; Maria Doyle, University of Limerick, Ireland; Lori Kern, Roswell Park Comprehensive Cancer Center, USA; Leo Lahti, University of Turku, Finland; Mengbo Li, Walter and Eliza Hall Institute of Medical Research, Australia; Estefania Mancini, Centre for Genomic Regulations, Spain; Jordana Muqanguzi, Northeastern University, USA; Kozo Nishida, RIKEN Center for Biosystems Dynamics Research, Japan; Nicole Ortogero, NanoString Technologies, USA; Stevie Pederson, Telethon Kids Institute, Australia; Johannes Rainer, Eurac Research, Italy; Janai Ravi, University of Colorado Anschutz, USA; Kevin Rue-Albrecht, University of Oxford, UK; Mike Smith, EMBL, Germany; Luyi Tian, Guangzhou Laboratory, China; Hedia Tnani, CNAG, Spain. Jiefei Wang, University of Texas Medical Branch, USA.

The **Scientific Advisory Board** provides oversight through yearly meetings. Current members include: Robert Gentleman, HMS; Rafael Irizarry, DFCI; Vince Carey, Brigham & Women's; Chris Wellington, NHGRI; Wolfgang Huber, EMBL; Martin Morgan, RPCCC; Benjamin Neale, Broad Institute; Mike Schatz, JHU; Aviv Regev, Genentech; Sandrine Dudoit, UC Berkeley; John Marioni, EMBL; Daniela Witten, U Washington; Barbara Engelhardt, Princeton; Susan Holmes, Stanford; Benilton S Carvalho, Universidade Estadual de Campinas; Jay Shendure, U Washington.