Bioconductor Annual Report (preliminary)

Martin Morgan
Roswell Park Comprehensive Cancer Center

July 31, 2018

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 1903 software packages for the analysis of data ranging from single-cell sequencing to flow cytometry.

1.1 Funding

Bioconductor funding is summarized in Table 1.

The project is primarily funded through National Human Genome Research Institute award U41HG004059 (Community Resource Project; Morgan PI, with Carey and Irizzary), 'Bioconductor: An Open Computing Resource for Genomics'. The grant expires February, 2021.

The project receives additional funding through U24CA180996 (Morgan MPI, with Waldron MPI, Carey, Risso), 'Cancer Genomics: Integrative and Scalable Solutions in *R* / *Bioconductor*'. This provides funding through 2024.

Table 1: *Bioconductor*-related funding

	Award	Start	End
NHGRI / NIH	U41HG004059	3/1/2016	2/28/2021
NCI / NIH	U24CA180996	9/1/2014	8/31/2024
NHGRI / NIH	U24HG010263	9/21/2018	6/30/2023
NCI / NIH	U01CA214846	5/1/2017	4/30/2020
Chan / Zuckerberg	Seed Networks	6/1/2019	5/31/2021
Chan / Zuckerberg	EOSS		

Table 2: 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			
2005	1.6	123	2009	2.4	320	2013	2.12	671	2017	3.5	1381			
	1.7	141		2.5	352		2.13	749		3.6	1473			

Bioconductor participates in the AnVIL project (U24HG010263) for 'Implementing the Genomic Data Science Analysis, Visualization And Informatics Lab-Space' provides opportunities for cloud based data access and computation. This provides funding through 2023.

Funding from the Chan / Zuckerberg foundation provides support for 8 *Bioconductor* research groups to develop software for access and representation of Human Cell Atlas single-cell data, methods and benchmarks for emerging and integrative data, and methods for scalable and performant analysis. This provides funding through May, 2021.

Funding supports 6 - 7 full-time personnel at RPCI, plus additional individuals at subcontract sites; see section 3.3.

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 2 summarizes growth in the number of packages hosted by *Bioconductor*, with 1903 software packages available in release 3.11. The project produces 962 '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 392 'experiment data' packages to provide heavily curated results for pedagogic and comparative purposes. We have standardized reproducible, cross-package protocols into 27 'workflow' packages.

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 50277 records in the AnnotationHub, and 2850 ExperimentHub records.

The number of distinct IP addresses downloading software continues to grow in an approximately exponential fashion (Figure 1).

1.3 Courses and Conferences

Our annual conferences include:

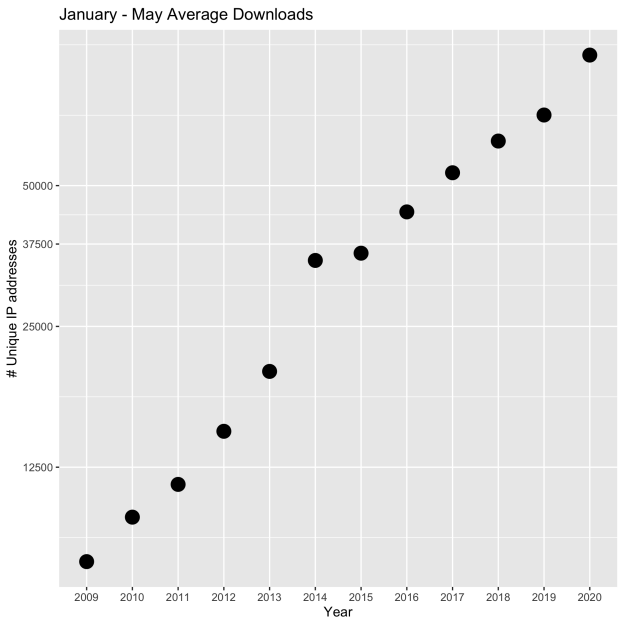

Figure 1: *Bioconductor* package download statistics, average number of unique downloads, first six months of each year.

- BioC2019 North American conference was held in New York City, reaching our capacity of 220+ participants. BioC2020 pivoted to a 5-day virtual conference, with reduced registration fee and participation reaching our maximum capacity of 500 attendees.
- BioC Asia, held in Sydney, Australia, December, 2019. The 2020 event will be held in Beijing, China, and will be a virtual event.
- European Bioconductor Meeting, held in Brussels, December, 2019. The 2020 event will be organized for Padua, Italy, and will be a virtual event.

A significant annual event, CSAMA, attracts approximately 80 participants to a week-long training course; it was held in July 2019, but was postponed due to COVID-19 restrictions for the 2020 season.

The course materials section of the web site summarizes material from these and some of the many other courses and conferences offered by *Bioconductor*.

1.4 Community Support

The *Bioconductor* support site has about 180 new 'top-level' posts and 584 comments or answers per month. There were about 17000 (Google analytics) sessions per week. Statistics are summarized in Table 3, with a decline in visitors and support site activity over the last year.

The decline in visitors and support site use is strongly influenced by a technical issue that unfortunately persisted for several months. We experienced a 'denial of service' attack in March of 2020. In responding to this, we disabled 'robot' (web crawler) access to the support site, so that google and other search engines did not index the site. Unfortunately, once recovering from the attack, we neglected to re-enable robots, so the support site disappeared from the search engines. Thus any new users are discovering the site through documentation rather than searches.

A secondary reason for changes in support site use may be increasing reliance on the *Bioconductor* community slack.

Table 3: Support site visitors from October, 2014. Users: registered users visiting during the reporting period; Visitors: Google analytics visitors during the reporting period. 2014-15 spans 10-months. Subsequent values are trailing 12 months from data of annual report.

Year	Users	Visitors	Posts	Replies
2014-15	2179	122,332	2169	6535
2015-16	3101	297,467	3359	10976
2016-17	3426	343,459	3346	13077
2017-18	4162	429,977	3354	9515
2018-19	6042	492,422	2873	8556
2019-20	3517	387,269	2180	7012

Table 4: Monthly average number of posts and number of unique authors for the *Bioconductor* 'devel' mail list from January, 2005.

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	164	53
2009	26	17	2015	142	43			
2010	30	18	2016	153	45			

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 1615 subscribers on this list (versus 1504 in the last report). Table 4 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. Discussion of the latter occurs below.

Web site access is summarized in Figure 2. The web site served 2.9M sessions (1.1M users) in the trailing 12 months (statistics from Google analytics). Visitors come from the United States (30%), China (15%), the United Kingdom (5.2%), Germany (4.9%), Japan, India, France, Canada, Spain, Australia, and 213 other countries. Unique visitors grew by 28%, substantially more than last year's 14% increase.

1.5 Publication

Bioconductor has become a vital software platform for the worldwide genomic research community, with more than 38,500 PubMedCentral full-text citations for 'Bioconductor'. Table 5 summarizes PubMed author / title /

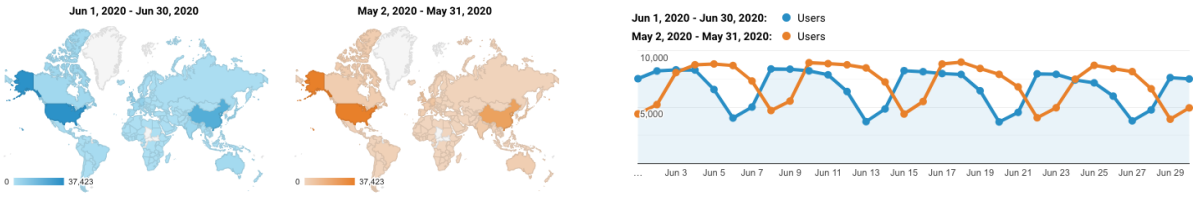

Figure 2: *Bioconductor* Access Statistics. Left: international visits, trailing 12 months. Right: Web site access, June 2020 (orange) and 2019 (blue).

Table 5: 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*	3224*
2005	19	2009	62	2013	2048	2017	3988		
2006	30	2010	52	2014	2401	2018	4610		

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.

2 New and Ongoing Accomplishments

2.1 Leadership structure & community engagement

The Technical Advisory Board (membership enumerated below) meets monthly to discuss technical issues important to establishing and maintaining project momentum. Over the past two years the structure has been formalized with a governance document and procedures to ensure influx of new participants through a public nomination process followed by election to a three year term. The July, 2020 meeting saw the departure of long-term board members Sean Davis and Michael Lawrence, and Matt Ritchie's transition to co-leadership of the newly-formed Community Advisory Board. New members include Drs. Michael Love (UNC Chapel Hill), Hector Corrada Bravo (Genentech), and Shila Ghazanfar (Cancer Research UK, Cambridge). Martin Morgan departed as chair of the executive, with Vincent Carey taking that role (previously, vice-chair), Levi Waldron becoming vice-chair (previously secretary), and Charlotte Soneson becoming secretary.

The Community Advisory Board was established this year to more directly address the training and outreach mission of *Bioconductor*. Motivation is several-fold. It is clear the project as a whole would benefit from attention focused on community engagement. The combined purview of the technical and community boards is too large for a single board. The second board expands leadership opportunities within the project.

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. Technical activities of the board can be supported by grant funding, particularly though not exclusively to support the core team, whereas activities of the community board may both generate revenue (e.g., from conferences) and expenditures in ways that are less consistent with current funding sources. The Bioconductor Foundation of NA may represent one mechanism for managing these financial resources, but this implies greater integration of the Foundation into the overall organization structure.

2.2 Containers and clouds

Involvement with the AnVIL project has helped to clarify container and cloud based strategies for *Bioconductor*, resulting in several interesting developments.

We have revised our docker strategy to provide images with necessary system dependencies, rather than pre-installed packages. This provides a flexible image that can be easily tailored to the diverse needs of our community. The images are built on community-standard 'rocker' base images. One consequence of a docker image is that fixed underlying system requirements allow Linux-based 'binary' package installation. Binary installations are much faster than source installations, allowing containers to be used effectively for tasks with relatively short life

cycles, e.g., during teaching, running workflows, etc. We have explored binary *Bioconductor* package repositories within AnVIL, and have to some extent had binary images thrust upon us by the adoption of the RStudio CRAN package manager by the underlying rocker container; there have been important lessons learned about providing robust and stable containers, including the need for much more extensive integration tests of our own containers.

Our involvement with the AnVIL project has provided opportunities to use our container knowledge to explore cloud-based computation. Again there are several insights from this, although the situation seems one more of promise that delivery at the moment. Access to large data resources, including data requiring authenticated access and to scalable computing capabilities seems very promising. The use of containers as a mechanism for delivering *R* / *Bioconductor*, including *RStudio*, moves the burden of 'system administration' from the user to the *Bioconductor* core team. Cost-management and the need to fund use through Google and credit card introduces many barriers to use by the *Bioconductor* community. Workshop infrastructure developed by Sean Davis for our just-concluded annual conference illustrates opportunities for scalability through the use of clouds and containers, with thousands of workshops launched with minimal end-user complaint and modest cost using Google Kubernettes Engine.

2.3 Software

BiocIO extracts from the *rtracklayer* package a common language and flexible infrastructure for importing and exporting file types, simplifying file access and focusing development efforts on more advanced file processing models.

GenomicRanges and friends represent a mature infrastructure for working with range-based and sequence data.

DelayedArray and the *HDF5Array* back-end provide a framework for managing large out-of-memory rectangular data representations.

BiocFileCache manages a cache of local or remote files.

AnnotationHub and *ExperimentHub* and supporting infrastructure play increasingly important roles in distribution of annotation and experiment results.

Incremental enhancements to *BiocParallel*, *GenomicFiles*, *BiocCheck*, *MultiAssayExperiment*, *SummarizedExperiment* and other core packages.

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.

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.

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.

2.5 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 3.2.
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.

3.3 Key Personnel

The **Core Development Team** are primarily employees of Roswell Park Cancer Institute, 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 Martin Morgan, Hervé Pagès, Marcel Ramos, Lori Shepherd, Nitesh Turaga, Daniel van Twisk, 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; Hector Corrada Bravo, Genentech Research and Early Development, USA; Laurent Gatto Institut de Duve, Belgium; Robert Gentleman, Computational Biology, 23andMe, 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; Aaron Lun, Genentech Research and Early Development, USA; Michael Love, University of North Carolina-Chapel Hill, USA; Martin Morgan, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA.

The recently formulated **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: Yagoub Adam, Covenant University, Nigeria; Benilton Carvalho, University of Campinas, Brazil; Leonardo Collado-Torres, Lieber Institute for Brain Development, USA; Aedin Culhane, Dana-Farber Cancer Institute, USA; Saskia Freytag, Harry Perkins Institute of Medical Research, Australia; Susan Holmes, Stanford, 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; Lori Shepherd, Roswell Park Comprehensive Cancer Center, USA.

The **Scientific Advisory Board** provides oversight through yearly meetings. Current members include: Robert Gentleman (Advisory Board Chair, 23andMe); Jenny Bryan (RStudio); Vincent Carey (Brigham & Women's); Valentina di Francesco (NHGRI); Wolfgang Huber (European Molecular Biology Laboratory); Rafael Irizarry (Dana Farber); Audrey Kauffmann (Novartis); Martin Morgan (Roswell Park); Benjamin Neale (Broad Institute); Mike Schatz (Johns Hopkins University); Jay Shendure (University of Washington); Levi Waldron (CUNY School of Public Health).