Bioconductor Annual Report

Martin Morgan

Roswell Park Comprehensive Cancer Center

June 26, 2019

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 1741 packages for the analysis of data ranging from 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 provides funding through 2021.

Table 1: *Bioconductor*-related funding

	Award	Start	End
NHGRI / NIH	U41HG004059	3/1/2016	2/28/2021
NCI / NIH	U24CA180996	9/1/2014	8/31/2019
NCI / NIH	U01CA214846	5/1/2017	4/30/2020
NHGRI / NIH	U24HG010263	9/1/2018	6/30/2023
Chan / Zuckerberg		7/1/2019	6/30/2022

The project receives additional funding through U24CA180996 (Morgan PI, with Carey, Hansen, Waldron), 'Cancer Genomics: Integrative and Scalable Solutions in *R* / *Bioconductor*'. This provides funding through 2019. A renewal, with Morgan and Waldron as MPI, was submitted 6/13/2019.

Carey receives funding through U01CA214846 for 'Accelerating cancer genomics with cloud-scale *Bioconductor*'.

Morgan, Carey, and Waldron are members of a large collaboration developing NHGRI cloud resources under U24HG010263 'Implementing the Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)' (James Taylor, PI). Funding is available through June, 2023.

Funding from the Chan / Zuckerberg foundation will provide support for an 8-member collaboration lead by Dr. Morgan to access and analysis of Human Cell Atlas data. Collaborators include Drs. Culhane, Finak, Hansen, Hicks, Huber, Risso, and Ritchie.

Funding supports 6 - 7 full-time personnel at RPCI, plus additional individuals at subcontract sites; see section 2.3.

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 1741 software packages available in release 3.9. The project produces 948 'annotation' packages to help researchers place analytic results into biological context. Annotation packages are currated resources derived from external data sources, and are updated at each release. The project also produces 371 'experiment data' packages to provide heavily currated 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 46259 records in the AnnotationHub, and 2329 ExperimentHub records.

The number of distinct IP addresses downloading software continues to grow in an approximately exponential fashion (Figure 1), surpassing 1/2 million unique IP addresses last year.

Some insight into new package dynamics comes from 'biocViews' terms authors apply to their package. The terms come from a controlled vocabulary arranged as a directed acyclic graph. Figure **??** summarizes change in four categories of biocViews terms. Transcriptomics,

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	2007	2.0	214	2012	2.10	554	2017	3.5	1381
	1.1	20		2.1	233		2.11	610		3.6	1473
2003	1.2	30	2008	2.2	260	2013	2.12	671	2018	3.7	1560
	1.3	49		2.3	294		2.13	749		3.8	1649
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			
2005	1.6	123	2010	2.6	389	2015	3.1	1024			
	1.7	141		2.7	419		3.2	1104			
2006	1.8	172	2011	2.8	467	2016	3.3	1211			
	1.9	188		2.9	517		3.4	1294			

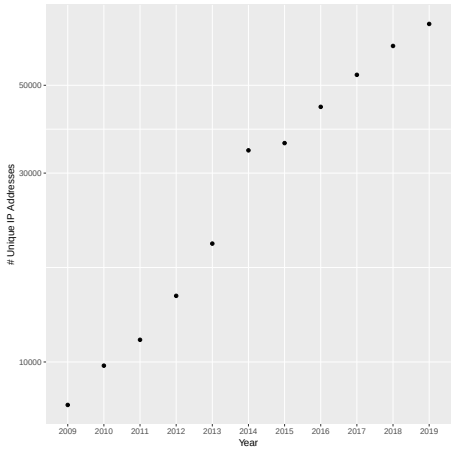

Figure 1: *Bioconductor* package download statistics, average number of unique downloads, first five months of each year

proteomics, metabolomics, and biomedical informatics research fields have all increased in representation; sequencing and single-cell, and mass-spectronomy technologies have increased representation, primarily at the expense of microarrays.

1.3 Courses and Conferences

Course and conference material and announcements for upcoming events are available. A partial list of courses and conferences with significant input from key *Bioconductor* personnel have been held in the following worldwide locations in the last year:

- Morgan, M.T., Waldron, L., Carey, V., Huber, W., 2019 (July) CSAMA 2018: Statistical Data Analysis for Genome-Scale Biology, various lecture and lab contributions in a week-long course. Italy.
- Morgan, M.T., 2019 (July) How *Bioconductor* advances science while contributing to the *R* language and community. *useR!* Keynote Address, Toulouse, France.
- Morgan, M.T., 2019 (July) Summer School in Bioinformatics, Brussels, Belgium.

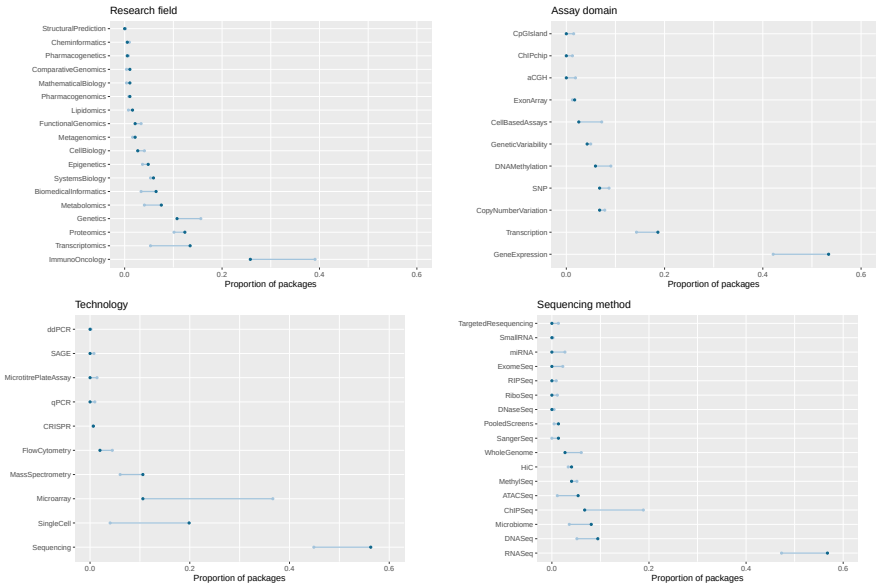

Figure 2: 'biocViews' term use for existing (light) and new (dark; contributed in the last year) packages

- Morgan, M.T., 2019 (May) Cancer Genomics: Integrative and Scalable Solution in R / Bioconductor. Informatics Technology for Cancer Research Annual Meeting, Salt Lake City.
- Carey, V.J., 2019 (May) Opening Apps and Notebooks for Cloud Scale Cancer Genomics with Bioconductor. Informatics Technology for Cancer Research Annual Meeting, Salt Lake City.
- Morgan, M.T., 2019 (April) Keynote address. Natal Bioinformatics Forum, Natal, Brazil.
- Morgan, M.T., Waldron, L., Carey, V., Huber, W. 2018 (December) Organization and other activites, BiocEurope, Munich, Germany.
- Morgan, M.T., 2018 (November) Workshop and lecture, BiocAsia, Melbourne, Australia.
- Morgan, M.T., 2018 (September). *Bioconductor* Train-the-Trainer. Jackson Labs, Bar Harbor, Maine.
- Morgan, M.T., 2018 (August) Workshop and Keynote Address, Latin American *R* / *Bioconductor* Workshop, Cuernavaca, Mexico

1.4 Community Support

The *Bioconductor* support site has about 240 new 'top-level' posts and 700 comments or answers per month. There are about 23600 (Google analytics) weekly sessions. Statistics are summarized in Table 3.

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

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 1504 subscribers on this list (versus 1387 in the last report). Table 4 lists the number of posts and number of unique authors per month as a monthly average since 2002.

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
2005	27	13	2013	97	34
2006	39	19	2014	139	41
2007	50	23	2015	142	43
2008	27	18	2016	153	45
2009	26	17	2017	137	45
2010	30	18	2018	162	48
2011	52	24	2019	139	49
2012	75	25			

Our community slack has 350 members, posting a total of 69725 messages on 39 public channels. This forum has provided an important avenue for collaboration on projects (software development, manuscript preparation, grants), as a forum for engaging the community, and as a resource for advanced user / developer support.

Web site access is summarized in Figure 3. The web site served 2.378M sessions (777,981 unique visitors) in the trailing 12 months (statistics from Google analytics). Visitors come from the United States (31%), China (12%), the United Kingdom (6%), Germany (5.1%), Japan, India, Canada, France, Spain, Italy, and 213 other countries. Unique visitors grew by 14%, similar to last year's.

1.5 Publication

Bioconductor has become a vital software platform for the worldwide genomic research community. Table 5 summarizes PubMed author / title / abstract or PubMedCentral full-text citations for 'Bioconductor'.

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

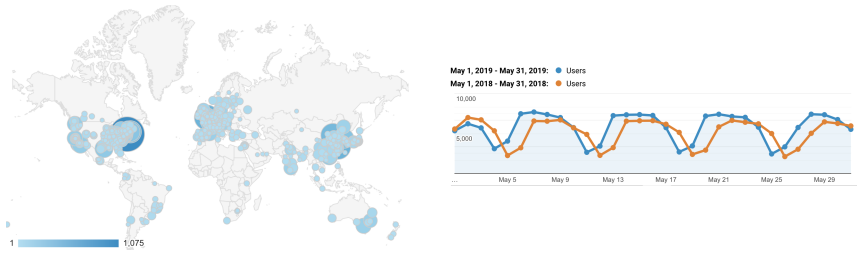

Figure 3: *Bioconductor* Access Statistics
Left: international visits, trailing 12 months. Right: Web site access, May 2018 (orange) and 2019 (blue).

Table 5: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January, 2003 – June, 2019

Year	N	Year	N	Year	N	Year	N	Year	N
2003	7	2007	44	2011	68	2015	3138	2019*	1902
2004	13	2008	52	2012	1386	2016	3415		
2005	19	2009	62	2013	2048	2017	3988		
2006	30	2010	52	2014	2401	2018	4610		

2 Core Tasks & Capabilities

2.1 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 2.2.
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.

2.2 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 other 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.

2.3 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, Qian Liu, and Kayla Morrell; Valerie Obenchain left the project during the last year. 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: Vincent Carey, Brigham & Women's; Aedin Culhane, Dana-Farber Cancer Institute; Sean Davis, National Cancer Institute; Robert Gentleman, 23andMe; Kasper Daniel Hansen, Bloomberg School of Public Health, Johns Hopkins University; Wolfgang Huber, European Molecular Biology Laboratory, Heidelberg, Germany; Rafael Irizarry, Dana-Farber Cancer Institute; Michael Lawrence, Genentech Research and Early Development; Matt Richie, Walter and Eliza Hall Institute of Medical Research, Australia; and Levi Waldron, CUNY School of Public Health at Hunter College, New York.

The **Scientific Advisory Board** provides oversight through yearly meetings. Current members include: Robert Gentleman (Advisory Board chair (23andMe); Jan Vitek (Northeastern University); Wolfgang Huber (European Molecular Biology Laboratory); Vincent Carey (Brigham & Womens); Raphael Irizzary (Dana Farber), James Taylor (Johns Hopkins University), Jenny Bryan (RStudio), Benjamin Neale (Broad Institute), and Valentina di Francesco (NHGRI).

3 Accomplishments and opportunities

Single cell sequencing is influencing core infrastructure in several ways. Effort has been extended to use file-based approaches to large data management, in mature software based on the *rhdf5* and *DelayedArray* framework as well as with attempts to track standard representations such as the 'loom' file format. A conceptually important goal is to retain a

'matrix-like' syntax of two-dimensional subsetting, and to continue developing core infrastructure classes like *SummarizedExperiment* and *MultiAssayExperiment* to require only this matrix-like interface for the data sets they contain. We have also tried to facilitate interoperability with single-cell resources, for instance developing and tracking Human Cell Atlas Data Coordinating Platform and matrix services.

Involvement in the AnVIL project (U24HG010263) has stimulated enhanced approaches to containerization and cloud computation. We have developed a docker image that contains 'most' of the system software required to build or use *Bioconductor* packages. The image is not unreasonably large, and the user can install and to the host (rather than container) file system for easy re-use across projects. Since the docker image provides a consistent binary environment, we have started to investigate management and distribution of pre-built binary packages. We have become much more familiar with the google computational cloud, complementing our current understanding of Amazon AWS, and extending to currently *ad hoc* approaches to scalability. One avenue of development has emphasized *kubernetes* and *helm*-based container orchestration for deployment of multiple *R* sessions communicating via *BiocParallel*.

We have formalized our monthly technical advisory board meetings with a governance document, meeting minutes, and call for new member nominations. This increases transparency of internal decision-making processes, and sets the stage for more structured approaches to delegation of responsibilities across the community.

Much of our core infrastructure (new package ingestion, 'devel' branches with twice-yearly releases, git-based version control, nightly build system, web site) remain largely unchanged. The support site has been updated to again align with the Biostars source from which it is derived; this provides better opportunities for support outside our team. Some of this infrastructure is not completely satisfactory. New package 'technical' review takes considerable time and is difficult to standardize. Our independent git management of each package is confusing to new developers, and our system of managing credentials has several weak points. The nightly build system has proven difficult to transition to new sites, a step that will become essential in the next year. The public web site requires aesthetic modernization as well as significant content management. A code of conduct needs to be in place to ensure that interactions in the community remain civil; this is especially true for slack, where many channels make for difficult oversight and where private messaging may mean inappropriate interactions can develop unchecked.

The project continues to produce 'annotation' resources, and to enable user access to online resources via packages such as *Biomart*. Our classic annotation packages are in need of revision, with updated data sources and more robust data processing. Unfortunately, this requires considerable scientific familiarity with contemporary data resources, and our team is not well-equipped for that. The *Biomart* package is very well maintained, but relies on commitment from the EBI for maintaining this resource. The traditional mechanism for disseminating annotation resources, via large data packages, should change to 'thin' packages providing versioning and provenance, and with large data resources managed under AnnotationHub; this transition in progress.