Bioconductor Annual Report

Martin Morgan
Roswell Park Cancer Institute

July 25, 2017

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 1383 packages for the analysis of data ranging from sequencing to flow cytometry.

1.1 Funding

Funding is summarized in Table 1.

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
EC-H2020	SOUND	9/1/2015	8/31/2018

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
	1.1	20		1.9	188		2.7	419		3.0	936
2003	1.2	30	2007	2.0	214	2011	2.8	467	2015	3.1	1024
	1.3	49		2.1	233		2.9	517		3.2	1104
2004	1.4	81	2008	2.2	260	2012	2.10	554	2016	3.3	1211
	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			

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 has been renewed through 2021.

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.

Carey receives funding through U01CA214846 for 'Accelerating cancer genomics with cloud-scale *Bioconductor*'. European Commission Horizon 2020 project 633974 (Huber, PI, with Morgan and others), 'SOUND: Statistical multi-Omics UNDerstanding of Patient Samples' has significant *R* / *Bioconductor* components.

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 1383 software packages available in release 3.5. The project produces 913 '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 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 42034 records in the current hub.

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

1.3 Courses and Conferences

Course and conference material and announcements for upcoming events are available. Courses and conferences

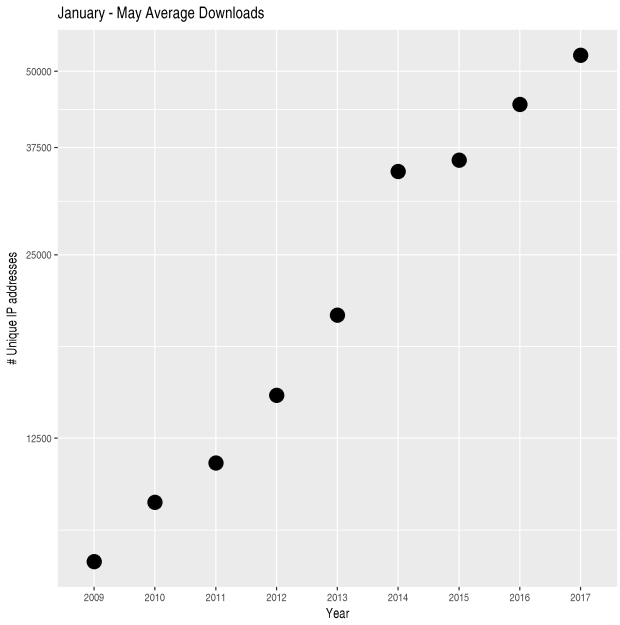

Figure 1: *Bioconductor* package download statistics, average number of unique downloads, first five months of each year.

with significant input from key *Bioconductor* personnel have been held in the following worldwide locations in the last year:

- Morgan, M.T., Waldon, L., Carey, V., 2016 (July). CSAMA 2016: Statistical Data Analysis for Genome-Scale Biology, various lecture and lab contributes.
- Morgan, M.T. 2016 (November). Bioconductor for Genomic Analysis, ABACBS, Brisbane, Australia.
- Morgan, M.T., 2016, (November) Annotation, Communication, and Performance; Introduction to Bioconductor. Technion-BKU, Haifa, Israel.
- Morgan, M.T., 2016 (December) Bioconductor for Omics analysis. Rochester Medical School.
- Morgan, M.T., 2017 (January). Introduction to R. Roswell Park Cancer Institute, Buffalo.
- Morgan, M.T., 2017 (January). R / Bioconductor for Omics analysis. University of Idaho.
- Morgan, M.T., and Shepherd L. 2017 (March). Introduction to R and Bioconductor. Moffit Cancer Research Center, Tampa.
- Morgan, M.T. 2017 (March). Good software: simple, tidy, rich. Boston, MA.
- Morgan, M.T., 2017 (May). Introduction to R and Bioconductor. Oklahoma Medical Research Facility, Oklahoma.
- Waldron, L., 2017 (May). Cancer Genomics: Integrative and Scalable Solutions in R / Bioconductor. UCSC, Santa Cruz, California.
- Morgan, M.T., Waldron, L., Carey, V, 2017 (June) CSAMA 2017: Statistical Data Analysis for Genome-Scale Biology, various lecture and lab contributions.
- Waldron, L., 2017 (June). Multi-assay experiments. The Centre for Integrative Biology at the University of Trento, Italy.

1.4 Community Support

The *Bioconductor* support site has about 278 new 'top-level' posts and 1090 comments or answers per month. The number of (Google analytics) weekly sessions are about 21000 per week in June, 2017. Statistics are summarized

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

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	2011	52	24
2006	39	19	2012	75	25
2007	50	23	2013	97	34
2008	27	18	2014	139	41
2009	26	17	2015	142	43
2010	30	18	2016	153	45

in Table 3. Mailing list statistics are provided in Table 4.

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 1265 subscribers on this list (versus 1132 in the last report). Table 4 lists the number of posts and number of unique authors per month as a monthly average since 2002.

Web site access is summarized in Figure 2. The web site served 1.907M sessions (630,380 unique visitors) in the trailing 12 months (statistics from Google analytics). Visitors come from the United States (33%), China (10.1%), the United Kingdom (7.1%), Germany (5.9%), Japan, India, Canada, France, Spain, Italy, and 213 other countries. China, India, and Japan all increased slightly in ranking. Unique visitors grew by 14%, substantially more than last year's 8% increase.

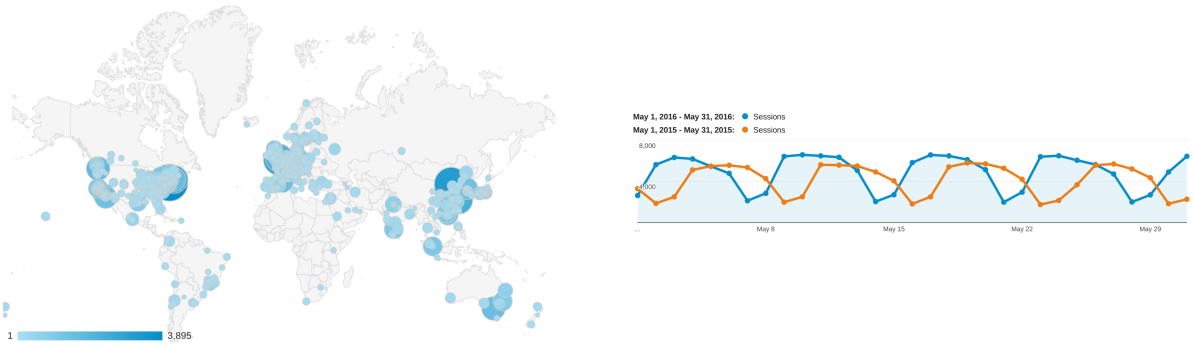

Figure 2: *Bioconductor* Access Statistics. Left: international visits, trailing 12 months. Right: Web site access, June 2016 (orange) and 2017 (blue).

Table 5: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January, 2003 – July, 2017.

Year	N	Year	N	Year	N	Year	N
2003	7	2007	44	2011	68	2015	3138
2004	13	2008	52	2012	1386	2016	3415
2005	19	2009	62	2013	2048	2017*	1631
2006	30	2010	52	2014	2401		

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.

2 New and Ongoing Accomplishments

2.1 Software

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.

Organism.dplyr provides a replacement to the 'Organism' packages integrating `TxDb` and `org` annotation packages. *AnnotationFilter* provides a new platform for consistent annotation resource queries.

BiocFileCache manages a cache of local or remote files.

RaggedExperiment and contributions to *MultiAssayExperiment* facilitate analysis of multiple assays on common samples.

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

2.2 Infrastructure

Version control transition from svn to git for package management is all but complete. Git repositories capture the full commit history of each package, including trunk and release branches.

New package contributions use Github and a public review process; reviews emphasize technical rather than scientific aspects of the software.

Virtualization Docker and Amazon Machine Instance images are available and current.

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

2.3 User Support

Support site has established itself as an important resource; it has gained a markdown-based editor.

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

Workflows provide cross-package training material and integrate with the F1000 *Bioconductor* channel.

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 the automated building and testing 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.
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 OS X. 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 and Linux machines are physical servers located at Roswell Park, the macOS machines are 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 (*italic*: currrent members) *Martin Morgan*, *Valerie Obenchain*, *Hervé Pagès*, *Marcel Ramos*, *Lori Shepherd*, Dan Tenenbaum, *Nitesh Turaga*, *Daniel van Twisk*, Greg Wargula.

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; 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).

4 Challenges and Opportunities

4.1 From the previous report

It is useful to summarize areas of challenge and opportunity in the last annual report.

Project relocation is largely complete. The group at Roswell Park is appropriately staffed. All resources (other than version control system) are located at Roswell Park or in the cloud.

Revision control, build system, and release strategies needs have been addressed by the imminent transition to git, coupled with new hardware, incremental changes to the build system, and continued rationalization of our daily, new package, and workflow build processes.

There are positive developments in terms of **project participation**. The support site has enabled less centralized support for the project, allowing strong communities to establish and develop. The public new package review process has lead to more consistent technical expectations and increased emphasis on inter-operability and re-use. A slack channel and activity around single cell experiments is helping to shape an important area of scientific development. Several funding opportunities have been pursued by U24 and U41 members; letters of support are regularly provided to investigators wishing to elaborate on new or established *Bioconductor* packages.

Cloud computing has not been addressed effectively in the project *per se*; Dr. Carey's recent award U01 award Accelerating Cancer Genomics with Cloud-Scale Bioconductor is an important step.

4.2 Scientific directions

Gene-level differential expression analysis has been a mainstay of the project since its inception, first with microarrays and more recently RNA-Seq. This is likely to remain an important core component, but the shift to ultra-fast alignment emphasizes the need for packages such as *tximport* that allow easy transition from alignment counts to *Bioconductor* analysis.

Changing technologies and research questions will likely result in increasingly common large single-cell experiments; *Bioconductor* is in a good position to capitalize on this direction, with existing packages, emerging infrastructure to support very large experiments, and an interactive developer community.

A particular strength of *Bioconductor* is the comparative ease of conducting integrative analyses, both within a single assay (e.g., from differential expression to annotation, gene set enrichment, pathway analysis, etc.) and between assay types (e.g., enable by the *MultiAssayExperiment* package).

The breadth of research questions addressed by *Bioconductor* packages continues to expand, with well-established communities working in DNA variants (called variants, copy number alterations, higher-order structure), flow cytometry, proteomics, and other domains.

4.3 Project and leadership development

The project serves several different purposes. It is at least (1) a vehicle for developing, distributing, and supporting leading edge statistical informatic software of broad relevance; (2) a repository for analytic methods and results of targeted relevance; and (3) a training resource for graduate and other students.

The size of the project means that considerable effort and resources are dedicated to maintenance activities.

Funding to date has largely been centralized primarily via the U41 and U24 (and to a lesser extent EC-H2020); more diversified efforts have been activities of core team members (e.g., Carey's U01), *ad hoc* participation in funding activities that overlap with the project, and letters to funding agencies in support of individual initiatives. The project has not successfully engaged in recent NIH initiatives such as the Data Commons or Cloud Pilot.

The U24 may be eligible for renewal in two years, and the U41 in four years. It is therefore an excellent time to identify scientific and leadership directions that reflect the reality of the project (significant ongoing maintenance) and the enthusiasm of project leaders (for novel statistical and informatic analysis) while leveraging the success of the project to ensure broad, contemporary funding.