Bioconductor Annual Report (preliminary)

Martin Morgan
Roswell Park Comprehensive Cancer Center

July 27, 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 1560 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 has been renewed 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	(7/1/2018)	(6/30/2023)
Chan / Zuckerberg		4/1/2018	3/30/2019
EC-H2020	SOUND	9/1/2015	8/31/2018

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 6/18/2018.

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.

A pending award supports Morgan, Carey, and Waldron as members of a large collaboration developing NHGRI cloud resources under ‘Implementing the Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)‘ (James Taylor, PI).

Funding from the Chan / Zuckerberg foundation provides modest support through small, one-year grants to individual *Bioconductor* collaborators, including Morgan, Carey, Irizarry, Huber, Hansen, Risso, Marioni, Gottardo, and others. Resources are directed toward various aspects of single-cell computing on Human Cell Atlas data and infrastructure.

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 1560 software packages available in release 3.7. The project produces 919 ‘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 342 ‘experiment data’ packages to provide heavily curated results for pedagogic and comparative purposes. We have standardized reproducible, cross-package protocols into 21 ‘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 44925 records in the AnnotationHub, and 1239 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

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

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			
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		3.6	1473			

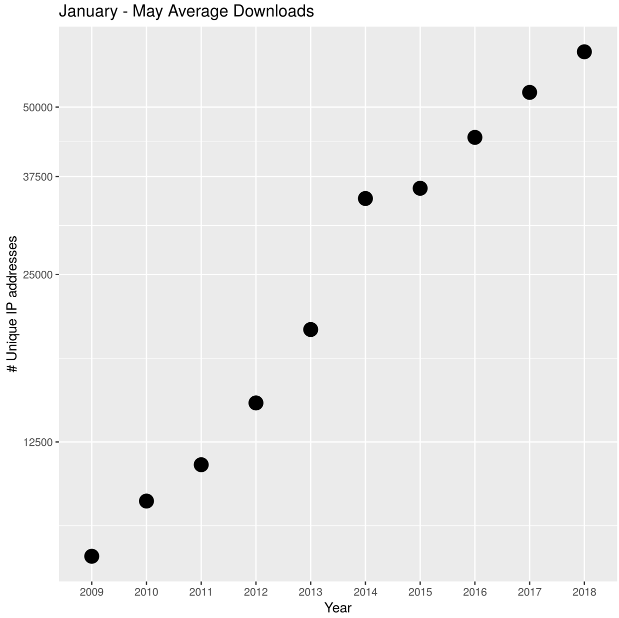

Figure 1: *Bioconductor* package download statistics, average number of unique downloads, first five months of each year.

- Morgan, M.T., Waldron, L., Carey, V, 2018 (July) CSAMA 2018: Statistical Data Analysis for Genome-Scale Biology, various lecture and lab contributions in a week-long course. Italy.
- Morgan, M.T., 2018 (June). Summer School of Advanced R for Bioinformatics. Visby, Sweden.
- Waldron, L., 2018 (May). "Cancer Genomics: Integrative and Scalable Solutions in R/Bioconductor," Informatics Technology for Cancer Research, National Cancer Institute, Bethesda, MD.
- Carey, VJ., 2018 (May). Semantically rich interfaces for cloud-scale computational biology. Northeastern University.
- Carey, VJ., 2018 (May) Wrangling cloud-scale data for cancer genomics. Bioconductor Meetup, Boston.
- Waldron, L., 2018 (March). Workshop, Bioinformatics Laboratory: Applied Statistics for High-throughput Biology, University of Verona Department of Computer Science, 15 participants.
- Geistlinger, L., 2018 (March). Workshop, Gene set analysis for RNA-seq and microarray gene expression data, Memorial Sloan Kettering Cancer Center
- Morgan, M.T., 2018 (February). Course: Using RPCCC's High Performance Computing Resources for Scientific Research. Roswell Park Cancer Care Center (2018). 5 x 1 hour sessions, 45 registered participants.

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

- Carey, VJ., 2018 (February). Semantically rich interfaces for cloud-scale genomics. Harvard School of Public Health, Boston, MA.
- Waldron, L., 2018 (February). Workshop, Analysis of RNA-seq and ChIP-seq data with R/Bioconductor, Brown University Center for Computational Biology of Human Disease.
- Morgan, M.T., 2018 (January). Workshop: Introduction to R, Roswell Park Cancer Institute.
- Morgan, M.T., Waldron, L., Carey, V, Hansen, K 2017 (Decemeber) Bioconductor European conference; organization and various presentations. Cambridge, UK.
- Waldron, L., 2017 (December). New York City R/Bioconductor for Genomics, "Microbiome Data Analysis," Memorial Sloan Kettering Cancer Center, Memorial Sloan Kettering Cancer Center, New York, NY, United States.
- Morgan, M.T., 2017 (November). Bioconductor master classes; Bioconductor Asia conference, Adelaide, Australia.
- Waldron, L., 2017 (November). Data Analysis & Bioinformatics User Group, "Multi-omics data representation and analysis with MultiAssayExperiment," Applied Bioinformatics Core at Weill Cornell Medicine, New York, NY, United States.
- Waldron, L., 2017 (September) New York City R/Bioconductor for Genomics, "Multi-omics infrastructure and data for R/Bioconductor," Memorial Sloan Kettering Cancer Center, Rockefeller Research Institute, New York, NY, United States.
- Morgan, M.T., 2017 (September). R and Bioconductor for Genomic Analysis. Ohio State University.
- Waldron, L., 2017 (July). Workshop, Multi-omics data representation and analysis with MultiAssayExperiment, BioC 2017.
- Waldron, L., Workshop, 2017 (July). Microbiome Data Analysis, BioC 2017.
- Waldron, L., 2017 (June). University of Trento - Center for Integrative Biology, "Multi-omics infrastructure and data for R/Bioconductor," Trento, Italy.
- Morgan, M.T., 2017 (June). Advanced R and Bioconductor. University of Zurich, Switzerland.
- Morgan, M.T., Waldron, L., Carey, V, Hansen, K 2017 (July) Bioconductor annual conference; organization and various presentations. Boston, MA.
- Morgan, M.T., Waldon, L., Carey, V., Huber 2017 (June). CSAMA 2017: Statistical Data Analysis for Genome-Scale Biology, various lecture and lab contributes.

1.4 Community Support

The *Bioconductor* support site has about 268 new 'top-level' posts and 1000 comments or answers per month. The number of (Google analytics) weekly sessions are about 25000 per week in June, 2018. Statistics are summarized in Table 3.

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 1387 subscribers on this list (versus 1265 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.

	Posts	Authors		Posts	Authors		Posts	Authors
Year	per month	per month	Year	per month	per month	Year	per month	per month
2005	27	13	2011	52	24	2017	137	45
2006	39	19	2012	75	25	2018	200	57
2007	50	23	2013	97	34			
2008	27	18	2014	139	41			
2009	26	17	2015	142	43			
2010	30	18	2016	153	45			

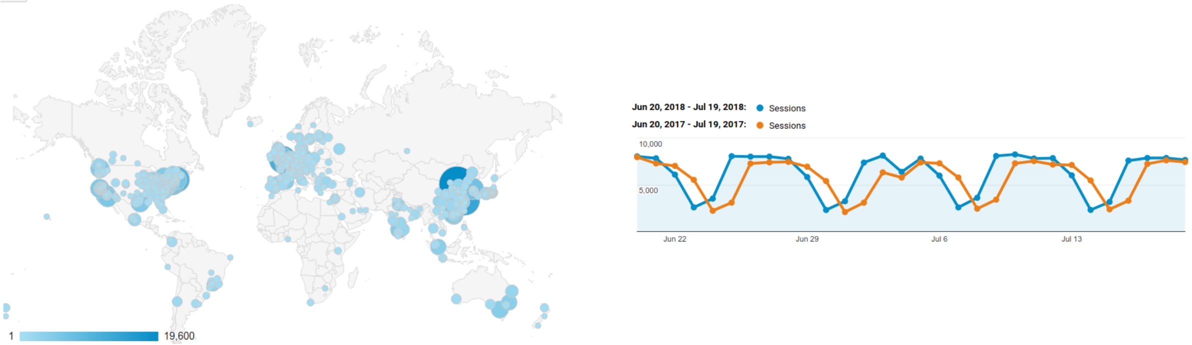

Figure 2: *Bioconductor* Access Statistics. Left: international visits, trailing 12 months. Right: Web site access, June 2016 (orange) and 2017 (blue).

Web site access is summarized in Figure 2. The web site served 2.225M sessions (697,559 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. China, India, and Japan all increased slightly in ranking. Unique visitors grew by 14%, substantially more than last year's 8% increase.

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.

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	3988
2006	30	2010	52	2014	2401	2018*	1758

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.

BiocFileCache manages a cache of local or remote files.

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

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

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

2.2 Infrastructure

Version control transition from svn to git for package management is 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. Almost all packages are reviewed by a core team member, sometimes soliciting input from a third party.

2.3 User Support

Support site has established itself as an important resource; it has gained a markdown-based editor. It had diverged significantly from Biostars, with little techinical capability to support the site on the core team. This has been largely mitigated by development of expertise, re-establishing contact with the Biostars author, and harmonizing our site with the Biostars code base.

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.

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 and Linux machines are physical servers located at Roswell Park, the macOS machines are rented via MacStatdium (currently moving to physical hardware at Roswell Park). 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, Valerie Obenchain, Hervé Pagès, Marcel Ramos, Lori Shepherd, Nitesh Turaga, Daniel van Twisk, and Qian Liu. 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 &amp; 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), and Jenny Bryan (RStudio).