# *Bioconductor* Annual Report, June 25, 2016

Martin Morgan
Roswell Park Cancer Institute

23 June, 2016

## 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 >1024 packages for the analysis of data ranging from sequencing to flow cytometry.

Table 1: *Bioconductor*-related funding

| | Award | Start | End |
|---|---|---|---|
| Active | | | |
| NHGRI / NIH | U41HG004059 | 3/1/2016 | 2/28/2021 |
| NCI / NIH | U24CA180996 | 9/1/2014 | 8/31/2019 |
| Participating | | | |
| 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 | | | |
| 2005 | 1.6 | 123 | 2009 | 2.4 | 320 | 2013 | 2.12 | 671 | | | |
| | 1.7 | 141 | | 2.5 | 352 | | 2.13 | 749 | | | |

## 1.1 Funding

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.

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. 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 7-8 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 1211 software packages available in release 3.1. The project produces 916 '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 year, the 'AnnotationHub' resource for serving and managing genome-scale annotation data, e.g., from the Roadmap Epigenomics project, NCBI, and Ensembl. There are 43720 records in the current hub.

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

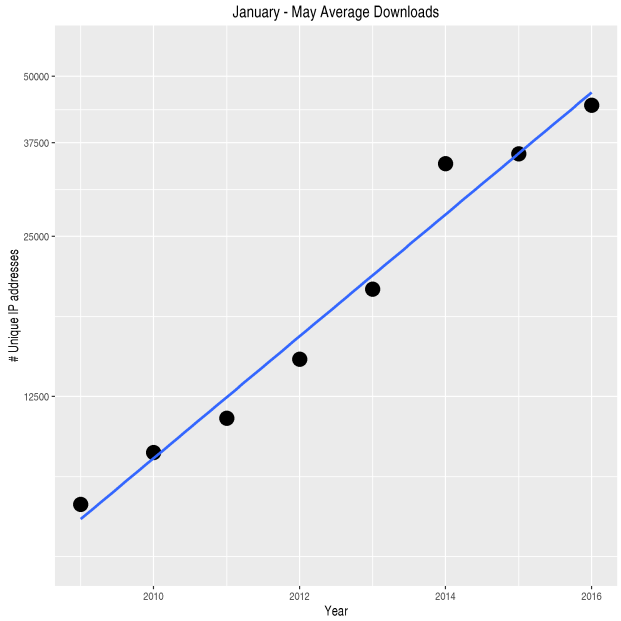


Figure 1: *Bioconductor* package download statistics, average number of unique downloads, first five months of each year.

## 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:

- BioC 2016 – June, Stanford University, Stanford, CA, USA.
- MOOC: Bioconductor for Genomic Data Science – June, Coursera.
- China R / Bioconductor session – May, Beijing, China.
- MOOC: Data Analysis for Life Sciences, module 6x: High-performance computing for reproducible genomics – March, EdEx.
- CDSE Days, Using R for High-Throughput (Genomic) Analysis – March,. University at Buffalo.
- Introduction to High Throughput DNA Sequence Data Analysis Using R / Bioconductor – April, Eastern North American Region International Biometric Society.
- *Bioconductor* European Developer Meeting – December, Cambridge, UK.
- Workshop: R / Bioconductor: Untangling Genomes – October, Montevideo, Uruguay.
- Introduction to R and Bioconductor for high-throughput genomic analysis – October, Lyon, France.
- First Asia-Pacific *Bioconductor* Developers Meeting – September, Tokyo, Japan.
- Computational Statistics for Genome Biology – June, Brixen / Bressanone, Italy.

Prominent talks and presentations include

- Workshop in Statistical Methods for Omics Data Integration and Analysis (Keynote Address, Morgan) – September, Valencia, Spain.
- Bioconductor for Integrative Cancer Genomic Analysis – April, CI4CC NCI-ITCR Workshop Presentation, Napa, CA.
- Bioconductor for Integrative Cancer Genomic Analysis – April, CBIIT Speaker Series.

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 |

Table 4: Monthly average number of posts and number of unique authors for the `bioconductor` 'devel' mail list from January, 2005 – December, 2015.

| 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 | | | |

### 1.4 Community Support

The project transitioned from a user mailing list to support site in October, 2014. There are about 185 new 'top-level' posts and 595 comments or answers per month. The number of (google analytics) weekly sessions have grown from about 3000 per week at introduction to about 11000 per week in July, 2015. Statistics are summarized in Table 3. Mailing list statistics are provided in Table 4.

We continue to provide bioc-devel, a mailing list forum for package contributors' questions and discussion relating to the development of *Bioconductor* packages. There are 1132 subscribers on this list (versus 1013 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 1.592M sessions (551,876 unique visitors) in the trailing 12 months (statistics from Google Analytics). Visitors come from the United States (33%), China (8.3%), the United Kingdom (7.1%), Germany (6.4%), France (3.0%), India, Japan, Canada, Spain, Italy, and 210 other countries. China, India, and Japan all increased slightly in ranking. Unique visitors grew by 8%, substantially less than last year's 26% 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.

## 2 New and Ongoing Accomplishments

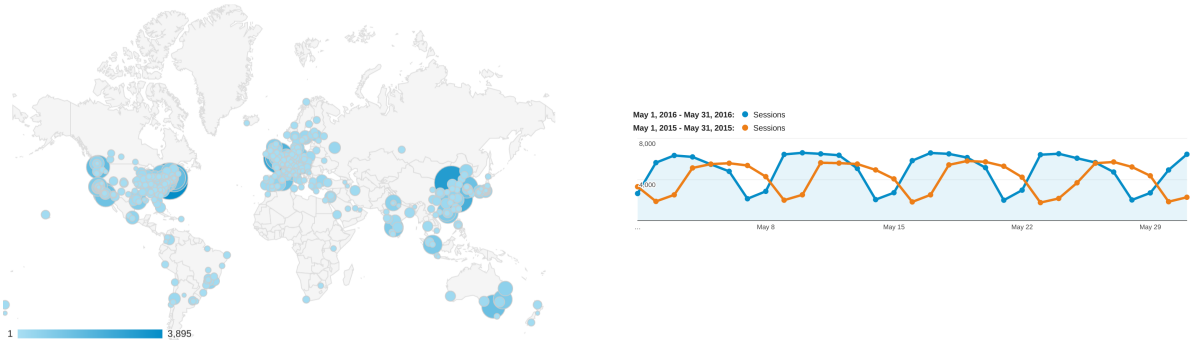


Figure 2: *Bioconductor* Access Statistics, 2015. Left: international visits. Right: Web site access, May 2015 (orange) and 2016 (blue).

Table 5: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January, 2003 – July, 2016.

| Year | N | Year | N | Year | N | Year | N |
|---|---|---|---|---|---|---|---|
| 2003 | 7 | 2007 | 44 | 2011 | 68 | 2015 | 3138 |
| 2004 | 13 | 2008 | 52 | 2012 | 1386 | 2016* | 1465 |
| 2005 | 19 | 2009 | 62 | 2013 | 2048 | | |
| 2006 | 30 | 2010 | 52 | 2014 | 2401 | | |

## 2.1 Software

**GenomicRanges** represents a mature infrastructure for working with sequence data. Implementation of 'nested containment lists' substantially enhances memory use and speed associated with a central operation (finding overlaps between millions of ranges).

**HDF5Array** coupled with development of the `GPos` class in *GenomicRanges* provides *Bioconductor* developers with access to on-disk representation of very large numeric data. Abstractions in the `DelayedArray` class allow lazy large data evaluation, providing the user with immediately feedback on the consequences of their code without requiring full evaluation of large data.

**AnnotationHub** continues to provide access to consortium-level resource summaries that require limited curation to be useful to end users. *ExperimentHub* is under active development as an improved mechanism for deploying data summarizing large experimental results that require considerable processing before being useful.

**BiocParallel** provide a consistent parallel computing interface across cores, computers, and clusters; development to include cloud-based back-ends is very relevant.

## 2.2 Infrastructure

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

**Github** hosts a mirror of our SVN repository of all *Bioconductor* packages. We have revised the git-svn bridge faciliites to ease developer contributions through github.

**Shields** on package landing pages (e.g., *Rsamtools*) provide users and developers with additional insight and inspiration on package use and robustness, including unit test and coverage metrics.

## 2.3 User Support

**Support site** replaces our user mailing list.

**Course Materials** organize and make much more accessible recent course and training material.
**Quarterly Newsletters** provide users and developers with insight into project developments.
**MOOCs** offered by Irizzary, Carey, and colleagues have reached 10,000's of people.
**biocViews** have been more heavily curated to enhance utility in package discovery.
**Videos** have been explored as a training mechanism.
**Workflows** provide cross-package training material; Huber has extended this concept with the recently launched 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 Seattle *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.
3. Software Development.
4. End-User and Developer Support.
5. New Package Submission. The *Bioconductor* project relies on technical review process of candidate packages to ensure they contain high-quality software. The Seattle *Bioconductor* team spends a considerable amount of time managing new contributions by previewing the software for quality, managing peers during the review process to ensure scientific relevance, and communicating with the software developers on what steps need to be taken for their contribution to be included within *Bioconductor*. From August, 2014 – July, 2015, approximately 291 software packages have been managed by the Seattle *Bioconductor* team.
6. 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. Due to quickly changing data standards, the maintenance of the code used to produce the annotation packages requires constant attention. Work during the recent release cycles has focused on flexible approaches to transitioning from gene-level annotations relevant for expression arrays to genome coordinate annotations that form the basis of sequence-based annotations.
7. Semi-Annual Releases.

## 3.2 Hardware and Infrastructure

The *Bioconductor* project provides packages for computing platforms common in the bioinformatics 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 web site, support site, AnnotationHub, and additional servers are hosted on virtual machines, some of which are Amazon machine instances. The build machines are heavily taxed, and the overall architecture of our build system (complete nightly builds) leave little room for growth.

## 3.3 Key Personnel

The **Core Development Team** are employees of the Fred Hutchinson Cancer Research Center, developing software and other infrastructure and ensuring day-to-day operation of the project. Core team members in the period covered by this report have incuded (*italic*: currrent members) *Martin Morgan*, Sonali Arora, Marc Carlson, Nathaniel Hayden, James Hester, Jim Java, Brian Long, *Valerie Obenchain*, *Hervé Pagès*, *Marcel Ramos*, *Lori Shepherd*, *Dan Tenenbaum*, and Paul Shannon.

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; obert 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: Simon Tavaré (Advisory Board chair; University of Southern California / Cambridge University); Robert Gentleman (23andMe); Paul Flicek (European Bioinformatics Institute); Simon Urbanek (AT&T Labs – Research); Wolfgang Huber (European Molecular Biology Laboratory); Vincent Carey (Brigham & Womens); Raphael Irizzary (Dana Farber).

# 4 Challenges and Opportunities

It is useful to summarize areas of challenge and opportunity in the last annual report; many of these remain prominent and imperfectly addressed, including: project relocation; cloud computing; project participation; and revision control and build system.

## 4.1 Project Relocation

The *Bioconductor* project relocated to Roswell Park Cancer Institute (RPCI) in Buffalo, NY in September, 2015. US grants are now administered by RPCI; the SOUND consortium EU grant is in the final stages of transfer. Project relocation has posed significant challenges in retention of key employees, recruitment of new staff, transitioning of physical equipment, fulfilling the complex administrative needs of the project. These challenges have been mitigated in part by an ongoing subcontract to FHCRC, which allows staff retention and access to compute resources, and by enabling remote employment.

## 4.2 Cloud Computing

The 'download to desktop / user account' model places significant burden on both users (struggling to install packages with idiosyncratic dependencies) and developers (needing to produce software that works across computing platforms). Additional issues involve data movement and storage. There are three directions implied by these considerations. The first involves continued elaboration of Docker and other portable software containers. The second involves use of these containers to deliver scalable computing facilities to appropriate segments of our user community, e.g., using the high-performance computing cluster at SUNY Buffalo.

Cloud computing exposes additional challenges to *Bioconductor*'s current computational model. (1) the need for *meaningful* access to modern cloud resources such as the the NCI Genomic Data Commons (see the initial development of the GenomicDataCommons package). (2) Access to cloud computing endpoints such as the Google bigquery table technology underlying some NCI Cancer Genomics Cloud projects, as illustrated by the

cgcR package, especially when integrated with interactive visualization and *Bioconductor* data representations. (3) development of *Bioconductor*-driven cloud-based work flows, as is being developed in the *sevenbridges* package.

## 4.3 Project Participation

There are three dimensions of project participation that represent growing points. The project's key strengths derive from its appeal to separate communities: statistics; computing; and biology.

The first challenge is to remain accessible to, and rewarding for, each of these communities, so that new package contributions remain on the leading edge of bioinformatics.

The second challenge reflects the diversity of domain areas in which *Bioconductor* has credible strength. Differential expression represents a primary focus, but there is considerable expertise in other areas of high throughput sequencing, as well as significant contributions in flow cytometry and proteomics. More generally there are *R* communities (e.g., ecology, phylogeny, rOpenSci) which offer the opportunity for considerable synergy. The challenge then is to engage and nurture these domains of expertise. Approaches include focused activities (e.g., facilitating flow or proteomics workshops), active engagement in relevant communities (e.g., advisory board members overlapping rOpenSci?), and traditional scientific grantsmanship (e.g., letters of support or collaborative proposals).

The third challenge involves transitions from user to developer, and from developer to thought leader. The latter transition is a particularly valuable opportunity, as evident at the *Bioconductor* annual conference where there were an intimidating group of graduate students, post-docs, and junior faculty making valuable contributions to *Bioconductor*. How is this group's enthusiasm and contribution to be marshalled into long-term and productive commitment to *Bioconductor*?

## 4.4 Bioconductor Revision Control, Build Systems, and Release Strategies

The need for a revised version control strategy and build system is now acute. Many of our commits are now from git(hub) repositories; the facilities to support this are very cumbersome and can generate significant work for developers trying to adopt this approach. A lesson learned from our current attempts is the need for 'expert'-level understanding of git, coupled with a thorough understanding of the *Bioconductor* versioning and build system. We anticipate introducing direct support for git-based repositories at the start of the next release cycle.

The project has three distinct build systems – nightly builds for the main repository, single-package builds to check new package submissions, and workflow builds dedicated to potentially long-running and resource-intensive workflows. Each system is complicated and, especially the latter two, prone to breakage that confuses the user while requiring manual intervention. In addition, expectations in the broader *R* development community place increasing pressure on development of build-on-commit rather than nightly builds. Our main strategy is (a) stabilize current build systems to a robust state; (b) simplify the single-package and workflow builders into a single code base, and (c) enhance the nightly build software to make more efficient use of computational resources. We believe that this will place us in a good position to more comprehensively address build-on-commit and more flexible approaches to builds.

The *Bioconductor* release model is different from the main *R* archive. This produces tension at several levels. Users must follow special instructions for *Bioconductor* package installation. Many user problems arise from incorrect installation. The strategy for archiving individual packages and hence reproducing work flows is different for CRAN and *Bioconductor* packages. In many ways the democratization of package dissemination enabled by github and tools like the *devtools* package exacerbate the challenges of providing users with a stable, robust, and repeatable computational environment.

## 4.5 Additional infrastructure needs

There are a number of directions for software development. A very incomplete list includes: effective work with on-disk data resources; coordinated multi-assay analysis; framework for interactive visualization.