Bioconductor Semi-Annual Report

Martin Morgan
Fred Hutchinson Cancer Research Center

July 2013

Contents

1 Project Scope **1**
1.1 Packages . . . 1
1.2 Courses and Conferences . . . 2
1.3 Community . . . 3
1.4 Publication . . . 3
1.5 Funding . . . 4

2 Core Tasks & Capabilities **4**
2.1 Automated Package Building and Testing . . . 4
2.2 Package submission management . . . 5
2.3 Annotation data package building . . . 5
2.4 Other Tasks . . . 5
2.5 Hardware . . . 5
2.6 Key Personnel . . . 5

A Appendix: Proposal Specific Aims **6**

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 approximately 12 years ago (Fall, 2001) by Dr. Robert Gentleman and others, and now consists of >670 packages for the analysis of data ranging from expression microarrays through next-generation sequencing.

1.1 Packages

R software packages represent the primary product of the Bioconductor project. Packages are produced by the Bioconductor team and from international contributors. Table 1 summarizes growth in the number of packages hosted by Bioconductor, with 671 packages available in the current release. The project also produces 668 '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.

Software packages on the Bioconductor web site, `http://bioconductor.org`, were downloaded by 143,754 unique non-FHCRC IP addresses between August, 2012 and July, 2013.

Table 1: Number of contributed packages included in each Bioconductor release. Releases occur twice per year.

Release		N	Release		N	Release		N
2002	1.0	15	2006	1.8	172	2010	2.6	389
	1.1	20		1.9	188		2.7	419
2003	1.2	30	2007	2.0	214	2011	2.8	467
	1.3	49		2.1	233		2.9	517
2004	1.4	81	2008	2.2	260	2012	2.10	554
	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			

1.2 Courses and Conferences

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

- *Practical Genomics: From Biology to Biostatistics* – Baltimore, MD – October 1-3, 2012 p
- *Advanced R / Bioconductor Programming* – Seattle, WA – October 15-16, 2012
- *Hands-on training at EBI – EMBO Practical Course: Analysis of High-Throughput Sequencing Data* – EMBL-EBI, Hinxton, UK – October 29-November 1, 2012
- *Intermediate R /Bioconductor for High-Throughput Sequence Analysis* – Seattle, WA – February 14-15, 2013
- *Boston Bioconductor Basics* – Boston, MA – 4-5 April 2013
- *Intermediate R / Bioconductor for High-Throughput Sequence Analysis* – Seattle, WA, USA 28 - 29 May 2013
- *CSAMA 2013 (Computational Statistics for Genome Biology)* – Brixen-Bressanone, Italy, 24 - 28 June 2013
- *R/Bioconductor for Analysis and Comprehension of High-throughput Genomic Data* Albacete, Spain, 10 - 12 July 2013

There are two prominent conferences organized to benefit the Bioconductor scientific community.

- The *European Bioconductor Developer Workshop* was held 13-14 December in Zurich, Switzerland. Participants were exposed to new developments in R and Bioconductor, with particularly stimulating flashlight talks from the broader Bioconductor community.
- *BioC2013 – Where Software and Biology Connect* was held in Seattle at the Fred Hutchinson Cancer Research Center on July 17-19, 2012. Over 105 registered scientists attended, including 75 during Developer Day. The conference consisted of 6 talks from leading researchers in computational biology, 6 short presentations from prominent members of the Bioconductor community, and 19 hands-on lab sessions presented by Bioconductor package developers. We provided travel expense and conference fee scholarships for attending the conference to *approx*10 students.

Table 2: Monthly average number of posts and number of unique authors for the `bioconductor` mail list from January, 2002 – January, **2012**.

Year	Posts per month	Authors per month	Year	Posts per month	Authors per month
2002	59	13	2008	424	83
2003	231	47	2009	450	86
2004	320	60	2010	504	170
2005	353	61	2011	467	166
2006	348	59	2012	597	195
2007	432	75	2013	600	

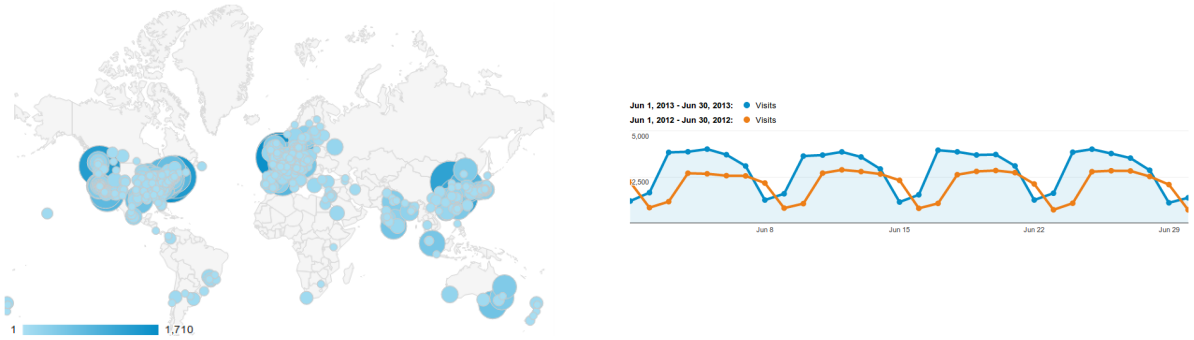

Figure 1: Bioconductor Access Statistics, 2013. Left: international visits. Right: Web site acccess, June 2012 (orange) and 2013 (blue).

1.3 Community

The project maintains two email lists:

- `bioconductor`[1] is a forum for user questions, project announcements, and general discussion of interest to the Bioconductor community. Subscribers: 3327.
- `bioc-devel`[2] is a forum for package contributors' questions and discussion relating to the development of Bioconductor packages. Subscribers: 717.

All lists provide a means of disseminating project news and a space for members of the community to share their knowledge about use of Bioconductor packages and best practices for data analysis. Table 2 lists the number of posts and number of unique authors as a monthly average since 2002.

Web site access is summarized in Figure 1. The web site received 926,208 visitors (318,094 unique visitors) from July 1, 2012, through June 30, 2013 (statistics from Google Analytics). Visitors come from the United States (327,570), the United Kingdom (70,408), Germany (67,347), China (55,031), Canada (32,126), France, Japan, India, Spain, Italy, and 178 other countries. Unique visitors grew by 20%.

1.4 Publication

Bioconductor has become a vital software platform for the worldwide genomic research community. There are 5174 Google Scholar citations of Gentleman et al. (2004); this is the fourth most accessed article of all

[1] `http://www.stat.math.ethz.ch/mailman/listinfo/bioconductor`
[2] `http://www.stat.math.ethz.ch/mailman/listinfo/bioc-devel`

Table 3: PubMed title and abstract or (2012 and later) PubMedCentral full text searches for "Bioconductor" on publications from January, 2003 – July, 2013.

Year	N	Year	N	Year	N
2003	7	2007	44	2011	68
2004	13	2008	52	2012*	1386
2005	19	2009	62	2013	669
2006	30	2010	52		

Table 4: Citations for select Bioconductor software packages as captured by Google scholar in July, 2013. 'Citation' may be pubmed id.

Package	Citation	N	Package	Citation	N
limma	Smyth (2005)	1788	biomaRt	16082012	291
vsn	12169536	1309	aCGH	16159913	266
affy	14960456	1153	eisa	12689096	213
xcms	16448051	782	MassSpecWavelet	16820428	207
DESeq	20979621	697	beadarray	17586828	169
globaltest	14693814	522	cellHTS2	16869968	147
edgeR	19910308	500	affylmGUI	16455752	126
GOstats	17098774	458	tilingArray	16787969	114
lumi	18467348	469	made4	15797915	109
limmaGUI	15297296	354	altcdfenvs	15310390	91
affycomp	14960458	297			

time from *Genome Biology*. Table 3 summarizes PubMed author / title / abstract (65) or PubMedCentral full-text citations for 'Bioconductor'.

Software packages within the Bioconductor project are cited in leading scientific journals. Table 4 contains citations captured in July, 2011 by Google scholar for select Bioconductor packages. The citations are either for the article with the associated PubMed ID or is a chapter in Gentleman et al. (2005).

1.5 Funding

The project is primarily funded through a National Human Genome Research Institute U41 (Community Resource Project). Current leadership includes Drs. Martin Morgan (PI; Fred Hutchinson Cancer Research Center), Vincent Carey (Brigham & Women's, Harvard Medical School), and Raphael Irizzary (Harvard School of Public Health). Additional key participants are listed in section 2.6.

2 Core Tasks & Capabilities

2.1 Automated 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 2.5.

2.2 Package submission management

The Bioconductor project relies on technical review process of candidate packages to ensure they contain high-quality software. It has achieved a virtuous cycle, where its success has brought in new scientific software developers, and they, in turn, have been contributing more and more to the Bioconductor project.

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 July, 2012 – January, 2013, 215 software packages have been managed by the Seattle Bioconductor team.

2.3 Annotation data package building

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.

2.4 Other Tasks

In addition to the tasks listed above, the Seattle Bioconductor team engages in the following auxiliary tasks:

1. Providing user and developer support on project mail lists.
2. Developing new functionality and improving architecture of key packages.
3. Orchestrating the Bioconductor releases that occur every six months.

2.5 Hardware

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 two Windows machines, two Linux machines, and two MacOS machines. The web site is hosted on an independent Linux machine. The build machines are heavily taxed, and the overall architecture of our build system (complete nightly builds) leave little room for growth.

2.6 Key Personnel

The Scientific Advisory Board for 2012-2013 includes: Simon Tavare (Advisory Board chair; University of Southern California / Cambridge University); Vivien Bonazzi (NHGRI), Robert Gentleman (Genentech); Paul Flicek (European Bioinformatics Institute); Simon Urbanek (AT&T Labs – Research); and Wolfgang Huber (European Molecular Biology Laboratory).

These individuals, all working at the Fred Hutchinson Cancer Research Center (FHCRC) in Seattle, Washington, played a central role in executing project objectives during 2011 and 2012: Martin Morgan, principal investigator; Marc Carlson, developer; Hervé Pagès, developer; Valerie Obenchain, developer; Dan Tenenbaum, developer; and Paul Shannon, developer.

Additional collaborations, sub-contracts, and leadership roles involve the following individuals: Vincent Carey, Harvard Medical School; Rafael Irizarry, Johns Hopkins University School of Hygiene and Public Health; Kasper Daniel Hansen, Johns Hopkins University; Michael Lawrence, Genentech; Sean Davis, National Institutes of Health; and James MacDonald, University of Michigan.

References

R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 2005.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5:R80, 2004.

Gordon K. Smyth. Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420. Springer, New York, 2005.

A Appendix: Proposal Specific Aims

Bioconductor is an established and successful open source software collection for analysis of high throughput genomic data. Important data types include sequence 'short reads', microarrays, images, and flow cytometry. Users create reproducible work flows essential for collaboration, scientific integrity, and analytic quality. Bioconductor successfully tracks developments in software engineering, statistical methodology, and biotechnology. Bioconductor's active developer community enables cost-effective development by scientists familiar with the biological implications of the data. Bioconductor package requirements provide standardization, enhancing end-user experience while encouraging software reuse and extension.

Enable Bioconductor Software Distribution and Use This aim emphasizes user access and developer support. 1. Extend an easily accessible repository of well-tested and curated analysis packages. Following our current successful model, packages will be contributed by the Bioconductor core team, and by independent self-identified collaborators. 2. Make analyses requiring specialized computational and statistical skills accessible to the scientific community. Activities include: (a) creating package vignettes to illustrate data analysis tasks; (b) expanding 'experimental data' packages so that data exemplars are immediately available in well-documented forms; (c) conducting short courses, frequently coordinated with major conferences; (d) organizing an annual conference and developer meetings; (e) participate in synergistic international activities; and (f) publish research on significant project contributions. 3. Provide technical and logistic support to a large developer community, especially those translating biological expertise to useful scientific software. Technical support includes assistance in software design and deployment, and provision of a multi-platform facility for package testing and building. Logistical support consists of creation, management, and operation of software distribution and quality assurance processes.

Develop Computational Analytic Facilities This aim addresses challenges to the use of Bioconductor for sequence and other very high throughput data types. 1. Processing very large data sets, addressed through: (a) exploiting multiple computation units, e.g., multiple cores, distributed computing, and cloud

computing; (b) transparent hierarchical (RAM, disk, network) memory management; and (c) stream-oriented processing. 2. Managing large experiment-wide data sets to reduce 'book-keeping' error while fostering reproducibility, by enhancing facilities to consistently bind metadata (experiment, sample, and analysis descriptions) to the underlying large-scale data. 3. Facilitating annotation and integrative analysis, by packaging genomic annotation and data resources (e.g., NCBI, UCSC, GEO, ArrayExpress, BioMart, SRA) for easy and flexible inter-operation with analytic work flows. 4. Representing data for specific application domains, for example variant whole-genome and multiple cancer genome representations. 5. Interoperability with external software. through: (a) integrating Bioconductor output with genome browser and other advanced genome-scale visualization tools as a way to make analytic results accessible to general users; (b) providing facilities for use of Bioconductor as an analytic engine in third-party commercial or open source software projects; and (c) orchestrating analysis across software products.

Contribute and Foster Statistical Methods for Genome-Scale Biology New methodology and infrastructure will be developed to promote reliable use of high-throughput technology in clinical settings, principally by leveraging massive public microarray archives to accurately fit models that distinguish biologic signals from artifacts such as differential probe affinity and reagent batch effects. Targets of this work are 'single-array normalization' gene expression 'barcodes' algorithms enabling rapid determination of tissue type and state from single array scans. Improved integrative analysis of transcript profiles and high-density genotypes will be supported through data structures and algorithms that exploit parallel computation, comparative research on techniques such a surrogate variable analysis that isolate components of transcriptome variation specifically subject to genetic control, and improvement of support for tools addressing transcriptional impacts of rare structural variants. The project will foster comparative evaluation of new methodologies through exemplar data sets and work flows that simplify conducting fair comparisons and calculation of relevant performance metrics.