Bioconductor Annual Report 2006

Seth Falcon
Fred Hutchinson Cancer Research Center

October 2006

Contents

1 Summary of Core Tasks and Challenges

1.1 Automated package building and testing

The Bioconductor project provides access to its packages through package repositories hosted on `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.

1.2 Annotation data package building

The Bioconductor project synthesizes genomic and proteomic information available in public data repositories in order to annotate the 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.

In order to synthesize data from the various public repositories, we must maintain automated tools that can parse the available information. Due to quickly changing data standards, the maintenance of the code used to produce the annotation packages requires constant attention.

We are also focusing resources on the underlying storage mechanism used for the annotation data packages. New high-throughput technologies such as SNP and exon arrays require significantly larger annotation libraries; the infrastructure requires improvement to support work with these emerging technologies.

1.3 Other Tasks

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

2 Size of Project

The Bioconductor project is comprised of R packages contributed by a worldwide bioinformatics community. There are currently 131 active developers and 203 contributed packages in Bioconductor's development repository. The project also maintains 231 annotation data packages that aid in the

analysis of data from microarray experiments. Table **??** tracks the growth of the project over the semi-annual releases.

Release	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9
Package Count	20	30	49	81	100	123	141	172	188

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

3 Bioconductor Electronic Mail Lists

The project maintains two email lists, `bioconductor`[1] and `bioc-devel`[2]. The `bioconductor` list is a forum for user questions, project announcements, and general discussion of interest to the Bioconductor community. As of October, 2006 the list has 1542 subscribers (individuals who receive mail from the list).

The `bioc-devel` list is a forum for package contributors' questions and discussion relating to the development of Bioconductor packages. As of October, 2006 this list has 287 subscribers.

Both 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 over the past four years.

Year	Posts/month	Authors/month
2002	60	13
2003	232	47
2004	320	60
2005	352	61
2006	356	68

Table 2: Monthly average number of posts and number of unique authors for the `bioconductor` mail list for the years 2002–2005.

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

4 The Bioconductor Website

The Bioconductor website, `http://bioconductor.org` averages over 17700 unique visitors and over 560GB of content per month. In September, 2006, the site served 554GB of content of which 514GB (92%) corresponded to package downloads. The Biobase package was downloaded by 2097 unique IP addresses in September, 2006. We host the website on a dual-Xeon 3.0GHz server with 2GB of RAM from Dell.

5 Package Building and Testing

The Bioconductor project is committed to providing packages for all computing platforms common in the bioinformatics community. We currently provide packages for Linux, Solaris, and most UNIX-like variants as well as 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 development repository. Table 3 provides details on the systems we currently have available for the nightly build.

Platform	CPU	RAM	Build Time (Hours)
Linux 64-bit	dual-Xeon 3.40 GHz	4 GB	5
Windows 32-bit	dual-Xeon 2.80 GHz	2 GB	9
Solaris 64-bit	4x - Sparc 1.3 GHz	16 GB	8

Table 3: Servers used to build and test Bioconductor packages along with the number of hours required for a build/test cycle of all packages.

6 Accomplishments

6.1 Papers Citing Bioconductor

Based on a PubMed search for "bioconductor", there have been 29 publications citing Bioconductor since October 2005. They are listed in the bibliography of this report.

6.2 Bioconductor Courses

The following Bioconductor training courses were held in 2006:

- Advanced R Programming Course, Seattle, Washington, USA, January 18th-20th 2006
- Bioconductor Short Course, Seattle, Washington, USA, April 12th-14th 2006
- Computational and Statistical Aspects of Microarray Analysis, Bressanone-Brixen, June 19th-23th 2006
- Workshops on S, R, and Bioconductor, Auckland, New Zealand, July 7th-8th 2006
- Boston Bioconductor Workshop, Harvard Medical School, Boston MA, July 13th-15th 2006
- BioC2006 Conference, Seattle, Washington, USA, August 3rd-4th 2006.
- Gene Expression Analysis with R/Bioconductor. Short Course on Mathematical Approaches to the Analysis of Complex Phenotypes. The Jackson Laboratory, Bar Harbor, Maine, USA, September 16th-22nd 2006.
- Bioconductor Short Course, Seattle, Washington, USA, October 9th-11th.

6.3 BioC2006 Conference

The Gentleman Lab organized a conference to highlight current Bioconductor developments and to provide a forum for discussing the use and design of software for analyzing data arising in biology with a focus on Bioconductor and genomic data.

The *BioC2006: Where Software and Biology Connect* conference was held in Seattle at the Fred Hutchinson Cancer Research Center on August 3–4, 2006. Over 100 scientists attended. The conference consisted of 11 talks from leading researchers in computational biology and 13 hands-on lab sessions presented by Bioconductor package developers.

Of the 12 attendees who responded to our conference evaluation survey, all 12 reported that they would consider coming again. Below are a few quotes from the survey responses:

What was your main reason for coming to BioC2006?

> I use BioC packages for most of the data analysis I do. I know some packages pretty well, but I wanted to learn about other

packages, as well as interact with people who are knowledgeable in the field.

Do you have any general comments about morning speakers?

It was good to see the new issues and areas in genomics, both from a researcher's perspective and from the developer's perspective. I thought it was a good mix between biologists and biostatisticians.

What did you like best about the conference?

Meeting and talking with the "names" on the BioC mailing list. The internet is great, but I just can't replace the ease of verbal communication, especially with a group. I had several important discussions which never would have happened on the mailing list.

Was two days enough time, too much time?

I would have liked it to be a day longer, even if the 3rd day was all labs. You really could only go to 2 of the 14 lab sessions, and I would have liked to gone to 5 or 6 of them. Having access to all the lab's materials on-line is helpful - I can attempt to go through them on my own.

7 Project Participants and Key Personnel

7.1 Gentleman Lab Members

These individuals, all working in the Gentleman Lab at the Fred Hutchinson Cancer Research Center in Seattle, Washington, played a central role in executing project objectives during 2006.

Seth Falcon Scientific programmer, project and release manager.

Martin Morgan Developer in charge of Biobase.

Nianhua Lee Developer in charge of annotation data package construction and data management.

Herve Pages Developer in charge of the build and test system.

7.2 Bioconductor Core Developers

The Bioconductor Core Developers guide the project through direct project development and participation in design discussions.

- Douglas Bates, University of Wisconsin, USA.
- Ben Bolstad, Division of Biostatistics, UC Berkeley, USA.
- Vince Carey, Harvard Medical School, USA.
- Tony Chiang, Fred Hutchinson Cancer Research Center, USA.
- Marcel Dettling, Federal Inst. Technology, Switzerland.
- Sandrine Dudoit, Division of Biostatistics, UC Berkeley, USA.
- Byron Ellis, Harvard Department of Statistics, USA.
- Seth Falcon, Fred Hutchinson Cancer Research Center, USA
- Laurent Gautier, France.
- Robert Gentleman, Fred Hutchinson Cancer Research Center, USA.
- Jeff Gentry, Dana-Farber Cancer Institute, USA.
- Kurt Hornik, Wirtschafts Universitat Wien, Austria.
- Torsten Hothorn, Institut fuer Medizininformatik, Biometrie und Epidemiologie, Germany.
- Wolfgang Huber, European Bioinformatics Institute, UK
- Stefano Iacus, Italy
- Rafael Irizarry, Department of Biostatistics (JHU), USA.
- Friedrich Leisch, Technische Universitat Wien, Austria.
- Li Long, Swiss Institute of Bioinformatics, Switzerland.
- Ting-Yuan Liu, Fred Hutchinson Cancer Research Center, USA.
- James MacDonald, University of Michigan, USA.
- Martin Maechler, Federal Inst. Technology, Switzerland.

- Crispin Miller, Paterson Institute Bioinformatics Group, UK.
- Herves Pages, Fred Hutchinson Cancer Research Center, USA.
- Anthony Rossini, Modeling and Simulation, Novartis Pharma AG, Switzerland
- Colin Smith, University of California, San Francisco, USA.
- Luke Tierney, University of Iowa, USA.
- Jean Yee Hwa Yang, Sydney, Australia.
- Jianhua (John) Zhang, Dana-Farber Cancer Institute, USA.

References

Adam Ameur, Vladimir Yankovski, Stefan Enroth, Ola Spjuth, and Jan Komorowski. The LCB Data Warehouse. *Bioinformatics*, 22(8):1024–1026, Apr 2006.

Suse Beyer, Yvonne Walter, Juergen Hellmann, Peter-Juergen Kramer, Annette Kopp-Schneider, Michaela Kroeger, and Carina Ittrich. Comparison of software tools to improve the detection of carcinogen induced changes in the rat liver proteome by analyzing SELDI-TOF-MS spectra. *J Proteome Res*, 5(2):254–261, Feb 2006.

Pan Du, Warren A Kibbe, and Simon M Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17):2059–2065, Sep 2006. Evaluation Studies.

M Dugas, F Weninger, S Merk, A Kohlmann, and T Haferlach. A generic concept for large-scale microarray analysis dedicated to medical diagnostics. *Methods Inf Med*, 45(2):146–152, 2006.

Hahne F, Arlt D, Sauermann M, Majety M, Poustka A, Wiemann S, and Huber W. Statistical methods and software for the analysis of high throughput reverse genetic assays using flow cytometry readouts. *Genome Biol*, 7(8):R77, Aug 2006a. JOURNAL ARTICLE.

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, and Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, Sep 2006b. JOURNAL ARTICLE.

Darlene R Goldstein. Partition resampling and extrapolation averaging: approximation methods for quantifying gene expression in large numbers of short oligonucleotide arrays. *Bioinformatics*, 22(19):2364–2372, Oct 2006.

Wolfgang Huber, Joern Toedling, and Lars M Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22 (16):1963–1970, Aug 2006.

Rafael A Irizarry, Zhijin Wu, and Harris A Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22(7):789–794, Apr 2006.

Lesley Jones, Darlene R Goldstein, Gareth Hughes, Andrew D Strand, Francois Collin, Stephen B Dunnett, Charles Kooperberg, Aaron Aragaki, James M Olson, Sarah J Augood, Richard L M Faull, Ruth Luthi-Carter, Valentina Moskvina, and Angela K Hodges. Assessment of the relationship between pre-chip and post-chip quality measures for Affymetrix GeneChip expression data. *BMC Bioinformatics*, 7:211, 2006. Evaluation Studies.

Macdonald JW and Ghosh D. COPA-cancer outlier profile analysis. *Bioinformatics*, Aug 2006. JOURNAL ARTICLE.

Richard E Kennedy, Robnet T Kerns, Xiangrong Kong, Kellie J Archer, and Michael F Miles. SScore: an R package for detecting differential gene expression without gene expression summaries. *Bioinformatics*, 22(10): 1272–1274, May 2006.

Eun-Kyung Lee, Sung-Gon Yi, and Taesung Park. arrayQCplot: software for checking the quality of microarray data. *Bioinformatics*, 22(18):2305–2307, Sep 2006.

Sophie Lemoine, Florence Combes, Nicolas Servant, and Stephane Le Crom. Goulphar: rapid access and expertise for standard two-color microarray normalization methods. *BMC Bioinformatics*, 7:467, 2006.

Claudio Lottaz, Xinan Yang, Stefanie Scheid, and Rainer Spang. OrderedList–a bioconductor package for detecting similarity in ordered gene lists. *Bioinformatics*, 22(18):2315–2316, Sep 2006.

Juan Jose Lozano and Susana G Kalko. AMarge: Automated Extensive Quality Assessment of Affymetrix chips. *Appl Bioinformatics*, 5(1):45–47, 2006. Evaluation Studies.

H C Lukaski. Assessing regional muscle mass with segmental measurements of bioelectrical impedance in obese women during weight loss. *Ann N Y Acad Sci*, 904:154–158, May 2000.

Boutros M, Bras L, and Huber W. Analysis of cell-based RNAi screens. *Genome Biol*, 7(7):R66, Jul 2006. JOURNAL ARTICLE.

U Mansmann, M Ruschhaupt, and W Huber. Reproducible statistical analysis in microarray profiling studies. *Methods Inf Med*, 45(2):139–145, 2006.

E S Motakis, G P Nason, P Fryzlewicz, and G A Rutter. Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach. *Bioinformatics*, 22(20):2547–2553, Oct 2006.

Pierre Neuvial, Philippe Hupe, Isabel Brito, Stephane Liva, Elodie Manie, Caroline Brennetot, Francois Radvanyi, Alain Aurias, and Emmanuel Barillot. Spatial normalization of array-CGH data. *BMC Bioinformatics*, 7:264, 2006.

Mattia Pelizzola, Norman Pavelka, Maria Foti, and Paola Ricciardi-Castagnoli. AMDA: an R package for the automated microarray data analysis. *BMC Bioinformatics*, 7:335, 2006.

Johannes Rainer, Fatima Sanchez-Cabo, Gernot Stocker, Alexander Sturn, and Zlatko Trajanoski. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res*, 34(Web Server issue):498–503, Jul 2006.

Mark Reimers and Vincent J Carey. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol*, 411:119–134, 2006.

Peter C Scacheri, Gregory E Crawford, and Sean Davis. Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. *Methods Enzymol*, 411:270–282, 2006.

James M Wettenhall, Ken M Simpson, Keith Satterley, and Gordon K Smyth. affylmGUI: a graphical user interface for linear modeling of single channel microarray data. *Bioinformatics*, 22(7):897–899, Apr 2006.

Hanni Willenbrock and Jane Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, 21(22):4084–4091, Nov 2005.

C L Wilson, A H Sims, A Howell, C J Miller, and R B Clarke. Effects of oestrogen on gene expression in epithelium and stroma of normal human breast tissue. *Endocr Relat Cancer*, 13(2):617–628, Jun 2006.

Xiaoqin Xia, Michael McClelland, and Yipeng Wang. WebArray: an online platform for microarray data analysis. *BMC Bioinformatics*, 6:306, 2005.