# Bioconductor Annual Report 2008

Patrick Aboyoun
Fred Hutchinson Cancer Research Center

May 2008

## 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 Package submission management

The Bioconductor project relies on a peer-review process of candidate package add-ons to ensure it grows containing high-quality, scientifically-relevant 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 has been spending 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, 2007 – May, 2008, over 56 software packages add-ons have been managed by the Seattle Bioconductor team, of which over 40 have been accepted for inclusion in Bioconductor.

## 1.3 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.

### 1.4 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 Size of Project

The Bioconductor project is comprised of R packages contributed by a worldwide bioinformatics community. There are currently 167 active developers and 272 contributed packages in Bioconductor's development repository. The project also maintains 375 annotation data packages that aid in the analysis of data from microarray experiments. Table 1 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 | 2.0 | 2.1 | 2.2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Package Count | 20 | 30 | 49 | 81 | 100 | 123 | 141 | 172 | 188 | 214 | 233 | 260 |

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 three email lists, `bioconductor`[1], `bioc-devel`[2], and `bioc-sig-sequencing`[3].

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

1. The `bioconductor` list is a forum for user questions, project announcements, and general discussion of interest to the Bioconductor community. As of May, 2008 the list has 1868 subscribers (individuals who receive mail from the list).

2. The `bioc-devel` list is a forum for package contributors' questions and discussion relating to the development of Bioconductor packages. As of May, 2008 this list has 432 subscribers.

3. The `bioc-sig-sequencing` list, started in February, 2008, is a forum for discussing the management and analysis of high-throughput short read data such as that from Solexa or 454 technologies. As of May, 2008 this list has 174 subscribers.

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 over the past six years.

| Year | Posts/month | Authors/month |
|---|---|---|
| 2002 | 59 | 13 |
| 2003 | 231 | 47 |
| 2004 | 320 | 60 |
| 2005 | 353 | 61 |
| 2006 | 348 | 59 |
| 2007 | 432 | 75 |
| 2008* | 380 | 112 |

Table 2: Monthly average number of posts and number of unique authors for the `bioconductor` mail list from January, 2002 – April, 2008.

## 4 The Bioconductor Website

The Bioconductor website, `http://bioconductor.org`, averaged over 11445 unique visitors and over 934GB of content per month in the year from May, 2007 to April, 2008. The most active month during this period was October 2007, where the site served 1459GB of content of which 1444GB (99%) corresponded to package downloads. The Biobase package was downloaded by

3401 unique IP addresses between May, 2007 and April, 2008. This website is hosted 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 source packages that can be installed on Linux, Solaris, 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 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 | 2x dual-core Xeon 3.00 GHz | 8 GB | 7h + 1h |
| Linux 32-bit | 2x dual-core Xeon 2.80 GHz | 4 GB | 11h |
| Windows 32-bit | 2x dual-core Xeon 3.00 GHz | 3 GB | 9h + 3h |
| OS X 32-bit | 2x dual-core Xeon 2.00 GHz | 3 GB | 6h30 + 1h30 |

Table 3: Servers used to build and test Bioconductor packages along with the number of hours required for a build/test cycle of all software packages (first number) and all experiment data packages (second number).

# 6 Accomplishments

## 6.1 Papers Citing Bioconductor

Bioconductor has become a vital software platform for the worldwide genomic research community. As of May, 2008, Google Scholar notes there are 970 scientific documents that cite the groundbreaking *Genome Biology* 2004 paper **Bioconductor: open software development for computational biology and bioinformatics**. These widespread citations has made the 2004 Bioconductor paper the third most accessed article of all time from *Genome Biology*.

Bioconductor citations in leading scientific journals have increased from January, 2003 to May, 2008. Table 4 contains the results of PubMed searches for "bioconductor" over different timeframes. It shows there have been at

least 136 journal citings from January, 2003 to May, 2008, with a little over 50% (72) being made in *Bioinformatics*. A sample of 60 publications citing Bioconductor in 2007 or 2008 are listed in the bibliography of this report.

| Publication | 2003 | 2004 | 2005 | 2006 | 2007 | 2008* |
|---|---|---|---|---|---|---|
| *Bioinformatics* | 3 | 8 | 13 | 14 | 22 | 12 |
| Other | 4 | 5 | 6 | 16 | 22 | 11 |
| Total | 7 | 13 | 19 | 30 | 44 | 23 |

Table 4: PubMed searches for "bioconductor" on publications from January, 2003 – May, 2008.

### 6.2 Bioconductor Courses

Bioconductor courses have been held in the following worldwide locations in 2007 and early 2008:

1. **Laussane Bioconductor Developer Meeting** – Laussane, Switzerland – April 24-25, 2008.
2. **Boston Bioconductor Intermediate Training** – Boston, MA – March 5-7, 2008.
3. **Advanced R for Bioinformatics** – Seattle, WA – February 13-15, 2008.
4. **Bioinformatics and Genetic Data Analysis Using R** – Seoul, South Korea – Dec 3, 3007.
5. **Introduction to R and Bioconductor** – Seattle, WA – November 28-30, 2007.
6. **Introductory level workshops on R, Bioconductor or Programming in R** by Thomas Girke, UC Riverside.
7. **Statistical Analysis of Microarray Expression Data with R and Bioconductor** – Copenhagen, DK – November 5-9, 2007.
8. **Bioconductor Advanced Course** – Chicago, IL – October 1-3, 2007.

9. **BioC2007: Where Software and Biology Connect** – Seattle, WA – August 6-7, 2007.

10. **Computational and Statistical Aspects of Microarray Analysis** – Bressanone-Brixen, Italy – June 18-22, 2007.

11. **Local Training** – Seattle, WA – May 2007.

12. **BioC Developers Meeting** – Lausanne, Switzerland – April 4-5, 2007.

13. **Advanced R Programming and Bioconductor** – Hinxton, UK – 30 March - 1 April 2007.

14. **Bioconductor Advanced Course** – Seattle, WA, USA – January 2007.

15. Bioconductor is a central component of the Cold Spring Harbor Lab summer course on **Integrative Data Analysis for High-throughput Biology** (13-27 July 2007).

### 6.3 Sponsorships

- We provided travel expense and conference fee scholarships for attending the BioC2007 conference to four BioC package developers and three students for a total of seven scholarships.

- We provided $1500USD for student travel expenses to the DSC2007 conference in Auckland, NZ.

- Refreshments for two evening lab sessions at Cold Sping Harbor course

- Support for a last-minute substitution speaker, Alexandre Morozov, of Siggia Lab at Rockefeller University, at Interface Conference 2007, Philadelphia.

### 6.4 BioC2007 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 **BioC2007: Where Software and Biology Connect** conference was held in Seattle at the Fred Hutchinson Cancer Research Center on August 6–7, 2007. Over 100 scientists attended. The conference consisted of 12 talks from leading researchers in computational biology and 11 hands-on lab sessions presented by Bioconductor package developers.

BioC2008 will take place in Seattle, August, 2008.

# 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 2007 and 2008.

**Patrick Aboyoun** Scientific programmer, build and test manager.

**Marc Carlson** Developer in charge of annotation data packages.

**Martin Morgan** Developer in charge of Biobase package and release manager for BioC 2.2.

**Herve Pages** Developer in charge of Biostrings package.

## 7.2 Harvard Medical School Members

**Vincent Carey** Co-investigator.

## 7.3 European Bioinformatics Institute Members

**Wolfgang Huber** Co-investigator.

## 7.4 Johns Hopkins University School of Hygiene and Public Health Members

**Rafael Irizarry** Co-investigator.

# References

O. Bembom, S. Keles, and M.J. van der Laan. Supervised detection of conserved motifs in DNA sequences with cosmo. *Stat Appl Genet Mol Biol*, 6:Article8, 2007.

G. Bontempi. A blocking strategy to improve gene selection for classification of gene expression data. *IEEE/ACM Trans Comput Biol Bioinform*, 4: 293–300, 2007.

J.M. Bowen, R.J. Gibson, A. Tsykin, A.M. Stringer, R.M. Logan, and D.M. Keefe. Gene expression analysis of multiple gastrointestinal regions reveals activation of common cell regulatory pathways following cytotoxic chemotherapy. *Int. J. Cancer*, 121:1847–1856, Oct 2007.

V.J. Carey, M. Morgan, S. Falcon, R. Lazarus, and R. Gentleman. GGtools: analysis of genetics of gene expression in bioconductor. *Bioinformatics*, 23:522–523, Feb 2007.

V.J. Carey, J. Gentry, R. Sarkar, D. Gentleman, and S. Ramaswamy. SGDI: system for genomic data integration. *Pac Symp Biocomput*, pages 141–152, 2008.

B. Carvalho, H. Bengtsson, T.P. Speed, and R.A. Irizarry. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, 8:485–499, Apr 2007.

T. Chiang, N. Li, S. Orchard, S. Kerrien, H. Hermjakob, R. Gentleman, and W. Huber. Rintact: enabling computational analysis of molecular interaction data from the IntAct repository. *Bioinformatics*, 24:1100–1101, Apr 2008.

H. Cho, Y.J. Kim, H.J. Jung, S.W. Lee, and J.W. Lee. OutlierD: an R package for outlier detection using quantile regression on mass spectrometry data. *Bioinformatics*, 24:882–884, Mar 2008.

H. Choi, R. Shen, A.M. Chinnaiyan, and D. Ghosh. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, 8:364, 2007.

G.V. Cohen Freue, Z. Hollander, E. Shen, R.H. Zamar, R. Balshaw, A. Scherer, B. McManus, P. Keown, W.R. McMaster, and R.T. Ng. MDQC: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, 23:3162–3169, Dec 2007.

D. Diez, R. Alvarez, and A. Dopazo. Codelink: an R package for analysis of GE healthcare gene expression bioarrays. *Bioinformatics*, 23:1168–1169, May 2007.

C.C. dos Santos, D. Okutani, P. Hu, B. Han, E. Crimi, X. He, S. Keshavjee, C. Greenwood, A.S. Slutsky, H. Zhang, and M. Liu. Differential gene profiling in acute lung injury identifies injury-specific gene expression. *Crit. Care Med.*, 36:855–865, Mar 2008.

P. Du, W.A. Kibbe, and S.M. Lin. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*, May 2008.

M.J. Dunning, M.L. Smith, M.E. Ritchie, and S. Tavar. beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, 23:2183–2184, Aug 2007.

R. Daz-Uriarte and O.M. Rueda. ADaCGH: A parallelized web-based application and R package for the analysis of aCGH data. *PLoS ONE*, 2: e737, 2007.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23:257–258, Jan 2007.

F. Ferrari, S. Bortoluzzi, A. Coppe, A. Sirota, M. Safran, M. Shmoish, S. Ferrari, D. Lancet, G.A. Danieli, and S. Bicciato. Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics*, 8:446, 2007.

H. Froehlich, M. Fellmann, H. Sueltmann, A. Poustka, and T. Beissbarth. Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinformatics*, 8:386, 2007.

H. Frhlich, M. Fellmann, H. Sltmann, A. Poustka, and T. Beissbarth. Estimating Large Scale Signaling Networks through Nested Effect Models with Intervention Effects from Microarray Data. *Bioinformatics*, Jan 2008.

J.J. Goeman and U. Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24:537–544, Feb 2008.

W. Gregory Alvord, J.A. Roayaei, O.A. Quiones, and K.T. Schneider. A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. *Brief. Bioinformatics*, 8:415–431, Nov 2007.

F. Hahne, A. Mehrle, D. Arlt, A. Poustka, S. Wiemann, and T. Beissbarth. Extending pathways based on gene lists using InterPro domain signatures. *BMC Bioinformatics*, 9:3, 2008.

C. Harbron, K.M. Chang, and M.C. South. RefPlus: an R package extending the RMA Algorithm. *Bioinformatics*, 23:2493–2494, Sep 2007.

A.D. Hershey, D. Burdine, C. Liu, T.G. Nick, D.L. Gilbert, and T.A. Glauser. Assessing quality and normalization of microarrays: case studies using neurological genomic data. *Acta Neurol. Scand.*, Jan 2008.

W. Huber, V.J. Carey, L. Long, S. Falcon, and R. Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8 Suppl 6:S8, 2007.

M. Hummel, R. Meister, and U. Mansmann. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, 24:78–85, Jan 2008.

S.M. Hunter, F.C. Mansergh, and M.J. Evans. Optimization of minuscule samples for use with cDNA microarrays. *J. Biochem. Biophys. Methods*, 70:1048–1058, Apr 2008.

A. Kuhn, R. Luthi-Carter, and M. Delorenzi. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC Bioinformatics*, 9:26, 2008.

T. Laderas and S. McWeeney. Consensus framework for exploring microarray data using multiple clustering methods. *OMICS*, 11:116–128, 2007.

N. Lama and M. Girolami. Vbmp: variational Bayesian Multinomial Probit Regression for multi-class classification in R. *Bioinformatics*, 24:135–136, Jan 2008.

N. Le Meur, A. Rossini, M. Gasparetto, C. Smith, R.R. Brinkman, and R. Gentleman. Data quality assessment of ungated flow cytometry data in high throughput experiments. *Cytometry A*, 71:393–403, Jun 2007.

S.M. Lin, P. Du, W. Huber, and W.A. Kibbe. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.*, 36:e11, Feb 2008.

J. Liu, J.M. Hughes-Oliver, and J.A. Menius. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics*, 23:1225–1234, May 2007.

C. Lottaz, J. Toedling, and R. Spang. Annotation-based distance measures for patient subgroup discovery in clinical microarray studies. *Bioinformatics*, 23:2256–2264, Sep 2007.

J. Lu, J.C. Lee, M.L. Salit, and M.C. Cam. Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays. *BMC Bioinformatics*, 8:108, 2007.

F. Markowetz, D. Kostka, O.G. Troyanskaya, and R. Spang. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics*, 23:i305–312, Jul 2007.

P.G. Martini, D.M. Taylor, J. Bienkowska, J. Jackson, G. McAllister, H. Keilhack, and R.K. Campbell. Comparative expression analysis of four breast cancer subtypes versus matched normal tissue from the same patients. *J. Steroid Biochem. Mol. Biol.*, Mar 2008.

N. Mascellani, X. Liu, S. Rossi, J. Marchesini, D. Valentini, D. Arcelli, C. Taccioli, M. Helmer Citterich, C.G. Liu, R. Evangelisti, G. Russo, J.M. Santos, C.M. Croce, and S. Volinia. Compatible solutes from hyperthermophiles improve the quality of DNA microarrays. *BMC Biotechnol.*, 7:82, 2007.

A. McQuillin, M. Rizig, and H.M. Gurling. A microarray gene expression study of the molecular pharmacology of lithium carbonate on mouse brain mRNA to understand the neurobiology of mood stabilization and treatment of bipolar affective disorder. *Pharmacogenet. Genomics*, 17:605–617, Aug 2007.

C. Murie and R. Nadon. A correction for estimating error when using the Local Pooled Error Statistical Test. *Bioinformatics*, May 2008.

E.F. Murphy, G.J. Hooiveld, M. Muller, R.A. Calogero, and K.D. Cashman. Conjugated linoleic acid alters global gene expression in human intestinal-like Caco-2 cells in an isomer-specific manner. *J. Nutr.*, 137:2359–2365, Nov 2007.

M.J. Okoniewski and C.J. Miller. Comprehensive analysis of affymetrix exon arrays using BioConductor. *PLoS Comput. Biol.*, 4:e6, Feb 2008.

M.J. Okoniewski, T. Yates, S. Dibben, and C.J. Miller. An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. *Genome Biol.*, 8:R79, 2007.

W. Raffelsberger, Y. Krause, L. Moulinier, D. Kieffer, A.L. Morand, L. Brino, and O. Poch. RReportGenerator: automatic reports from routine statistical analysis using R. *Bioinformatics*, 24:276–278, Jan 2008.

G. Rigaill, P. Hup, A. Almeida, P. La Rosa, J.P. Meyniel, C. Decraene, and E. Barillot. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. *Bioinformatics*, 24:768–774, Mar 2008.

M.E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G.K. Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23:2700–2707, Oct 2007.

R. Sanges, F. Cordero, and R.A. Calogero. oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. *Bioinformatics*, 23:3406–3408, Dec 2007.

D. Sarkar, N. Le Meur, and R. Gentleman. Using flowViz to visualize flow cytometry data. *Bioinformatics*, 24:878–879, Mar 2008.

R.B. Scharpf, J.C. Ting, J. Pevsner, and I. Ruczinski. SNPchip: R classes and methods for SNP array data. *Bioinformatics*, 23:627–628, Mar 2007.

D. Scholtens, T. Chiang, W. Huber, and R. Gentleman. Estimating node degree in bait-prey graphs. *Bioinformatics*, 24:218–224, Jan 2008.

H. Schwender and K. Ickstadt. Empirical Bayes analysis of single nucleotide polymorphisms. *BMC Bioinformatics*, 9:144, 2008.

D. Sean and P.S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23:1846–1847, Jul 2007.

W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23:1164–1167, May 2007.

M.A. Stalteri and A.P. Harrison. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8: 13, 2007.

J. Toedling, O. Skylar, O. Sklyar, T. Krueger, J.J. Fischer, S. Sperling, and W. Huber. Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, 8:221, 2007.

E. Turro, N. Bochkina, A.M. Hein, and S. Richardson. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC Bioinformatics*, 8:439, 2007.

E.S. Venkatraman and A.B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23:657–663, Mar 2007.

T. Yates, M.J. Okoniewski, and C.J. Miller. X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res.*, 36:D780–786, Jan 2008.

H. Zhao, K. Engelen, B. De Moor, and K. Marchal. CALIB: a Bioconductor package for estimating absolute expression levels from two-color microarray data. *Bioinformatics*, 23:1700–1701, Jul 2007.

D. Zhu, Y. Li, and H. Li. Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data. *Bioinformatics*, 23:2298–2305, Sep 2007.