Bioconductor Annual Report 2011

Martin Morgan
Fred Hutchinson Cancer Research Center

July 2011

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 approximately 10 years ago (Fall, 2001) by Dr. Robert Gentleman and others, and now consists of >460 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 or are from international contributors. Table 1 summarizes growth in the number of packages hosted by

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			
2004	1.4	81	2008	2.2	260			
	1.5	100		2.3	294			
2005	1.6	123	2009	2.4	320			
	1.7	141		2.5	352			

Bioconductor, with 467 packages available in the current release. The project also produces 517 '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 82,921 unique non-FHCRC IP addresses between Mar, 2010 and February, 2011. This is up from 72,245 the previous year.

1.2 Courses and Conferences

Courses with significant input from key Bioconductor personnel have been held in the following worldwide locations in 2010 / 2011:

- *Introduction to R and Bioconductor* – Seattle, WA – December 9-10, 2010.
- *Bioconductor Overview Course* – Boston, MA – January 14, 2011.
- *International Workshop on Bioinformatics* – Port of Spain, Trinidad – January 19-21, 2011.
- *Advanced R Programming* – Seattle, WA – February 17-18, 2011.
- *Advanced RNA-Seq and ChiP-Seq data analysis* – Heidelberg, Germany – June 7-9, 2010.
- *Computational Statistics for Genome Biology 2010* – Bressanone-Brixen, Italy – June 13-18, 2010.

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

- *BioC2010 – Where Software and Biology Connect* was held in Seattle at the Fred Hutchinson Cancer Research Center on July 29–30, 2010. Over 105 scientists attended. The conference consisted of 8 talks from

leading researchers in computational biology and 16 hands-on lab sessions presented by Bioconductor package developers. We also provided travel expense and conference fee scholarships for attending the conference to *approx*20 students.

- The *European Bioconductor Developer Meeting* was held 17-18 November in Heidelberg, Germany. Fifty-one participants were exposed to new developments in R and Bioconductor, with particularly stimulating flashlight talks from the broader Bioconductor community.

1.3 Community

The project maintains four email lists:

- `bioconductor`[1] is a forum for user questions, project announcements, and general discussion of interest to the Bioconductor community. Subscribers: 2814.
- `bioc-devel`[2] is a forum for package contributors' questions and discussion relating to the development of Bioconductor packages. Subscribers: 611.
- `bioc-sig-sequencing`[3] is a forum for discussing the management and analysis of high-throughput short read data such as that from Solexa or 454 technologies. Subscribers: 742.
- `bioc-sig-proteomics`[4] started recently, in response to a developer request. Subscribers: 37.

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 eight years. The apparent decline in posts in 2011 is seen in other archives (e.g., posts to the main R mailing list declined from $\approx$ 3470 in 2010 to $\approx$ 2894 in 2011), and may reflect emergence of alternative resources (e.g., StackOverflow, SEQanswers).

The web site was revised in August, 2010. The updated site has received 481,250 total visitors (184,500 absolute unique visitors); there were 27,100 visitors in August, 2010, versus 51,298 in June, 2011. Visitors come from the United States (175,667), Germany (37,448), the United Kingdom (36,738), China (24,936), Canada (17,556), France, Italy, Spain, Japan, India, and 172 other countries.

1.4 Publication

Bioconductor has become a vital software platform for the worldwide genomic research community. There are 3136 Google Scholar citations of Gentleman

[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`
[4] `http://www.stat.math.ethz.ch/mailman/listinfo/bioc-sig-proteomics`

Table 2: Monthly average number of posts and number of unique authors for the `bioconductor` mail list from January, 2002 – July, 2011.

Year	Posts per month	Authors per month	Year	Posts per month	Authors per month
2002	59	13	2007	432	75
2003	231	47	2008	424	83
2004	320	60	2009	450	86
2005	353	61	2010	504	170
2006	348	59	2011	467	N/A

Table 3: PubMed searches for "bioconductor" on publications from January, 2003 – Feburary, 2011.

Year	N	Year	N	Year	N
2003	7	2006	30	2009	62
2004	13	2007	44	2010	52
2005	19	2008	52	2011	42[5]

et al. (2004); this is the third most accessed article of all time from *Genome Biology*. Table 3 summarizes PubMed citations for 'Bioconductor' (total: 321). Citations from 2011 are in the bibliography.

Software packages within the Bioconductor project also have been 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 P41 (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 (Department of Biostatistics, Johns Hopkins University). Significant additional leadership is provided by Drs. Robert Gentleman (Genentech) and Wolfgang Huber (EMBL). Additional key participants are listed in section 2.6.

Table 4: Citations for select Bioconductor software packages as captured by Google scholar in July, 2011. 'Citation' may be pubmed id.

Package	Citation	N	Package	Citation	N
limma	Smyth (2005)	903	MassSpecWavelet	16820428	121
vsn	12169536	898	lumi	18467348	131
affy	14960456	701	cellHTS2	16869968	82
xcms	16448051	370	affylmGUI	16455752	82
globaltest	14693814	329	beadarray	17586828	81
affycomp	14960458	252	altcdfenvs	15310390	75
limmaGUI	15297296	222	tilingArray	16787969	72
aCGH	16159913	212	made4	15797915	64
GOstats	17098774	224	edgeR	19910308	60
biomaRt	16082012	190	DESeq	20979621	40
eisa	12689096	170			

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 it containing 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 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 February, 2010 – February, 2011, 113 software packages have been managed by the Seattle Bioconductor team, of which 76 have been accepted for inclusion in Bioconductor.

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 three 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 2010-2011 includes: Simon Tavare (Advisory Board chair; University of Southern California); Robert Gentleman (Genentech); Paul Flicek (European Bioinformatics Institute); Walter Ruzzo (University of Washington); 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 2010 and 2011: Martin Morgan, principal investigator; Marc Carlson, developer; Hervé Pagès, developer; Nishant Gopalakrishnan, developer; Valerie Obenchain, developer; Dan Tenenbaum, developer; and Chao-Jen Wong, 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; Sean Davis, National Institutes of Health; and James MacDonald, University of Michigan.

References

T. Adamusiak, T. Burdett, N. Kurbatova, K. Joeri van der Velde, N. Abeygunawardena, D. Antonakaki, M. Kapushesky, H. Parkinson, and M. A. Swertz. OntoCAT – simple ontology search and integration in Java, R and REST/JavaScript. *BMC Bioinformatics*, 12:218, 2011.

N. Aghaeepour, R. Nikolic, H. H. Hoos, and R. R. Brinkman. Rapid cell population identification in flow cytometry data. *Cytometry A*, 79:6–13, Jan 2011.

S. Bauer, P. N. Robinson, and J. Gagneur. Model-based gene set analysis for Bioconductor. *Bioinformatics*, 27:1882–1883, Jul 2011.

D. Beck, M. Settles, and J. A. Foster. OTUbase: an R infrastructure package for operational taxonomic unit data. *Bioinformatics*, 27:1700–1701, Jun 2011.

H. Bolouri, R. Dulepet, and M. Angerman. resource-sharing for R and Bioconductor. *Bioinformatics*, Jun 2011.

J. Cairns, C. Spyrou, R. Stark, M. L. Smith, A. G. Lynch, and S. Tavare. BayesPeak–an R package for analysing ChIP-seq data. *Bioinformatics*, 27: 713–714, Mar 2011.

S. Chateauvieux, S. Eifes, F. Morceau, C. Grigorakaki, M. Schnekenburger, E. Henry, M. Dicato, and M. Diederich. Valproic acid perturbs hematopoietic homeostasis by inhibition of erythroid differentiation and activation of the myelo-monocytic pathway. *Biochem. Pharmacol.*, 81:498–509, Feb 2011.

T. M. Che, R. W. Johnson, K. W. Kelley, W. G. Van Alstine, K. A. Dawson, C. A. Moran, and J. E. Pettigrew. Mannan oligosaccharide modulates gene expression profile in pigs experimentally infected with porcine reproductive and respiratory syndrome virus. *J Anim Sci*, May 2011.

L. M. Crowther, S. C. Wang, N. A. Eriksson, S. A. Myers, L. A. Murray, and G. E. Muscat. Chicken ovalbumin upstream promoter-transcription factor II regulates nuclear receptor, myogenic, and metabolic gene expression in skeletal muscle cells. *Physiol. Genomics*, 43:213–227, Feb 2011.

E. J. Devitt, K. A. Power, M. W. Lawless, J. A. Browne, P. O. Gaora, W. M. Gallagher, and J. Crowe. Early proteomic analysis may allow noninvasive identification of hepatitis C response to treatment with pegylated interferon -2b and ribavirin. *Eur J Gastroenterol Hepatol*, 23:177–183, Feb 2011.

F. Ferrari, A. Solari, C. Battaglia, and S. Bicciato. PREDA: an R-package to identify regional variations in genomic data. *Bioinformatics*, Jul 2011.

O. Flores and M. Orozco. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*, 27:2149–2150, Aug 2011.

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.

A. Goncalves, A. Tikhonov, A. Brazma, and M. Kapushesky. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, 27:867–869, Mar 2011.

A. M. Hess, A. N. Prasad, A. Ptitsyn, G. D. Ebel, K. E. Olson, C. Barbacioru, C. Monighetti, and C. L. Campbell. Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. *BMC Microbiol.*, 11:45, 2011.

A. Honkela, P. Gao, J. Ropponen, M. Rattray, and N. D. Lawrence. tigre: Transcription factor inference through gaussian process reconstruction of expression for bioconductor. *Bioinformatics*, 27:1026–1027, Apr 2011.

S. K. Houston, Y. Pina, J. Clarke, T. Koru-Sengul, W. K. Scott, L. Nathanson, A. C. Schefler, and T. G. Murray. Regional and Temporal Differences in Gene Expression of LHBETATAG Retinoblastoma Tumors. *Invest. Ophthalmol. Vis. Sci.*, 52:5359–5368, 2011.

M. Hummel, S. Bonnin, E. Lowy, and G. Roma. TEQC: an R package for quality control in target capture experiments. *Bioinformatics*, 27:1316–1317, May 2011.

H. U. Klein, C. Bartenhagen, A. Kohlmann, V. Grossmann, C. Ruckert, T. Haferlach, and M. Dugas. R453Plus1Toolbox: an R/Bioconductor package for analyzing Roche 454 Sequencing data. *Bioinformatics*, 27:1162–1163, Apr 2011.

N. Kurbatova, T. Adamusiak, P. Kurnosov, M. A. Swertz, and M. Kapushesky. ontoCAT: an R package for ontology traversal and search. *Bioinformatics*, Jun 2011.

D. Lai, X. Yang, G. Wu, Y. Liu, and C. Nardini. Inference of gene networks–application to Bifidobacterium. *Bioinformatics*, 27:232–237, Jan 2011.

H. Landmark-H?yvik, V. Dumeaux, K. V. Reinertsen, H. Edvardsen, S. D. Fossa, and A. L. B?rresen-Dale. Blood gene expression profiling of breast cancer survivors experiencing fibrosis. *Int. J. Radiat. Oncol. Biol. Phys.*, 79: 875–883, Mar 2011.

O. Larsson, N. Sonenberg, and R. Nadon. anota: Analysis of differential translation in genome-wide studies. *Bioinformatics*, 27:1440–1441, May 2011.

A. Le?niewska and M. J. Okoniewski. rnaSeqMap: a Bioconductor package for RNA sequencing data exploration. *BMC Bioinformatics*, 12:200, 2011.

P. Lopez-Romero. Pre-processing and differential expression analysis of Agilent microRNA arrays using the AgiMicroRna Bioconductor library. *BMC Genomics*, 12:64, 2011.

J. C. Mar, C. A. Wells, and J. Quackenbush. Defining an informativeness metric for clustering gene expression data. *Bioinformatics*, 27:1094–1100, Apr 2011.

E. Mercier, A. Droit, L. Li, G. Robertson, X. Zhang, and R. Gottardo. An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS ONE*, 6:e16432, 2011.

S. K. Mohapatra and A. Krishnan. Microarray data analysis. *Methods Mol. Biol.*, 678:27–43, 2011.

M. Mondal, B. Schilling, J. Folger, J. P. Steibel, H. Buchnick, Y. Zalman, J. J. Ireland, R. Meidan, and G. W. Smith. Deciphering the luteal transcriptome: potential mechanisms mediating stage-specific luteolytic response of the corpus luteum to prostaglandin F. *Physiol. Genomics*, 43:447–456, Apr 2011.

I. Ostrovnaya, V. E. Seshan, A. B. Olshen, and C. B. Begg. Clonality: an R package for testing clonal relatedness of two tumors from the same patient based on their genomicprofiles. *Bioinformatics*, 27:1698–1699, Jun 2011.

C. C. Overall, D. A. Carr, E. S. Tabari, K. J. Thompson, and J. W. Weller. ArrayInitiative - a tool that simplifies creating custom Affymetrix CDFs. *BMC Bioinformatics*, 12:136, 2011.

S. Pounds, C. L. Gao, R. A. Johnson, K. D. Wright, H. Poppleton, D. Finkelstein, S. E. Leary, and R. J. Gilbertson. A procedure to statistically evaluate agreement of differential expression for cross-species genomics. *Bioinformatics*, 27:2098–2103, Aug 2011.

F. Ribalet, D. M. Schruth, and E. V. Armbrust. flowPhyto: enabling automated analysis of microscopic algae from continuous flow cytometric data. *Bioinformatics*, 27:732–733, Mar 2011.

A. E. Ross, L. Marchionni, T. M. Phillips, R. M. Miller, P. J. Hurley, B. W. Simons, A. H. Salmasi, A. J. Schaeffer, J. P. Gearhart, and E. M. Schaeffer. Molecular effects of genistein on male urethral development. *J. Urol.*, 185: 1894–1898, May 2011a.

A. E. Ross, L. Marchionni, M. Vuica-Ross, C. Cheadle, J. Fan, D. M. Berman, and E. M. Schaeffer. Gene expression pathways of high grade localized prostate cancer. *Prostate*, Feb 2011b.

M. Sankar, K. S. Osmont, J. Rolcik, B. Gujas, D. Tarkowska, M. Strnad, I. Xenarios, and C. S. Hardtke. A qualitative continuous model of cellular auxin and brassinosteroid signaling and their crosstalk. *Bioinformatics*, 27:1404–1412, May 2011.

M. Sarvari, E. Hrabovszky, I. Kallo, N. Solymosi, K. Toth, I. Liko, J. Szeles, S. Maho, B. Molnar, and Z. Liposits. Estrogens regulate neuroinflammatory genes via estrogen receptors alpha and beta in the frontal cortex of middle-aged female rats. *J Neuroinflammation*, 8:82, Jul 2011.

R. B. Scharpf, I. Ruczinski, B. Carvalho, B. Doan, A. Chakravarti, and R. A. Irizarry. A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics*, 12:33–50, Jan 2011.

M. A. Shah, R. Khanin, L. Tang, Y. Y. Janjigian, D. S. Klimstra, H. Gerdes, and D. P. Kelsen. Molecular classification of gastric cancer: a new paradigm. *Clin. Cancer Res.*, 17:2693–2701, May 2011.

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.

X. H. Sun, Y. B. Guo, N. Liu, L. Ma, and Q. K. Deng. [Design and realization of a microarray data analysis platform]. *Nan Fang Yi Ke Da Xue Xue Bao*, 31:610–614, Apr 2011.

E. Szczurek, F. Markowetz, I. Gat-Viks, P. Biecek, J. Tiuryn, and M. Vingron. Deregulation upon DNA damage revealed by joint analysis of context-specific perturbation data. *BMC Bioinformatics*, 12:249, 2011.

X. Wang, C. Terfve, J. C. Rose, and F. Markowetz. HTSanalyzeR: an R/Bioconductor package for integrated network analysis of high-throughput screens. *Bioinformatics*, 27:879–880, Mar 2011.

B. Zacher, P. Torkler, and A. Tresch. Analysis of Affymetrix ChIP-chip data using starr and R/Bioconductor. *Cold Spring Harb Protoc*, 2011, May 2011.

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 multiplatform 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 multi-

ple 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.