

## New Developments in Bioconductor

Martin Morgan, Fred Hutchinson Cancer Research Center,  
Seattle, WA, USA, [mtmorgan@fhcrc.org](mailto:mtmorgan@fhcrc.org)

Vince Carey, Brigham and Women's Hospital, Harvard Medical  
School, Boston, MA, USA, [stvjc@channing.harvard.edu](mailto:stvjc@channing.harvard.edu)

17-18 November, 2010

# Acknowledgements

Hervé Pagès *Biostrings*, sequence infrastructure.

Marc Carlson Annotations, *GenomicFeatures*, sequence infrastructure.

Nishant Gopalakrishnan New packages, flow cytometry, *graph*, sequence infrastructure.

Dan Tenenbaum Web site, build system.

Valerie Obenchain Sequence infrastructure and analysis.

# Project Highlights: Analysis and Comprehension of High Throughput Genetic Data

- ▶ Sequencing and microarrays; flow cytometry; ...
- ▶ > 400 software packages, 500 annotation packages, 270 developers
- ▶ About 3000 mailing list subscribers, 9000 web site visits per week, 5000 *Biobase* downloads per month.

## Google Scholar citations

Bioconductor	2594
<i>limma</i>	754
<i>affy</i>	540
...	
<i>edgeR</i>	24
<i>DESeq</i>	7

# Web Site

- ▶ Installation
- ▶ BiocViews, package descriptions, vignettes
- ▶ Workflows
- ▶ News groups, FAQ
- ▶ Course and conference material

# Developer Resources

- ▶ New package submission: preview, addition to 'development' branch, incorporated in next release
- ▶ Each package under version control, authors mutually responsible for content and maintenance.
- ▶ The *Bioconductor* build system.

# Plans

- ▶ Significant package-level clean-up: maintainers should expect more activity this release cycle – redundant / obsolete package deprecation, conformance to guidelines, suitability.
- ▶ Annotations: genomic coordinates, local serialization, diverse data sources.
- ▶ Sequence analysis: representation of variants, 'fine scale' local alignment, ...
- ▶ Other: integrated / distributed deployment?

# (FHCRC) Infrastructure for Sequence Analysis

Input and Output *rtracklayer*, *Rsamtools*, *ShortRead*.

Sequence manipulation *Biostrings*.

Range-based manipulations *IRanges*, *GenomicRanges*

Annotations *GenomicFeatures*, *AnnotationDbi*, *BSgenome*.

# Sequence Analysis

25 additional packages

**ChIP-seq** *BayesPeak, CSAR, ChIPpeakAnno, ChIPseqR, ChIPsim, PICS, chipseq*

**RNA-seq** *DEGseq, DESeq, Genominator, baySeq, edgeR, rnaSeqMap, goseq* also *gage*

**Diverse** infrastructure *genomeIntervals, girafe*; base calling: *Rolexa*; visualization: *HilbertVis HilbertVisGUI*; motif: *MotIV, rGADEM*; domain-specific *MEDIPS, OTUbase, R453Plus1Toolbox* database: *SRAdb, oneChannelGUI* smRNA: *segmentSeq*



# Favorites

## *Rsamtools*

- ▶ BAM as a common representation of aligned reads.
- ▶ Fast, indexed, flexible.

## *SummarizedExperiment* in *GenomicRanges*.

- ▶ Derived data, like *ExpressionSet* from *Biobase*.

# Resources

- ▶ Web site: <http://bioconductor.org>.
- ▶ Mailing lists:  
<http://bioconductor.org/help/mailing-list/>. `bioc` for general questions, `bioc-devel` for questions about package development, `bioc-sig-seq` for sequence-specific questions.