Bioconductor: accomplishments and opportunities

Martin Morgan
mtmorgan@fhcrc.org
Bioconductor / Fred Hutchinson Cancer Research Center
Seattle, WA, USA

28 July 2009

Accomplishments: the numbers

Community: activity this year

- 60 new packages
- 14,000 unique visitors per month
- 34,354 unique Biobase downloads / year
- 2050 mailing list subscribers (plus 480 to `bioc-devel`, 400 to `bioc-sig-seq`

Outreach

- 8 Bioconductor courses from 'us'
- International activities in Switzerland, Germany, Italy
- Nearly 100 conference attendees

Science

- 76 new pubmed citations (January 2008-May 2009)
- 1450 citations of the original Bioconductor paper

Accomplishments: recent contributions

Microarrays

- methylumi, AffyTiling, crlmm, betr, ...

Pathways, graphs, and networks

- GOSemSim, KEGGgraph, SPIA, RpsiXML, ...

Flow cytometry

- flowMerge, flowFP, flowStats, ...

Sequencing

- ...

Accomplishments: sequencing

Released packages

- ShortRead, Biostrings, IRanges, genomeIntervals, HilbertViz, HilbertVizGUI, BSgenome, rtracklayer,

Packages in the next release

- chipseq: tools implementing an approach to analysis of ChIP-seq analysis
- Rolexa: probabilistic base calling, quality checks, and diagnostics for Illumina GA II sequencing platform
- Additional packages not yet through review
- ...

Representing sequence data

Third party input: ShortRead

- e.g., *AlignedRead*, containing reads, quality scores, ids, chromosome, position, strand, and other information

Sequences: Biostrings, BSgenome

- *DNAString*, *DNAStringSet*: one or many strings

Streamlined: IRanges

- Run-length encoding, e.g., *coverage* vectors
- *RangedData*: spaces (e.g., chromosomes / contigs) and ranges (e.g., island extent, in genome coordinates)

Who is Bioconductor?

Community

- Nearly 100 conference participants, 210 different package authors, innumerable mailing list contributors, ...

FHCRC Development and support

- Patrick Abouyoun, Hervé Pagès, Marc Carlson, Chao-Jen Wong, Nishant Gopalakrishnan
- Michael Lawrence, Florian Hahne, Deepayan Sarkar, Zhizhen Zhao

Leadership

- James MacDonald (U. Mich), Sean Davis (NIH)
- Raphael Irrizary (JHU), Vince Carey (Harvard University Medical School), Wolfgang Huber (EBI, Hiedelberg)
- Robert Gentleman (FHCRC / Genentech)

Major opportunities for development

1. Ongoing microarray support
2. Graphs: Rgraphviz, large graph representation, statistical analysis
3. Sequences: 'domain' development; infrastructure; large volume data; annotation
4. Community contributions

Sequence analysis

- Domains: ChIP-seq; RNA-seq; quality assessment; . . .
- Infrastructure: Biostrings and IRanges; experiment-level data objects, like *ExpressionSet*
- Large-volume (e.g., 1000 genomes) data: storage, representation, manipulation, access
- Sequence-appropriate annotation: genome coordinates; transcript-level; *transparent* integration
- Third-party integration, e.g., the Cancer Genome Atlas, caBIG

Future community contributions — ???

- Join us for Developer Day (Wednesday)
- Talk with us about your ideas and challenges!

Key resources

- `http://bioconductor.org`
  - Especially 'Software' link and 'Workflows' tab
  - `http://bioconductor.org/install` for installation
- Bioconductor mailing list for all questions
  - Include output of `sessionInfo()`
  - Short, reproducible examples help to convey the problem, and easily identify a solution