

The Bioconductor Project for Reproducible Analysis of High Throughput Genomic Data

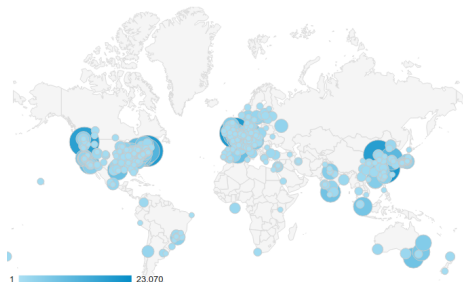
Martin Morgan (mtmorgan@fredhutch.org)
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

12 January 2015

Outline

1. Project status
2. Recent highlights
3. Future directions
4. Why *Bioconductor*?

Project status (December, 2014)



2014 Web site visitors, by city

Users

- ▶ 320,000 unique IP address package downloads / year
- ▶ 1,300 support site contributors / year, 8,200 visitors / month
- ▶ 10,500 PubMed Central mentions of 'Bioconductor';
≈ 22,000 citations to *Bioconductor* packages
- ▶ At least 12 of 15 initial TCGA publications

Project status (December, 2014)

Packages & developers

- ▶ 935 release packages (vs. 750 last year)
- ▶ > 900 bioc-devel subscribers
- ▶ 100's of maintainers in N. America, Europe, & world-wide.

Funding

- ▶ US NHGRI/NIH U41 (since 2005) – Bioconductor: An Open Computing Resource for Genomics (Morgan, Carey, Irizzary)
- ▶ US NSF BIGDATA – Scalable Statistical Computing for Emerging Omics Data Streams (Morgan, Carey, Huber, Taylor)
- ▶ US NCI U24: *Bioconductor* for Integrative Analysis for Cancer Genomics (Morgan, Carey, Hansen, Waldron)
- ▶ EC – coming soon?!

Recent highlights

Web site

- ▶ [biocViews](#) curation & use (e.g., docker images)
- ▶ Newsletters.
- ▶ Indexed [course material](#), recent [literature citations](#), [videos](#) (contributions welcome)!

Infrastructure

- ▶ <https://support.bioconductor.org>.
- ▶ Docker & Amazon machine instances ([AMIs](#), e.g., *StarCluster*)
- ▶ *AnnotationHub*: file-based and other annotation resources.
- ▶ *BiocParallel*, *GenomicFiles*: processing files \times ranges.
- ▶ *BiocStyle*, especially *rmarkdown*-flavored vignettes.
- ▶ `findOverlaps`: nested containment list implementation – 3-10 \times faster, up to 1/4th memory use.
- ▶ `biocLite()` support for github repositories.

Recent highlights

Contributed packages

- ▶ !
- ▶ !
- ▶ !
- ▶ !
- ▶ !
- ▶!

Recent highlights: *AnnotationHub*

File-based resources, e.g., UCSC *liftOver* files

```
## hg19SNPs <- GRanges(...)
library(AnnotationHub)
hub <- AnnotationHub()
chain <- query(hub, 'hg19ToHg38')[[1]]
hg38SNPs <- liftOver(hg19SNPs, chain)
```

Annotation-style resources, e.g., *grasp2*

```
library(grasp2db) # Annotation package,
                  # 6 Gb AnnotationHub resource
d <- GRASP2()     # dplyr instance
hispanic <- tbl(d, "count") %>%
  filter(Population=="Hispanic")
semi_join(tbl(d, "variant"), hispanic)
```

Recent highlights: docker

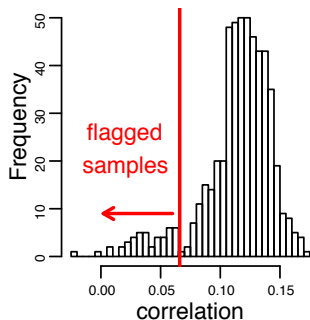
- ▶ <http://bioconductor.org/help/docker/>
- ▶ Reproducibility, ease of use, convenience

```
docker run --rm -ti --name  
    devel bioconductor/devel_base R
```


Future directions

General

- ▶ Integrative analysis.
- ▶ 'Populations', e.g., single cells; tumors; many samples.
- ▶ Statistically informed actionable insights; visual and interactive.



TCGA ovarian expression / CN correlation; Levi Waldron

Specific

- ▶ Allow *SummarizedExperiment* `rowData()` without ranges.
- ▶ Update *Rsamtools* via *Rhtslib*
- ▶ `mapTo*`: DNA / Protein integration; individual genomes
- ▶ *BiocParallel* logging, ease-of-use.
- ▶ *methods* performance improvements

Why *Bioconductor*?

A *community* of users and developers.

- ▶ Extensive & interoperable
- ▶ Statistical (volume, technology, experimental design, population samples)
- ▶ Reproducible: long-term, multi-participant science
- ▶ Leading edge: embrace novel technologies and analysis
- ▶ Accessible: affordable, transparent, usable (e.g., vignettes & man pages)

Huber et al., Orchestrating high-throughput genomic analysis with *Bioconductor*. *Nature Methods*: soon!

Why *Bioconductor*?

More than a software archive.

- ▶ Build on relevant software, e.g.,
 - ▶ *GenomicRanges* for efficient interoperability; *ExpressionSet* / *SummarizedExperiment* for genetic / phenotypic integration...
 - ▶ I/O via *rtracklayer*, *Rsamtools*, *illuminaio*, ...
 - ▶ Resource access via *biomaRt*, *GEOquery*, ...
- ▶ Commit to long-term support
 - ▶ e.g., *affy* in use 10 years after introduction.
 - ▶ Comprehensive documentation coupled with traditional scientific publications
 - ▶ Engage users via support forum, foster productive collaborations
- ▶ Enable transitions
 - ▶ User to developer
 - ▶ Student to professional

Driving principle: analysis & comprehension of high-throughput genomic data

Future events

- ▶ Computational Statistics for Genome Biology (CSAMA), 15-19 June, Brixen / Bressanone, Italy
- ▶ *useR!*, 1-3 July, Aalborg, Denmark
- ▶ BioC 2015, 20 - 22 July, Seattle, WA USA

Acknowledgments

Core (Seattle): Sonali Arora, Marc Carlson, Nate Hayden, Valerie Obenchain, Hervé Pagès, Paul Shannon, Dan Tenenbaum.

Technical Advisory Board: Vincent Carey, Aedin Culhane, Sean Davis, Robert Gentleman, Kasper Hansen, Wolfgang Huber, Rafael Irizarry, Levi Waldron.

Scientific Advisory Board: Paul Flicek, Simon Tavaré, Simon Urbanek.