Bioc 2017: Where Software and Biology Connect

Martin Morgan

Roswell Park Cancer Institute
Buffalo, NY, USA
martin.morgan@roswellpark.org

26 July 2017

Welcome!

Special thanks

- Erica Fieck, David Nunes, Pam Jarrett, Meeghan Becker.
- Fast talkers, workshop contributors, scholarship recipients.

Conference home page

- `https://bioconductor.org/bioc2017`
- Includes link to Developer Day schedule, `https://goo.gl/8oRmsp`.

Twitter: #bioc2017

Sponsors

Today

Informal and flexible

- Want to do something different? Say so!

Schedule `https://goo.gl/8oRmsp`

- Introduction / project overview / group activity
- Lightning talks (two parallel sessions)
- Workshops (bring your laptops!) / Birds-of-a-feather (discussion-oriented)
- Panel discussion

Coffee breaks & box lunches; no evening acivities

Tommorrow: 8:30 am, Jimmy Fund Auditorium

Logistics

Posters

- Please leave at the registration desk, today if possible, Thursday morning at the latest.

Conference Amazon Machine Instances

- Start yours today at courses.bioconductor.org
- Username: ...; password: ...
- 'This site can't be reached...' ?? Reload page (AMI is still spinning up)
- 'We're sorry, but something went wrong' ?? Re-submit request for AMI or notify Lori (Amazon ran out of cloud!)

Bioconductor – introduction

Statistical analysis and comprehension of high-throughput genomic data
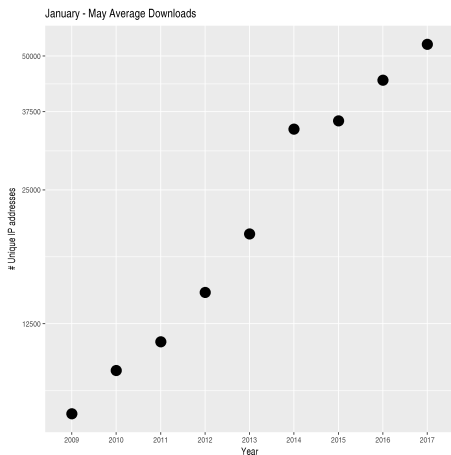
- Started 2001
- 1383 software packages
- > 900 distinct maintainers

Widely used, highly respected

- > 20000 PubMedCentral full-text citations

Supported

- https://bioconductor.org
- https://support.bioconductor.org

January - May Average Downloads

(scatter plot: # Unique IP addresses vs Year, 2009–2017)

Bioconductor – introduction

- Federally funded through NIH (NHGRI, NCI), EU, . . .
- International relevance

Important themes

- Statistical programming
- Leading-edge
- Reproducible
- Integrative
- Documented
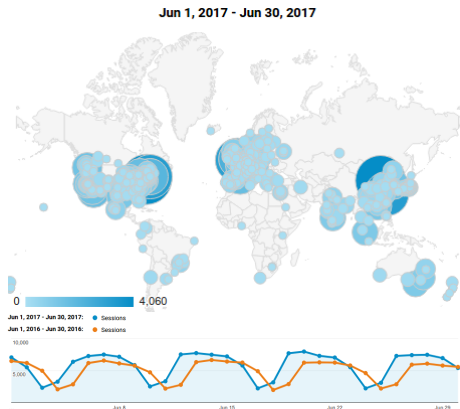
Web site visitors

Bioconductor – core team

Valerie Obenchain *VariantAnnotation*, *GenomicFiles*, *GRanges* infrastructure; nightly builds; AWS administration.

Hervé Pagès *GRanges* infrastructure; *Biostrings*, *DelayedArray* & friends; nightly builds.

Marcel Ramos *MultiAssayExperiment* (with Levi Waldron); *SOUNDBoard*.

Lori Shepherd *BiocFileCache*, 'single package builder', docker and AMI solutions.

Nitesh Turaga Transition to git version control.

Daniel van Twisk *Organism.dplyr* / *AnnotationFilter*.

Bioconductor – friends of the core team

Andrzej Olés *BiocStyle*, workflows, build system (chef, git integration).
Mike Smith *biomaRt*, *rhdf5* / *Rhdf5lib*.
Lukas Shiffer Support site markdown!

Organism.dplyr (Yubo Cheng, Daniel van Twisk) – motivation

- *TxDb.**, *org.** annotation package integration not entirely satisfying.
- Databases underly these resources, why not expose in a 'modern' (i.e., *dplyr*) way?

```
library(Organism.dplyr)
src_organism("TxDb.Hsapiens.UCSC.hg38.knownGene")

## src:  sqlite 3.19.3 [/home/mtmorgan/.cache/BiocFileCache/5b
## tbls: id, id_accession, id_go, id_go_all, id_omim_pm,
##   id_protein, id_transcript, ranges_cds, ranges_exon,
##   ranges_gene, ranges_tx
```

- See *Organism.dplyr* and *AnnotationFilter* vignettes for details.

Organism.dplyr – insights

Exposing sql tables is really helpful.

- Easy to see and manipulate data.

A pure *dplyr* approach is too low-level

- Common queries require complicated joins.
- Useful 'verbs' like `select()`, `mapIds()` are and extended on top of *dplyr*.
- Classes like *GRanges* are incredibly useful, even if superficial similarity to a `tibble`.

Ici c'est ne pas une pipe

- `tbl(src, "ranges_tx") %>% makeGRangesFromDataFrame()`
- Endomorphisms, consistency, and restricted vocabulary really help users.

TENxGenomics (github only)

- E.g., single-cell RNA-seq, 30,000 genes by 1.3 million samples.
- On-disk representation in hdf5.
- Convenient in-memory 'matrix' abstraction for subsetting, etc.; easy input of manageable subset.
- https://github.com/mtmorgan/TENxGenomics

```
> basename(fl)
[1] "1M_neurons_filtered_gene_bc_matrices_h5.h5"
> (tenx <- TENxGenomics(fl))
class: TENxGenomics
h5path: ./1M_neurons_filtered_gene_bc_matrices_h5.h5
dim(): 27998 x 1306127
> tenk <- tenx[, sample(ncol(tenx), 10000)] ## fast
> m <- as.matrix(tenk)                       ## manageable
> se = SummarizedExperiment(list(tenx))      ## rich
```

Slack

Want to join? Look for an invitation on the bioc-devel mailing list next week.

Acknowledgments