

# The *Bioconductor* Project: Current Status

Martin Morgan

Roswell Park Cancer Institute  
Buffalo, NY, USA  
[martin.morgan@roswellpark.org](mailto:martin.morgan@roswellpark.org)

4 November 2016



<https://bioconductor.org>

<https://support.bioconductor.org>

Analysis and comprehension of high-throughput genomic data.

- Started 2002
- 1296 *R* packages – developed by 'us' and user-contributed.

Well-used and respected.

- 43k unique IP downloads / month.
- 17,000 PubMedCentral citations.

# State of the project

- Packages
- Users
- Web & support sites
- Training & meetings
- Release & devel builders
- Funding
- Governance: (annual) Scientific Advisory Board; (monthly) Technical Advisory Board

# Recent developments

- New package reviews
- *ExperimentHub* and *AnnotationHub*
- Large data representation: *HDF5Array*
- (Sneak peak) *Organism.dplyr*

## HDF5Array

```
library(HDF5Array)    # available in release & devel
n = 10000; m = 1000; # very large size
h5 = HDF5Array(matrix(rnorm(n * m), n))
h5 + h5               # 'delayed' computation
library(SummarizedExperiment)
SummarizedExperiment(h5) # rich context
```

## Sneak peak: *Organism.dplyr*

```
> library(Organism.dplyr) # not yet publicly available
> src = src_ucsc("Homo sapiens") # any org.* + TxDb.*
using org.Hs.eg.db, TxDb.Hsapiens.UCSC.hg38.knownGene
> src
src:  sqlite 3.8.6 [/home/mtmorgan/organism_dplyr.sqlite]
tbls: id, id_accession, id_go, id_go_all, id_omim_pm,
      id_protein, id_transcript, ranges_cds, ranges_exon,
      ranges_gene, ranges_tx
> tbl(src, 'id') %>% filter(symbol == 'BRCA1') %>%
  select(ensembl, symbol, genename)
> exons(src, filter=list(symbol='BRCA1')) # GRanges
> exons_tbl(src, filter=list(symbol='BRCA1')) # tibble
```

# Programming best practices

- Reuse & interoperability
- Correct, robust, efficient (vectorized) code; *BiocParallel*
- Documentation: classic or *roxygen2*
- Testing: *RUnit* or *testthat*
- Classic, tidy, and semantically rich data

## Correct, robust, efficient...

```
f = function(n) {
  x = integer(0)
  for (i in 1:n)
    x = c(x, i)
  x
}
microbenchmark(f(1000),
  f(10000), f(100000))

f1 = function(n) {
  x = integer(n)
  for (i in 1:n)
    x[i] = i
  x
}

f2 = function(n)
  vapply(1:n, c, integer(1))

f3 = function(n)
  seq_len(n)

## correct
identical(f(100), f3(100))

## robust!
f(0); f3(0)

## efficient
system.time(f3(1e9))
```



# Classic, tidy, rich: RNA-seq count data

## Classic

- Sample  $\times$  (phenotype + expression) Feature `data.frame`

## Tidy

- 'Melt' expression values to two long columns, replicated phenotype columns. End result: long data frame.

## Rich, e.g., `SummarizedExperiment`

- Phenotype and expression data manipulated in a coordinated fashion but stored separately.

## Classic, tidy, rich: RNA-seq count data

```
df0 <- as.data.frame(list(mean=colMeans(classic[, -(1:22)])))
df1 <- tidy %>% group_by(probeset) %>%
  summarize(mean=mean(exprs))
df2 <- as.data.frame(list(mean=rowMeans(assay(rich))))
ggplot(df1, aes(mean)) + geom_density()
```

# Classic, tidy, rich: RNA-seq count data

## Vocabulary

- Classic: extensive
- Tidy: restricted endomorphisms
- Rich: extensive, meaningful

## Constraints (e.g., probes & samples)

- Tidy: implicit
- Classic, Rich: explicit

## Flexibility

- Classic, tidy: general-purpose
- Rich: specialized

## Programming contract

- Classic, tidy: limited
- Rich: strict

## Lessons learned / best practices

- Considerable value in semantically rich structures
- Current implementations trade-off user and developer convenience
- Endomorphism, simple vocabulary, consistent paradigm aid use

# Future challenges

- Git
- Cloud. Possible visions:
  - ▶ As now, but 'in the cloud'
  - ▶ Integrated with 'third party' compute efforts, e.g., NCI, NIH in the United States

## Acknowledgments

Core team (current & recent): Yubo Cheng, Valerie Obenchain, Hervé Pagès, Marcel Ramos, Lori Shepherd, Dan Tenenbaum, Greg Wargula.

Technical advisory board: Vincent Carey, Kasper Hansen, Wolfgang Huber, Robert Gentleman, Rafael Irizzary, Levi Waldron, Michael Lawrence, Sean Davis, Aedin Culhane

Scientific advisory board: Simon Tavaré (CRUK), Paul Flicek (EMBL/EBI), Simon Urbanek (AT&T), Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Rafael Irizzary (Dana Farber), Robert Gentleman (23andMe)

Research reported in this presentation was supported by the National Human Genome Research Institute and the National Cancer Institute of the National Institutes of Health under award numbers U41HG004059 and U24CA180996. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.