Good software: simple, tidy, rich

Martin Morgan

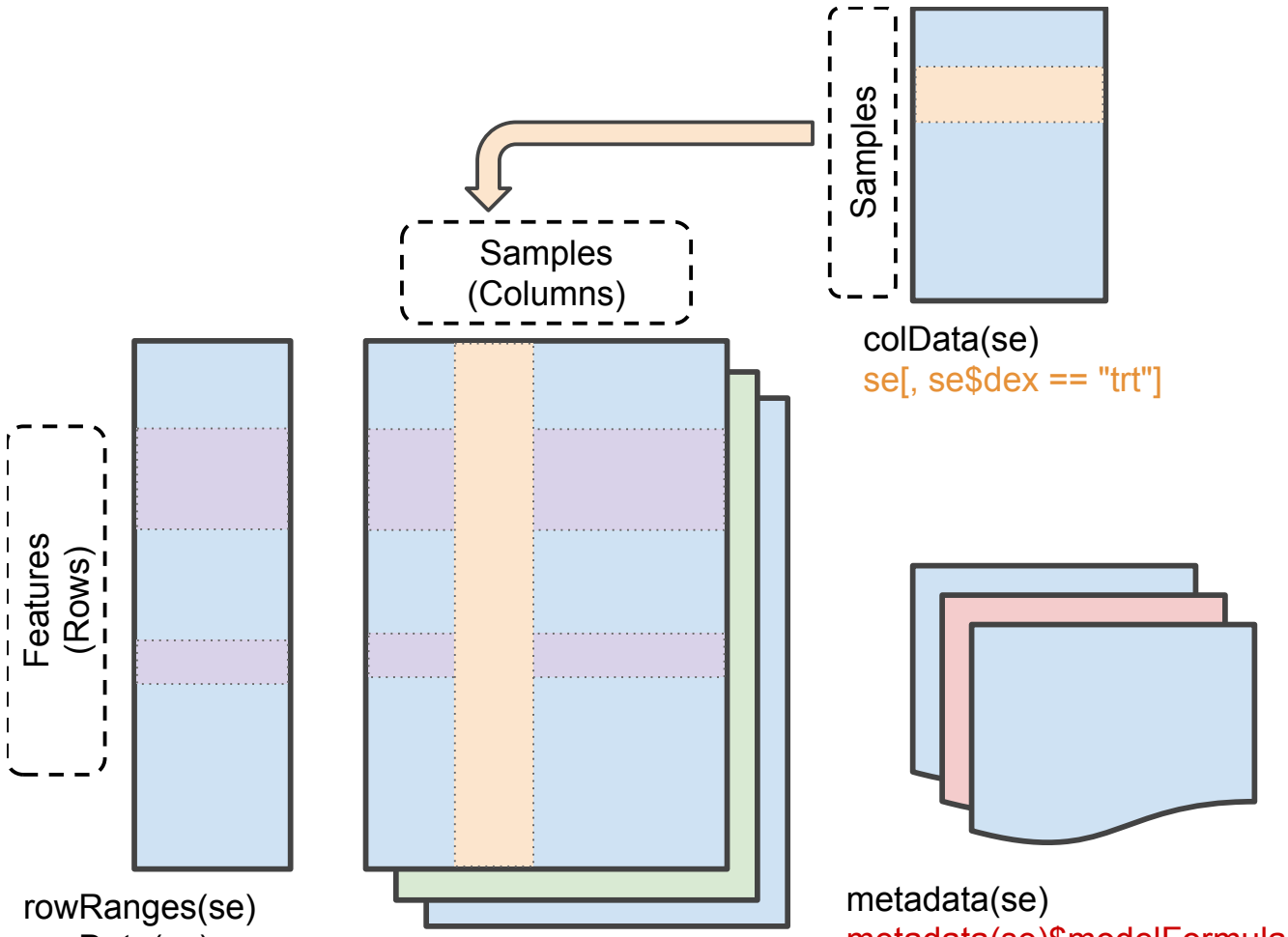
Roswell Park Cancer Institute
Buffalo, NY, USA
martin.morgan@roswellpark.org

20 April 2017

R / Bioconductor

`https://bioconductor.org`
`https://support.bioconductor.org`

1. Statistical analysis and comprehension of high-throughput genomic data.
2. Open – public version control
3. Reproducibile research – vignettes, 'experiment data' packages, release / devel branches, ...
4. Interoperable – package re-use, data structures, ...
5. Usable – documentation, support, ...


Samples
Samples
(Columns)
colData(se)
se[, se$dex == "trt"]
Features
(Rows)
rowRanges(se)
rowData(se)
subsetByOverlaps(se, roi)
assays(se)
assay(se, n = 2)
assay(subsetByOverlaps(se, roi))
assay(se[, se$dex == "trt"])
metadata(se)
metadata(se)$modelFormula

Simple, tidy, rich: RNA-seq count data

Vocabulary

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

Constraints (e.g., probes & samples)

- Tidy: implicit
- Simple, Rich: explicit

Flexibility

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

Programming contract

- Simple, 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

Pretty big data

- 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
```

Opportunities

Programmer analysts

- `https://roswellpark.org/careers`
- Programmer / Analyst – *R* software development 4924
- Senior Programmer Analyst – cloud / new-age sys. admin 4932

Bioconductor annual conference

- Boston, July 26 (D-day) – 28.
- `https://bioconductor.org/BioC2017`

Acknowledgments

Core team (current & recent): Yubo Cheng, Valerie Obenchain, Hervé Pagès, Marcel Ramos, Lori Shepherd, Nitesh Turaga.

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 Tavare (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.