Containers for Experimental and Integrative Data

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center

13-14 December 2012
What we want?

Hsu et al. 2007 J Clin Oncol 25: 4350-4357 (retracted)

- Provenance
  - Sample and row ‘metadata’
- Book-keeping, e.g., during subset
- Integration
  - With annotation resources
  - With GenomicRanges
- Re-use
What we want?

- Provenance
  - Sample and row ‘metadata’
- Book-keeping, e.g., during subset
- Integration
  - With annotation resources
  - With GenomicRanges
- Re-use

Baggerly & Coombes 2009 Ann Appl Stat 3: 1309-1334
What we have?

- **SummarizedExperiment**
  - Range-based rows; `IRanges` data structures
- **eSet**-derived
  - E.g., *DESeq countDataSet*
- Other, e.g., *edgeR*
  - Simple lists wrapped as S4 classes
- ... 
- **BamViews**
Design

SummarizedExperiment
- Experiment data
- Regions of interest
- Samples
- Assay(s)

Assays implemented to avoid unnecessary copy
Design

SummarizedExperiment

- Experiment data
- Regions of interest
- Samples
- Assay(s)

Assays implemented to avoid unnecessary copy.
Design

- list
- DataFrame
- SimpleList of arrays
- GRanges / GRangesList

**SummarizedExperiment**
- Experiment data
- Regions of interest
- Samples
- Assay(s)

*Assays* implemented to avoid unnecessary copy
Use & re-use

Use

▶ Accessors, rowData(se)
▶ Subset, se[, se$Treatment == "ChIP"]
▶ Annotation, seqinfo(se), mcols(se)
▶ Overlap, e.g., subsetByOverlaps to select rows within regions of interest

> roi <- GRanges("chr1", IRanges(1, 2e6))
> subsetByOverlaps(se, roi)

Re-use

▶ easyRNASeq, ggbio, Gviz, …
▶ VariantAnnotation VCF class; minfi, …
Limitations and Alternatives

*SummarizedExperiment*

- Ranges required? Not really, but a bit of a hack (*GRangesList* as *rowData*).
- Rectangular; not suitable for ‘ragged’ data
- Equal-sized arrays as assays
- In-memory

*eSet*-derived

- No ranges, so harder to integrate.
- Inherits sub-optimal representations, e.g., *annotatedDataFrame* rather than *DataFrame*

Other, e.g., *edgeR*

- Simple, but limited interoperability with *Bioconductor* resources
Ideas and needs?