Cancer Genomics: Integrative and Scalable Solutions in \textit{R} / \textit{Bioconductor}

Martin Morgan

\texttt{martin.morgan@roswellpark.org}

13 June 2016
Bioconductor

Statistical analysis and comprehension of high-throughput genomic data

Established 2002, widely used and well-respected

- [https://bioconductor.org/](https://bioconductor.org/)
- [https://support.bioconductor.org](https://support.bioconductor.org)

BioC 2016
Join us for our annual conference BioC 2016: Where Software and Biology Connect, this year at Stanford University, June 24 (Developer Day), 25, and 26.

About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1211 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News
- Bioconductor 3.3 is available.
- Bioconductor F1000 Research Channel launched.
- Orchestrating high-throughput genomic analysis with Bioconductor (abstract) and other recent literature.
- Read our latest newsletter and course material.
- Use the support site to get help installing, learning and using Bioconductor.

Install
Get started with Bioconductor
- Install Bioconductor
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn
Master Bioconductor tools
- Courses
- Support site
- Package vignettes
- Literature citations
- Common workflow
- FAQ
- Community resources
- Videos

Develop
Contribute to Bioconductor
- Use RForApache dependency
- Use RForDocker dependency
- GitHub
- New package submission
- Developers
- Build reports

Tweets by @Bioconductor
Common *Bioconductor* tasks and packages

Differential gene expression

- RNA-seq: `DESeq2`, `edgeR`, `scde`, …
- Microarray: `limma`

Gene regulation

- ChIP-seq: `csaw`, `DiffBind`
- Methylation arrays: `minfi`, `missMethyl`
- Gene set enrichment: `topGO`, `limma`

Working with called variants

- `VariantAnnotation`
- `VariantFiltering`

Flow cytometry

- `flowCore`

Data access

- `GEOquery / SRAdb`
- `TCGAbiolinks`
- `AnnotationHub / ExperimentHub`

Annotation resources

- Identifier, gene model, and sequence packages: `org.*`, `TxDb.*`, `BSgenome.*`
- Online queries: `biomaRt`, …

Visualization

- `Gviz`, `ComplexHeatmap`, `ggtree`, …

Many other packages!
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   ○ In-memory -- `MultiAssayExperiment`
   ○ On-disk (coming soon...)

2. Easy access to high-quality curated consortium-scale data
   ○ `AnnotationHub`
   ○ `ExperimentHub`
   ○ Emerging resources

3. Scalable performance
   ○ Large data representation -- `GenomicRanges, HDF5Array`
   ○ Core, cluster, cloud -- `BiocParallel`
   ○ Interactive and batch-iterative
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   - In-memory -- `MultiAssayExperiment`
   - On-disk

2. Easy access to high-quality curated consortium-scale data
   - `AnnotationHub`
   - `ExperimentHub`
   - Emerging resources

3. Scalable performance
   - Large data representation -- `GenomicRanges`, `HDF5Array`
   - Core, cluster, cloud -- `BiocParallel`
   - Interactive and batch-iterative

```r
> ##
> ## Multi-assay experiments, 'devel' only
> ##
> library(MultiAssayExperiment)
> ovarian
MultiAssayExperiment with 13 experiments
Containing an Elist class object of length 13
[8] mRNAArray: ExpressionSet - 18632 rows, 575 columns
[9] miRNAArray: ExpressionSet - 821 rows, 573 columns
[10] RPPAArray: ExpressionSet - 208 rows, 427 columns
[12] gistica: ExpressionSet - 24776 rows, 573 columns

To access slots use:
Elist() - to obtain the "Elist" of experiment instances
pData() - for the primary/phenotype "DataFrame"
sampleMap() - for the sample availability "DataFrame"
metadata() - for the metadata object of 'ANY' class
See also: subsetByAssay(), subsetByRow(), subsetByColumn()
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   - In-memory -- *MultiAssayExperiment*
   - On-disk

2. Easy access to high-quality curated consortium-scale data
   - *AnnotationHub*
   - *ExperimentHub*
   - Emerging resources

3. Scalable performance
   - Large data representation -- *GenomicRanges*, *HDF5Array*
   - Core, cluster, cloud -- *BiocParallel*
   - Interactive and batch-iterative

```r
> library(AnnotationHub)
> (hub <- AnnotationHub())
AnnotationHub with 43720 records
...
> hub["AH30903"]$title
[1] "E129-H3K4me1.narrowPeak.gz"
> hub["AH30903"]
...

> library(ExperimentHub)
> hub <- ExperimentHub()
> tcga <- hub["EH1"]
see ?GSE62944 and browseVignettes('GSE62944') for documentation
loading from cache '/home/mtmorgan/.ExperimentHub/1'
> table(tcga$CancerType)
BLC A BRCA COAD GBM HNSC KICH KIRC KIRP LAML LGG LIHC LUAD LUSC
273 1082 468 170 481 66 540 226 164 528 212 514 490
OV PRAD READ SKCM STAD THCA UCEC
344 423 164 373 146 506 536
> tcga[, tcga$CancerType == "OV"]
ExpressionSet (storageMode: lockedEnvironment)
...
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   - In-memory -- `MultiAssayExperiment`
   - On-disk

2. Easy access to high-quality curated consortium-scale data
   - `AnnotationHub`
   - `ExperimentHub`
   - Emerging resources

3. Scalable performance
   - Large data representation -- `GenomicRanges`, `HDF5Array`
   - Core, cluster, cloud -- `BiocParallel`
   - Interactive and batch-iterative

```r
> ## Emerging resources
> library(GenomicDataCommons)  # not yet public
> endpoints()
available endpoints:
   status, projects, cases, files, annotations, data, manifest,
   slicing, submission
> files()
class: files_list
cases: 10
names:
   6acecceb-7d71-4c50-bd76-781dffe13060,
   8361f2f1-8d30-444b-be70-6aa3e7557c8b,
   3414c8e2-21f4-41b6-ba1d-cfa7e83f30f7, ...
   dd7c8b01-6173-40a1-abc1-04d195d7cee7,
   43eb57cc-7e99-40f0-83ca-9a0208de0531
```
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   - In-memory -- MultiAssayExperiment
   - On-disk

2. Easy access to high-quality curated consortium-scale data
   - AnnotationHub
   - ExperimentHub
   - Emerging resources

3. Scalable performance
   - Large data representation -- GenomicRanges, HDF5Array
   - Core, cluster, cloud -- BiocParallel
   - Interactive and batch-iterative

```r
# Large data representation and processing
library(GenomicRanges)
gpos <- GPos(seqinfo(bfl)["chr14"])

# Parallel evaluation -- cores, clusters, clouds
library(BiocParallel)
library(RNAseqData.HNRNPC.bam.chr14)
bfl <- BamFileList(RNAseqData.HNRNPC.bam.chr14_BAMFILES)
cvg <- bplapply(bfl, coverage)
mcols(gpos) <- DataFrame(lapply(cvg, "[", "chr14"))
gpos
```

GPos object with 107349540 positions and 8 metadata columns:

```r
seqnames       pos strand | ERR127306 ERR127307
<Rle> <integer>  <Rle> |     <Rle>     <Rle>
[1]    chr14         1      * |         0         0
...
[107349540]    chr14 107349540      * |         0         0
```
Bioconductor and ITCR

Three directions to further enable cancer genomics research

1. Multi-assay data representations
   - In-memory -- MultiAssayExperiment
   - On-disk

2. Easy access to high-quality curated consortium-scale data
   - AnnotationHub
   - ExperimentHub
   - Emerging resources

3. Scalable performance
   - Large data representation -- GenomicRanges, HDF5Array
   - Core, cluster, cloud -- BiocParallel
   - Interactive and batch-iterative

```r
> ## Integration with existing 'containers'
> 
> library(SummarizedExperiment)
> gpos <- GPos(seqinfo(bfl)["chr14"])
> df <- DataFrame(lapply(cvg, "[[", "chr14"))
> (se <- SummarizedExperiment(list(cvg=df), rowData=gpos))
> class: RangedSummarizedExperiment
dim: 107349540 8
> metadata(0):
> assays(1): cvg
> rownames: NULL
> rowData names(0):
> colnames(8): ERR127306 ERR127307 ... ERR127304 ERR127305
> colData names(0):
```
Learn, use, and contribute to **Bioconductor**

**Learn**
- Package vignettes (e.g., `DESeq2`)
- **Training material**
- **Scientific literature**

**Use**
- **Package discovery**
- **Support site**

**Contribute**
- Developer resources
- Submission and technical review process
Acknowledgements

**Core team** (current & recent): Valerie Obenchain, Herve Pages, Dan Tenenbaum, Lori Shepherd, Marcel Ramos, Jim Hester, Jim Java, Brian Long, Sonali Arora, Nate Hayden, Paul Shannon, Marc Carlson

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

**Scientific advisory board**: Simon Tavare (CRUK), Paul Fliceck (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.