R / Bioconductor for 'Omics Analysis

Martin Morgan

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

31 October 2016
Introduction

Analysis and comprehension of high-throughput genomic data.

- Started 2002
- 1295 packages – developed by ‘us’ and user-contributed.

Well-used and respected.

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

https://bioconductor.org
https://support.bioconductor.org
Scope

Based on the $R$ programming language.

- Intrinsically statistical nature of data.
- Flexible analysis options for new or customized types of analysis.
- ‘Old-school’ scripts for reproducibility; modern graphical interfaces for easy use.

Domains of application.

- Sequencing: differential expression, ChIP-seq, variants, gene set enrichment, . . .
- Microarrays: methylation, expression, copy number, . . .
- Flow cytometry, proteomics, . . .
Install, learn, use, develop

Install
- Install R
- Install RStudio
- Bioconductor

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

Use
- Create bioinformatic solutions with Bioconductor
- Software, Annotation, and Experiment packages
- Amazon Machine Image
- Latest release announcement
- Support site

Develop
- Contribute to Bioconductor
- Developer resources
- Use Bio "devel"
- ‘Devel’ Software, Annotation and Experiment packages
- Package guidelines
- New package submission
- Build reports

---

1 https://bioconductor.org
2 https://support.bioconductor.org

R / Bioconductor for 'Omics Analysis
R: base packages

```r
x <- rnorm(1000)
y <- x + rnorm(1000, sd=.5)
df <- data.frame(X=x, Y=y)
fit <- lm(Y ~ X, df)
anova(fit)
```

```
## Analysis of Variance Table

## Response: Y

##              Df Sum Sq Mean Sq F value Pr(>F)
## X             1 925.99 925.99 3557.7 < 2.2e-16 ***
## Residuals    998 259.76  0.26
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
R: contributed packages

```r
library(ggplot2)
ggplot(df, aes(x=x, y=y)) +
geom_point() +
stat_smooth(method="lm")
```
Learn & use

- **biocViews**³
- Workflows⁴, F1000
- Landing pages⁵
  - Description
  - Installation
  - Documentation
- Vignettes⁶

³https://bioconductor.org/packages/release
⁴http://bioconductor.org/help/workflows
⁵e.g., https://bioconductor.org/packages/edgeR
⁶e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf
Learn & use

- **biocViews**
- Workflows, F1000
- Landing pages
  - Description
  - Installation
  - Documentation
- Vignettes

---

3. [https://bioconductor.org/packages/release](https://bioconductor.org/packages/release)


5. e.g., [https://bioconductor.org/packages/edgeR](https://bioconductor.org/packages/edgeR)

Bioconductor provides software to help analyze diverse high-throughput genomic data. Common workflows include:

### Basic Workflows

- **Sequence Analysis** Import fasta, fastq, BAM, gff, bed, wig, and other sequence formats. Trim, transform, align, and manipulate sequences. Perform quality assessment, ChIP-seq, differential expression, RNA-seq, and other workflows. Access the Sequence Read Archive.

- **Oligonucleotide Arrays** Import Affymetrix, Illumina, Nimblegen, Agilent, and other platforms. Perform quality assessment, normalization, differential expression, clustering, classification, gene set enrichment, genetical genomics and other workflows for expression, exon, copy number, SNP, methylation and other assays. Access GEO, ArrayExpress, Biomart, UCSC, and other community resources.

- **Annotation Resources** Introduction to using gene, pathway, gene ontology, homology annotations and the AnnotationHub. Access GO, KEGG, NCBI, Biomart, UCSC, vendor, and other sources.

- **Annotating Genomic Ranges** Represent common sequence data types (e.g., from BAM, gff, bed, and wig files) as genomic ranges for simple and advanced range-based queries.

- **Annotating Genomic Variants** Read and write VCF files. Identify structural location of variants and compute amino acid coding changes for non-synonymous variants. Use SIFT and PolyPhen database packages to predict consequence of amino acid coding changes.

- **Changing genomic coordinate systems with rtracklayer::liftOver** The liftOver facilities developed in conjunction with the UCSC browser track infrastructure are available for transforming data in GRanges formats. This is illustrated here with an image of the NHGRI GWAS catalog that is, as of Oct. 31 2014, distributed with coordinates defined by NCBI build hg38.

### Advanced Workflows

---

3. https://bioconductor.org/packages/release
5. e.g., https://bioconductor.org/packages/edgeR
Learn & use

- biocViews³
- Workflows⁴, F1000
- Landing pages⁵
  - Description
  - Installation
  - Documentation
- Vignettes⁶

---

edgeR

Empirical Analysis of Digital Gene Expression Data in R

Bioconductor version: Release (3.4)

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce counts, including ChIP-seq, SAGE and CAGE.

Author: Yunshun Chen <yuchen at wehi.edu.au>, Aaron Lun <alun at wehi.edu.au>, Davis McCarthy <dmccarthy at wehi.edu.au>, Xiaobei Zhou <xiaobei.zhou at uzh.ch>, Mark Robinson <mark.robinson at imls.uzh.ch>, Gordon Smyth <smyth at wehi.edu.au>

Maintainer: Yunshun Chen <yuchen at wehi.edu.au>, Aaron Lun <alun at wehi.edu.au>, Mark Robinson <mark.robinson at imls.uzh.ch>, Davis McCarthy <dmccarthy at wehi.edu.au>, Gordon Smyth <smyth at wehi.edu.au>

Citation (from within R, enter citation("edgeR")):


³https://bioconductor.org/packages/release
⁴http://bioconductor.org/help/workflows
⁵e.g., https://bioconductor.org/packages/edgeR
⁶e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf
Bioconductor

Learn & use
- biocViews\(^3\)
- Workflows\(^4\), F1000
- Landing pages\(^5\)
  - Description
  - Installation
  - Documentation
- Vignettes\(^6\)

\(^3\)https://bioconductor.org/packages/release
\(^4\)http://bioconductor.org/help/workflows
\(^5\)e.g., https://bioconductor.org/packages/edgeR
\(^6\)e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf
Learn & use

- biocViews
- Workflows, F1000
- Landing pages
  - Description
  - Installation
  - Documentation
- Vignettes

---

3 https://bioconductor.org/packages/release
4 http://bioconductor.org/help/workflows
5 e.g., https://bioconductor.org/packages/edgeR
6 e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/
   inst/doc/DESeq2.pdf
Learn & use

- **biocViews**[^3]
- **Workflows**, F1000
- **Landing pages**[^5]
  - Description
  - Installation
  - Documentation
- **Vignettes**[^6]

[^3]: https://bioconductor.org/packages/release
[^4]: http://bioconductor.org/help/workflows
[^5]: e.g., https://bioconductor.org/packages/edgeR
Bioconductor

Input: description of experimental design and summary of read counts overlapping regions of interest.

```r
pdata <- read.table("pdata.tab")  # Plain text files
assay <- read.table("assay.tab")

library(DESeq2)
.dds <- DESeqDataSetFromMatrix(assay, pdata, ~ cell + dex)
result(DESeq(dds))
```

Output: top table of differentially expressed genes, log fold change, adjusted $P$-value, etc.
A typical work flow: RNA-seq

Research question
- Designed experiment
- Gene-level differential expression
- RNA-seq data

Data processing steps
- Quality assessment.
- Alignment and summary to count table.
- Assessment of differential expression.
- Results placed in context, e.g., gene set enrichment.

http://bio.lundberg.gu.se/courses/vt13/rnaseq.html
Pre-processing, alignment

Pre-processing
- FASTQ file read quality assessment

Alignment & summary (traditional)
- Full alignment to BAM files, summarizing gene or transcript abundance, e.g., Bowtie / tophat / cufflinks; RSEM; Rsubread
- Summarize to gene-level count tables or estimates of abundance
- **Counts** are important: information about statistical uncertainty of estimate

Alignment & summary (contemporary)
- Approximate alignment directly to count tables of transcripts or genes, e.g., kallisto\(^7\), salmon\(^8\)

\(^7\)https://pachterlab.github.io/kallisto/
\(^8\)http://salmon.readthedocs.io/en/latest/salmon.html
Differential expression

- E.g., limma, edgeR, DESeq2

```r
library(tximport)
df <- read.table("pdata.tab")
## tx2gene: see tximport vignette
txi <- tximport(df$files, type="kallisto", tx2gene=tx2gene)

library(DESeq2)
dds <- DESeqDataSetFromMatrix(txi, samples, ~ cell + dex)
result(DESeq(dds))
```

- Account for library size differences (normalization)
- Apply sophisticated statistical model (negative binomial)
- Moderate test statistics (helps with small sample size)
- Performant, tested, correct.
Analysis & comprehension

Annotation packages
- Packages, e.g., org.*: symbol mapping; BSgenome.*: genome sequence; TxDB.*: gene models
- Query web services, e.g., biomaRt, uniprot.ws, KEGGREST, ...
- AnnotationHub: consortium and other large-scale results

Gene set & pathway analysis
- limma fry(); pathview; ReactomePA

Visualization
- Gviz, ComplexHeatmap, ...
Analysis & comprehension

Annotation packages
- Packages, e.g., org.*: symbol mapping; BSgenome.*: genome sequence; TxDb.*: gene models
- Query web services, e.g., biomaRt, uniprot.ws, KEGGREST, ...
- AnnotationHub: consortium and other large-scale results

Gene set & pathway analysis
- limma f r y (); pathview; ReactomePA

Visualization
- Gviz, ComplexHeatmap, ...

> grtrack <- GeneRegionTrack(geneModels, genome = gen, + chromosome = chr, name = "Gene Model")
> plotTracks(list(itrack, gtrack, atrack, grtrack))
Analysis & comprehension

Annotation packages

- Packages, e.g., `org.*`: symbol mapping; `BSgenome.*`: genome sequence; `TxDb.*`: gene models

- Query web services, e.g., `biomaRt`, `uniprot.ws`, `KEGGREST`, ...

- `AnnotationHub`: consortium and other large-scale results

Gene set & pathway analysis

- `limma`, `fry()`, `pathview`, `ReactomePA`

Visualization

- `Gviz`, `ComplexHeatmap`, ...

---

*R / Bioconductor* for 'Omic Analysis

A typical work flow
Exploratory 'omics

Gene differential expression
- RNA-seq – DESeq2, edgeR, limma voom()
- Microarray – limma
- Single-cell – scde

Gene regulation
- ChIP-seq – csaw, DiffBind
- Methylation arrays – missMethyl, minfi
- Gene sets and pathways – topGO, limma, ReactomePA

Variants
- SNPs – VariantAnnotation, VariantFiltering
- Copy number
- Structural – InteractionSet

Flow cytometry
- flowCore & 41 other packages

Proteomics
- mzR, xcms, and 90 other packages
Key classes

GenomicRanges

- Genomic coordinates to represent data (e.g., aligned reads) and annotations (e.g., genes, binding sites).
- `findOverlaps()` and friends.

SummarizedExperiment

- Coordinate ‘assay’ data with row (feature) and column (sample) information.
Key classes

**GenomicRanges**
- Genomic coordinates to represent data (e.g., aligned reads) and annotations (e.g., genes, binding sites).
- `findOverlaps()` and friends.

**SummarizedExperiment**
- Coordinate ‘assay’ data with row (feature) and column (sample) information.
Big data

GenomicFiles
- Management of file collections, e.g., VCF, BAM, BED.

BiocParallel
- Parallel evaluation on cores, clusters, clouds.

HDF5Array
- On-disk storage.
- Delayed evaluation.
- Incorporates into SummarizedExperiment.

Key strategies
- Efficient R code
- Restriction to data of interest
- Chunk-wise iteration through large data
From student to developer

A common transition

- Naive users become proficient while developing domain expertise that they share with others in their lab or more broadly
- Share via packages!

Resources

- Learning: course material, videos, workflows, vignettes.
- Using: vignettes, help pages, support site.
- Developing: Wicham’s *R Packages*\textsuperscript{9}, *Bioconductor* developer resources\textsuperscript{10}, bioc-devel mailing list

\textsuperscript{9}http://r-pkgs.had.co.nz/
\textsuperscript{10}http://bioconductor.org/developers/
Developer

Really easy!

- Use `devtools` to create() a package
- Add functions to the R directory
- Add documentation with `roxygen2`
- Add 'markdown' vignettes using `knitr`

Best practices

- `build()`, `check()`, `install()`
- Version control – github
- Unit tests, e.g., using `testthat`
- ‘Continuous integration’
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 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.