R / *Bioconductor* for 'Omics Analysis

Martin Morgan

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

30 January 2017

Introduction

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

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.

Scope

Based on the *R* programming language.

- Intrinsically statistical nature of data.
- Flexiblity 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, SNP, expression, copy number, ...
- Flow cytometry, proteomics, ...

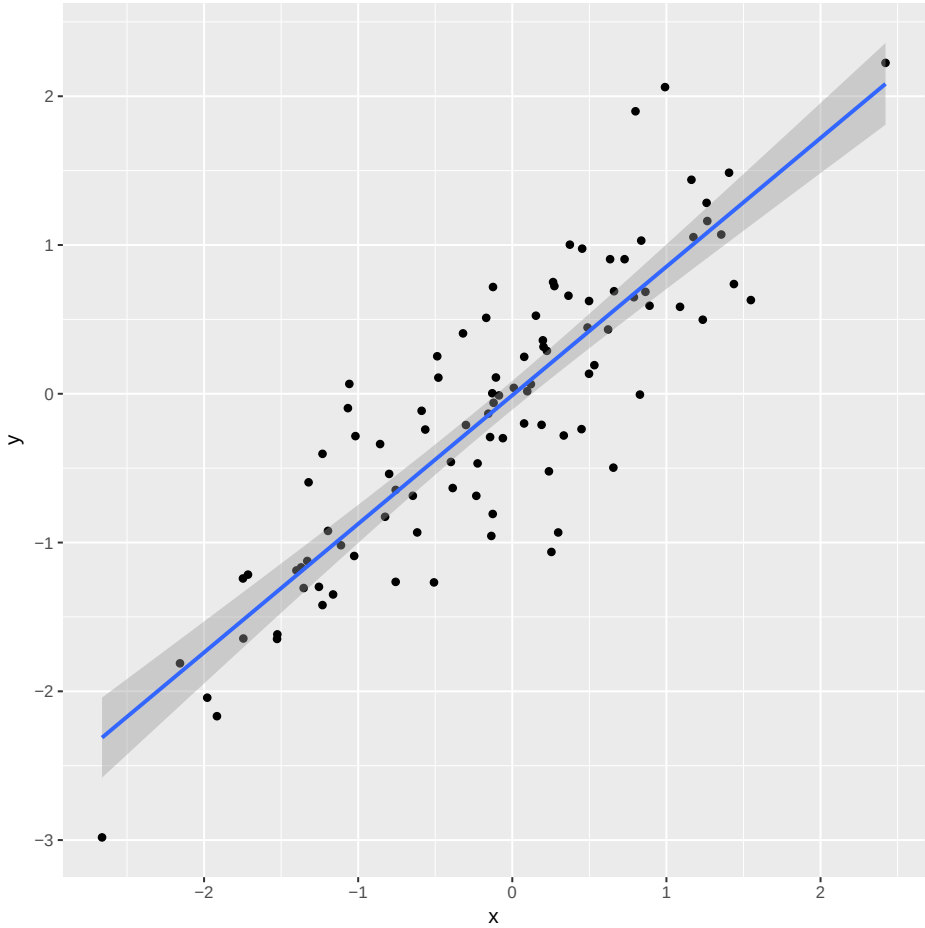
R: base packages

```
x <- rnorm(100)
y <- x + rnorm(100, 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 68.495  68.495  293.66 < 2.2e-16 ***
## Residuals 98 22.858   0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R: contributed packages

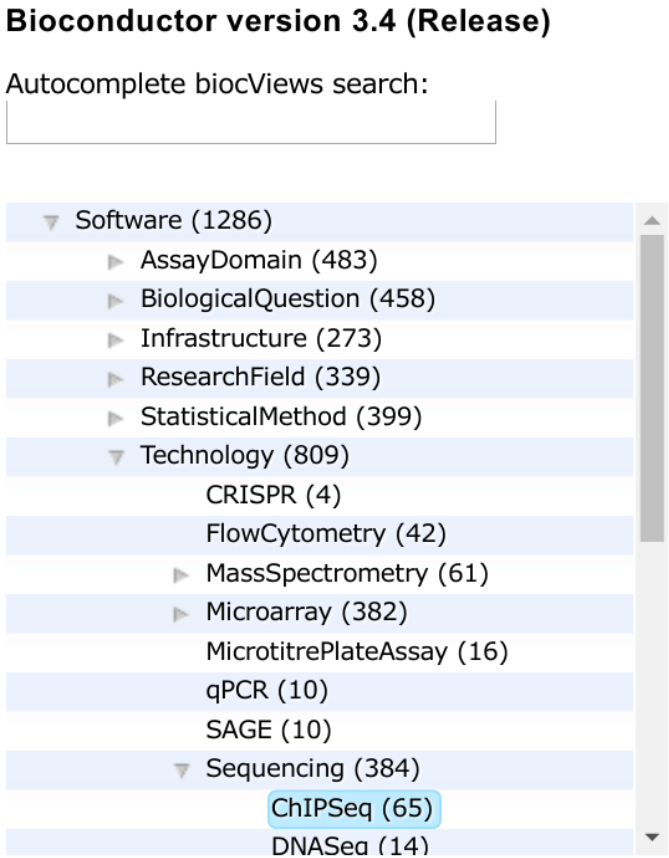
```
library(ggplot2)
ggplot(df, aes(x=x, y=y)) +
    geom_point() +
    stat_smooth(method="lm")
```



Bioconductor

Learn & use

- **biocViews**[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- Vignettes[3]
- Workflows[4], F1000 channel



[1] https://bioconductor.org/packages/release

[2] e.g., https://bioconductor.org/packages/edgeR

[3] e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

[4] http://bioconductor.org/help/workflows

Bioconductor

Learn & use

- **biocViews**[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- Vignettes[3]
- Workflows[4], F1000 channel

Packages found under ChIPSeq:

Show All entries Search table:

Package	Maintainer	Title
ALDEx2	Greg Gloor	Analysis Of Differential Abundance Taking Sample Variation Into Account
BaalChIP	Ines de Santiago	BaalChIP: Bayesian analysis of allele-specific transcription factor binding in cancer genomes
BayesPeak	Jonathan Cairns	Bayesian Analysis of ChIP-seq Data
ChIPComp	Li Chen	Quantitative comparison of multiple ChIP-seq datasets
ChIPpeakAnno	Lihua Julie Zhu, Jianhong Ou	Batch annotation of the peaks identified from either ChIP-seq, ChIP-chip experiments or any experiments resulted in large number of chromosome ranges
ChIPQC	Tom Carroll, Rory Stark	Quality metrics for ChIPseq data
ChIPseeker	Guangchuang Yu	ChIPseeker for ChIP peak Annotation, Comparison, and Visualization
chipseq	Bioconductor Package Maintainer	chipseq: A package for analyzing chipseq data
ChIPseqR	Peter Humburg	Identifying Protein Binding Sites in High-Throughput Sequencing Data
ChIPsim	Peter Humburg	Simulation of ChIP-seq experiments
ChIPXpress	George Wu	ChIPXpress: enhanced transcription factor target gene identification from ChIP-seq and ChIP-chip data using publicly available gene expression profiles
chromstaR	Aaron Taudt	Combinatorial and Differential Chromatin State Analysis for ChIP-Seq Data

[1] https://bioconductor.org/packages/release

[2] e.g., https://bioconductor.org/packages/edgeR

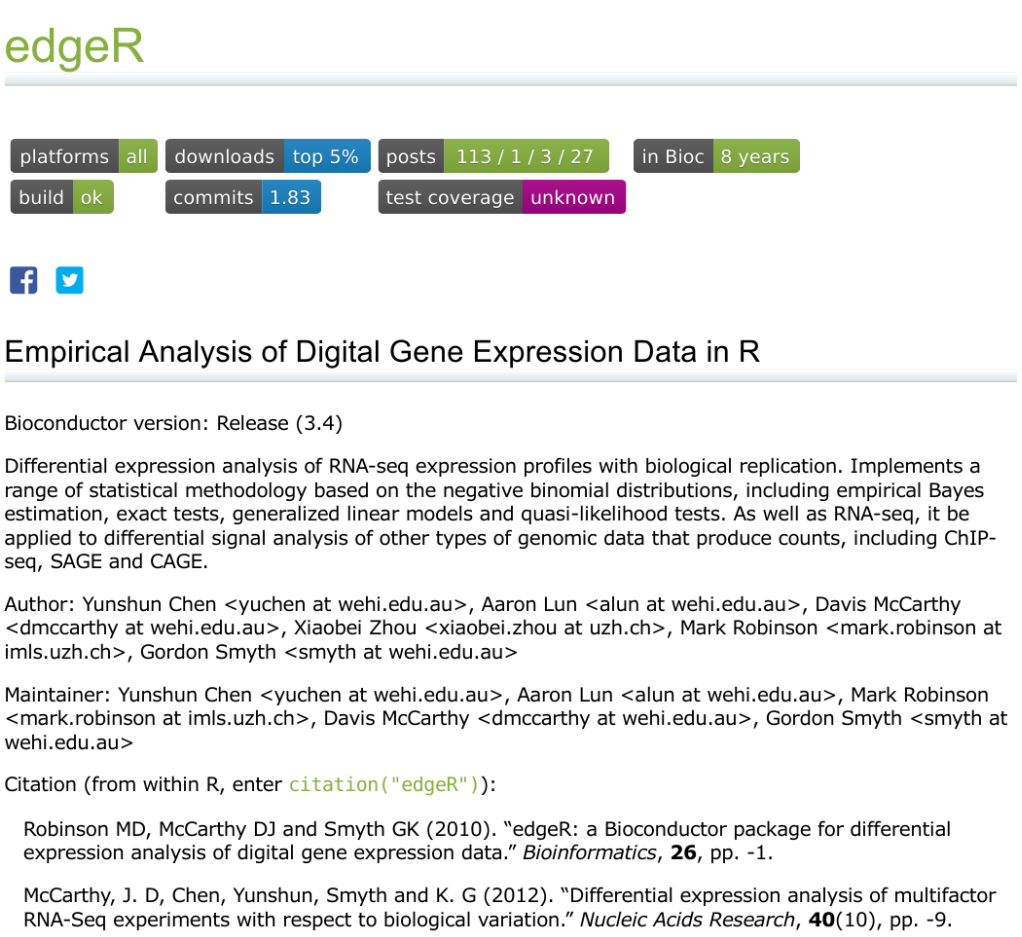
[3] e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

[4] http://bioconductor.org/help/workflows

Bioconductor

Learn & use

- biocViews[1]
- **Landing pages**[2]
  - Description
  - Installation
  - Documentation
- Vignettes[3]
- Workflows[4], F1000 channel



edgeR

platforms all | downloads top 5% | posts 113 / 1 / 3 / 27 | in Bioc 8 years
build ok | commits 1.83 | test coverage unknown

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")`):

Robinson MD, McCarthy DJ and Smyth GK (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics*, **26**, pp. -1.

McCarthy, J. D, Chen, Yunshun, Smyth and K. G (2012). "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation." *Nucleic Acids Research*, **40**(10), pp. -9.

[1] `https://bioconductor.org/packages/release`

[2] e.g., `https://bioconductor.org/packages/edgeR`

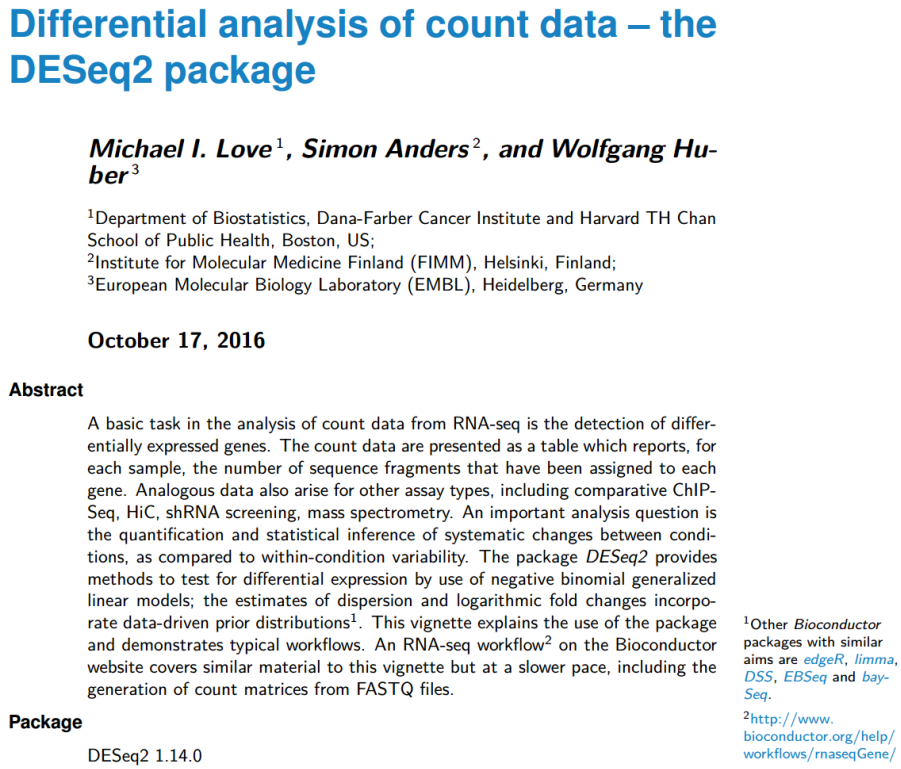
[3] e.g., `https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf`

[4] `http://bioconductor.org/help/workflows`

Bioconductor

Learn & use

- biocViews[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- **Vignettes**[3]
- Workflows[4], F1000 channel



Differential analysis of count data – the DESeq2 package

***Michael I. Love*[1], *Simon Anders*[2], *and Wolfgang Huber*[3]**

[1]Department of Biostatistics, Dana-Farber Cancer Institute and Harvard TH Chan School of Public Health, Boston, US;
[2]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland;
[3]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

October 17, 2016

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package *DESeq2* provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions[1]. This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow[2] on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files.

Package

DESeq2 1.14.0

[1]Other *Bioconductor* packages with similar aims are *edgeR*, *limma*, *DSS*, *EBSeq* and *baySeq*.

[2]http://www.bioconductor.org/help/workflows/rnaseqGene/

[1] `https://bioconductor.org/packages/release`

[2] e.g., `https://bioconductor.org/packages/edgeR`

[3] e.g., `https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf`

[4] `http://bioconductor.org/help/workflows`

Bioconductor

Learn & use

- biocViews[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- **Vignettes**[3]
- Workflows[4], F1000 channel

Contents

[1] https://bioconductor.org/packages/release

[2] e.g., https://bioconductor.org/packages/edgeR

[3] e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

[4] http://bioconductor.org/help/workflows

Bioconductor

Learn & use

- biocViews[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- **Vignettes**[3]
- Workflows[4], F1000 channel

2.2.1 Heatmap of the count matrix

To explore a count matrix, it is often instructive to look at it as a heatmap. Below we show how to produce such a heatmap for various transformations of the data.

```
library("pheatmap")
select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE)[1:20]

nt <- normTransform(dds) # defaults to log2(x+1)
log2.norm.counts <- assay(nt)[select,]
df <- as.data.frame(colData(dds)[,c("condition","type")])
pheatmap(log2.norm.counts, cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
pheatmap(assay(rld)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
pheatmap(assay(vsd)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df)
```

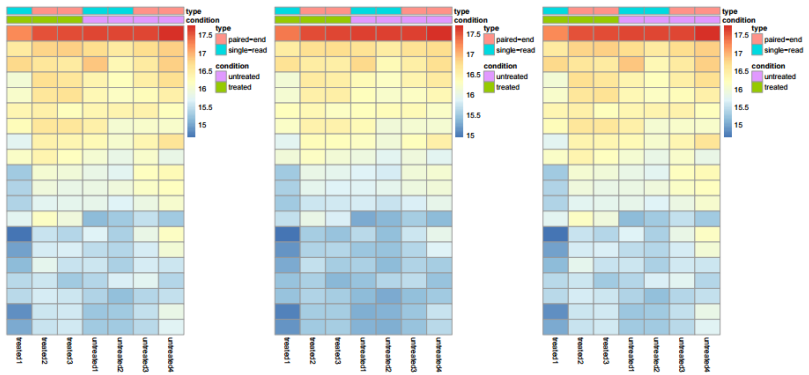


Figure 5: Heatmaps showing the expression data of the 20 most highly expressed genes. The data is of log2 normalized counts (left), from regularized log transformation (center) and from variance stabilizing transformation (right).

[1] https://bioconductor.org/packages/release

[2] e.g., https://bioconductor.org/packages/edgeR

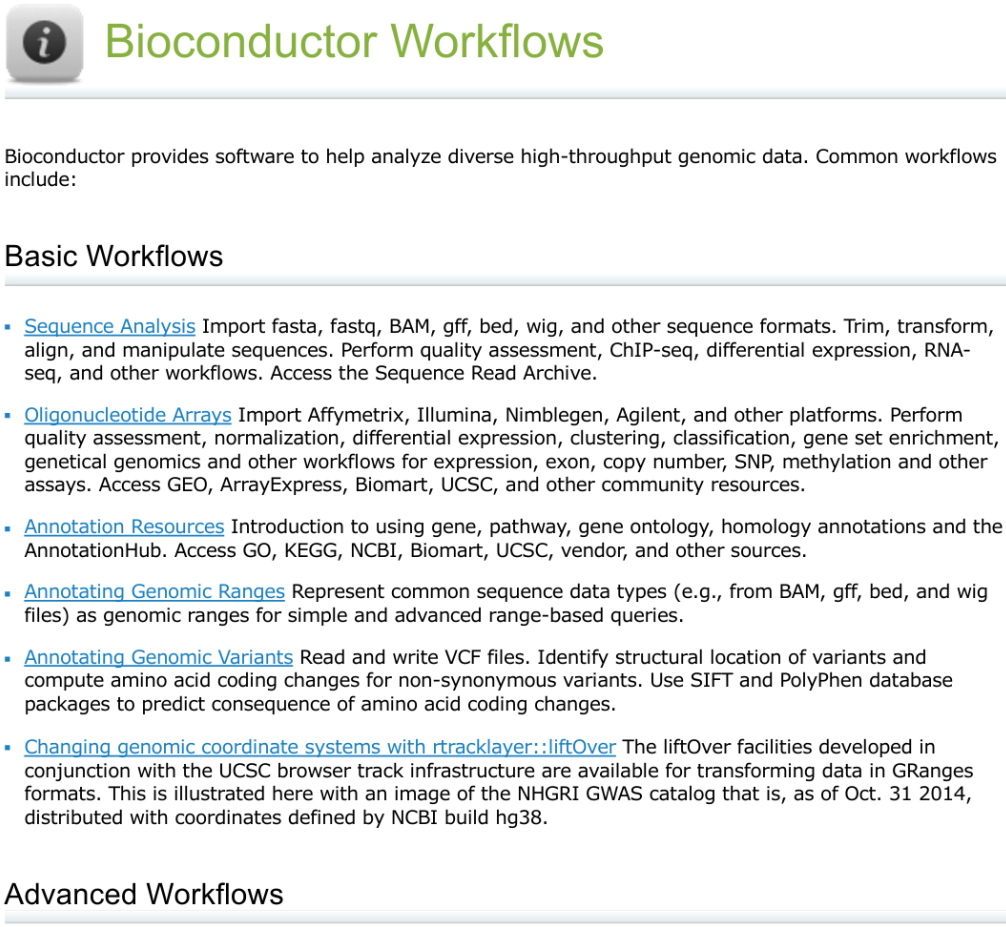
[3] e.g., https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

[4] http://bioconductor.org/help/workflows

Bioconductor

Learn & use

- biocViews[1]
- Landing pages[2]
  - Description
  - Installation
  - Documentation
- Vignettes[3]
- **Workflows**[4], F1000 channel



Bioconductor Workflows

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

[1] `https://bioconductor.org/packages/release`

[2] e.g., `https://bioconductor.org/packages/edgeR`

[3] e.g., `https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf`

[4] `http://bioconductor.org/help/workflows`

Bioconductor

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

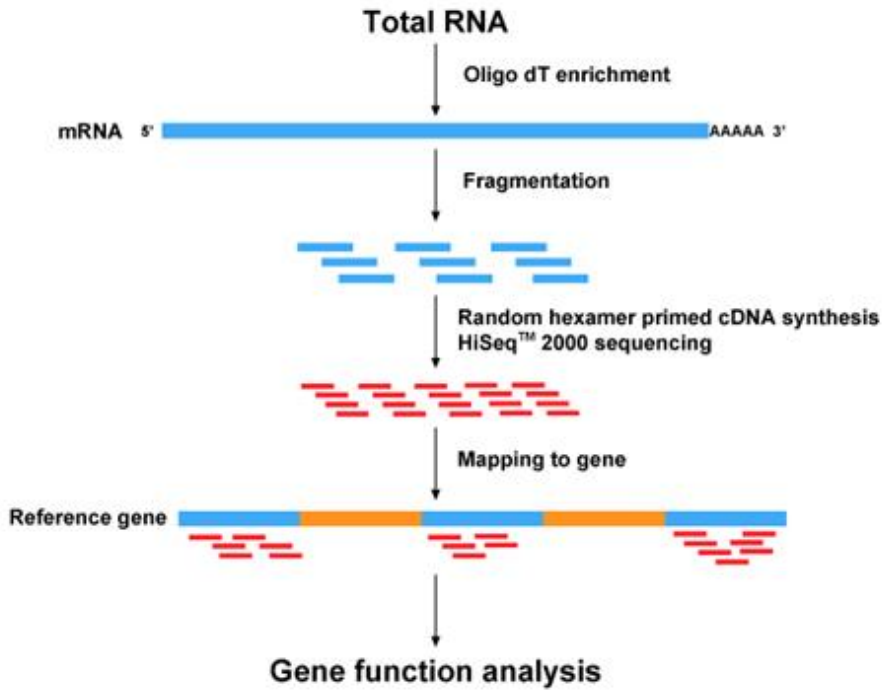
```
assay <- read.table("assay.tab")  # Plain text files
pdata <- read.table("pdata.tab")

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

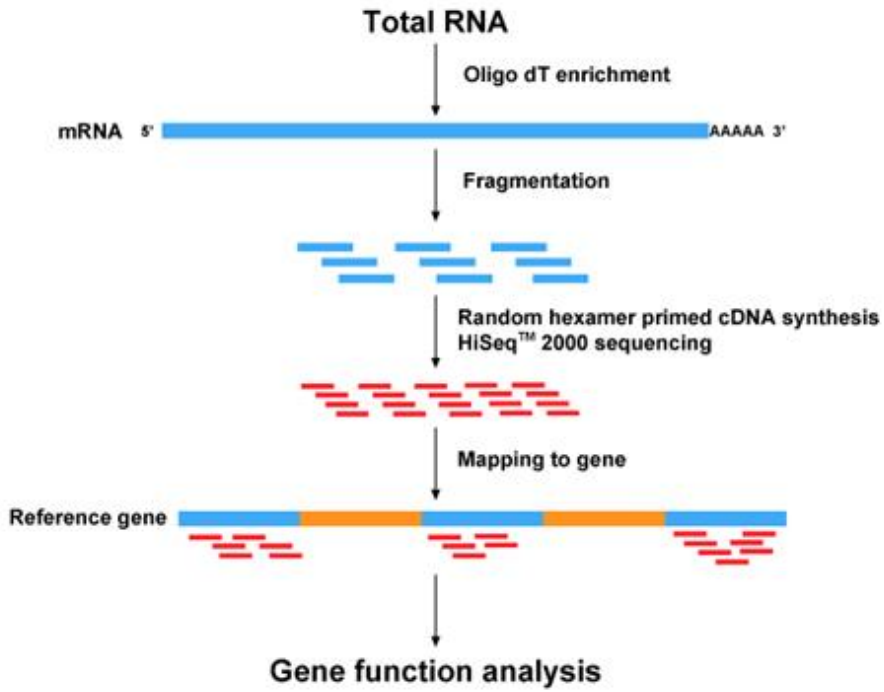
1. Experimental design
2. Wet-lab
3. **Sequencing; QC** – FASTQ
4. Alignment – BAM
5. Data reduction – count tables
6. Statistical analysis
7. Comprehension



`http://bio.lundberg.gu.se/courses/vt13/rnaseq.html`

A typical work flow: RNA-seq

1. Experimental design
2. Wet-lab
3. Sequencing; QC – FASTQ
4. **Alignment** – BAM
5. **Data reduction** – count tables
6. Statistical analysis
7. Comprehension



`http://bio.lundberg.gu.se/courses/vt13/rnaseq.html`

A typical work flow: RNA-seq

1. Experimental design
2. Wet-lab
3. Sequencing; QC – FASTQ
4. **Psuedo-alignment** – count tables
5. Statistical analysis
6. Comprehension

kallisto[5], salmon[6], ...

- *Very* fast
- Very memory efficient
- Good enough for many applications

Bioconductor

- *tximport*
- *limma* `voom()`

[5] https://pachterlab.github.io/kallisto/
[6] http://salmon.readthedocs.io/

A typical work flow: RNA-seq

1. Experimental design
2. Wet-lab
3. Sequencing; QC – FASTQ
4. Psuedo-alignment – count tables
5. **Statistical analysis**
6. **Comprehension**

- *DESeq2*, *edgeR*
- Gene set / pathway analysis
- Annotation & visualization

Differential expression

limma, *edgeR*, *DESeq2*

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

- Batch effects (e.g., surrogate variable analysis)
- Library size differences (robust normalization)
- Appropriate statistical model (negative binomial)
- Moderated, data-driven parameter estimates (shared design; small sample size)
- Multiple testing (independent hypothesis weighting)

Interoperability & reproducibility: 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.

```
> gr = exons(TxDb.Hsapiens.UCSC.hg19.knownGene); gr
GRanges with 289969 ranges and 1 metadata column:
            seqnames              ranges strand |   exon_id
               <Rle>           <IRanges>  <Rle> | <integer>
        [1]     chr1       [11874, 12227]      + |         1
        [2]     chr1       [12595, 12721]      + |         2
        [3]     chr1       [12613, 12721]      + |         3
        ...      ...                  ...    ... ...       ...
   [289967]     chrY [59358329, 59359508]      - |    277748
   [289968]     chrY [59360007, 59360115]      - |    277749
   [289969]     chrY [59360501, 59360854]      - |    277750
   ---
   seqinfo: 93 sequences (1 circular) from hg19 genome
```

GRanges
```
length(gr); gr[1:5]
seqnames(gr)
start(gr)
end(gr)
width(gr)
strand(gr)
```

DataFrame
```
mcols(gr)
gr$exon_id
```

Seqinfo
```
seqlevels(gr)
seqlengths(gr)
genome(gr)
```

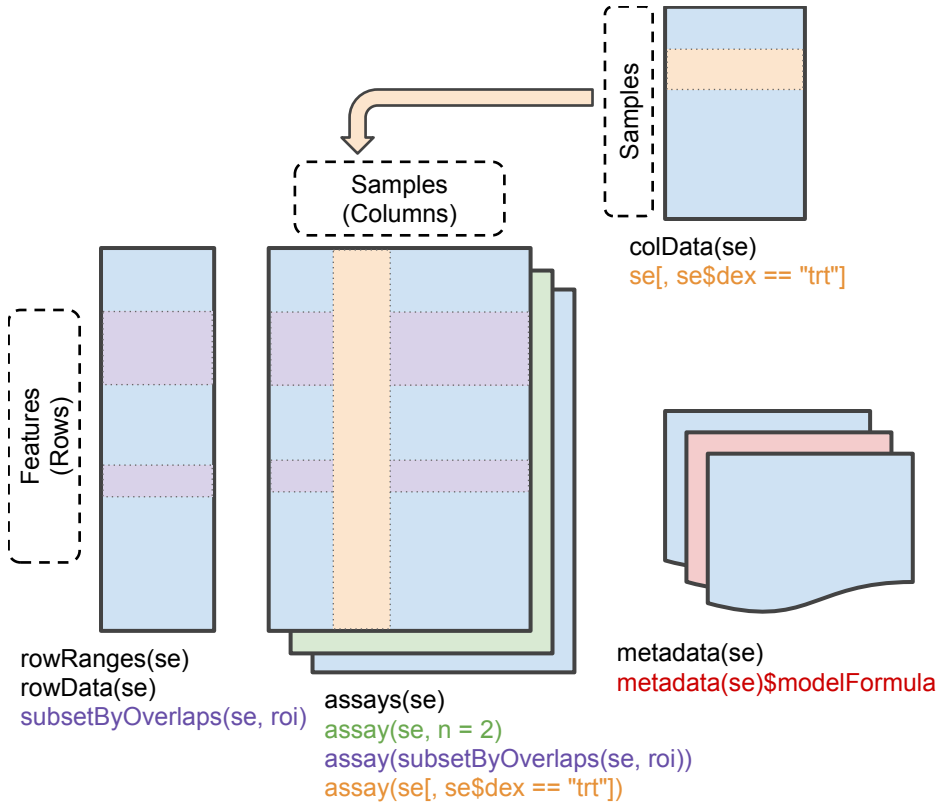
Interoperability & reproducibility: classes

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.



Classic, tidy, rich: RNA-seq count data

Classic

- Sample x (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

```
## Manipulate, e.g., mean expression of each gene

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

## Visualize

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

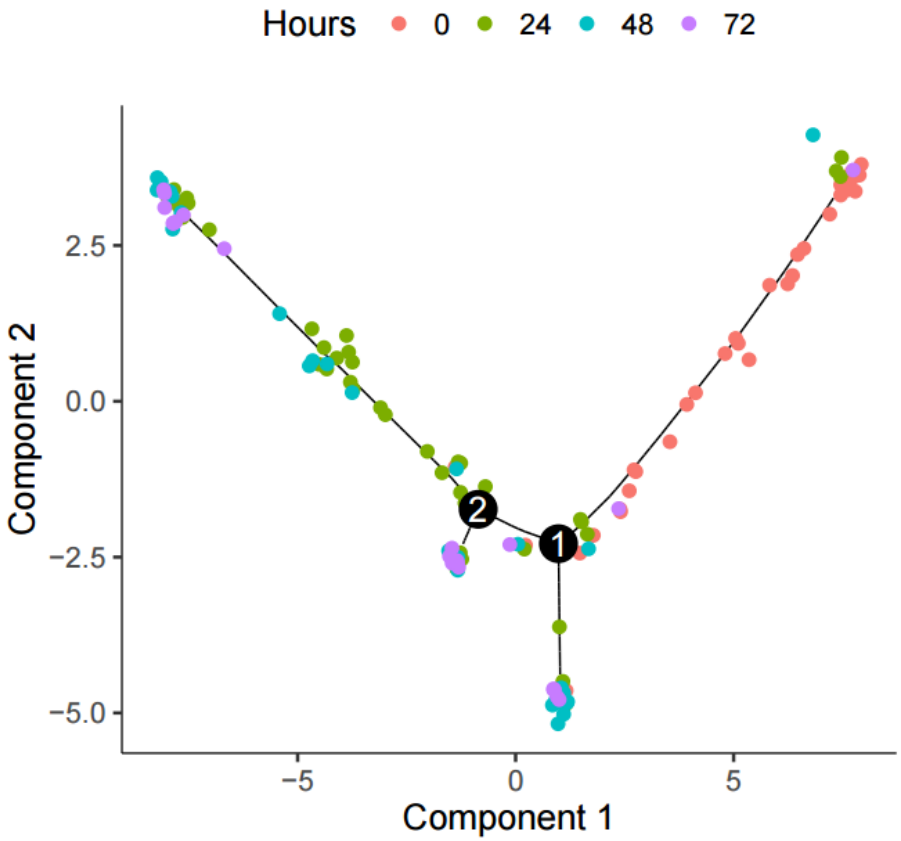
Programming contract

- Classic, tidy: limited
- Rich: strict

Lessons learned / best practices

- Considerable value in semantically rich structures
- Endomorphism, simple vocabulary, consistent paradigm aid use

Single-cell analysis



- Large & sparse
  - Outlier detection
  - Zero-inflated models
  - E.g., *MAST*
- Challenging
  - E.g., developmental trajectories

Trapnel et al.[5]

[5] `http://bioconductor.org/packages/monocle`

Comprehension

Gene set & pathway analysis

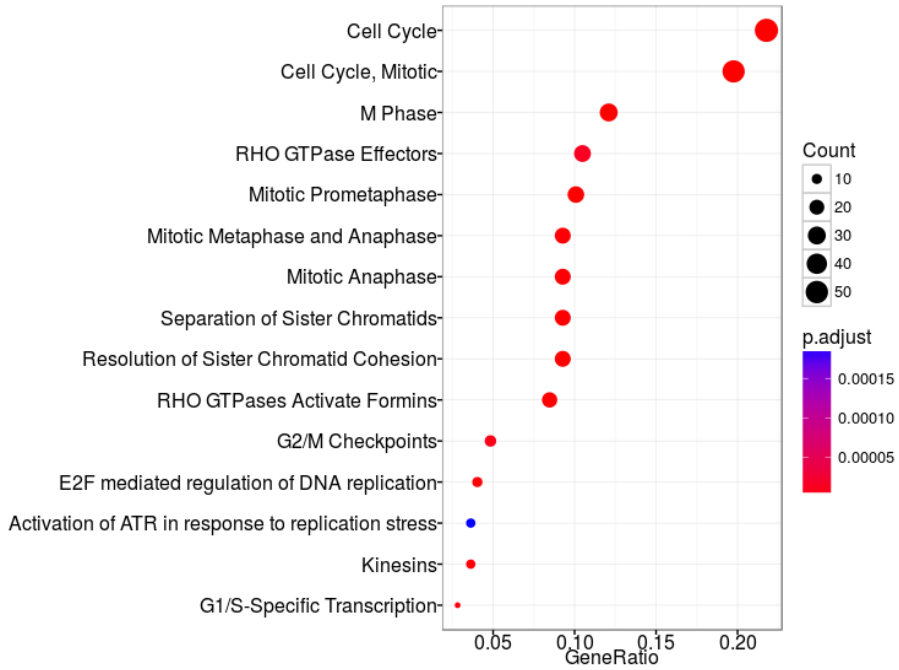
- *limma* `fry()`; *pathview*; **ReactomePA**

Visualization

- *Gviz*, *ComplexHeatmap*, ...

Communication

- Reports; interactive apps
- Statistical nuance, especially uncertainty, multiple testing



Comprehension

Gene set & pathway analysis

- *limma* `fry()`; *pathview*; *ReactomePA*

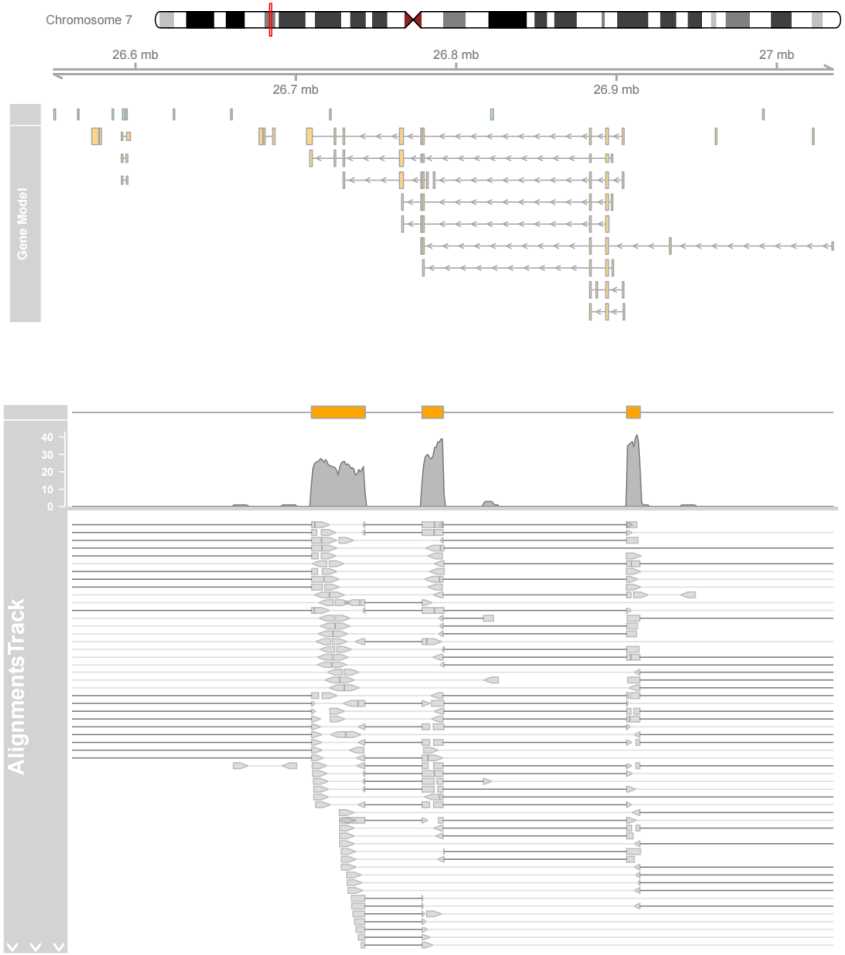
Visualization

- **Gviz**, *ComplexHeatmap*, . . .

Communication

- Reports; interactive apps
- Statistical nuance, especially uncertainty, multiple testing

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



Comprehension

Gene set & pathway analysis

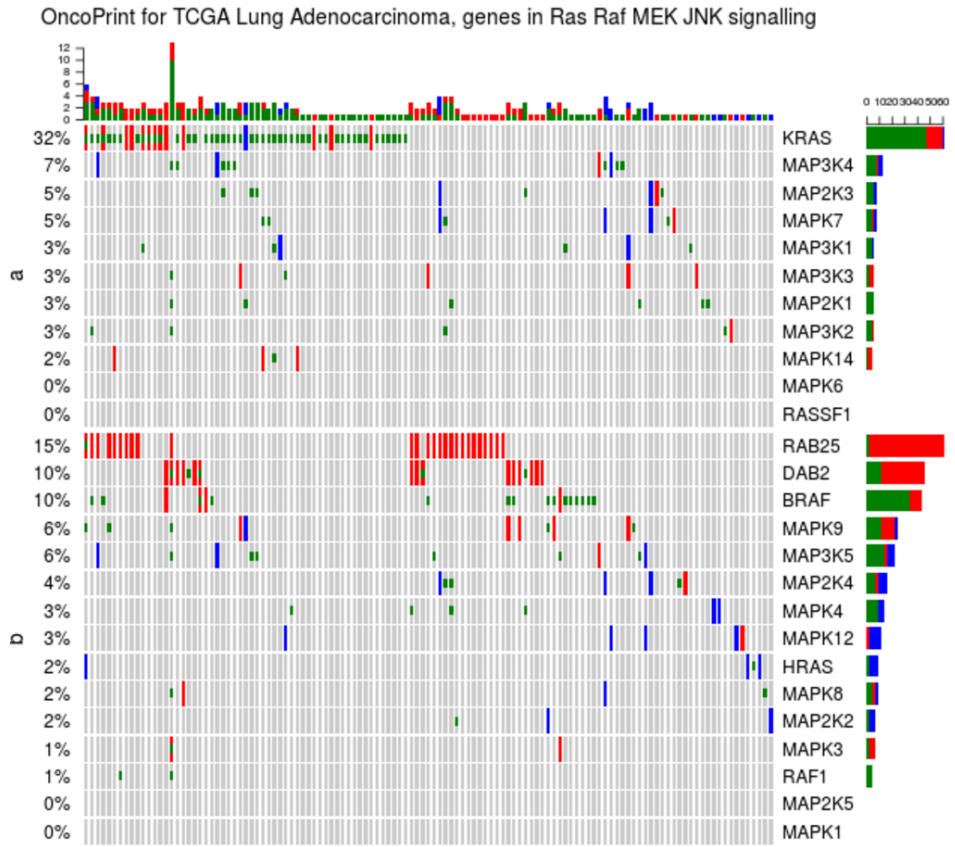
- *limma* `fry()`; *pathview*; *ReactomePA*

Visualization

- *Gviz*, **ComplexHeatmap**, ...

Communication

- Reports; interactive apps
- Statistical nuance, especially uncertainty, multiple testing



Comprehension

Gene set & pathway analysis

- *limma* `fry()`; *pathview*; *ReactomePA*

Visualization

- *Gviz*, *ComplexHeatmap*, ...

Communication

- Reports; interactive apps
- **Statistical nuance**, especially uncertainty, multiple testing

Multi-'omic integration

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*

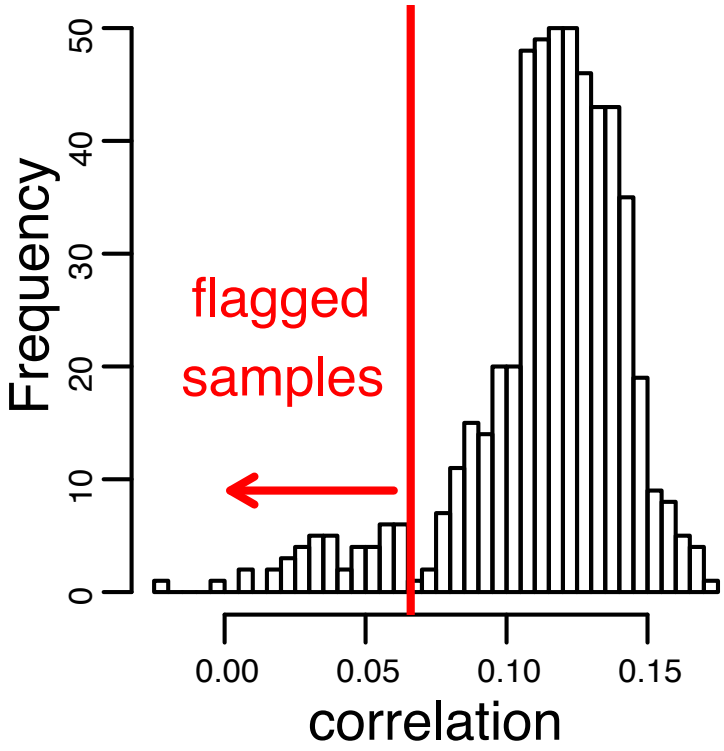
Flow cytometry

- *flowCore* & 41 other packages

Proteomics

- *mzR*, *xcms*, and 90 other packages

Multi-'omic integration



MultiAssayExperiment

- Easily manage multiple assays on overlapping samples

ExperimentHub

- Curated, summarized, large-scale experiment data (e.g., GEO RNA-Seq; HMP, TCGA) for incorporation in local analysis

Big data

Key strategies

- Efficient *R* code
- Restriction to data of interest
- Chunk-wise iteration through large data

GenomicFiles

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

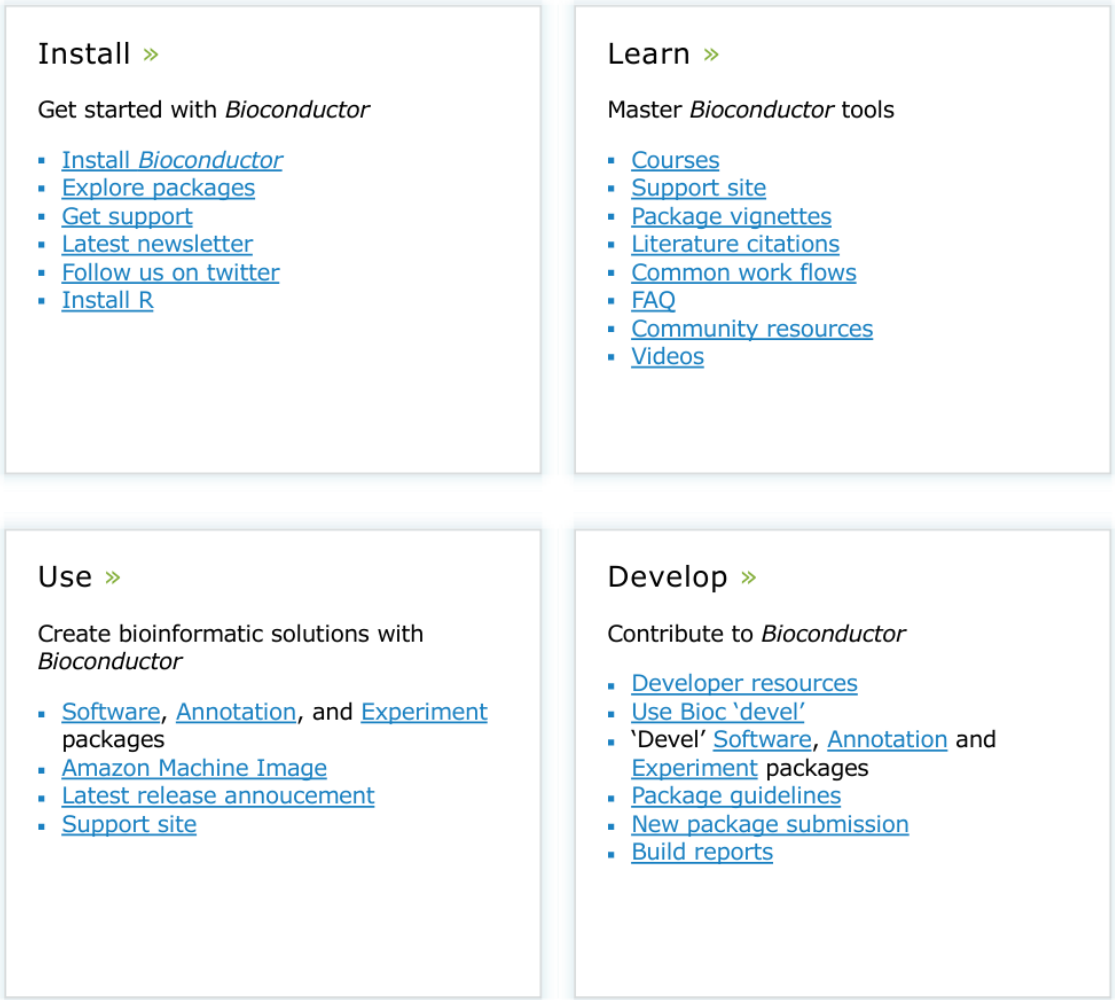
BiocParallel

- Parallel evaluation on cores, clusters, clouds.

HDF5Array

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

Install, learn, use, develop



Install[6]

- *R*, *RStudio*, *Bioconductor*

Learn

- Courses, vignettes, workflows

Use

- Vignettes, manuals, support site[7]

Develop

[6] `https://bioconductor.org`
[7] `https://support.bioconductor.org`

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
- Really easy!

Best practices

- *devtools* `create()`, `build()`, `check()`, `install()`
- Version control – github
- Unit tests, e.g., using *testthat*
- 'Continuous integration'

Core team jobs!

- Scientific Programmer / Analyst – core packages; *R* and *C* algorithms.
- Senior Programmer / Analyst – system / cloud management.
- `https://support.bioconductor.org/p/91548/`

Acknowledgments

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