Bioconductor for high-throughput genetic data

分析处理高通量基因数据: Bioconductor

Martin Morgan
martin.morgan@roswellpark.org
29 May 2016
Bioconductor

Statistical analysis and comprehension of high-throughput genomic data

- Sequence (RNA-seq, ChIP-seq, Variants, ...)
- Microarrays (expression, methylation, SNP, copy number, ...)
- Flow cytometry
- Proteomics
- Image analysis
- ...

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

1211 Software Packages  1211个软件包

- All packages  
  [https://bioconductor.org/packages](https://bioconductor.org/packages)
- Example package:  GenomicRanges

Features  特征

- Vignettes and support site for learning
- Stable ‘release’ branch for users; reproducible research
- ‘Devel’ branch for new features & packages
- Classes for inter-operability between packages

source(“https://bioconductor.org/biocLite.R”)
biocLite(“GenomicRanges”)

[https://bioconductor.org/packages/release/BiocViews.html](https://bioconductor.org/packages/release/BiocViews.html)
Bioconductor

1211 Software Packages  1211个软件包

- All packages
  https://bioconductor.org/packages
- Example package: GenomicRanges

Features  特征

- Vignettes and support site for learning
- Stable ‘release’ branch for users; reproducible research
- ‘Devel’ branch for new features & packages
- Classes for inter-operability between packages

source(“https://bioconductor.org/biocLite.R”)
biocLite(“GenomicRanges”)
Learn

Vignettes 文档说明
- Each package!

Courses 相关课程

Support site

ComplexHeatmap

Making Complex Heatmaps

Bioconductor version: Release (3.3)

Complex heatmaps are efficient to visualize associations between different sources of data sets and reveal potential structures. Here the ComplexHeatmap package provides a highly flexible way to arrange multiple heatmaps and supports self-defined annotation graphics.

Author: Zuguang Gu
Maintainer: Zuguang Gu <z.gu at dkfz.de>
Citation (from within R, enter citation("ComplexHeatmap"));


Installation

To install this package, start R and enter:

```r
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
bioCLite("ComplexHeatmap")
```
Differential analysis of count data – the DESeq2 package

Michael I. Love1, Simon Anders2,3, Wolfgang Huber3

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

May 15, 2016

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differently 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 distributions1. This vignette explains the use of the package and demonstrates

Differential analysis of count data – the DESeq2 package

DESeq2 version: 1.12

1 Standard workflow

1.1 Quick start

Here we show the most basic steps for a differential expression analysis. These steps require you have a "BatchSummarizedExperiment" object `se` which contains the counts and information about samples. The design indicates that we want to measure the effect of condition, controlling for batch differences. The two factor variables `batch` and `condition` should be columns of `colData(se).

```r
dds <- DESeqDataSet(se, design = ~ batch * condition)
`dds` <- DESeq(dds)
res <- results(dds, contrast = c("condition","trt","con"))
```

If you have a count matrix and sample information table, the first line would use `DESeqDataSetFromMatrix` instead of `DESeqDataSet`, as shown in Section 1.3.3.
Learn

Vignettes  文档说明
- Each package!

Courses  相关课程

Support site  支持网站

### Courses & Conferences

Bioconductor provides training in computational and statistical methods for the analysis of genomic data. You are welcome to use material from previous courses. However, you may not include these in separately published works (articles, books, websites). When using all or parts of the Bioconductor course materials (slides, vignettes, scripts) please cite the authors and refer your audience to the Bioconductor website.

*Upcoming events* are advertised 6 to 8 weeks in advance.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Title</th>
<th>Course</th>
<th>Materials</th>
<th>Date</th>
<th>BioC/R Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>A1: R Intro, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>A2: Input and Manipulation, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>A3: Statistical Analysis, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>A4: Visualization, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>B1: Bioconductor Intro, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>B3: RNASeq Workflow, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>B4: Next Steps, Martin Morgan, Lori Shepherd</td>
<td>BioIntroRCP</td>
<td>HTML R Rmd</td>
<td>2016-05-16</td>
<td>3.3/3.3</td>
</tr>
<tr>
<td>Introduction</td>
<td>Introduction to High Throughput DNA Sequence Data Analysis Using R / Bioconductor, Martin Morgan</td>
<td>FNAR 2016</td>
<td>Rmd R github</td>
<td>2016-04-08</td>
<td>3.2/3.2</td>
</tr>
</tbody>
</table>
Learn

Vignettes  文档说明  
- Each package!

Courses  相关课程

Support site  支持网站

### Bioconductor

**Learn**

Vignettes  文档说明  
- Each package!

Courses  相关课程

Support site  支持网站
## Learn

### Essential Software 基础软件包

- **Biostrings**
- **GenomicRanges**
- **SummarizedExperiment**

### Annotation Resources  注释资源

- **Packages**
  - org.*
  - TxDb.*
  - BSgenome.*

- **On-line**
  - biomaRt
  - AnnotationHub

```r
> 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

# seqlengths:
#    chr1    chr2    ...    chrUn_g1000249
# 249250621 243199373 ... 38502
```

- 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)
Learn

Essential Software 基础软件包

- **Biostrings**
- **GenomicRanges**
- **SummarizedExperiment**

Annotation Resources 注释资源

- Packages
  - org.*
  - TxDb.*
  - BSgenome.*
- On-line
  - [biomaRt](https://bioconductor.org/biocgua/)
  - [AnnotationHub](https://annotationhub.org/)

```r
rowRanges(se)
rowData(se)
subsetByOverlaps(se, roi)
as assays(se)
  assay(se, n = 2)
  assay(subsetByOverlaps(se, roi))
  assay(subset(se, select = dex == "trt"))
```
Learn

Essential Software

- Biostrings
- GenomicRanges
- SummarizedExperiment

Annotation Resources

- Packages
  - Identifier mapping
  - Gene models
  - Genome sequence
- On-line
  - biomaRt
  - AnnotationHub
Use

```r
> dds = DESeq(DESeqDataSet(airway, ~ cell + dex))
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing
> plotMA(dds, alpha=.01)
```

---

```r
1 2 title: "Bioinformatics Report"
3 author: "Martin Morgan"
4 date: "March 12, 2016"
5 output: html_document
6
7
8 # Introduction
9
10 It is very easy to create advanced reports in a fully reproducible way. Reports can be distributed as PDF, HTML, ...
11
12 Reports can include figures, tables, analytic results, interactive applets, etc.
13
14 """"{r, warning=FALSE, message=FALSE, echo=FALSE}
15 ## R code evaluated when report produced
16 library(DESeq2)
17 library(airway)  # example data set
18 data(airway)
19 dds = DESeq(DESeqDataSet(airway, ~ cell + dex))
20 plotMA(dds, alpha=.01)
```

---

```
mean of normalized counts
```

```
log fold change
```
Use

Gene expression 基因表达

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

Gene regulation 基因调控

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

Variants 变异

- VariantAnnotation
- VariantFiltering

Flow cytometry 流式细胞仪

- flowCore

Data access 数据访问

- biomaRt
- GEOquery / SRAdb
- TCGAbioliinks
- AnnotationHub / ExperimentHub

Visualization 可视化

- Gviz, ComplexHeatmap, ggbio, ggtree, …

Many other packages!

> plotTracks(list(itchack, gtrack, atrack, grtrack))
Develop

Create a package

- Use existing classes -- **DNAStringSet**, **GRanges**, **SummarizedExperiment**
- Use existing packages -- **rtracklayer**, **Rsamtools**, **Biostrings**, …
- Full vignette and examples
- [Unit tests](#) and other best practices

Submit to **Bioconductor**

- Technical review
- Long-term [support](#) & maintenance
- Introduced to ‘devel’ branch, new release every April and October
Acknowledgements

Core team (current & recent)

- Valerie Obenchain, Herve Pages, Dan Tenenbaum, Lori Shepherd, Marcel Ramos, Brian Long, Jim Hester, Jim Java, 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 Flicek (EMBL/EBI), Simon Urbanek (AT&T), Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Raphael Irizzary, 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.
Resources

Web sites

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

Events

- Annual Conference, Stanford, CA, USA, 24-26 June
- Bioconductor Asia-Pacific Conference, Brisbane, Australia 3-4 November
- Bioconductor European Developer Workshop, Basel, Switzerland, 6-7 December