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
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

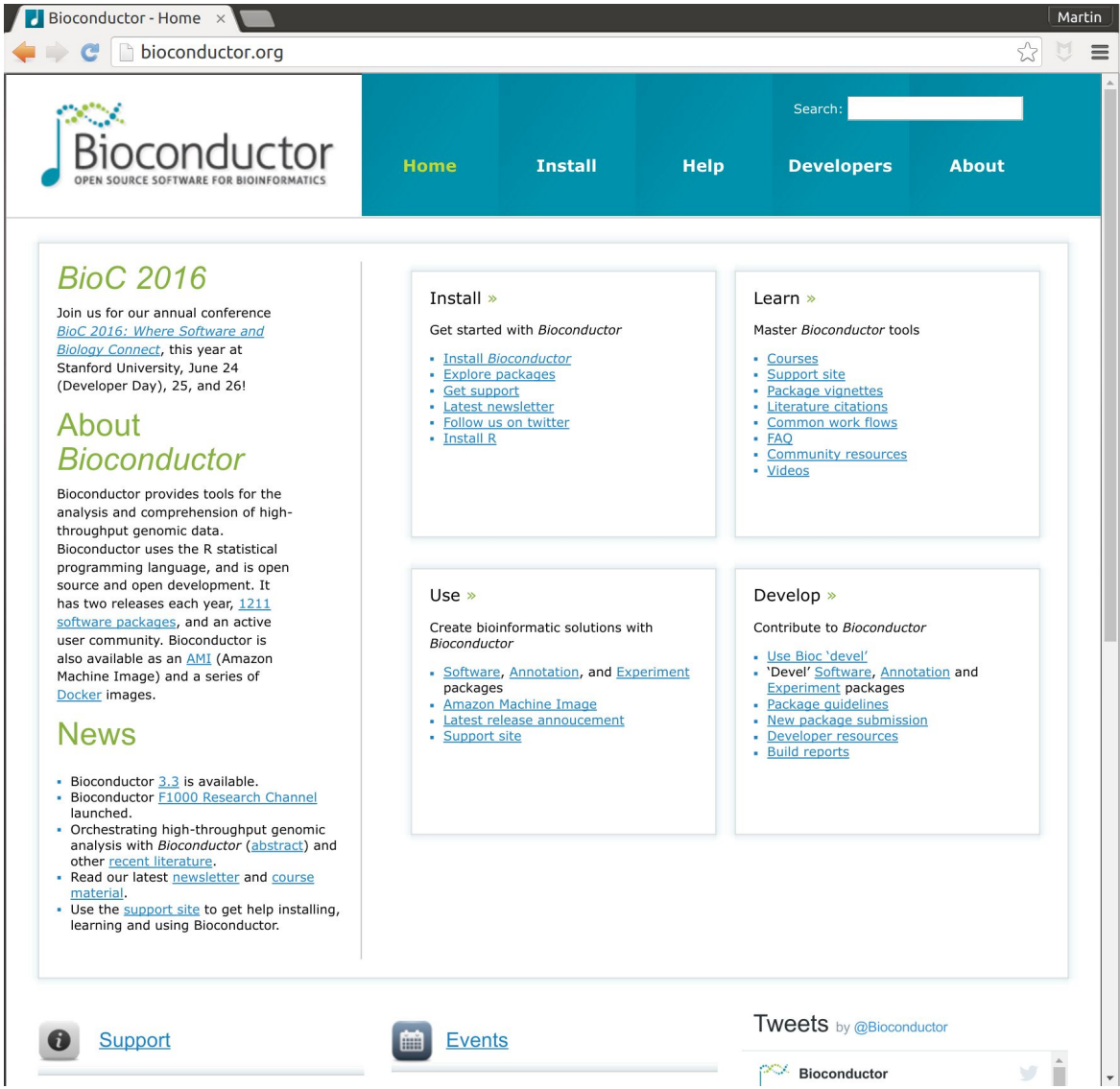
ROSWELL PARK CANCER INSTITUTE

Bioconductor

Statistical analysis and comprehension of high-throughput genomic data
统计学方法分析理解高通量基因 组数据

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

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



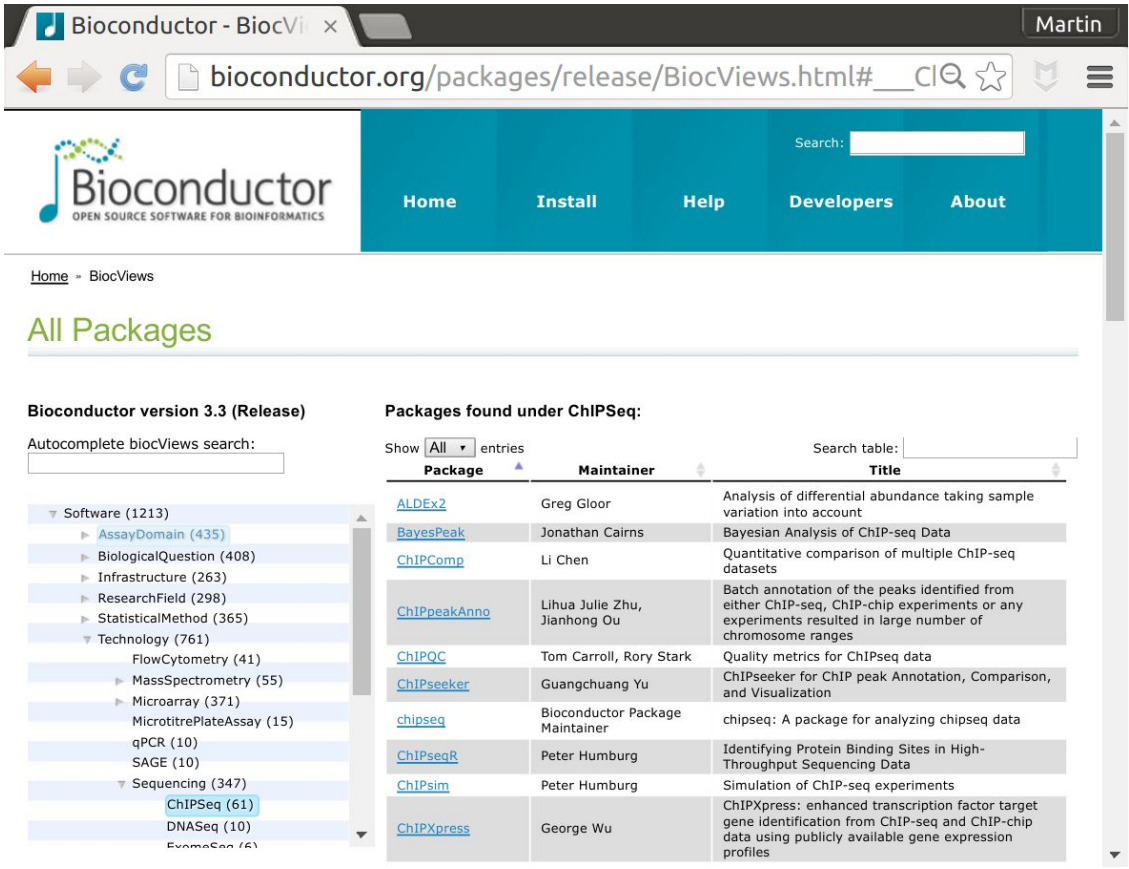
Bioconductor

1211 Software Packages　1211个软件包

- **All packages** https://bioconductor.org/packages
- 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”)
```

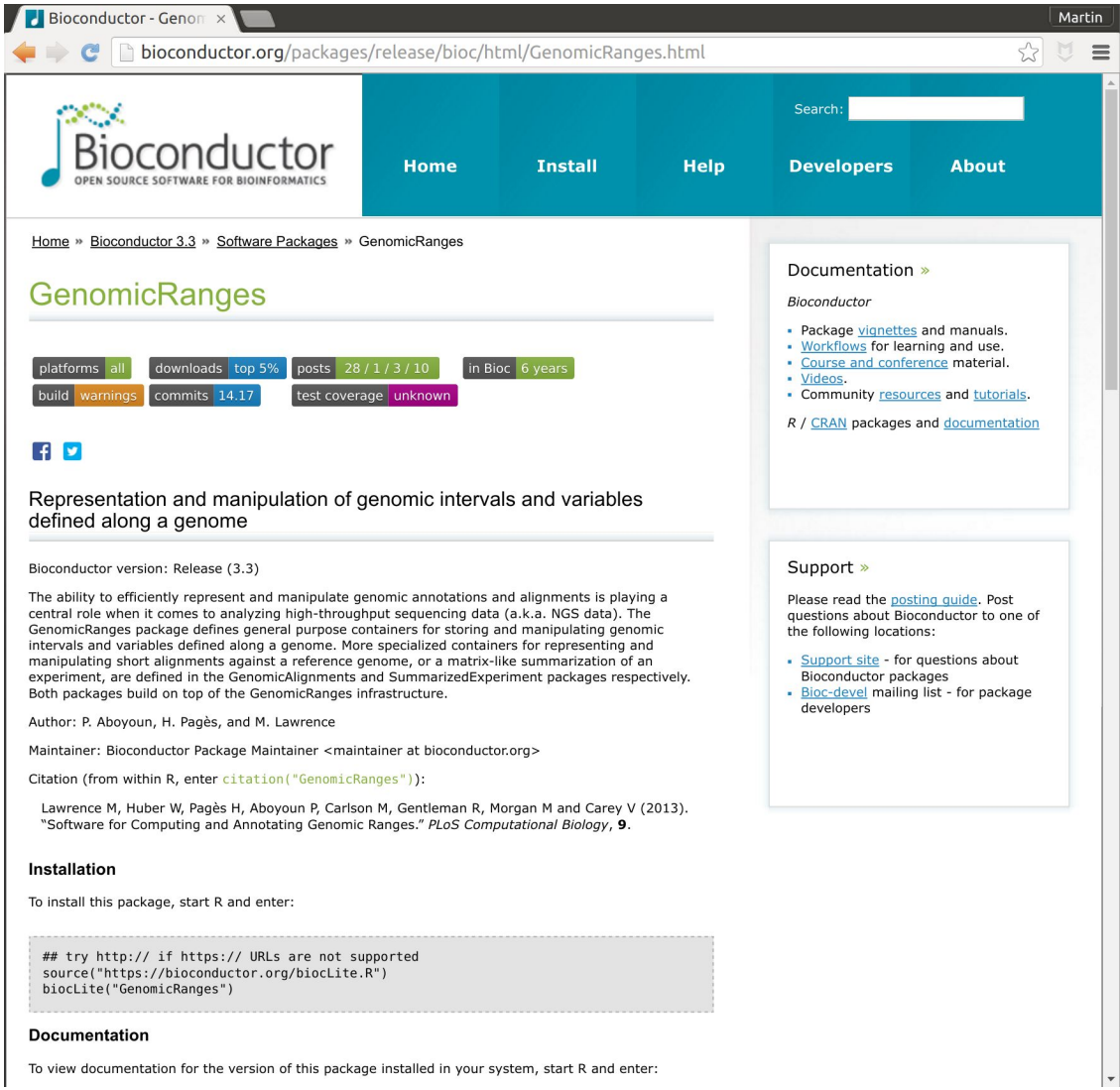
Bioconductor

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”)
```

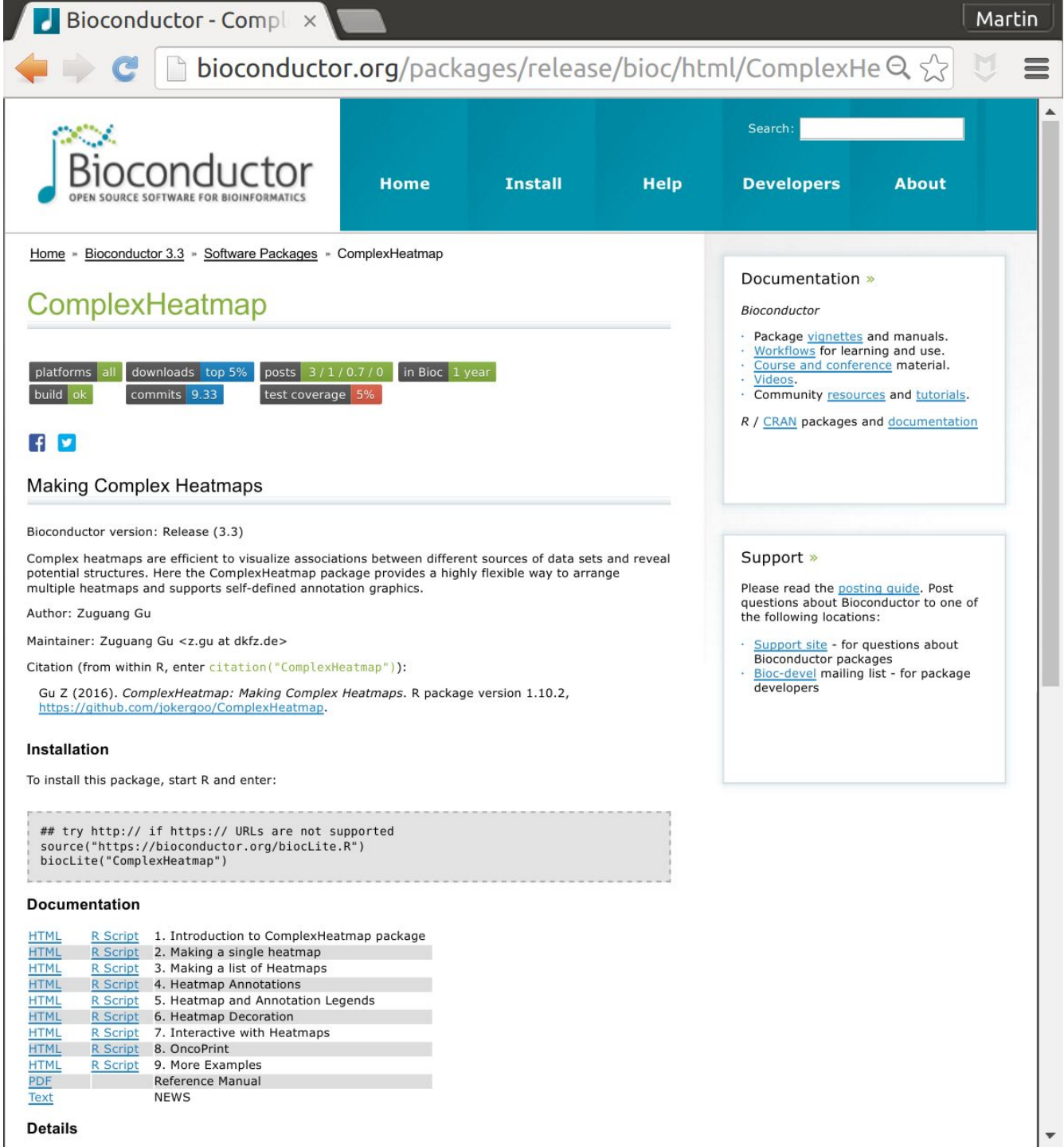
Learn

Vignettes 文档说明

- Each package!

Courses 相关课程

Support site



Learn

Vignettes 文档说明

- Each package!

Courses 相关课程

Support site 支持网站

Differential analysis of count data – the DESeq2 package 2

Contents

Differential analysis of count data – the DESeq2 package

Michael I. Love[1], Simon Anders[2,3], 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

May 15, 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 RN
vignette but at a slower pa

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 *RangedSummarizedExperiment* 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)`.

```
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.

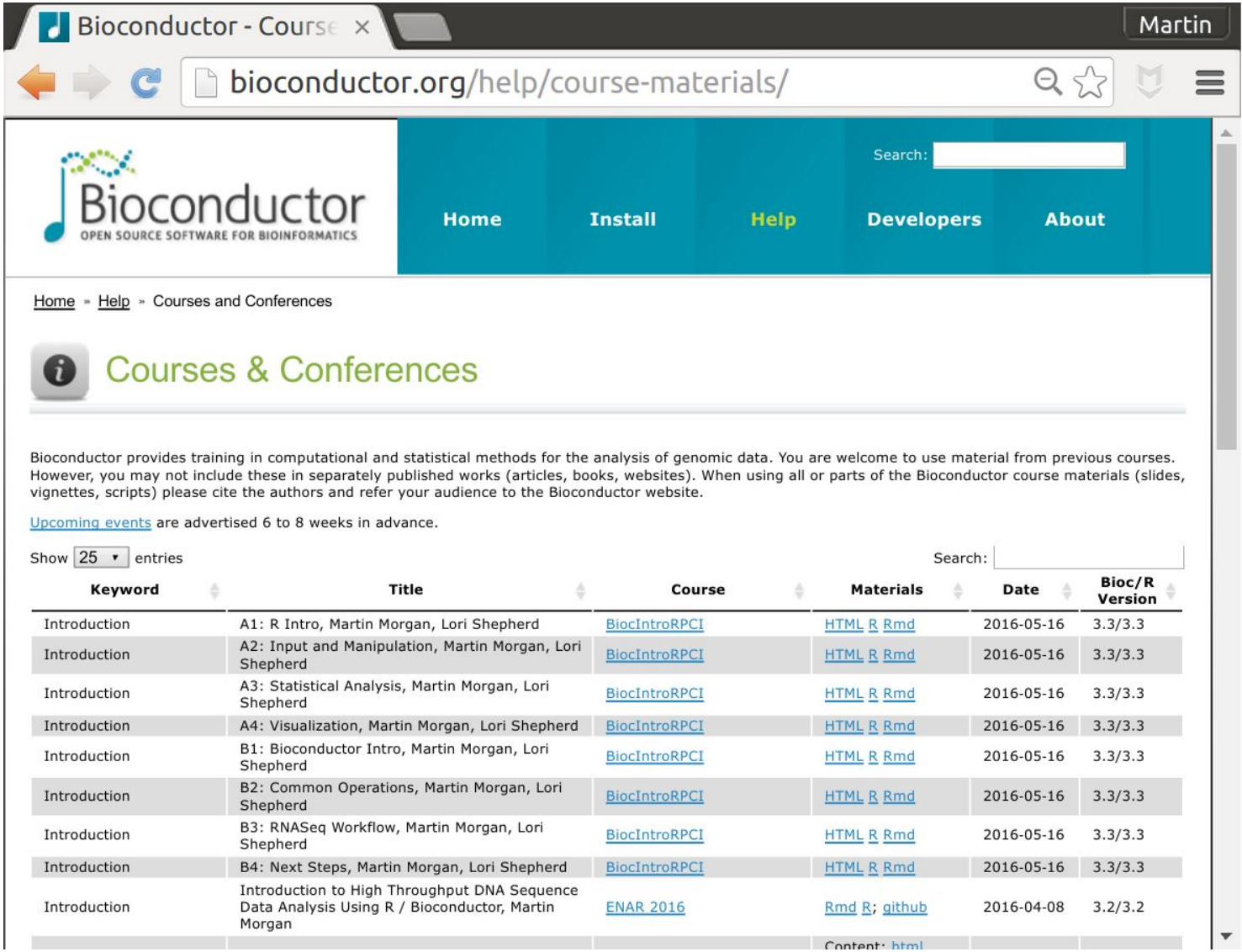
Learn

Vignettes 文档说明

- Each package!

Courses 相关课程

Support site 支持网站



Keyword	Title	Course	Materials	Date	Bioc/R Version
Introduction	A1: R Intro, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	A2: Input and Manipulation, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	A3: Statistical Analysis, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	A4: Visualization, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	B1: Bioconductor Intro, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	B2: Common Operations, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	B3: RNASeq Workflow, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	B4: Next Steps, Martin Morgan, Lori Shepherd	BiocIntroRPCI	HTML R Rmd	2016-05-16	3.3/3.3
Introduction	Introduction to High Throughput DNA Sequence Data Analysis Using R / Bioconductor, Martin Morgan	ENAR 2016	Rmd R; github	2016-04-08	3.2/3.2

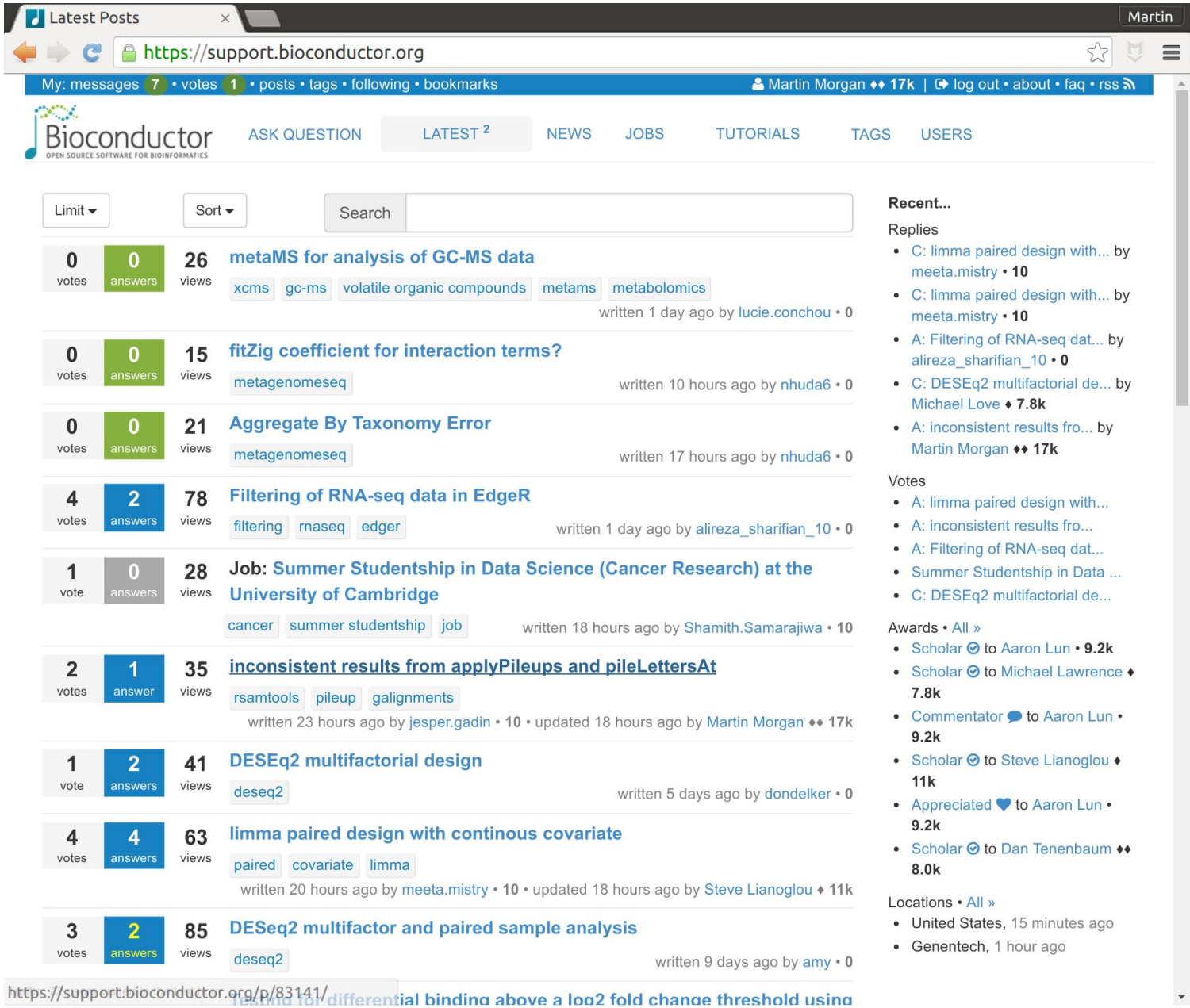
Learn

Vignettes 文档说明

- Each package!

Courses 相关课程

Support site 支持网站



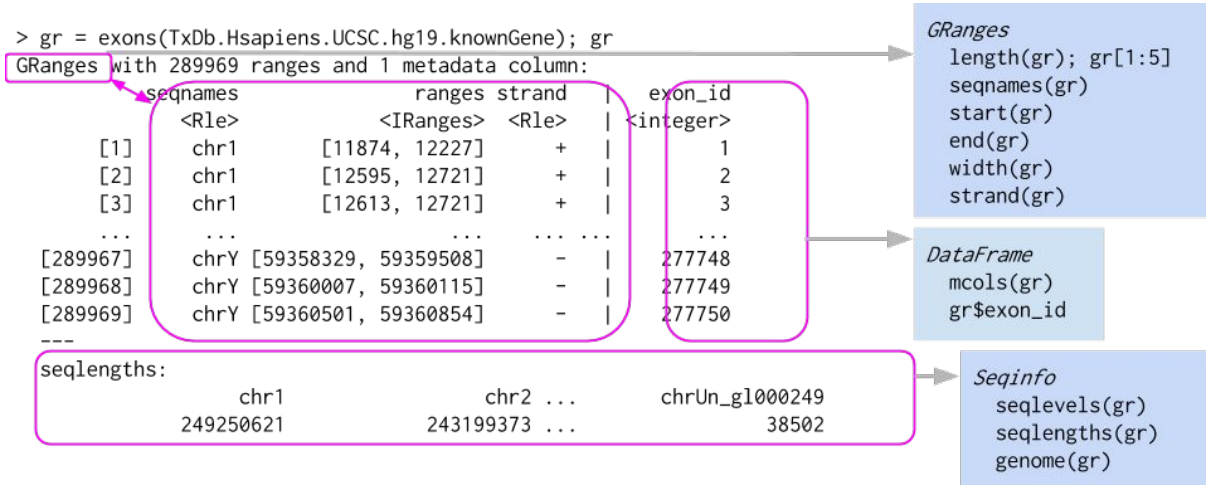
Learn

Essential Software　基础软件包

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

Annotation Resources　注释资源

- Packages
  - org.*
  - TxDb.*
  - BSgenome.*
- On-line
  - biomaRt
  - AnnotationHub



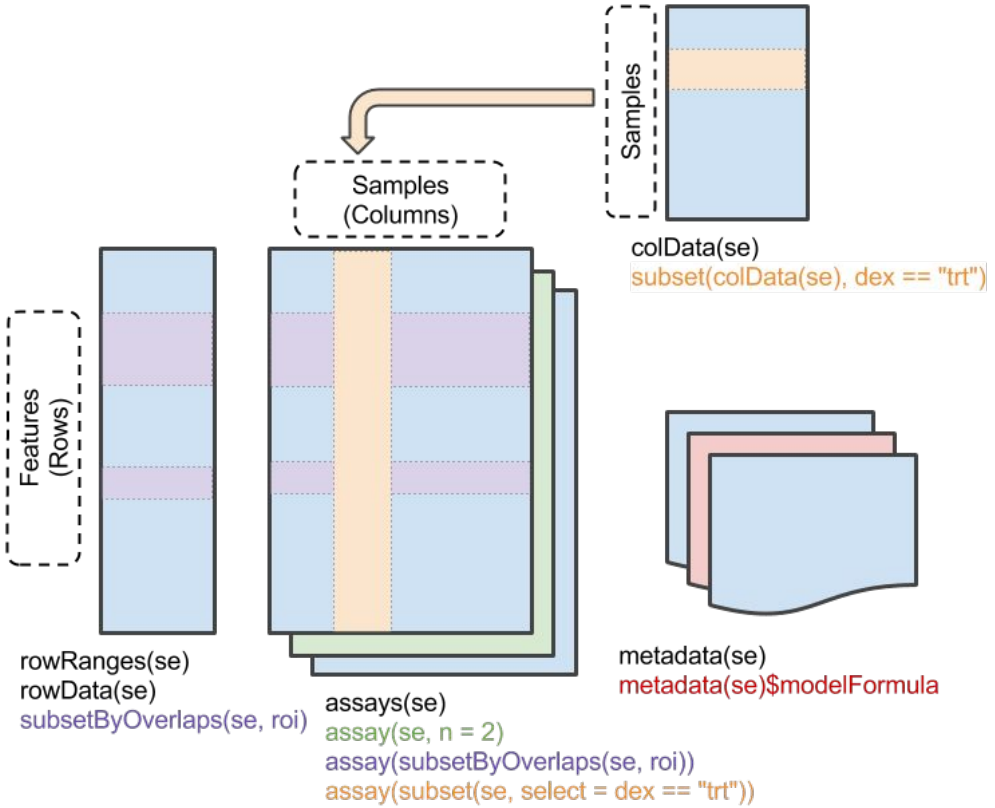
Learn

Essential Software　基础软件包

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

Annotation Resources　注释资源

- Packages
  - org.*
  - TxDb.*
  - BSgenome.*
- On-line
  - biomaRt
  - AnnotationHub



Learn

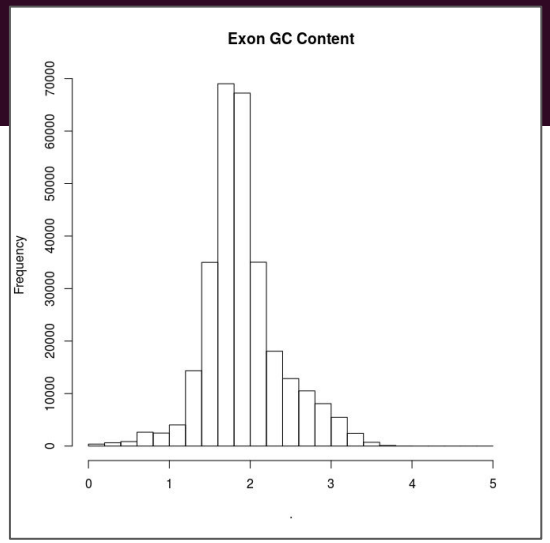
Essential Software　基础软件包

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

Annotation Resources 注释资源

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

```
> library(Homo.sapiens)
> head(id <- select(Homo.sapiens, "BRCA1", c("GENENAME", "TXID"), "SYMBOL"), 3)
'select()' returned 1:many mapping between keys and columns
  SYMBOL        GENENAME  TXID
1  BRCA1 breast cancer 1 63594
2  BRCA1 breast cancer 1 63595
3  BRCA1 breast cancer 1 63596
> brca1tx <- exonsBy(Homo.sapiens, "tx")[id$TXID]
> library(BSgenome.Hsapiens.UCSC.hg19)
> head(dna <- getSeq(BSgenome.Hsapiens.UCSC.hg19, brca1tx))
DNAStringSetList of length 6
[["63594"]] TTCATTGGAACAGAAAGAAATGGATTTATCTGCTCTTCGCGTTGAAGAAGTACAAAATGTCATTA...
[["63595"]] AGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGAGTCC...
[["63596"]] AGATAACTGGGCCCCTGCGCTCAGGAGGCCTTCACCCTCTGCTCTGGGTAAAGGTAGTAGAGTCC...
[["63597"]] CTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAATTACAGATAAATTAAAAC...
[["63598"]] CTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACGGAAAAGCGCGGGAATTACAGATAAATTAAAAC...
[["63599"]] GTACCTTGATTTCGTATTCTGAGAGGCTGCTGCTTAGCGGTAGCCCCTTGGTTTCCGTGGCAACG...
> library(magrittr)
> getSeq(BSgenome.Hsapiens.UCSC.hg19, exons(Homo.sapiens))  %>%
+     letterFrequency("GC") %>% log10 %>% hist(main="Exon GC Content")
>
```



Use

File Edit Code View Plots Session Build Debug Tools Help

Console | R Markdown

~/a/bioC/Docs/Talks/CI4CC/

```
> 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)
```

Report.Rmd

Knit HTML | Run | Chunks

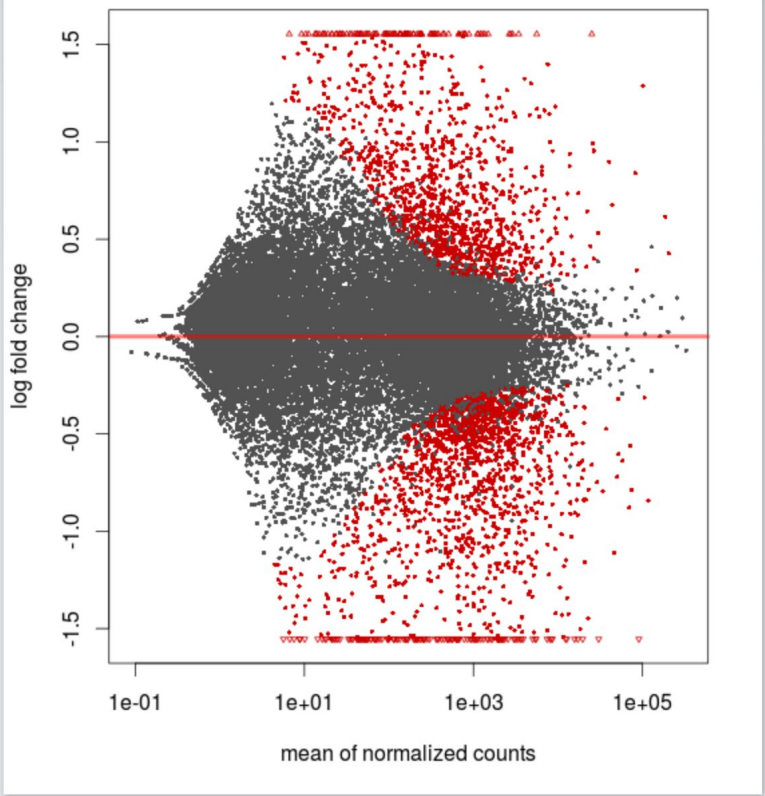
```
1  ---
2  title: "Bioinformatics Report"
3  author: "Martin Morgan"
4  date: "March 12, 2016"
5  output: html_document
6  ---
7
8  # Introduction
9
10 It is very easy to create advanced reports in a fully
   reproducible way. Reports can be distributed as PDF, HTML, ...
11
12 Reports can include 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)
21 ```
```

17:21 | Chunk 1 | R Markdown

Environment | History

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



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*

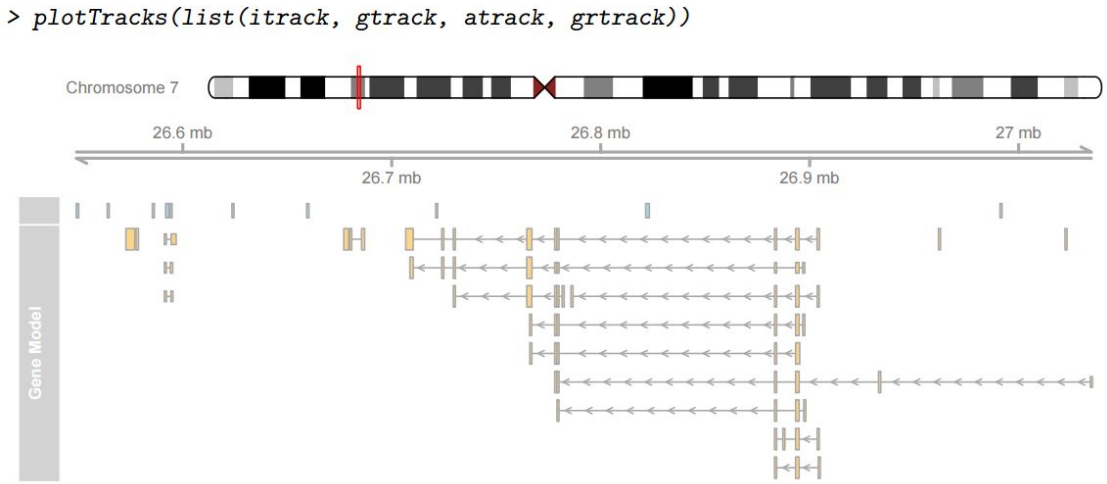
Data access　数据访问

- *biomaRt*
- *GEOquery / SRAdb*
- *TCGAbiolinks*
- *AnnotationHub / ExperimentHub*

Visualization　可视化

- *Gviz, ComplexHeatmap, ggbio, ggtree, ...*

Many other packages!



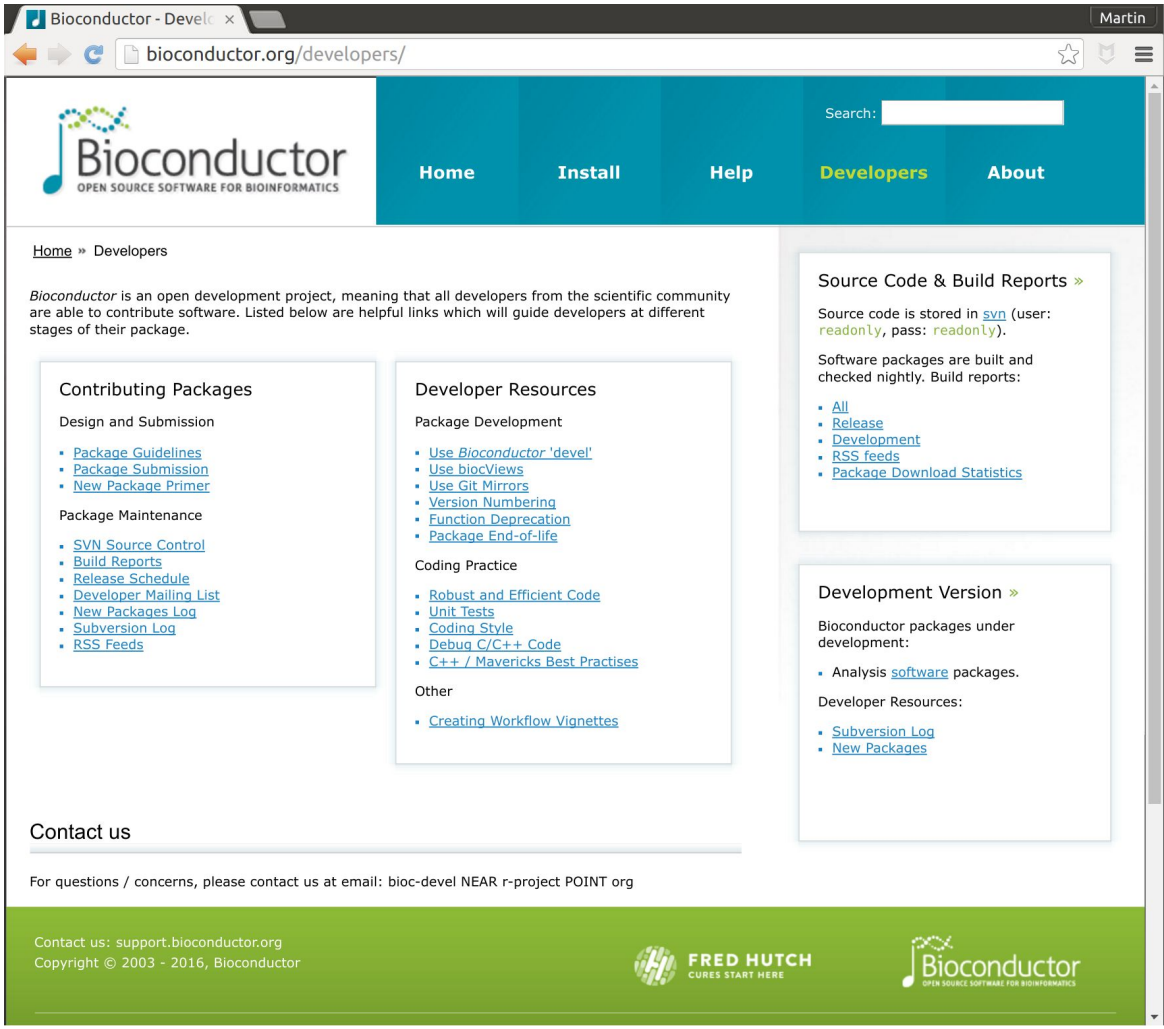
Develop

Create a package　创建包

- 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*　提交

- 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)

National Human Genome Research Institute
Advancing human health through genomics research

NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

SOUND

FRED HUTCH
40 YEARS OF CURES 1975–2015

ROSWELL PARK CANCER INSTITUTE

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