

# Cancer Genomics: Integrative and Scalable Solutions in *R* / *Bioconductor*

Martin Morgan

[martin.morgan@roswellpark.org](mailto: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://support.bioconductor.org>

The screenshot shows the Bioconductor website homepage. The browser address bar displays "bioconductor.org". The page features a teal header with the Bioconductor logo and navigation links: Home, Install, Help, Developers, and About. A search bar is located in the top right corner. The main content area is divided into several sections:

- BioC 2016**: A section for the annual conference, mentioning "BioC 2016: Where Software and Biology Connect" at Stanford University.
- About Bioconductor**: A section describing the tools for genomic data analysis, mentioning the use of R and the availability of software packages, AMI, and Docker images.
- News**: A section with a list of recent updates, including the availability of Bioconductor 3.3 and the launch of the F1000 Research Channel.
- Install**: A section titled "Get started with Bioconductor" with links for installing Bioconductor, exploring packages, getting support, and following on Twitter.
- Learn**: A section titled "Master Bioconductor tools" with links for courses, support sites, package vignettes, literature citations, common work flows, FAQ, and community resources.
- Use**: A section titled "Create bioinformatic solutions with Bioconductor" with links for software, annotation, experiment packages, Amazon Machine Image, latest release announcements, and support sites.
- Develop**: A section titled "Contribute to Bioconductor" with links for using 'devel', 'devel' software, annotation and experiment packages, package guidelines, new package submissions, developer resources, and build reports.

At the bottom of the page, there are links for "Support" and "Events", and a "Tweets" section for @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

```
> ##
> ## Multi-assay experiments, 'devel' only
> ##
> library(MultiAssayExperiment)
> ovarian
MultiAssayExperiment with 13 experiments
Containing an Elist class object of length 13
[1] RNASeqGene: ExpressionSet - 19990 rows, 299 columns
[2] RNASeq2GeneNorm: ExpressionSet - 20501 rows, 307 columns
[3] miRNASeqGene: ExpressionSet - 705 rows, 461 columns
[4] CNASNP: RangedRaggedAssay - 873768 rows, 1145 columns
[5] CNVSNP: RangedRaggedAssay - 254437 rows, 1141 columns
[6] CNACGH: RangedRaggedAssay - 126479 rows, 472 columns
[7] Methylation: ExpressionSet - 27578 rows, 591 columns
[8] mRNAArray: ExpressionSet - 18632 rows, 575 columns
[9] miRNAArray: ExpressionSet - 821 rows, 573 columns
[10] RPPAArray: ExpressionSet - 208 rows, 427 columns
[11] Mutations: RangedRaggedAssay - 20219 rows, 316 columns
[12] gistica: ExpressionSet - 24776 rows, 573 columns
[13] gistict: 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

```
> 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)

BLCA 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

```
> ##
> ## 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:
  6aceceb-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

```
> ##
> ## Large data representation and processing
> ##
> library(GenomicRanges)
> gpos <- GPos(seqinfo(bf1)["chr14"])

> ##
> ## Parallel evaluation -- cores, clusters, clouds
> ##
> library(BiocParallel)
> library(RNAseqData.HNRNPC.bam.chr14)
> bf1 <- BamFileList(RNAseqData.HNRNPC.bam.chr14_BAMFILES)
> cvg <- bplapply(bf1, coverage)
> mcols(gpos) <- DataFrame(lapply(cvg, "[[", "chr14"))
> gpos
GPos object with 107349540 positions and 8 metadata columns:
      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

```
> ##
> ## Integration with existing 'containers'
> ##
> library(SummarizedExperiment)
> gpos <- GPos(seqinfo(bf1)["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

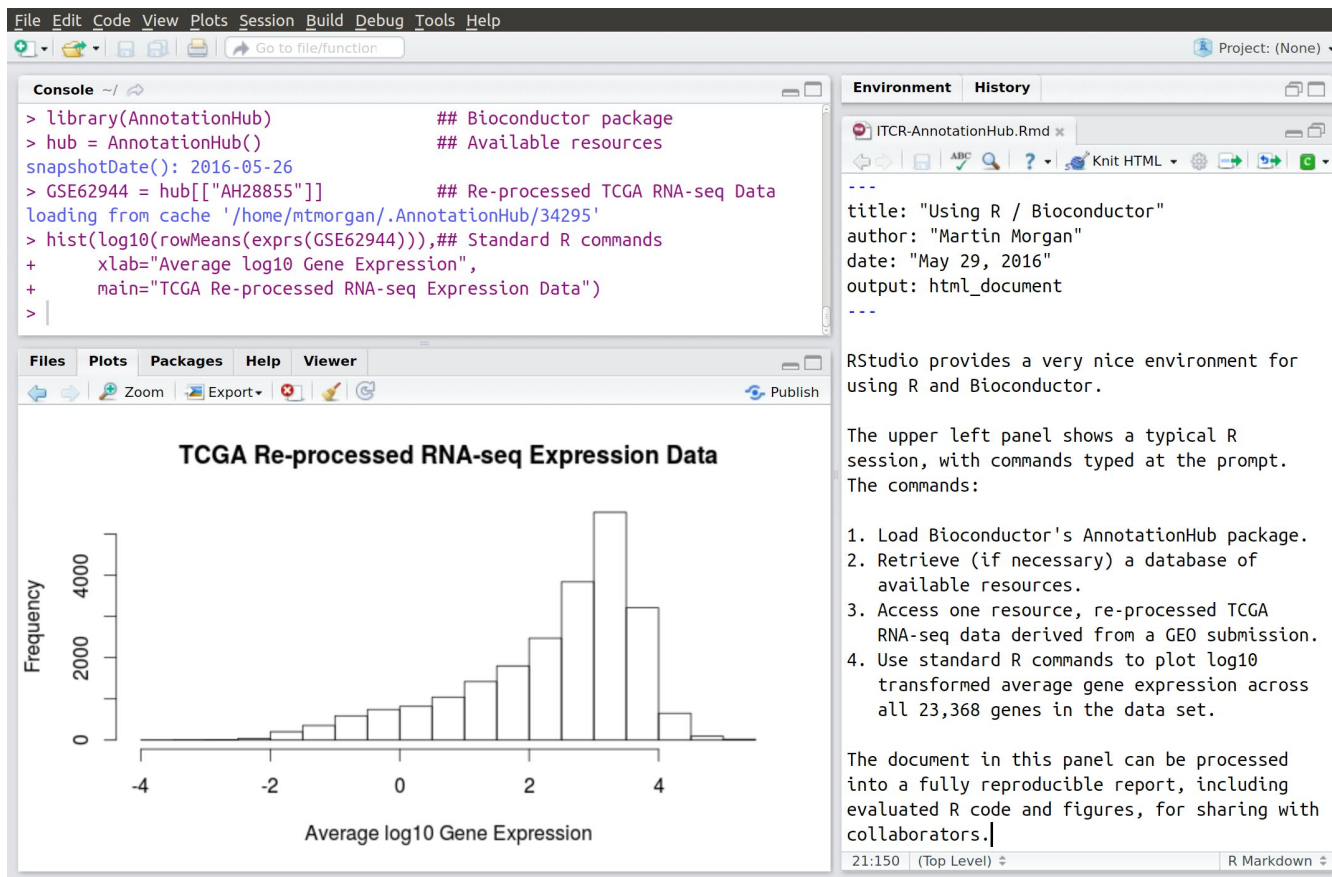
- Package vignettes (e.g., [DESeq2](#))
- [Training material](#)
- [Scientific literature](#)

## Use

- [Package discovery](#)
- [Support site](#)

## Contribute

- Developer resources
- Submission and technical review process



The screenshot displays the RStudio interface. The top-left pane shows the R console with the following code and output:

```
> library(AnnotationHub)           ## Bioconductor package
> hub = AnnotationHub()           ## Available resources
snapshotDate(): 2016-05-26
> GSE62944 = hub[["AH28855"]]     ## Re-processed TCGA RNA-seq Data
loading from cache '/home/mtmorgan/.AnnotationHub/34295'
> hist(log10(rowMeans(exprs(GSE62944))), ## Standard R commands
+       xlab="Average log10 Gene Expression",
+       main="TCGA Re-processed RNA-seq Expression Data")
> |
```

The bottom-left pane displays a histogram titled "TCGA Re-processed RNA-seq Expression Data". The x-axis is labeled "Average log10 Gene Expression" and ranges from -4 to 4. The y-axis is labeled "Frequency" and ranges from 0 to 4000. The histogram shows a distribution of average log10 gene expression values across 23,368 genes, with a peak frequency of approximately 4000 at an average log10 expression of about 3.5.

The right-hand pane shows the R Markdown document content:

```
---
title: "Using R / Bioconductor"
author: "Martin Morgan"
date: "May 29, 2016"
output: html_document
---
```

Below the document content, there is a paragraph of text:

RStudio provides a very nice environment for using R and Bioconductor.

The upper left panel shows a typical R session, with commands typed at the prompt. The commands:

1. Load Bioconductor's AnnotationHub package.
2. Retrieve (if necessary) a database of available resources.
3. Access one resource, re-processed TCGA RNA-seq data derived from a GEO submission.
4. Use standard R commands to plot log<sub>10</sub> transformed average gene expression across all 23,368 genes in the data set.

The document in this panel can be processed into a fully reproducible report, including evaluated R code and figures, for sharing with collaborators.

The bottom status bar shows the time 21:150, the level (Top Level), and the output format (R Markdown).

# 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 Flicek (EMBL/EBI), Simon Urbanek (AT&T), Vincent Carey (Brigham & Women's), Wolfgang Huber (EBI), Rafael Irizzary (Dana Farber), Robert Gentleman (23andMe)



SOUND



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.