Introduction to R and Bioconductor

Martin Morgan\textsuperscript{1}

June 23 – 28, 2013
Overview

R
1. Vectors, data frames, arrays
2. Statistical concepts
3. Functions
4. Packages
5. From script to collaboration
6. HELP!

Bioconductor
1. A short history
2. Sequence analysis
3. Classes, generics and methods
4. HELP!
Vectors, data frames, arrays

```r
> x <- c(1, 2, 3, 4, 5)
> x
[1] 1 2 3 4 5
> log(x)
[1] 0.0000000 0.6931472 1.0986123 1.3862944 1.6094379
> x[c(3, 2)]
[1] 3 2
> x[3:2]
[1] 3 2
> x[c(TRUE, FALSE)]
[1] 1 3 5
```
Vectors, data frames, arrays

```r
> df <- data.frame(
+   age = c(17, 23, 32, 37),
+   sex = c("Male", "Female", "Female", "Male"))
> df

   age sex
1   17  Male
2  23  Female
3  32  Female
4  37   Male

> df[df$age < 30 & df$sex == "Male", ]

   age sex
1   17  Male
```
Vectors, data frames, arrays

```r
> m <- matrix(1:8, 2, 4)
> m

[1,]   1   3   5   7
[2,]   2   4   6   8

> log(m)

[1,] 0.0000000 1.098612 1.609438 1.945910
[2,] 0.6931472 1.386294 1.791759 2.079442

> rowSums(m)

[1] 16 20
```
> df$sex

[1] Male  Female  Female  Male
Levels: Female  Male

> df$height <- c(180, 172, NA, 177)
> df

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>Male</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>Female</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>Female</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>Male</td>
</tr>
</tbody>
</table>
Functions

dir, read.table (and friends), scan List files, input data.
c, factor, data.frame, matrix Create objects.
summary, table, xtabs Summarize, cross-tabulate variables.
t.test, aov, lm, anova, chisq.test Compare groups.
dist, hclust Cluster data.
   plot Visualize data.
lapply, sapply, mapply, aggregate Apply a function to each element of a list or vector.
   ls, str List objects and their structure.
library, search Attach to or describe the library search path.
Packages

**Base**  base, stats, graphics, ...

**Recommended**  lattice, ...

**Contributed**  data.table, XML, biglm, ...

- Install a contributed package
  ```
  > source("http://bioconductor.org/biocLite.R")
  > biocLite("data.table")
  ```

- Use during a session
  ```
  > library(lattice)
  ```
From script to collaboration

Increasingly complicated tasks

1. Commands entered at the prompt
2. Scripts and functions saved in `.R` files
3. `.R` files and documentation ordered in packages
   
   % dir MyPackage
   
   data/ NAMESPACE DESCRIPTION man/ R/ vignettes/

4. Packages shared with colleagues

Net result: sophisticated, highly reproducible research
HELP!

> help.start()
> ?t.test
> vignette("datatable-faq")

- Books and training resources
- $R$ web site and mailing list\(^2\)
- StackOverflow\(^3\)

\(^2\)http://r-project.org
\(^3\)http://stackoverflow.com/questions/tagged/r
Bioconductor: A short history

Then

➤ Founded 2001
➤ Analysis and *comprehension* of high throughput genomic data
➤ Initial focus: microarrays
➤ Reproducibility, statistical analysis, computation

Now

➤ > 670 software packages
➤ World-wide contributions
➤ Sequence analysis, microarrays, systems biology, flow cytometry, ...
About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 671 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).

Use Bioconductor for...

Microarrays
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.

Transcription Factors
Find candidate binding sites for known transcription factors via sequence matching.

Counting Reads for Differential Expression
The parathyroidSE ExperimentData package and vignette illustrates how to count reads and perform other common operations required for differential expression analysis.

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.

Mailing Lists

Events

BioC2013
18 - 19 July 2013 — Seattle, WA, USA

CSAMA 2013 (Computational Statistics for Genome Biology)
24 - 28 June 2013 — Brixen-Bressanone, Italy

Tweets

Follow @Bioconductor

bioconductor.org/packages/2.13...spliceR
Classification of alternative splicing and...
Bioconductor: A short history
## Sequence Analysis (many packages!)

<table>
<thead>
<tr>
<th>I/O</th>
<th>ShortRead (FASTQ), Rsamtools (BAM), rtracklayer (GFF, WIG, BED), VariantAnnotation (VCF), ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representation</td>
<td>IRanges, GenomicRanges, GenomicFeatures, Biostrings, BSgenome, ...</td>
</tr>
<tr>
<td>Annotation</td>
<td>GenomicFeatures, biomaRt, AnnotationHub, ...</td>
</tr>
<tr>
<td>Alignment</td>
<td>gmapR, Rsubread, Biostrings, ...</td>
</tr>
<tr>
<td>Visualization</td>
<td>ggbioc, Gviz, ...</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>DESeq, edgeR, DEXSeq, cummeRbund, ...</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>DiffBind, ChIPpeakAnno, ...</td>
</tr>
<tr>
<td>Variants</td>
<td>VariantAnnotation, VariantTools, ...</td>
</tr>
<tr>
<td>Motifs</td>
<td>MotifDb, seqLogo, ...</td>
</tr>
<tr>
<td>Work flows</td>
<td>QuasR, ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Classes, generics and methods

```r
> library(ShortRead)
> fq <-
+  readFastq("~/BigData/fastq/SRR031724_1_subset.fastq")
> fq          # 'S4' class

class: ShortReadQ
length: 1000000 reads; width: 37 cycles

> sread(fq)   # 'generic' and 'method'; another S4 class

A DNAStringSet instance of length 1000000
width seq
[1] 37 GTTTTGTCCAAGTTCTGGTAGCTGAATCCTGGGGCGC
[2] 37 GTTGTCGCATTCCTTACTCTCATTCGGGAATTCTGTT
[3] 37 GAATTTTTTGAGAGCGAAATGATAGCCGATGCCCTGA
... ... ...
[1000000] 37 GAAGTCGGTACCCTCGAACAGAGAGTCGATCTCAATG
```
HELP!

- Use tab completion to find help on generics and methods
  
  ```
  > ?sread
  > ?"reverseComplement<tab>"  # tab key for completions!
  > ?"reverseComplement,DNAStringSet-method"
  ```

- Discover available functions and their definition.
  
  ```
  > showMethods("reverseComplement")
  > showMethods(class="DNAStringSet", where=search())
  > selectMethod("reverseComplement", "DNAStringSet")
  ```
Acknowledgements

**Bioconductor core**
- Vince Carey
- Wolfgang Huber
- Robert Gentleman
- Rafael Irizzary
- Sean Davis
- Kasper Hansen

**Bioconductor team**
- Marc Carlson (annotation)
- Valerie Obenchain (variants, ranges)
- Hervé Pagès (ranges, strings)
- Paul Shannon (systems biology)
- Dan Tenenbaum (web, build)

And the *Bioconductor* community!