Statistical LeaRning

Katja Nowick

Bioinformatics group,

Markus Kreuz

IMISE

What is R?

- **Programming/scripting language**
- **Comprehensive statistical environment**
- **Strength**: statistical data analysis

  \+ graphical display

Why use R?

- **It's free**!
- Runs on a **variety of platforms** including Windows, Unix and MacOS.
- Complicated bioinformatics analyses made easy by a huge collection of **packages in Bioconductor**
- **Potential to implement automated workflows**
- **Big datasets**
- **Advanced statistical routines**
- **State-of-the-art graphics** capabilities

How to obtain and install R?

- R can be **downloaded** from the Comprehensive R Archive Network (CRAN): http://cran.r-project.org/
- **Installation instructions** depend on your operating system and should be accessible from the R download page for you operating system.
- For our course, R is already installed

  We use **R-studio** as programming environment

750 packages in Bioconductor

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » BiocViews

All Packages

Bioconductor version 2.13 (Release)

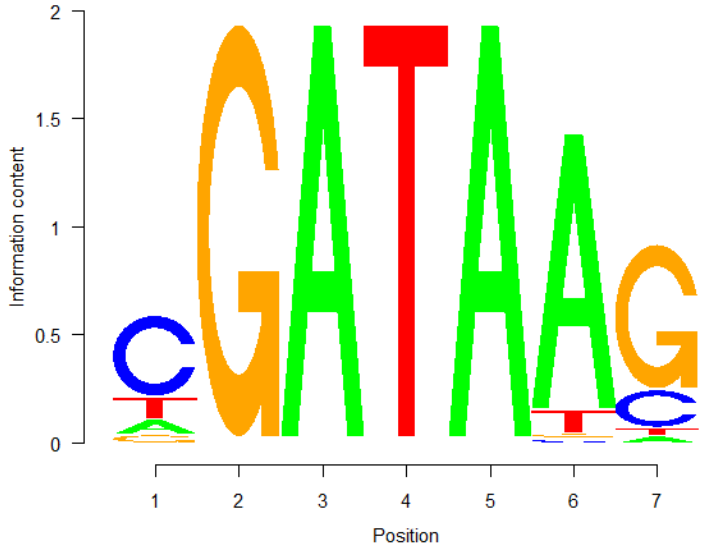
- Software (750)
  - Annotation (99)
  - AssayDomains (302)
  - AssayTechnologies (450)
  - Bioinformatics (507)
  - BiologicalDomains (132)
  - Infrastructure (201)
- AnnotationData (697)
- ExperimentData (180)

Packages

Package	Maintainer	Title
a4	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Umbrella Package
a4Base	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Base Package
a4Classif	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Classification Package
a4Core	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Core Package
a4Preproc	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Preprocessing Package
a4Reporting	Tobias Verbeke, Willem Ligtenberg	Automated Affymetrix Array Analysis Reporting Package
ABarray	Yongming Andrew Sun	Microarray QA and statistical data analysis for Applied Biosystems Genome Survey Microrarray (AB1700) gene expression data.
aCGH	Peter Dimitrov	Classes and functions for Array Comparative Genomic Hybridization data.
ACME	Sean Davis	Algorithms for Calculating Microarray Enrichment (ACME)
ADaCGH2	Ramon Diaz-Uriarte	Analysis of big data from aCGH experiments using parallel computing and ff objects
adSplit	Claudio Lottaz	Annotation-Driven Clustering
affxparser	Kasper Daniel Hansen	Affymetrix File Parsing SDK

Binding site detection

Finding binding sites for a transcription factor in the promoter of a gene

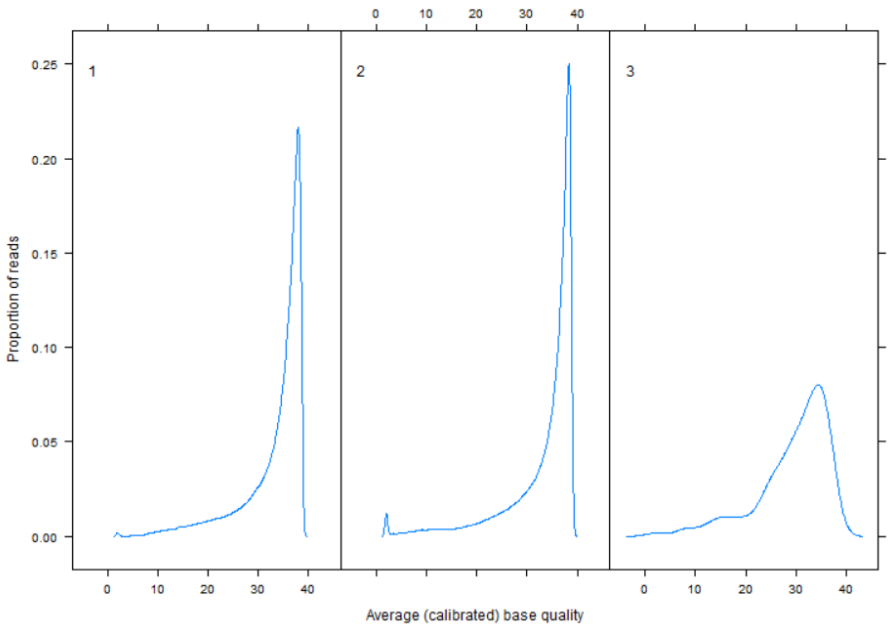


With only 8 lines of code:

```
query(MotifDb, "DAL80")
pfm.dal80.jaspar = query(MotifDb, "DAL80")[[1]]
seqLogo(pfm.dal80.jaspar)
dal1 = "YIR027C"
chromosomal.loc = transcriptsBy(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene, by = "gene")[dal1]
promoter.dal1 = getPromoterSeq(chromosomal.loc, Scerevisiae, upstream=1000, downstream=0)
pcm.dal80.jaspar = round(100 * pfm.dal80.jaspar)
matchPWM(pcm.dal80.jaspar, unlist(promoter.dal1)[[1]], "90%")
```

Quality assessment of NGS data

From a directory of FastQ files
to a full quality report:



<u>With 6 lines of code:</u>

```
files = list.files("fastq", full=TRUE)
names(files) = sub(".fastq", "", basename(files))
qas = lapply(seq_along(files),
             function(i, files) qa(readFastq(files[i]), names(files)[i]), files)
qa <- do.call(rbind, qas)
save(qa, file=file.path("output", "qa.rda"))
browseURL(report(qa))
```

Finding help

- R **mailing lists**: https://stat.ethz.ch/mailman/listinfo/
- **Manuals** and **FAQs**: http://www.r-project.org/
- Selected **tutorials**:
  - http://www.math.ilstu.edu/dhkim/Rstuff/Rtutor.html
  - http://www.statmethods.net/index.html
  - http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_BioCondManual.html

Goals for the next 5 x 5 hours

Dec 3rd:	Introduction to R
Dec 10th:	Statistics and Graphics
Dec 17th:	A small programming project
Jan 14th:	Analysis of gene expression data
Jan 21st:	Clustering and Gene Ontology

Goals for the first 5 hours

- R-Studio
- R as a calculator (interactive R)
- Variables: numeric, character, arrays, vectors, matrices
- Loops
- Apply
- Conditional executions (if-else-statements)
- Write your own functions

Multiple exercises in between

Goals for second 5 hours

- R packages
- Help pages
- Some more on functions
- Graphics
- Statistical tests

Multiple exercises in between

Optional for today

- If you know already R -

Conways Spiel des Lebens

Das Spiel des Lebens (engl. Conway's Game of Life) ist ein vom Mathematiker John Horton Conway 1970 entworfenes System, basierend auf einem zweidimensionalen zellulären Automaten. Es ist eine einfache und bis heute populäre Umsetzung der Automaten-Theorie von Stanisław Marcin Ulam (Quelle Wikipedia).
Gespielt wird auf einem schachbrettartigen Feld beliebiger Größe. Jedes Feld kann mit einer lebenden Zelle besetzt oder frei sein.

Der Folgezustand einer Spielsituation leitet sich durch folgende Regeln ab:

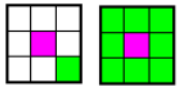
1) Eine tote/leere Zelle mit genau drei lebenden Nachbarn wird in der Folgegeneration neu geboren.
   Beispiel:

   (rot=betrachtete leere Zelle; grün=lebende Zelle; weiß leere Zelle)

2) Eine lebende Zelle mit zwei oder drei lebenden Nachbarn bleibt in der Folgegeneration lebend.
   Beispiel:

   (magenta=betrachtete lebende Zelle; grün=lebende Zelle; weiß leere Zelle)

3) Alle Zellen die 1) und 2) nicht erfüllen sterben ab
   Beispiele:



Aufgabenstellung:

Implementiere einen Game of Life Simulator in R. Der Simulator soll folgende Funktionalität umfassen:

1) Kreieren einer zufälligen Start-Spielsituation (Vorgabe der Dimensionen des Spielfeldes sowie des Anteils an lebenden Zellen).
2) Generieren der Nachfolgenenden Konfiguration des Spielfeldes
3) Graphische Ausgabe des Spielfeldes.