

How to convince Biologists to use R

Hans-Rudolf Hotz (hrh@fmi.ch)
Friedrich Miescher Institute for Biomedical Research
Basel, Switzerland

How to convince Biologists to use R

....you force them to use R

Hans-Rudolf Hotz (hrh@fmi.ch)
Friedrich Miescher Institute for Biomedical Research
Basel, Switzerland

background

Friedrich Miescher Institute

- part of the Novartis Research Foundation
- affiliated institute of Basel University

314 employees

(incl. 96 PhD students, 95 Post Docs)

Epigenetics

(8 research groups)

Growth Control

(7 research groups)

Neurobiology

(8 research groups)

Technology Platforms

Computational Biology – Cell Sorting – Imaging and Microscopy –
Functional Genomics – Histology – Mass Spectrometry – Protein Structure

FMI

Friedrich Miescher Institute
for Biomedical Research

background

Computational Biology Platform

five people supporting the wet lab scientists

dealing with all kind of data

~75 % of our work originates from NGS data

~75 % of our work is done with R (it used to be Perl)

FMI

Friedrich Miescher Institute
for Biomedical Research

How to convince Biologists to use R
....you force them to use R

....you help them to use R

....they are forced to use R

FMI

Friedrich Miescher Institute
for Biomedical Research

.....they are forced to use R

**Biological research has changed from studying
a single gene/phenotype to genome wide analysis**



a lot of (numeric) data



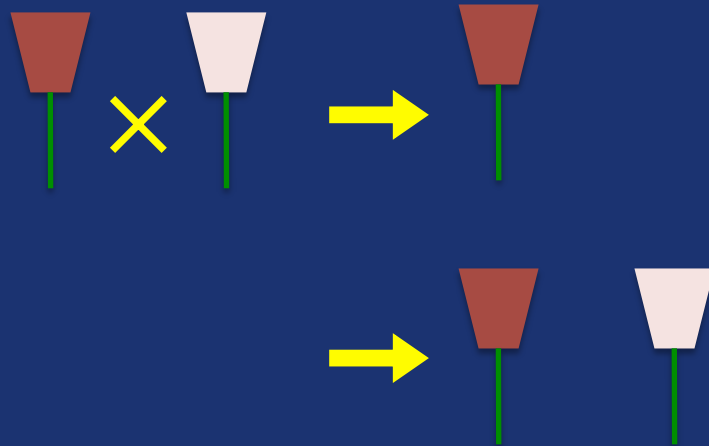
need for new software

FMI

Friedrich Miescher Institute
for Biomedical Research

Biological reserach

the classical experiment:



F1: 302:0 150:0

F2: 351:109 287:96

→ F2: 3:1

Biologists don't need R

FMI

Friedrich Miescher Institute
for Biomedical Research

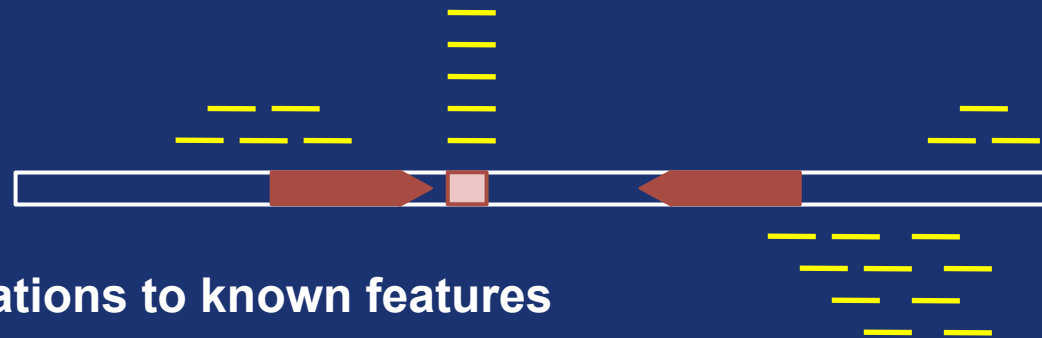
Biological reserach

the modern experiment:

100 million reads
from a DNA sequencer
(each 100 nucleotides long)

GATCGTGCCACC.....
GTGGCGAATGAT.....
CATCGTNCCACC.....
.....

align to the genome



compare the locations to known features
identify new features

***Biologists need big computers
....and the right software!***

FMI

Friedrich Miescher Institute
for Biomedical Research

Software and Tools for Biologists

- commercial software expensive, not up-to-date
- public website limited data volume
- open source tools not very flexible or
address only single steps
- self written software Biologists are not IT people
- “Bio* projects”

FMI

Friedrich Miescher Institute
for Biomedical Research

“Bio* projects

BioPerl, Biopython, BioJava.....Bioconductor

- open source
- ongoing development
- ‘easy’ to use
- can act as a ‘bridge’ to connect different open source tools / different data sets

FMI

Friedrich Miescher Institute
for Biomedical Research

Bioconductor <http://www.bioconductor.org/>

“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, more than 460 packages, and an active user community.”

Depending on the nature of your research,
Bioconductor might be the only available choice.

....they are forced to use R

FMI

Friedrich Miescher Institute
for Biomedical Research

Biological reserach

use of Bioconductor

100 million reads
from a DNA sequencer
(each 100 nucleotides long)



ShortRead, Biostrings

```
GATCGTGCCACC.....  
GTGGCGAATGAT.....  
CATCGTNCCACC.....  
.....
```

Rsamtools

align to the genome

DESeq, edgeR

IRanges, GenomicRanges

biomaRt, rtracklayer

compare the locations to known features
identify new features



chipseq

....you help them to use R

FMI

Friedrich Miescher Institute
for Biomedical Research

....you help them to use R

- running training courses
 - introductory
 - advanced
 - 'RLunch'
- we don't provide any help for Excel (you force them to use R)

FMI

Friedrich Miescher Institute
for Biomedical Research

introductory training course

- 4 x ~2.5 hours plus 'homework/exercises'
- based on "Introductory Statistics with R"
by Peter Dalgaard
- offered every ~9 months
- 25 - 30 participants (split in two groups)

FMI

Friedrich Miescher Institute
for Biomedical Research

session 1

Installing and running R

```
> hist(rnorm(1000))
```

Interacting with R

```
> 2 + 2
```

```
> x <- 2 + 2
```

Introducing the concept of vectors

```
> weight <- c(60, 72, 57, 72)
> mean(weight)
> weight + 1
```

*participants are not familiar with
working on the command line*

FMI

Friedrich Miescher Institute
for Biomedical Research

session 2

data types (numeric - logical - character - factor)

vector manipulation

higher data structures (matrix - data.frame - list)

```
> numbs <- c(1,2,3,4,5)
> chars <- as.character(numbs)
> as.numeric(chars) %% 2 == 0
[1] FALSE TRUE FALSE TRUE FALSE
```

*participants are struggling
participants don't see the relevance
to their work*

FMI

Friedrich Miescher Institute
for Biomedical Research

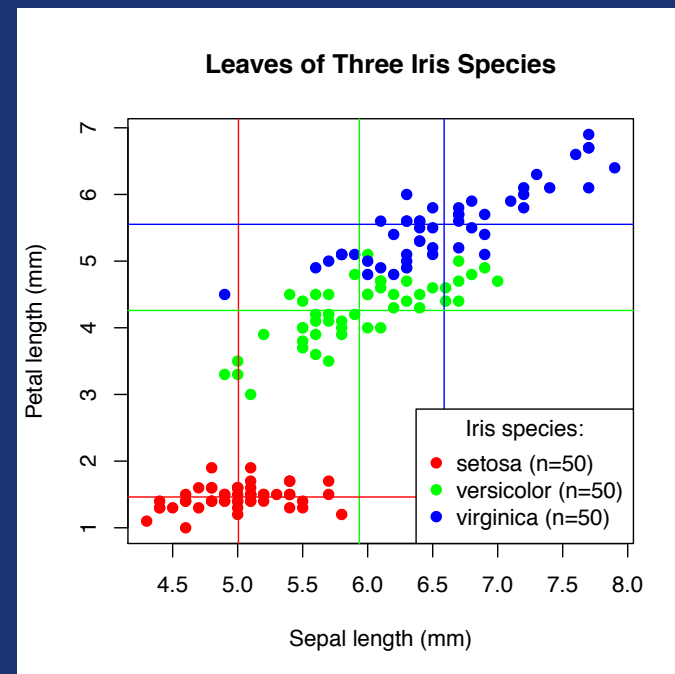
session 3

reading from text files
writing to text files

```
> read.delim("iris.txt")
```

visual data exploration

storing plots as image files



*participants are starting to see the light.....
..... or we lose them completely*

FMI

Friedrich Miescher Institute
for Biomedical Research

session 4

statistical hypothesis testing

fisher.test()
cor()
shapiro.test()
t.test()
wilcox.test()
phyper()
etc.

```
> D <- data.frame(DNAmet=c(T,F,T,T,T,T,F,T,F,F,T,F,F,F,F,F,F),  
                  H3acet=c(T,T,F,T,T,T,F,T,F,F,F,F,F,F,F,F))  
> ctD <- table(D)  
> ctD  
      H3acet  
DNAmet FALSE TRUE  
FALSE     9     1  
TRUE      1     5  
  
> fisher.test(ctD)  
  
      Fisher's Exact Test for Count Data  
  
data:  ctD  
p-value = 0.0345  
...
```

a simple Monte Carlo simulation

participants realize they lack basic statistics

FMI

Friedrich Miescher Institute
for Biomedical Research

after the introductory training course

do we have 30 new R experts, now? **no !**

but:

- next time they knock at our door we don't have to start at point zero.
- they will go on statistic training
- they want to know more
 - advanced training course
 - bi-weekly 'RLunch'

FMI

Friedrich Miescher Institute
for Biomedical Research

advanced training course

- 3 extra sessions
- working with Bioconductor packages

Using R for Short-Read Analysis (2 sessions)

Biostrings, IRanges, GenomicRanges, Rsamtools

Affymetrix microarray data analysis (1 session)

affy, limma

participants are overwhelmed

FMI

Friedrich Miescher Institute
for Biomedical Research

RLunch

- someone is presenting a 'problem' or 'solution'
- introduction of (new) Bioconductor packages



mixture between a 'journal club'
and a 'progress report'

FMI

Friedrich Miescher Institute
for Biomedical Research

How to convince Biologists to use R

....you help them to use R

....they are forced to use R

FMI

Friedrich Miescher Institute
for Biomedical Research

Acknowledgment

Computational Biology

- Michael Stadler (has written the original course material)
- Anita Lerch, Lukas Burger, Dimos Gaidatzis

Functional Genomics

- Tim Roloff

Epigenetics

- Robert Ivanek

FMI

Friedrich Miescher Institute
for Biomedical Research