"Developing" a PhD-Thesis in Perl, R and LaTeX

- and all the code involved -

Emanuel Heitlinger

November 19, 2009

R, Noweb/Sweave and Emacs-ESS

- Donald E. Knuth[1]: "The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer"

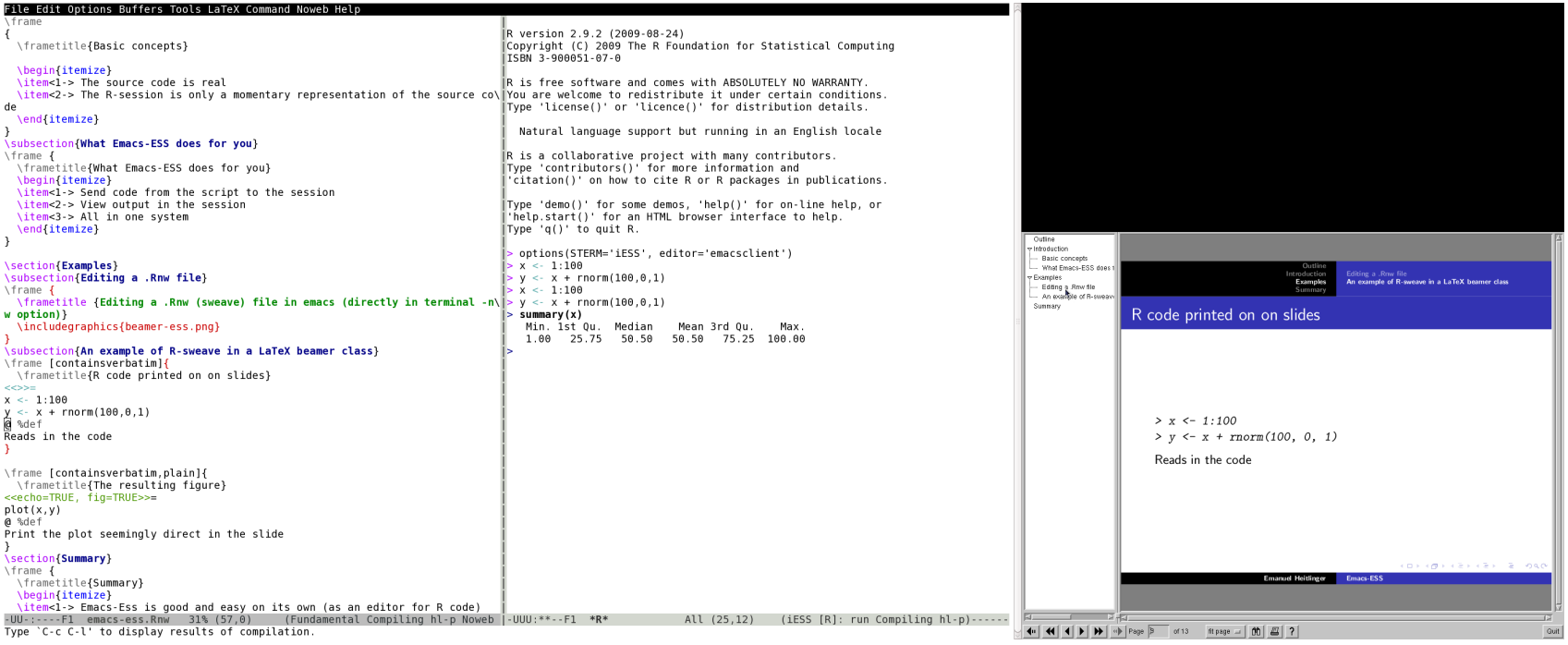
R, Noweb/Sweave and Emacs-ESS

- Donald E. Knuth[1]: "The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer"
- Friedrich Leisch: Integration into R -> Sweave[2]

R, Noweb/Sweave and Emacs-ESS

- Donald E. Knuth[1]: "The main idea is to treat a program as a piece of literature, addressed to human beings rather than to a computer"
- Friedrich Leisch: Integration into R -> Sweave[2]
- Emacs-ESS[3] is a very powerful editor for R-scripts and the .Rnw files of Sweave

Editing a .Rnw (sweave) file in Emacs



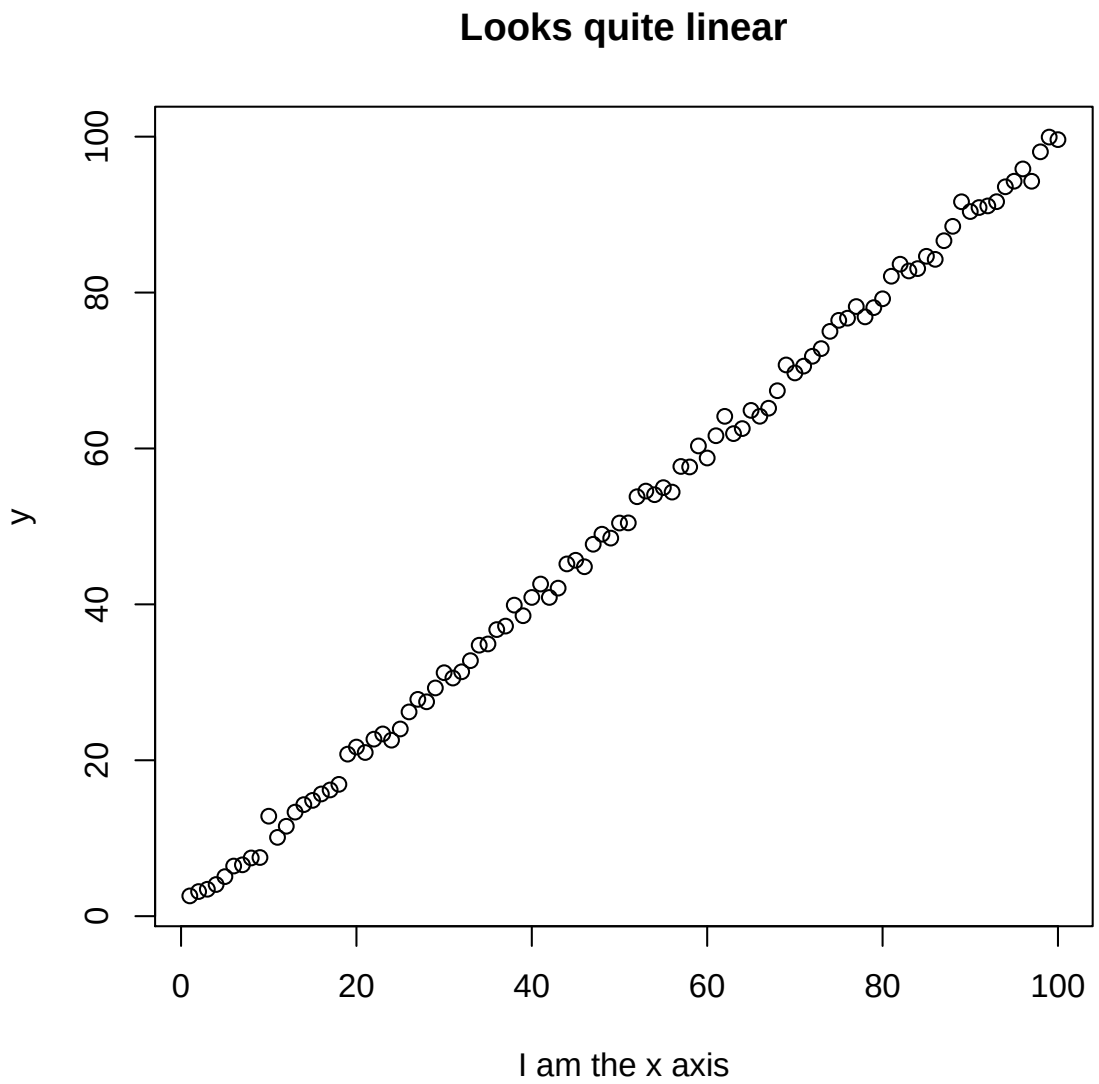
R code printed on on slides

```
> x <- 1:100
> y <- x + rnorm(100, 0, 1)
> x
```

```
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13
 [19]  19  20  21  22  23  24  25  26  27  28  29  30  31
 [37]  37  38  39  40  41  42  43  44  45  46  47  48  49
 [55]  55  56  57  58  59  60  61  62  63  64  65  66  67
 [73]  73  74  75  76  77  78  79  80  81  82  83  84  85
 [91]  91  92  93  94  95  96  97  98  99 100
```

Reads in the code

The resulting figure



Print the plot seemingly direct in the slide

An example involving perl

```
> S <- as.data.frame(read.delim(pipe("./pilot.pl"),
+      sep = ",", header = FALSE, as.is = TRUE))
```

This R function is calling a "custom for this analysis" perl-script and stores the output in a R-object. (The pilot.pl scipt calls istself multi purpose perl-scripts and blastall)

An example involving perl

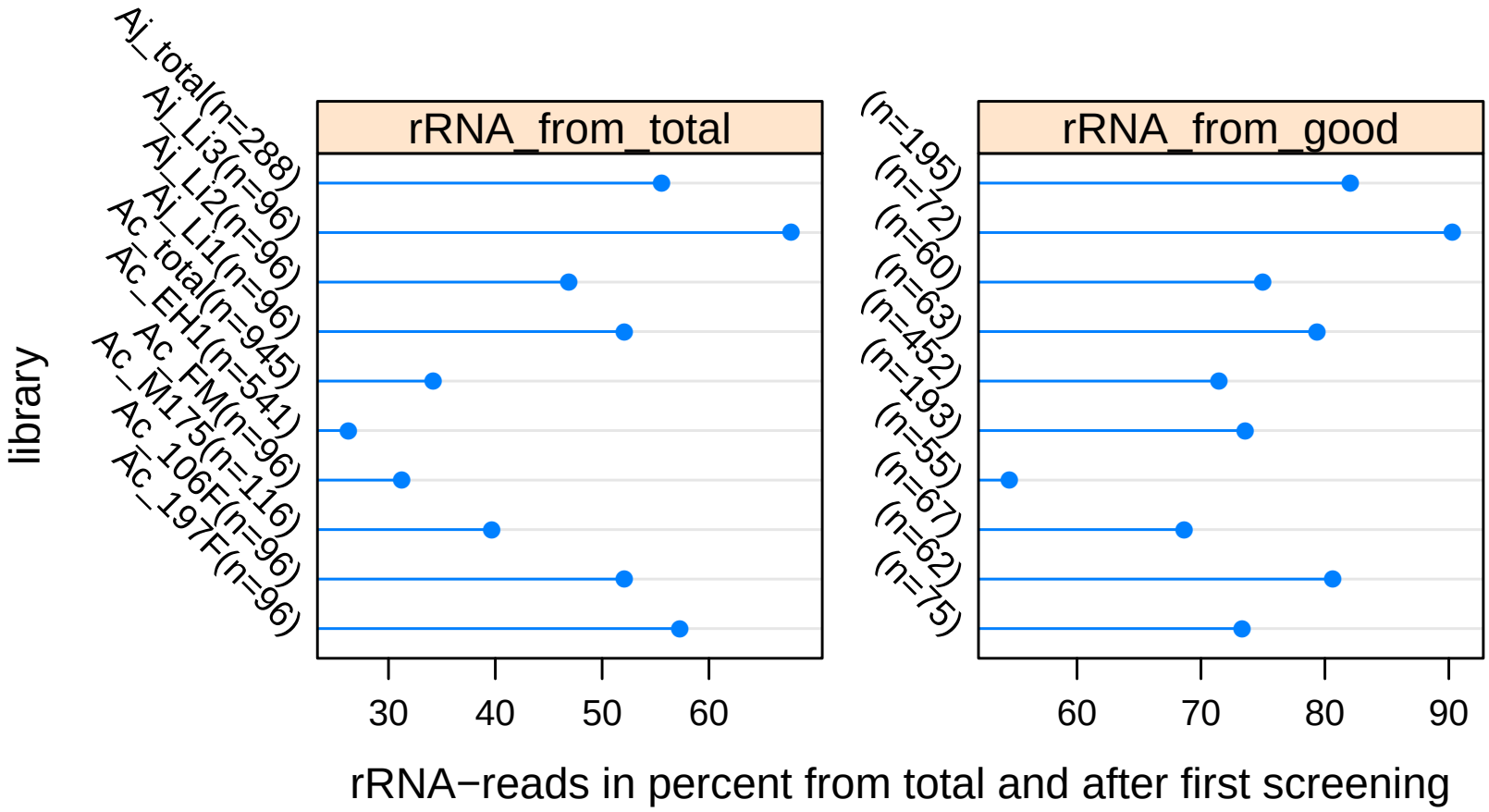
After some R-code (not included here but in the source-file) we can do

```
> xtable(An, caption = "Screening statistics",
+     display = c("s", rep("d", times = 5)),
+     label = "tab:num")
```

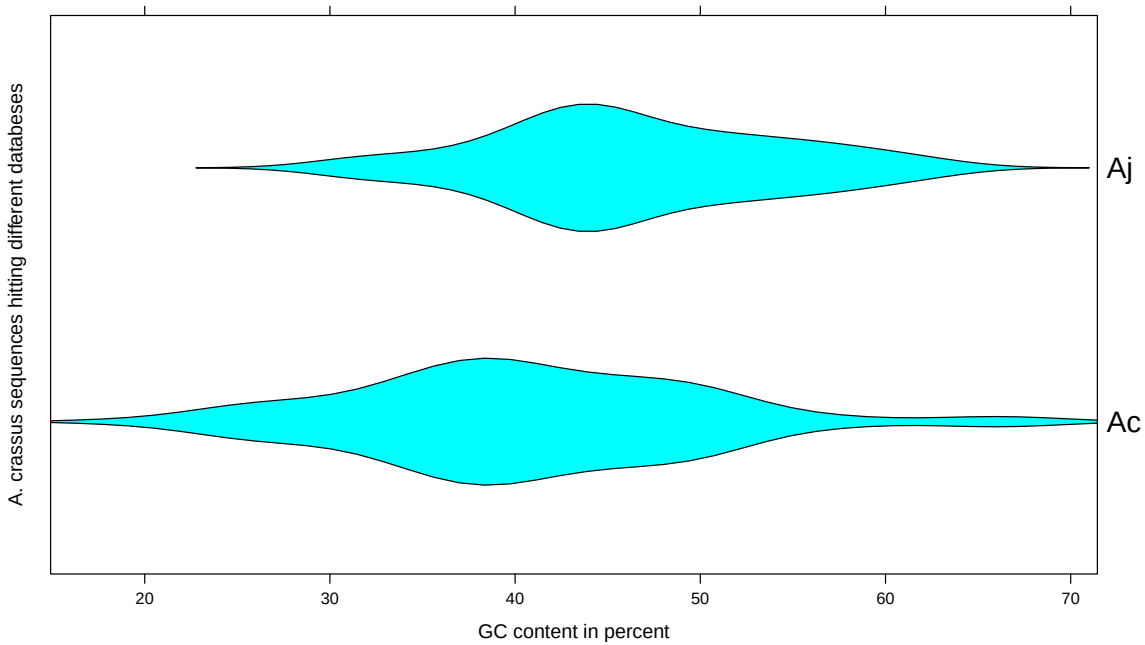
	short	poly	rRNA	fishpep	good2
Ac_197F(n=96)	4	17	55	4	16
Ac_106F(n=96)	25	9	50	2	10
Ac_M175(n=116)	30	19	46	2	18
Ac_FM(n=96)	12	29	30	5	20
Ac_EH1(n=541)	297	51	142	15	36
Ac_total(n=945)	368	125	323	28	100
Aj_Li1(n=96)	10	23	50		13
Aj_Li2(n=96)	10	26	45		15
Aj_Li3(n=96)	9	15	65		6
Aj_total(n=288)	29	64	160		34

Table: Screening statistics

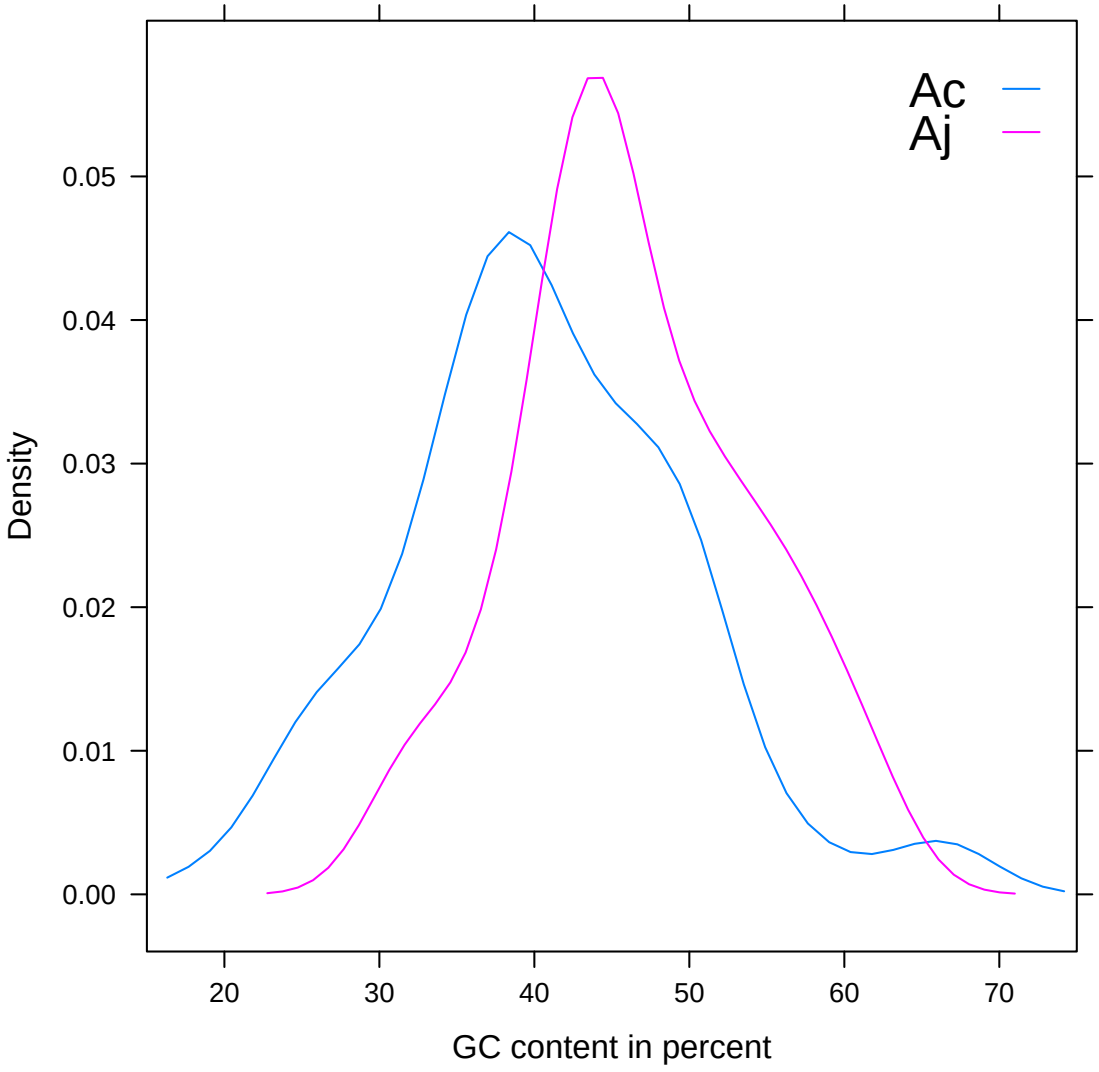
An example involving perl



An example involving perl



An example involving perl



A Next Gen example

- Much more data, but everything less complicated because R-bioconductor instead of perl is used

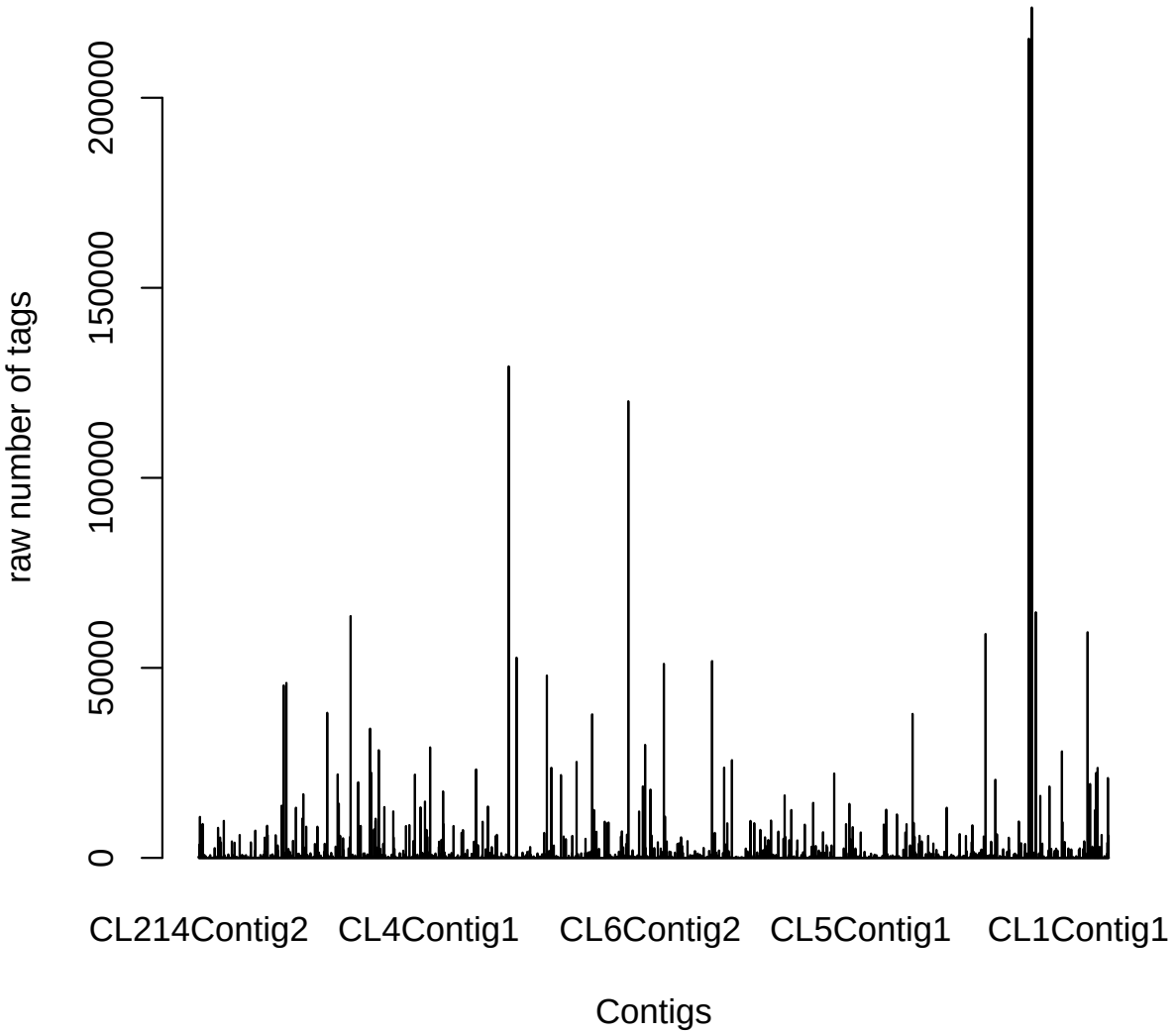
A Next Gen example

- Much more data, but everything less complicated because R-bioconductor instead of perl is used
- I just read in 6201930 tags from a tag-sequencing experiment. 3494662 are mapped to 4077 tgicl- cotigs from 454 using maq. 2836611 to the plus strand 658051 to the minus strand

A Next Gen example

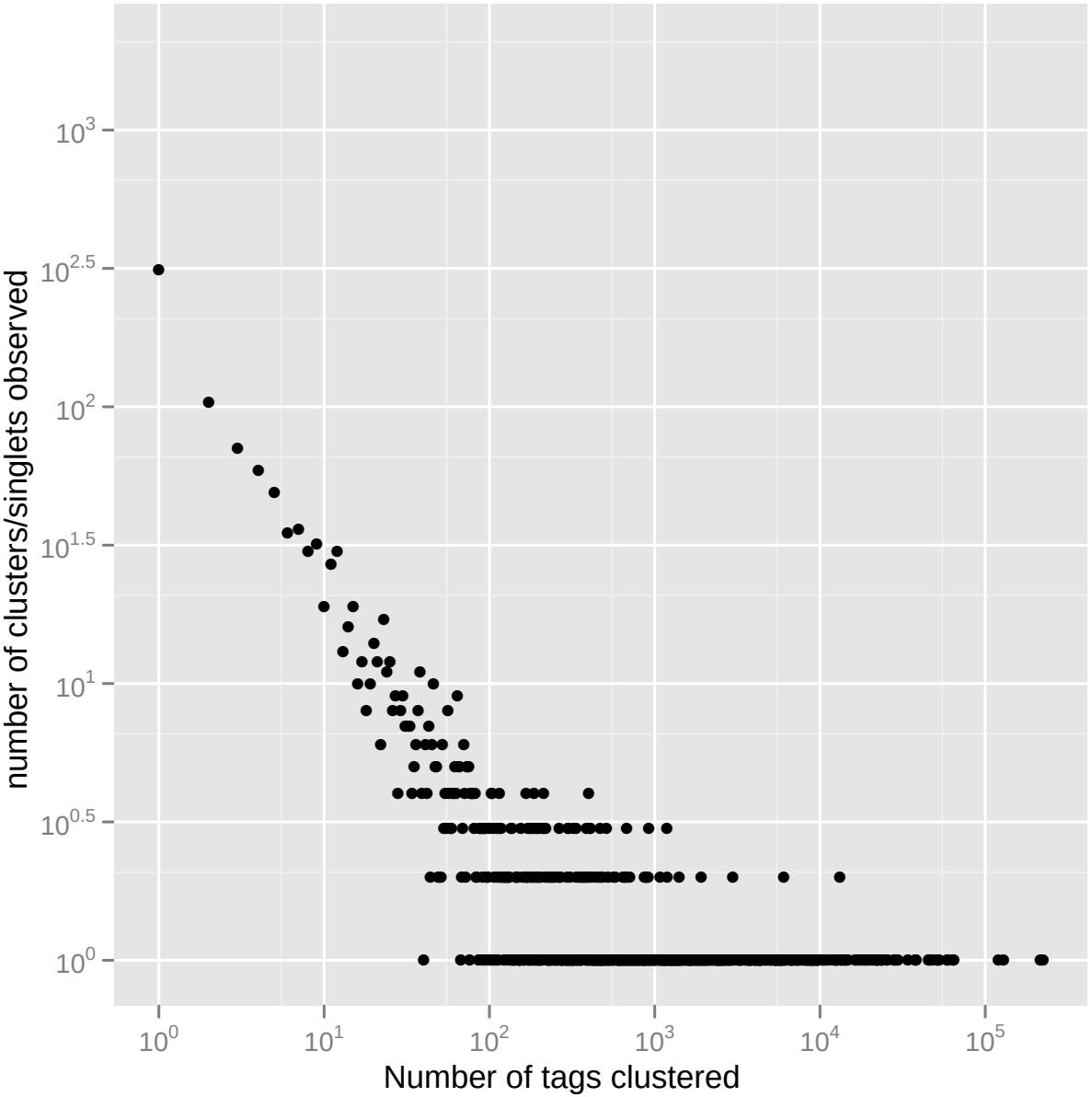
- Much more data, but everything less complicated because R-bioconductor instead of perl is used
- I just read in 6201930 tags from a tag-sequencing experiment. 3494662 are mapped to 4077 tgicl- cotigs from 454 using maq. 2836611 to the plus strand 658051 to the minus strand
- All numbers in this slide (expect 454 ;)) are literally coming from the programm! This is possible because one can include not only figures and tables but als short R-expressions compiling to numbers in text.

A Next Gen example



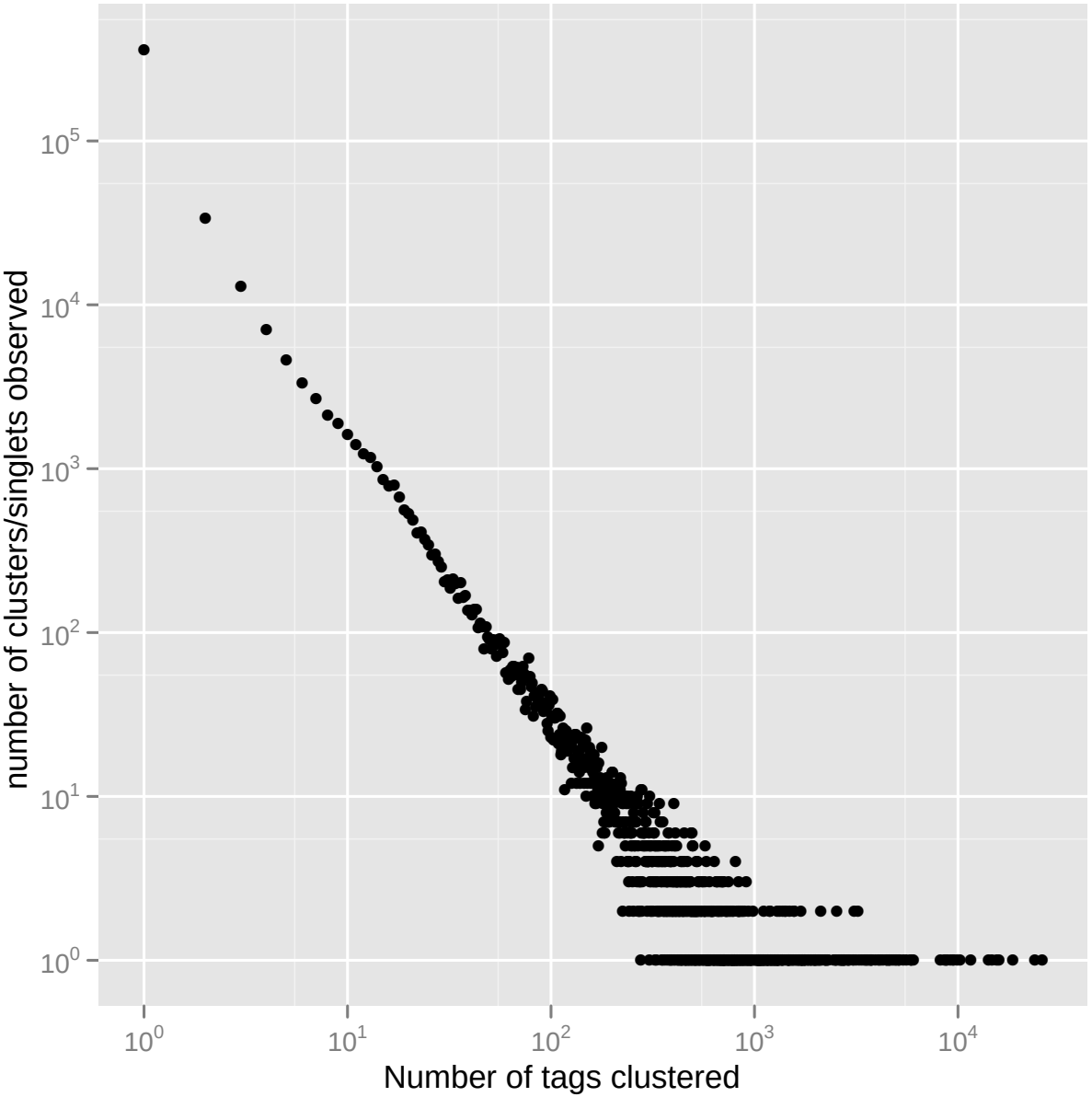
Of the 4077 contigs 2123 had at least one tag mapping 1954 had no tag.

A Next Gen example



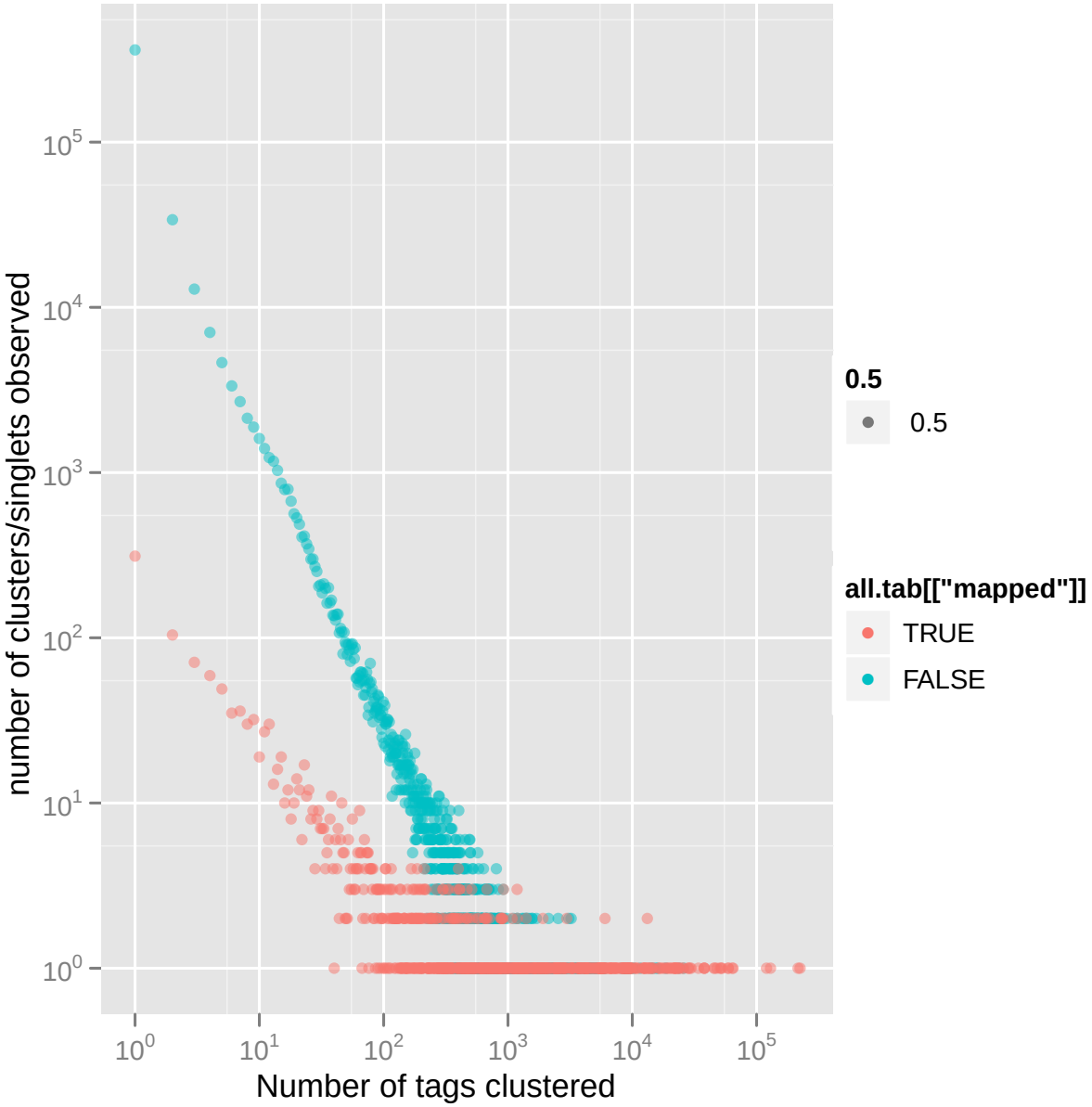
Of the 4077 contigs 2123 had at least one tag mapping 1954 had no tag.

A Next Gen example



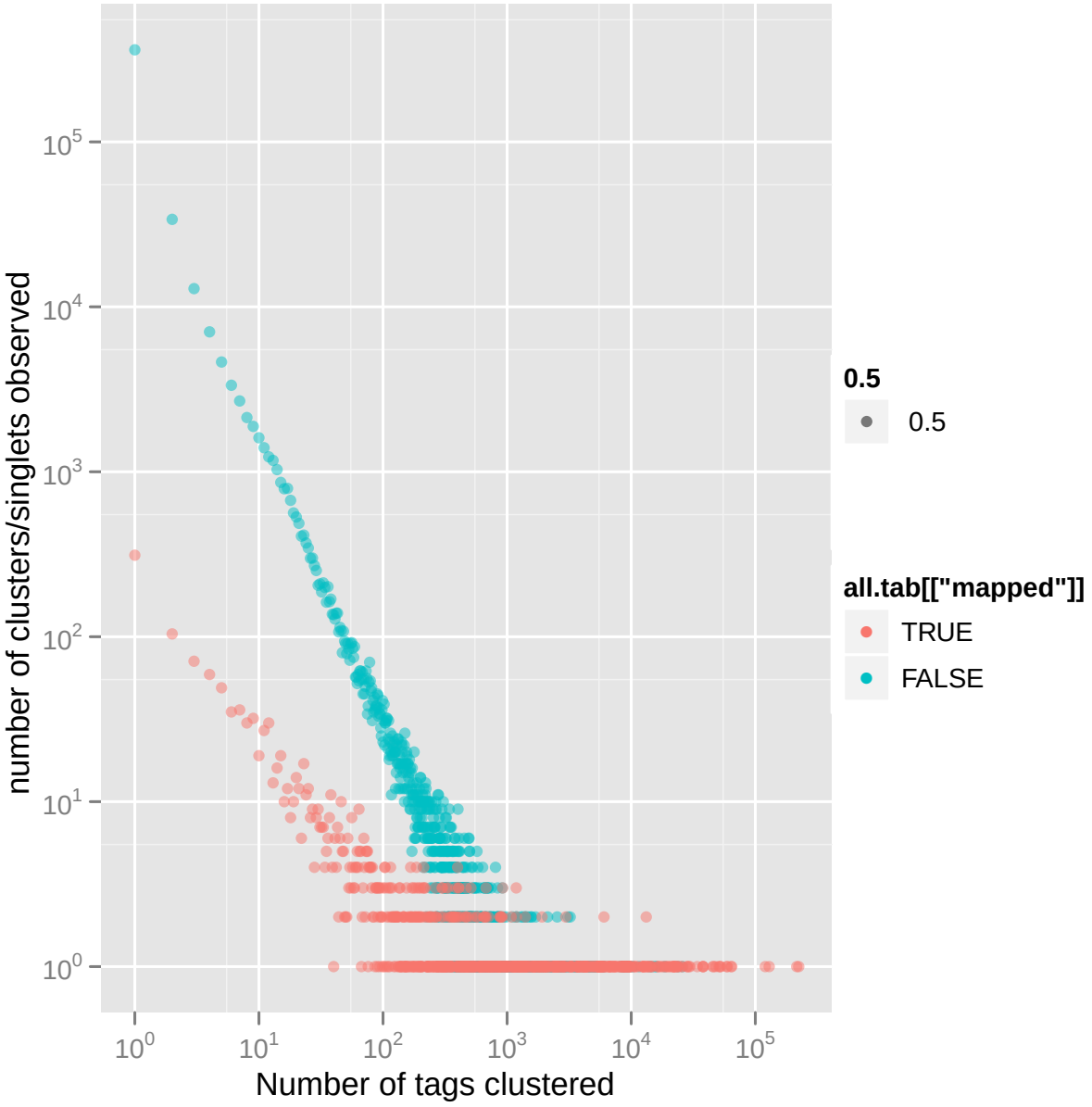
From the 2707268 sequences without hits in the 454 contigs 361468 were singletons. The remaining formed 91470 clusters.

A Next Gen example



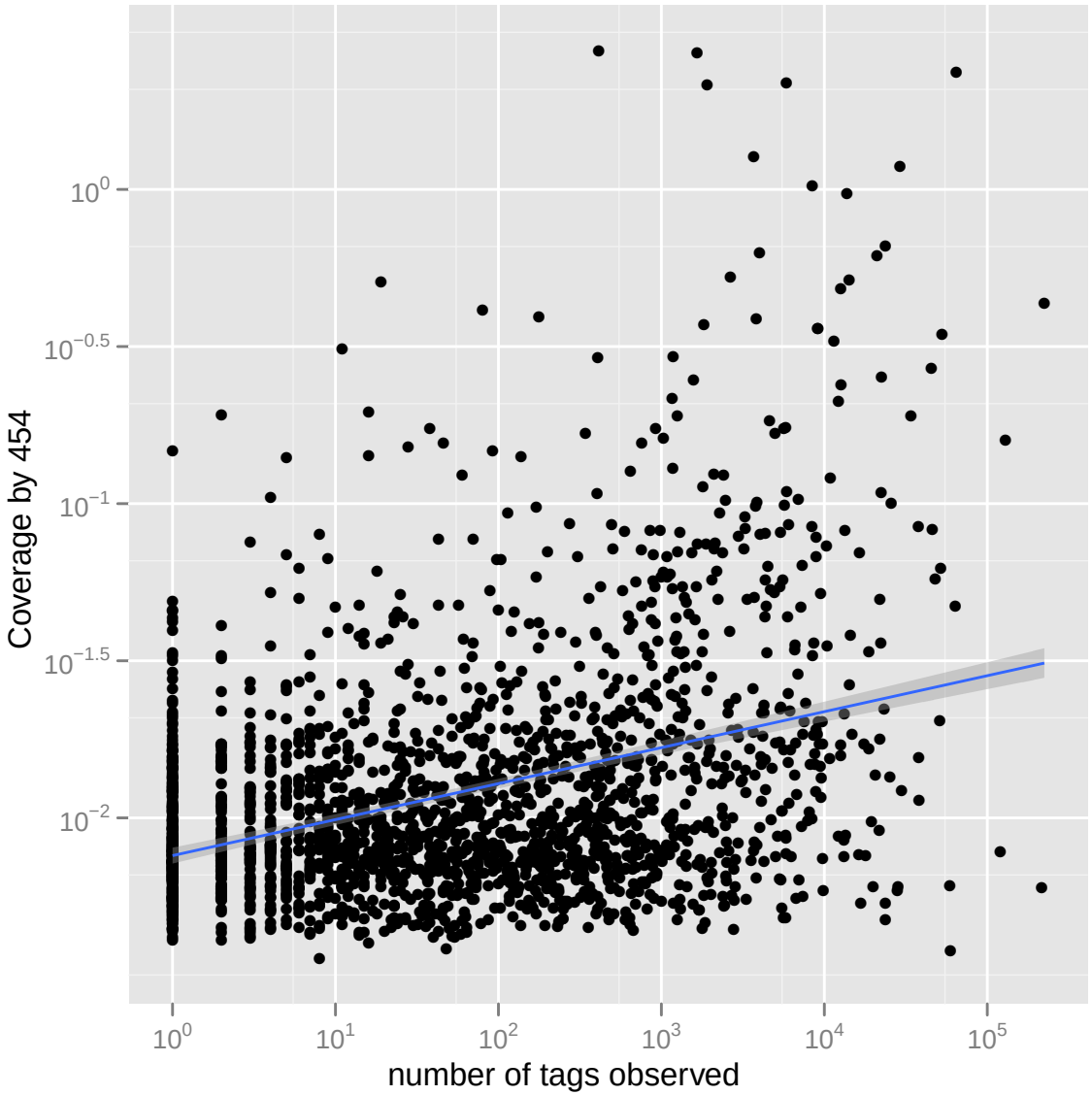
- Of the 4077 contigs 2123 had at least one tag mapping 1954 had no tag.

A Next Gen example



- Of the 4077 contigs 2123 had at least one tag mapping 1954 had no tag.
- From the 2707268 sequences without hits in the 454 contigs 361468 were singletons. The remaining formed 91470 clusters.

A Next Gen example



Caching data

Caching[4] of computations allows R-objects to be stored "on disk". If the object does not change it does not have to be computed again.

To Do

- Minor issues with syntax highlighting etc...

To Do

- Minor issues with syntax highlighting etc...
- Investigate ways to work on client but execute on server (would allow the editor to be used in X-window)

To Do

- Minor issues with syntax highlighting etc...
- Investigate ways to work on client but execute on server (would allow the editor to be used in X-window)
- Caching of R-objects produced by perl-scripts must be made aware of changes in perl-scripts

To Do

- Minor issues with syntax highlighting etc...
- Investigate ways to work on client but execute on server (would allow the editor to be used in X-window)
- Caching of R-objects produced by perl-scripts must be made aware of changes in perl-scripts

To Do

- Minor issues with syntax highlighting etc...
- Investigate ways to work on client but execute on server (would allow the editor to be used in X-window)
- Caching of R-objects produced by perl-scripts must be made aware of changes in perl-scripts
- Investigate sharing of Cache (between machines and files) - Maybe not even desirable

Summary

- Literate programming using R sweave can be used to clue togther code- and markup- language

Summary

- Literate programming using R sweave can be used to clue togther code- and markup- language
- When the system is used as a general working environment it needs some commitment

Summary

- Literate programming using R sweave can be used to clue togther code- and markup- language
- When the system is used as a general working environment it needs some commitment
- Caching of computations opens the way for large projects, where R code executes other time consuming scripts

Summary

- Literate programming using R sweave can be used to clue togther code- and markup- language
- When the system is used as a general working environment it needs some commitment
- Caching of computations opens the way for large projects, where R code executes other time consuming scripts

Summary

- Literate programming using R sweave can be used to clue togther code- and markup- language
- When the system is used as a general working environment it needs some commitment
- Caching of computations opens the way for large projects, where R code executes other time consuming scripts
- Caching in combination with a VCS may make possible a whole PhD-thesis using Next-Gen being written in Noweb/Sweave?


Knuth DE: *Literate Programming*. Stanford, California: Center for the Study of Language and Information 1992, [http://www-cs-faculty.stanford.edu/~knuth/lp.html].


Leisch F: **Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis**. In *Compstat 2002 — Proceedings in Computational Statistics*. Edited by Härdle W, Rönz B, Physica Verlag, Heidelberg 2002:575–580, [http://www.stat.uni-muenchen.de/~leisch/Sweave]. [ISBN 3-7908-1517-9].


Rossini A, Heiberger R, Sparapani R, Maechler M, Hornik K: **Emacs Speaks Statistics: A Multiplatform, Multipackage Development Environment for Statistical Analysis**. *Journal of Computational and Graphical Statistics* 2004, **13**:247–261, [http://ess.r-project.org/index.php?Section=home].


Falcon S: **Caching code chunks in dynamic documents**. *Computational Statistics* 2009, **24**(2):255–261, [http://www.springerlink.com/content/55411257n1473414].