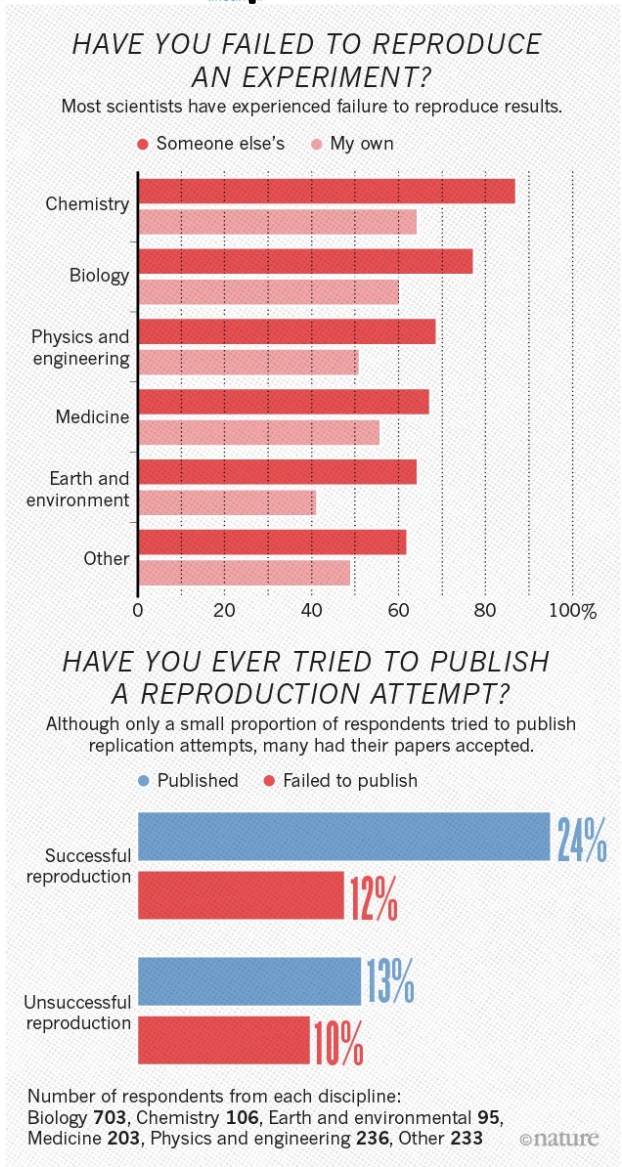
Reproducible Research

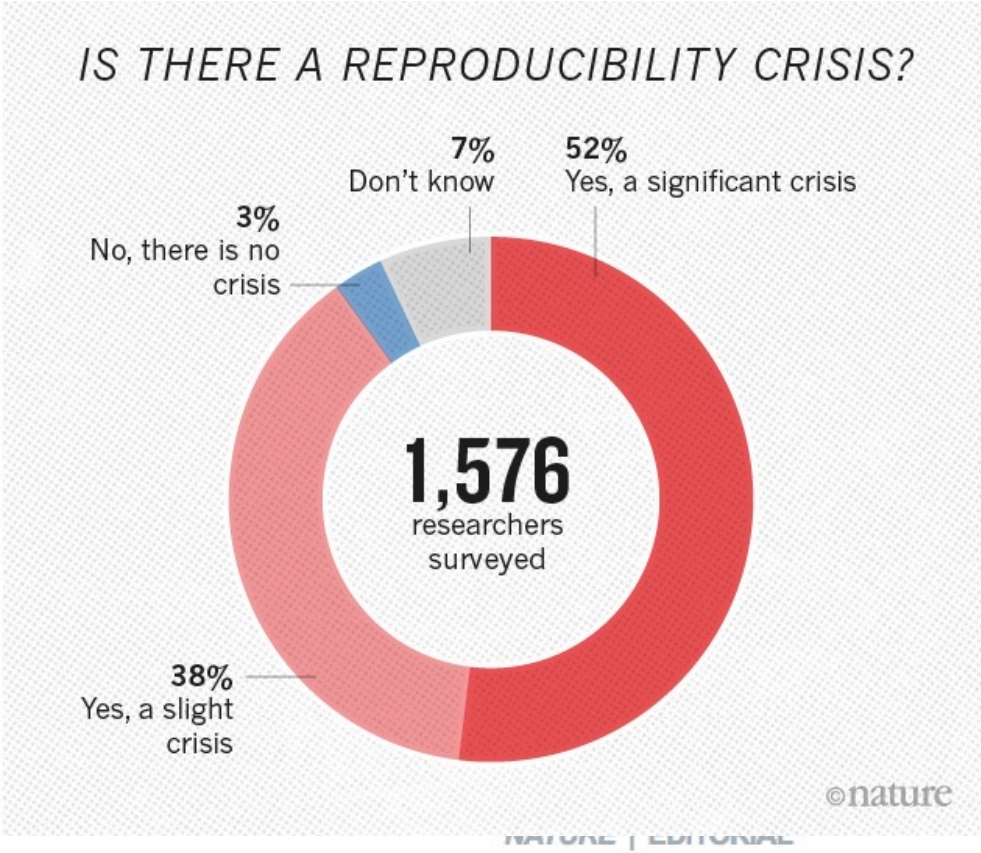
Niklaus Zemp

24 June 2021

Genetic Diversity Centre (GDC)
Bioinformatics
ETH Zurich

Reproducible research





- Imprecision
- Outdated tools
- Outlier removal
- Conceptual flaw
- Manipulating data

NATURE | EDITORIAL

Reality check on reproducibility

A survey of *Nature* readers revealed a high level of concern about the problem of irreproducible results. Researchers, funders and journals need to work together to make research more reliable.

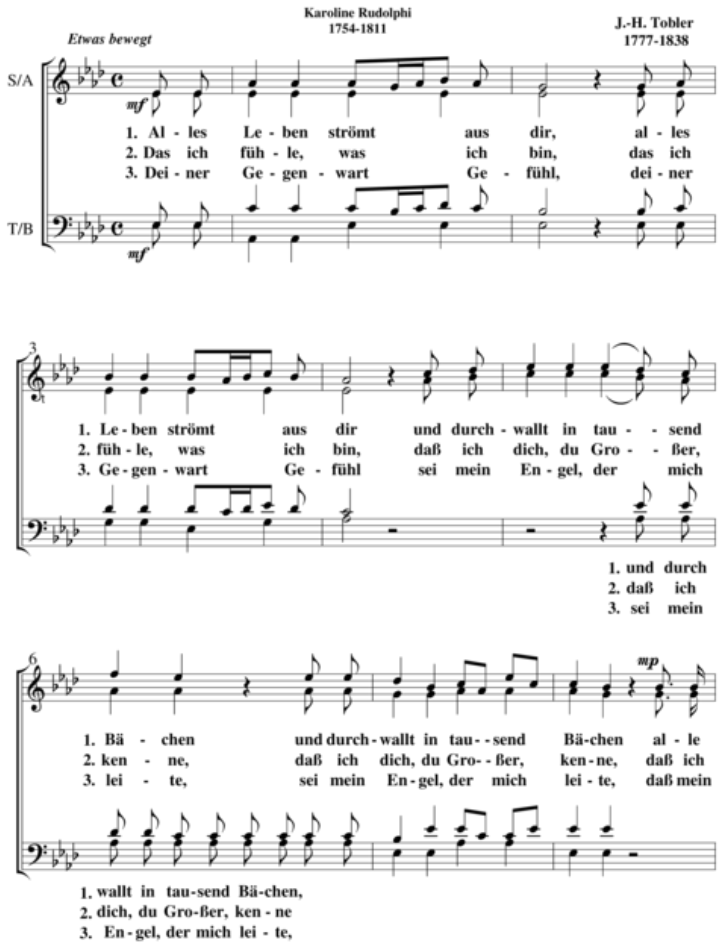
Recipe

Alles Leben strömt aus dir

Karoline Rudolphi
1754-1811

J.-H. Tobler
1777-1838

Etwas bewegt



Ingredients

3 fresh red chillies

2 onions

4 cloves of garlic

4 large plum tomatoes

1 bunch of fresh coriander

4 large free-range chicken legs, skin on

olive oil

2 teaspoons garam masala

1 tablespoon crumbled dried curry leaves

1 tablespoon mustard seeds

2 tablespoons white wine vinegar

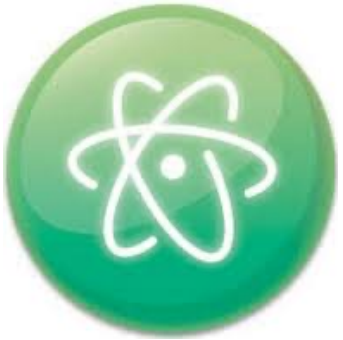
fat-free natural yoghurt

Method

1. Halve the chillies (deseed if you like), peel and finely slice the oni
   the garlic. Quarter the plum tomatoes, and pick the coriander lea
2. Rub the chicken legs all over with a drizzle of oil and the garam m
   large non-stick ovenproof pan.
3. Add another drizzle of oil and fry the chicken over a medium hea
   until lovely and dark golden all over. Be brave and let it get really
   make such a difference to the end result if you get it right at this s
   fat.
4. At this point, preheat the oven to 180ºC/350ºF/gas 4.
5. Next, add the curry leaves, mustard seeds, chillies, onion and garl
   stirring often, for 5 minutes, then add the tomatoes and white wi

   Transfer the pan to the oven. Cook, uncovered, for 50 minutes, or
   cooked through and falling off the bone.
6. Pop the pan on the hob and reduce the liquid until sticky. Scatter
   and serve with the cooling yoghurt. Delicious with rice or couscou

Scientific recipe



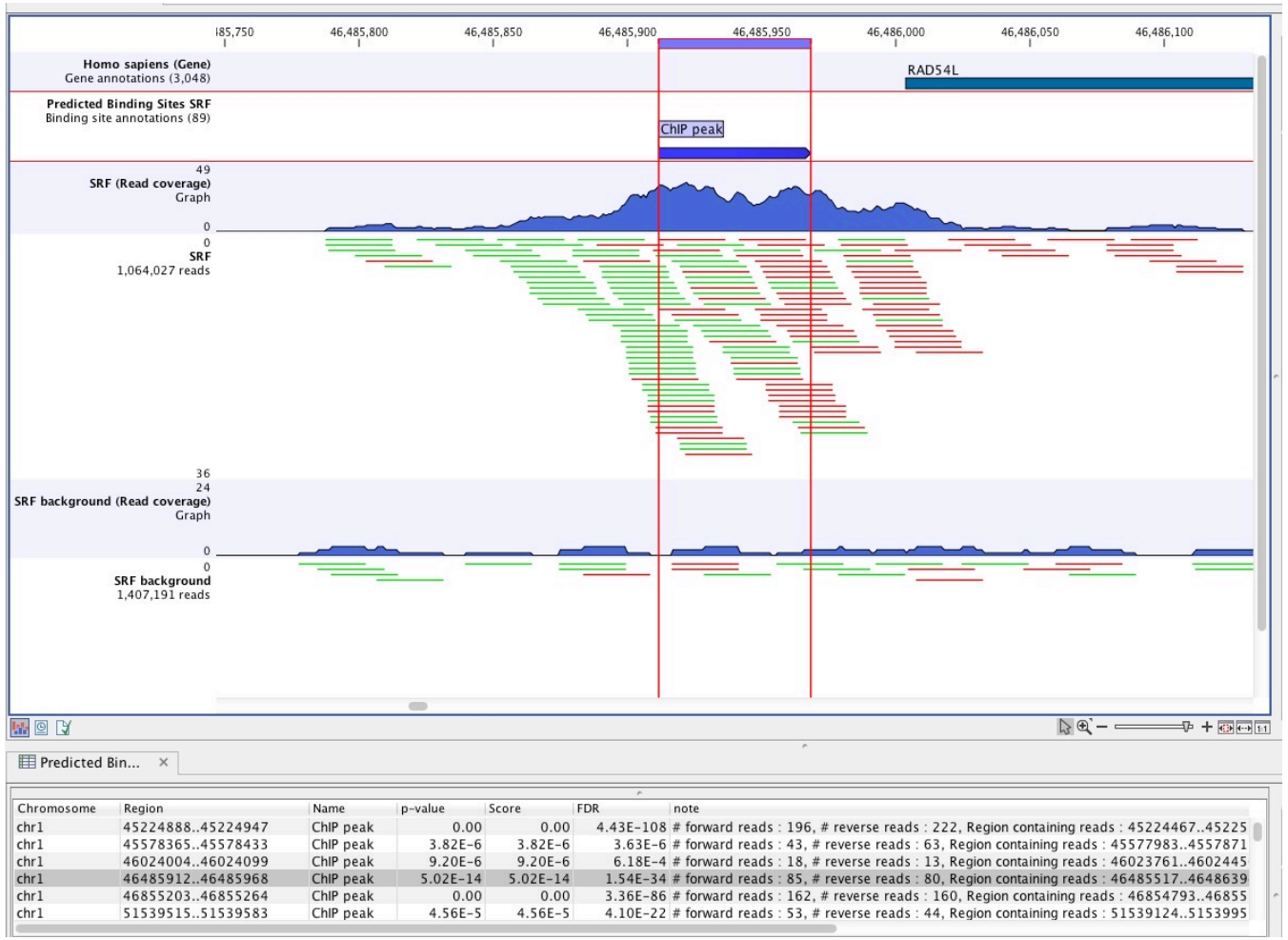
Log-file

```
################################################
Nik Zemp, niklaus.zemp@env.ethz.ch, GDC, ETH Zurich
################################################
#####ddRAD log file, Ivo Widmer, p432
################################################

#######################################################
####Download data
#######################################################
module load eth_proxy
bsub -n1 -W 4:00 -R "rusage[mem=1000]" \
"wget -r --no-parent --reject="index.htm*" http://gc3fstorage.u
bsub -n1 -W 4:00 -R "rusage[mem=1000]" \
"wget -r --no-parent --reject="index.htm*" http://gc3fstorage.u
bsub -n1 -W 4:00 -R "rusage[mem=1000]" \
"wget -r --no-parent --reject="index.htm*" http://gc3fstorage.u
bsub -n1 -W 4:00 -R "rusage[mem=1000]" \
```

script

```
#!/bin/bash
#BSUB -J "processradtags"
#BSUB -R "rusage[mem=10000]"
#BSUB -n 1
#BSUB -W 24:00


module load gcc/4.9.2 gdc perl/5.18.4
export PATH=$PATH:/cluster/project/gdc/shared/tools/stacks-1.48
#module load gcc/4.8.2 gdc perl/5.18.4 stacks/1.40
source /cluster/apps/gdc/perl5/etc/bashrc



mkdir samples

process_radtags -i gzfastq -f /cluster/project/gdc/people/buckleyj/Hiseq_
```

GUI tools



Chromosome	Region	Name	p-value	Score	FDR	note
chr1	45224888..45224947	ChIP peak	0.00	0.00	4.43E-108	# forward reads : 196, # reverse reads : 222, Region containing reads : 45224467..45225
chr1	45578365..45578433	ChIP peak	3.82E-6	3.82E-6	3.63E-6	# forward reads : 43, # reverse reads : 63, Region containing reads : 45577983..4557871
chr1	46024004..46024099	ChIP peak	9.20E-6	9.20E-6	6.18E-4	# forward reads : 18, # reverse reads : 13, Region containing reads : 46023761..4602445
chr1	46485912..46485968	ChIP peak	5.02E-14	5.02E-14	1.54E-34	# forward reads : 85, # reverse reads : 80, Region containing reads : 46485517..4648639
chr1	46855203..46855264	ChIP peak	0.00	0.00	3.36E-86	# forward reads : 162, # reverse reads : 160, Region containing reads : 46854793..46855
chr1	51539515..51539583	ChIP peak	4.56E-5	4.56E-5	4.10E-22	# forward reads : 53, # reverse reads : 44, Region containing reads : 51539124..5153995

geneious

CLC Genomics Workbench

Command-line version

bwa mem Ref reads > alignment.sam

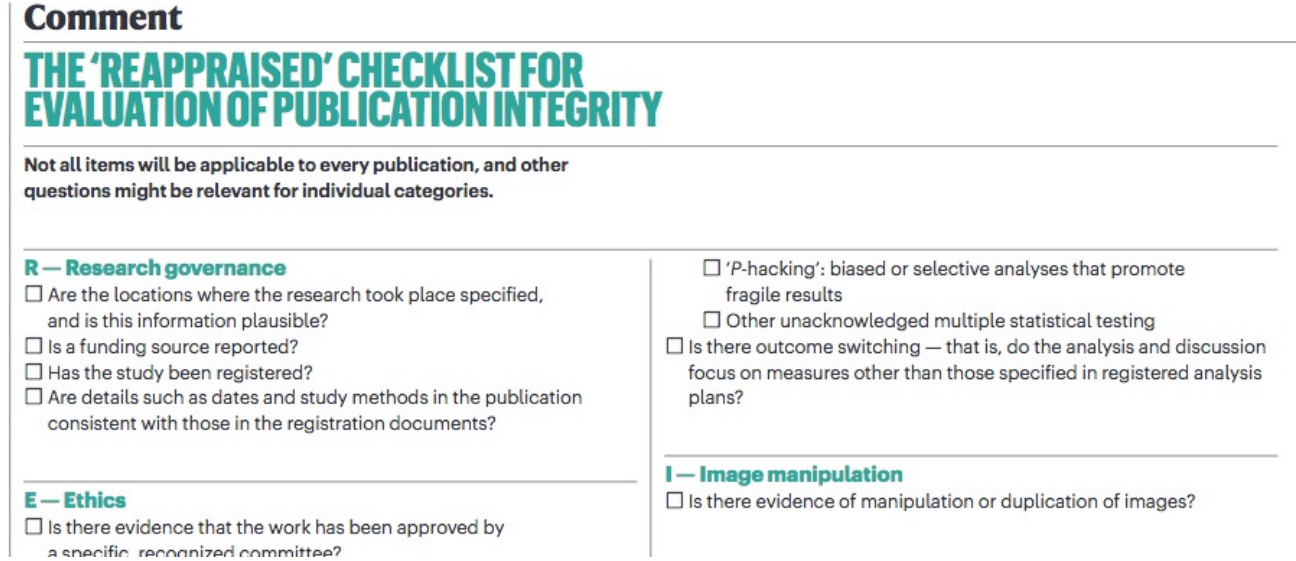
Scientific recipe

- (original) author or source
- your name
- date
- version of the tool
- version of the script
- reproducible code with comments (more comments than code)
- use style guides
- syntax coloring
- use always the same file names and structure

Publication

- keep all scripts and raw data
- use default settings or mention if not
- provide version information
- provide commands in supplementary material
- deposit scripts on github/gitlab
- Check-list

https://www.nature.com/articles/d41586-019-03959-6



Comment

THE 'REAPPRAISED' CHECKLIST FOR EVALUATION OF PUBLICATION INTEGRITY

Not all items will be applicable to every publication, and other questions might be relevant for individual categories.

R — Research governance

- ☐ Are the locations where the research took place specified, and is this information plausible?
- ☐ Is a funding source reported?
- ☐ Has the study been registered?
- ☐ Are details such as dates and study methods in the publication consistent with those in the registration documents?

E — Ethics

- ☐ Is there evidence that the work has been approved by a specific, recognized committee?
- ☐ 'P-hacking': biased or selective analyses that promote fragile results
- ☐ Other unacknowledged multiple statistical testing
- ☐ Is there outcome switching — that is, do the analysis and discussion focus on measures other than those specified in registered analysis plans?

I — Image manipulation

- ☐ Is there evidence of manipulation or duplication of images?

Example

Do not get fooled by nice figures!

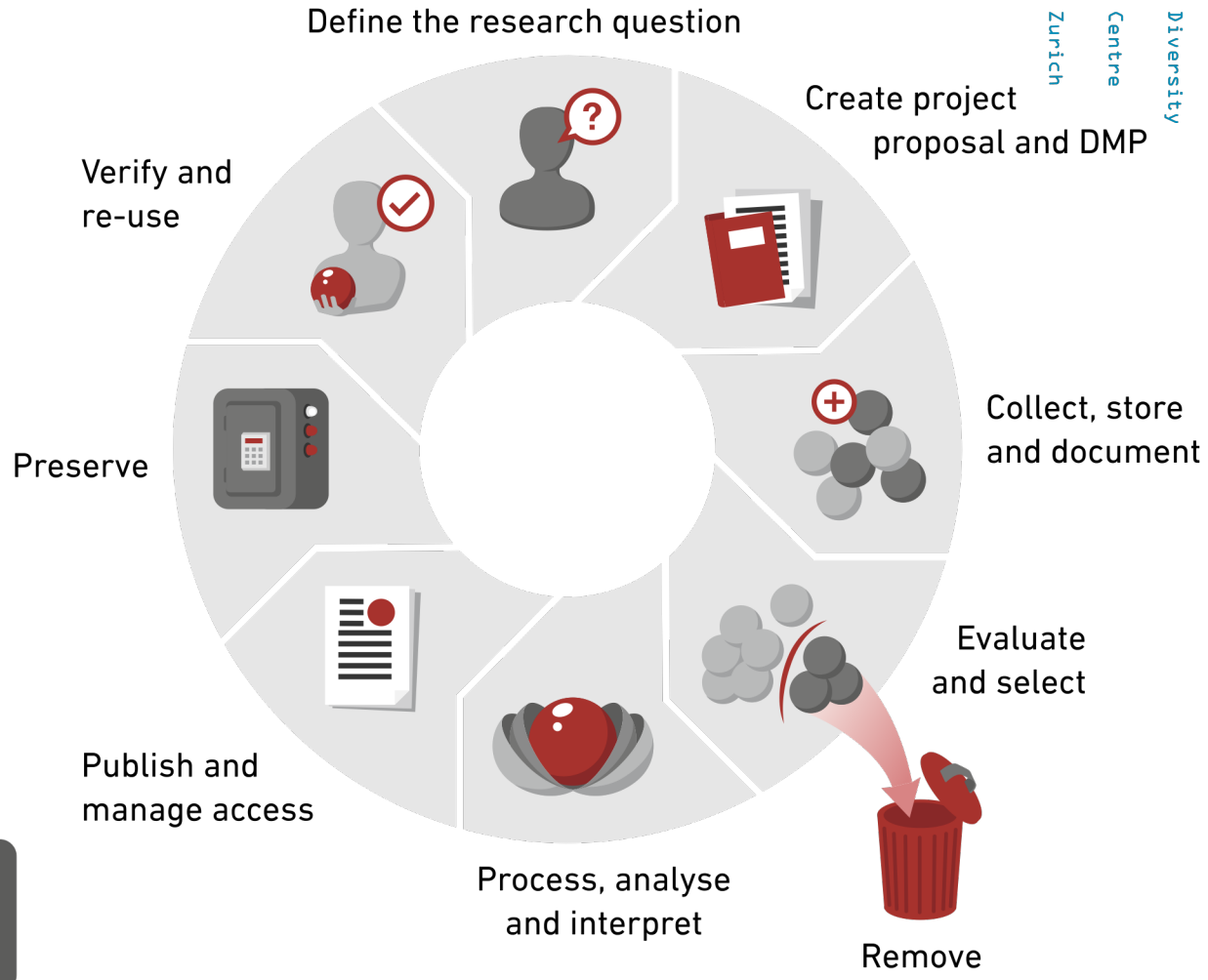
Transcriptome assembly and polymorphism detection [redacted]

Carlsbad, CA, USA). One sequencing run was carried out in an Illumina platform through 100 bp paired-end reads. Trimming was carried out with software Trimmomatic (Bolger *et al.*, 2014). Then, *S. ciliata* transcriptome was aligned with the genome of *Silene latifolia* (GenBank reference: GCA_900095335.1) using BWA software (Li and Durbin, 2010).

Herbarium sheet	Mapping (%)	Total sequences ($\times 10^6$)	H_O	F_i	Tajima's π^a
MA880768	45	27.2	0.249	0.025	0.364×10^{-3}
MA880763	43	34.6	0.286	0.053	0.358×10^{-3}
MA880766	44.6	27.5	0.120	0.120	0.326×10^{-3}
MA880765	37.9	26.2	0.270	0.105	0.350×10^{-3}
MA880769	40	34.6	0.309	−0.026	0.383×10^{-3}
MA880764	38.29	34.5	0.288	0.045	0.351×10^{-3}

Data management planning	Active data management	Publishing and registering	Long-term preservation

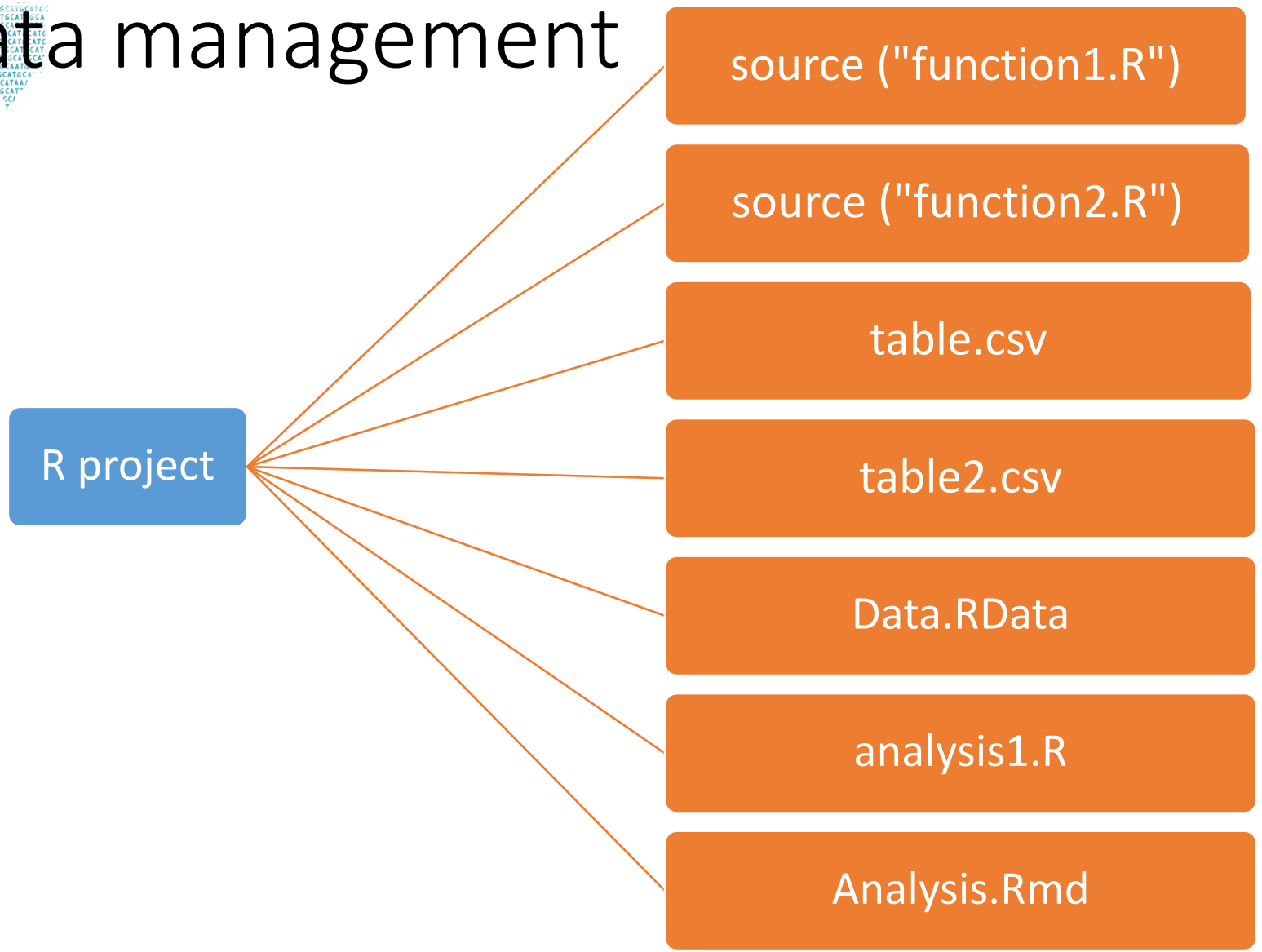
Standards and guidelines of the research funding agencies and the ETH Zurich



ETH Library

Genetic Diversity Centre Zurich

Data management



R with style

The tidyverse style guide

Hadley Wickham

https://style.tidyverse.org/index.html

Google R-style

https://google.github.io/styleguide/Rguide.xml

Styler

styler

The goal of styler is to provide non-invasive pretty-printing of R source code while adhering to the tidyverse formatting rules. styler can be customized to format code according to other style guides too.

Installation

You can install the package from CRAN:

```
install.packages("styler")
```

Internal variables and functions

```
2  ##internal variables
3  pi<-1
4  mean<-c(1,2)
5  norm <- 1
6
```

Some examples

```
# Good
day_one
day_1

# Bad
DayOne
dayone
```

```
# Good
x <- 5

# Bad
x = 5
```

```
# Good
if (y < 0) {
  stop("Y is negative")
}
```

```
# Bad
if (y < 0) stop("Y is negative")
```

```
# Good
"Text"
'Text with "quotes"'
'<a href="http://style.tidyverse.org">A link</a>'

# Bad
'Text'
'Text with "double" and \'single\' quotes'
```

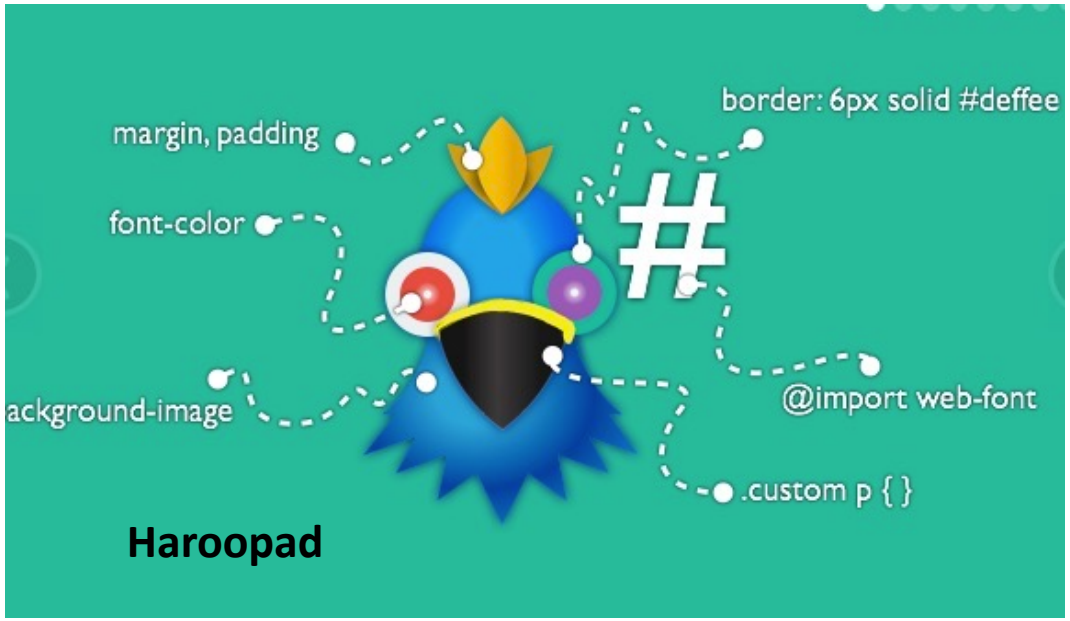
Reproducible Research

Markdown

Markdown

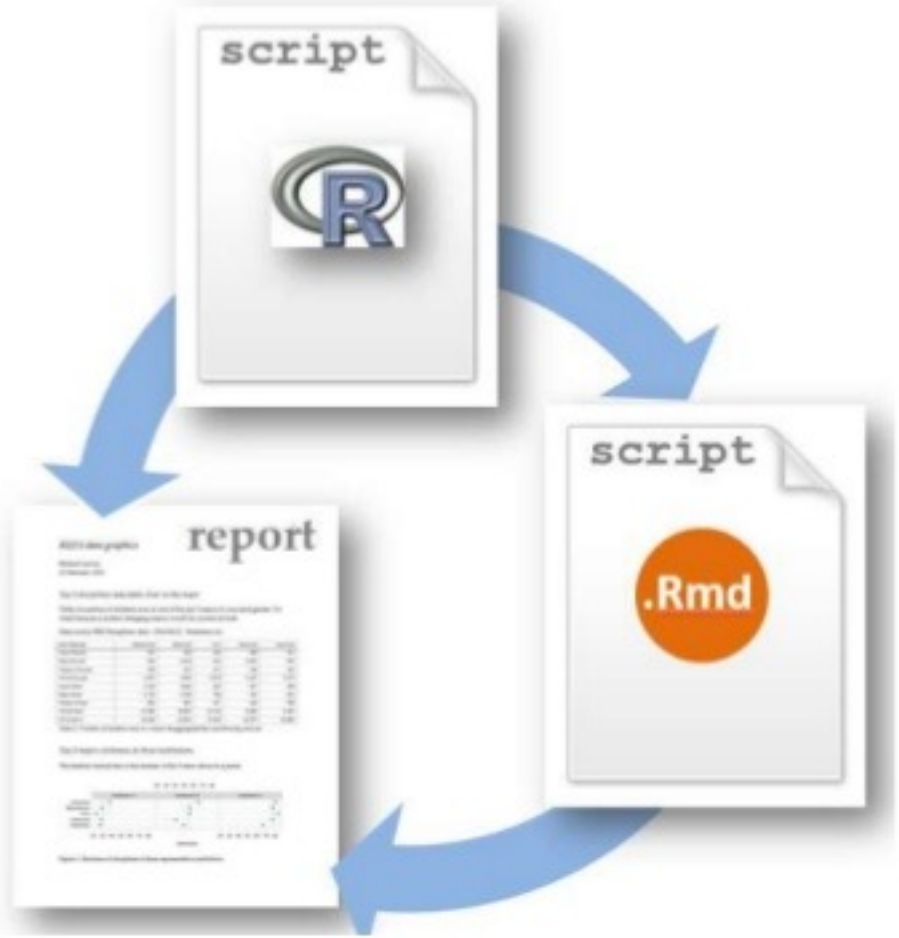
Slightly modified based on dDocent Version 2.6.1; overlapping paired-end reads; September 2018; Nik Zemp based on the Tutorial

```Bash
bsub -n 2 -W 4:00 -R "rusage[mem=5000]" -Is bash

module load gcc/4.8.2 gdc python/2.7.11 java/1.8.0_73 perl/5.18.4 freebayes/0.9.20 trimmomatic/0.35 bwa/0.7.12 cd-hit/
export PATH="$PATH:/cluster/project/gdc/shared/tools/pear-0.9.6-bin-64"
export PATH="/cluster/project/gdc/shared/tools/seqtk:$PATH"
```



Cheat sheet

Reproducible Research



rmarkdown

https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf

Reproducible Research
RegEx

Replace pattern

>E1_L96

AATTATTACTTTATGACACT

>E2_L119

CGAATTCGTCATTTGAAACCGATTCGG
CTAGAATT

>E4_L96

TTTTACTTACATGGTGAAAAAATAGAAT
ACGTATTCTCTGCCAAGATTCATTAACT
CAAAAGAGAATTTTTTGAGTTAATGCA
GAGGATACGAATT

>E1

AATTATTACTTTATGACACT

>E2

CGAATTCGTCATTTGAAACCGATTCGG
CTAGAATT

>E4

TTTTACTTACATGGTGAAAAAATAGAAT
ACGTATTCTCTGCCAAGATTCATTAACT
CAAAAGAGAATTTTTTGAGTTAATGCA
GAGGATACGAATT

Regular expression (RegEx)

Groups and Ranges	
.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Ranges are inclusive.

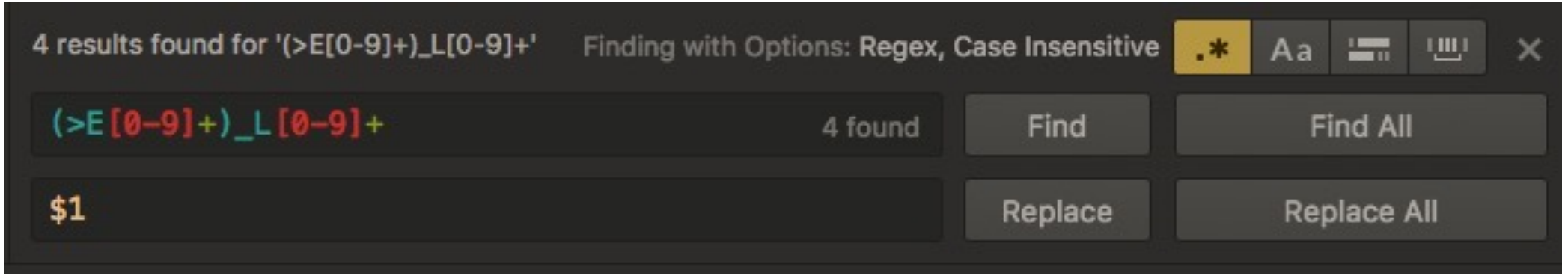
Quantifiers			
*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

Add a ? to a quantifier to make it ungreedy.

Anchors	
^	Start of string, or start of line in multi-line pattern
\A	Start of string
$	End of string, or end of line in multi-line pattern
\Z	End of string
\b	Word boundary
\B	Not word boundary
\<	Start of word
\>	End of word

RegEx with Atom

```
>E1_L96
AATTATTACTTTATGACACTGACACTGACACTGACACTGACATACAGAAATGAATTAAGTCAAGAACCAAAGCGGAGGAAGCGCTTCTAGAGAATT
>E2_L119
CGAATTCGTCATTTGAAACCGATTCAATGAGTTTTAGACTTGAGTTCACGAAGAAGTTTAATGAACTTAAAAACACCCTAGTTCTACTCTTCAAAT
>E4_L96
TTTTACTTACATGGTGAAAAAATAGAATACGTATTCTCTGCCAAGATTCATTAACTCAAAAGAGAATTTTTTGAGTTAATGCAGAGGATACGAATT
>E14_L135
CGAATGTCTCCTGGGACTTCTTGGTAGCTTGACCTTCATTCCACATCGTCTGCATTAACTCAAAAGAGTATCTTTTGAGTTAATACCTATCGGCTG
```



(>E[0-9]+)_L[0-9]+

Take home message

- Do reproducible research
- Markdown is helpful
- Remember RegEx