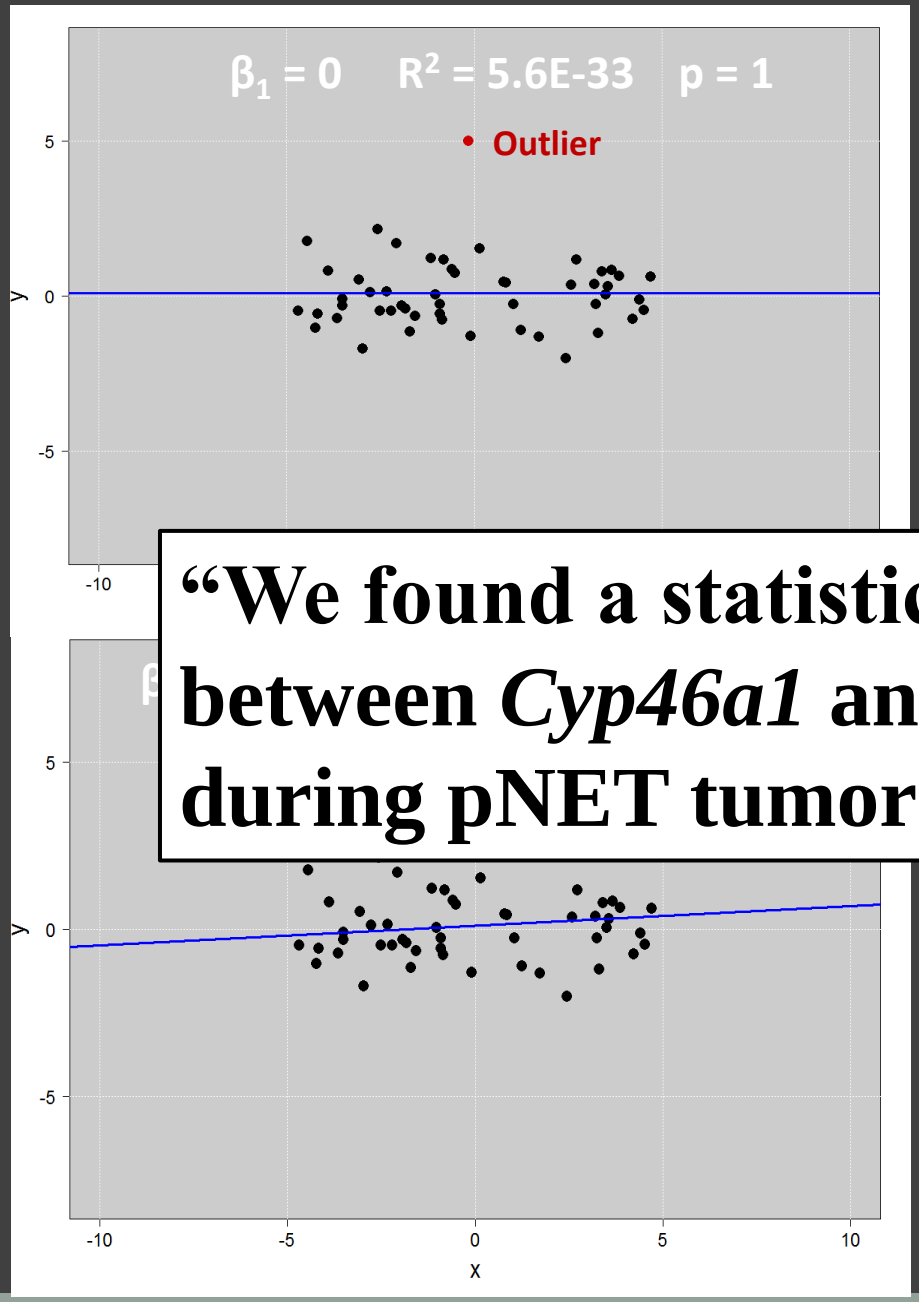
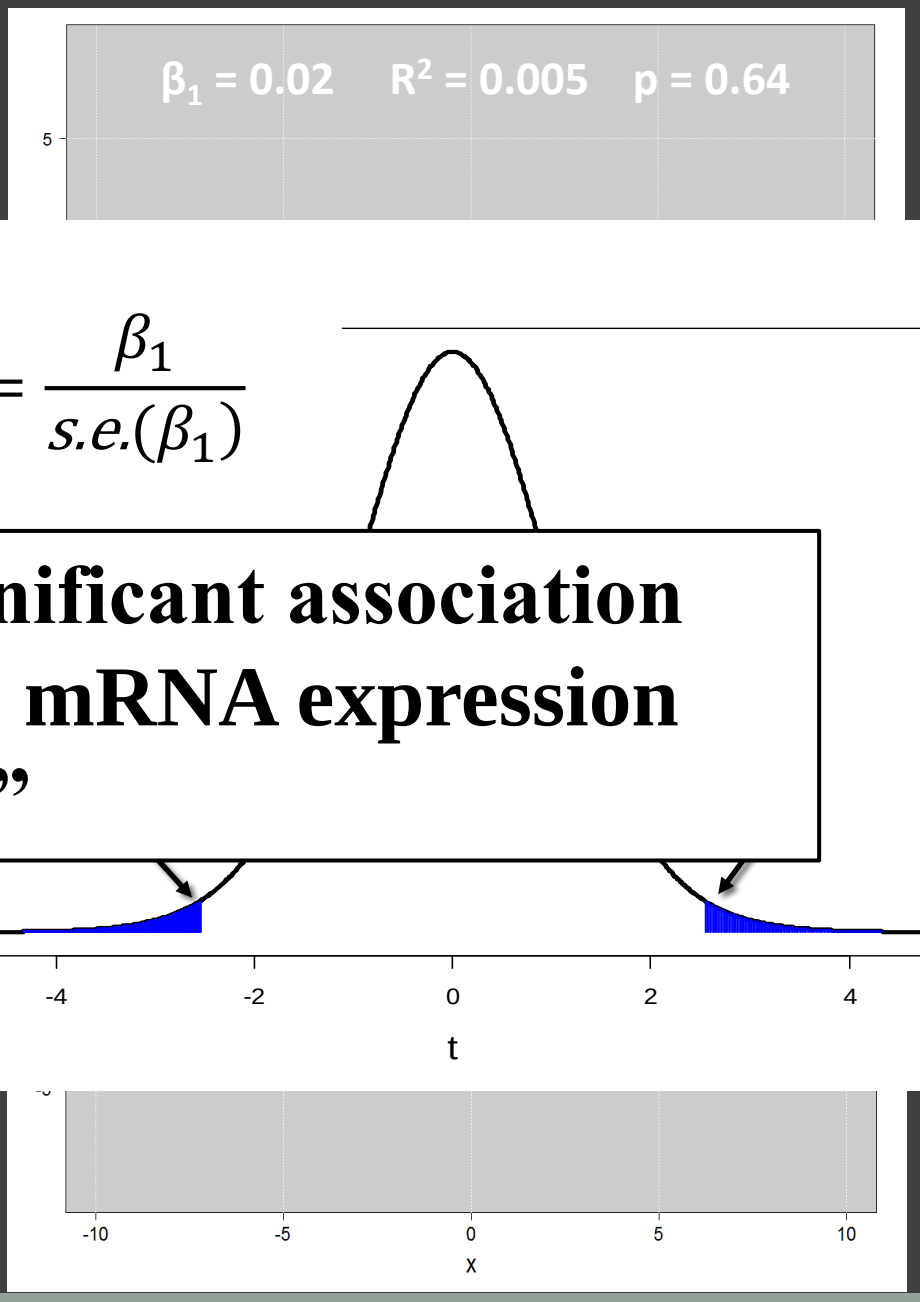
Meaningful and reproducible statistics:

Does my data hold what it promises?

Andrej-Nikolai Spiess, PhD
Dpt. Of Molecular Andrology,
University Hospital Hamburg-Eppendorf





$$t = \frac{\beta_1}{s.e.(\beta_1)}$$

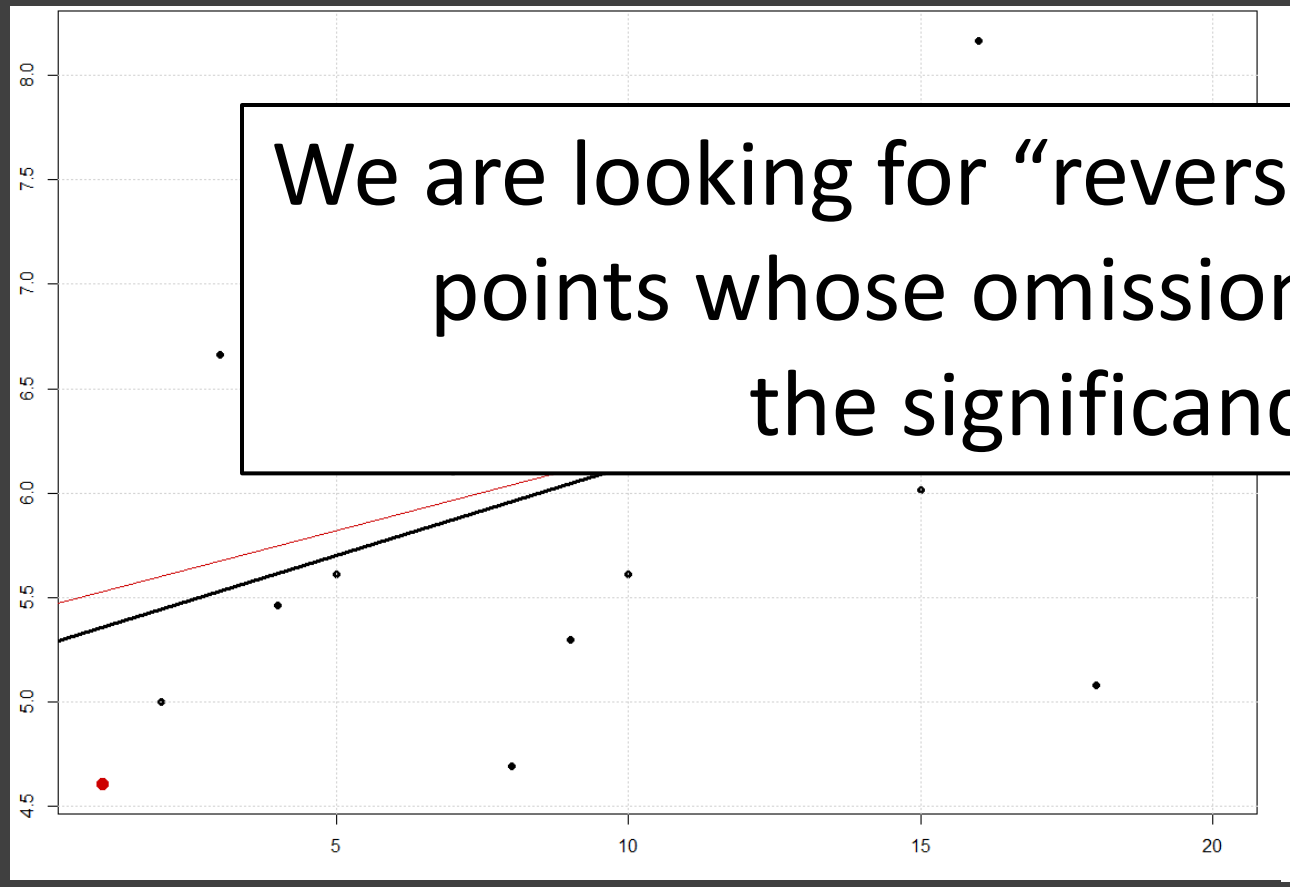
"We found a statistically significant association between *Cyp46a1* and *Abcg1* mRNA expression during pNET tumorigenesis"

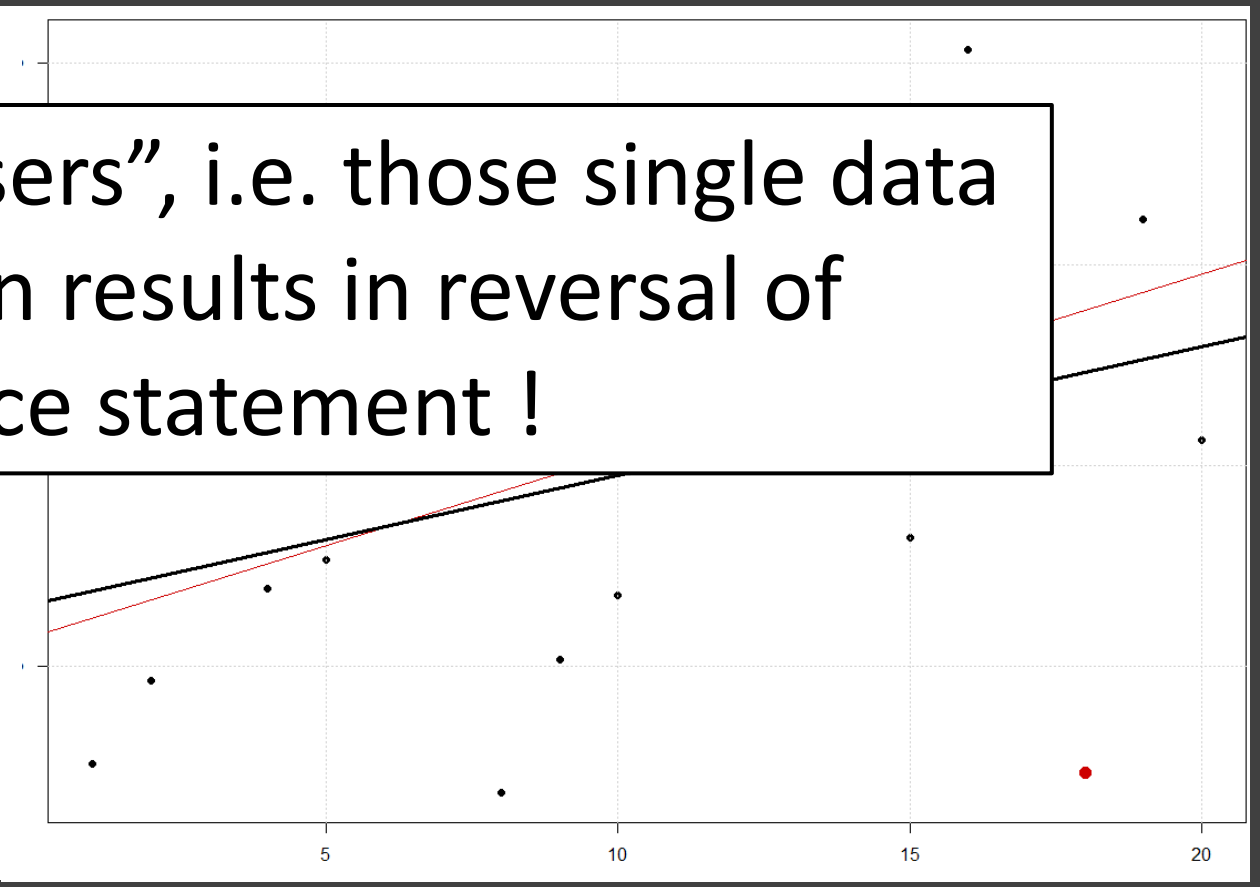
0.05

$p = 0.02$ → $p = 0.063$

0.05

$p = 0.012$ ← $p = 0.115$





We are looking for “reversers”, i.e. those single data points whose omission results in reversal of the significance statement !

	Formula	Cut-off	
Leverage	$h_{ii} = x_i'(X'X)^{-1}x_i$	$h_{ii} > 2\bar{h} = \frac{2k}{n}$	Outlyingness on the X-axis
Cook's D	$D_i = \frac{(b_i - b_{(i)})'X'X(b_i - b_{(i)})}{}$	$D_i > F_{0.5}(k, n-k)$	Moved vector of fitted values
DFBETAS	DFE		ange of regression efficients
	$\sqrt{S_{(i)}C_{jj}}$		
	$\sqrt{S_{(i)}^2 h_{ii}}$		
Covratio	$COVRATIO_i = \frac{}{MS_{res}^k\left(\frac{}{1-h_{ii}}\right)}$	$COVRATIO_i > 1 + \frac{3k}{n}$	Precision of estimation
Stud. Res.	$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}(1 + x_i^T(X_{(i)}^T X_{(i)})x_i)^{1/2}}$	$t_i > Q_t(0.975, n - p - 1)$	Outlyingness on the Y-axis

None of these measures relates to changes in the *p*-value !!

$$\Delta p = p_{\beta 1} - p_{\beta 1(i)} = 2\left(1 - P_t(t_{\beta 1}, \nu)\right) - 2\left(1 - P_t(t_{\beta 1(i)}, \nu - 1)\right)$$

$$\alpha \in \left[p_{\beta 1}, p_{\beta 1(i)}\right]$$

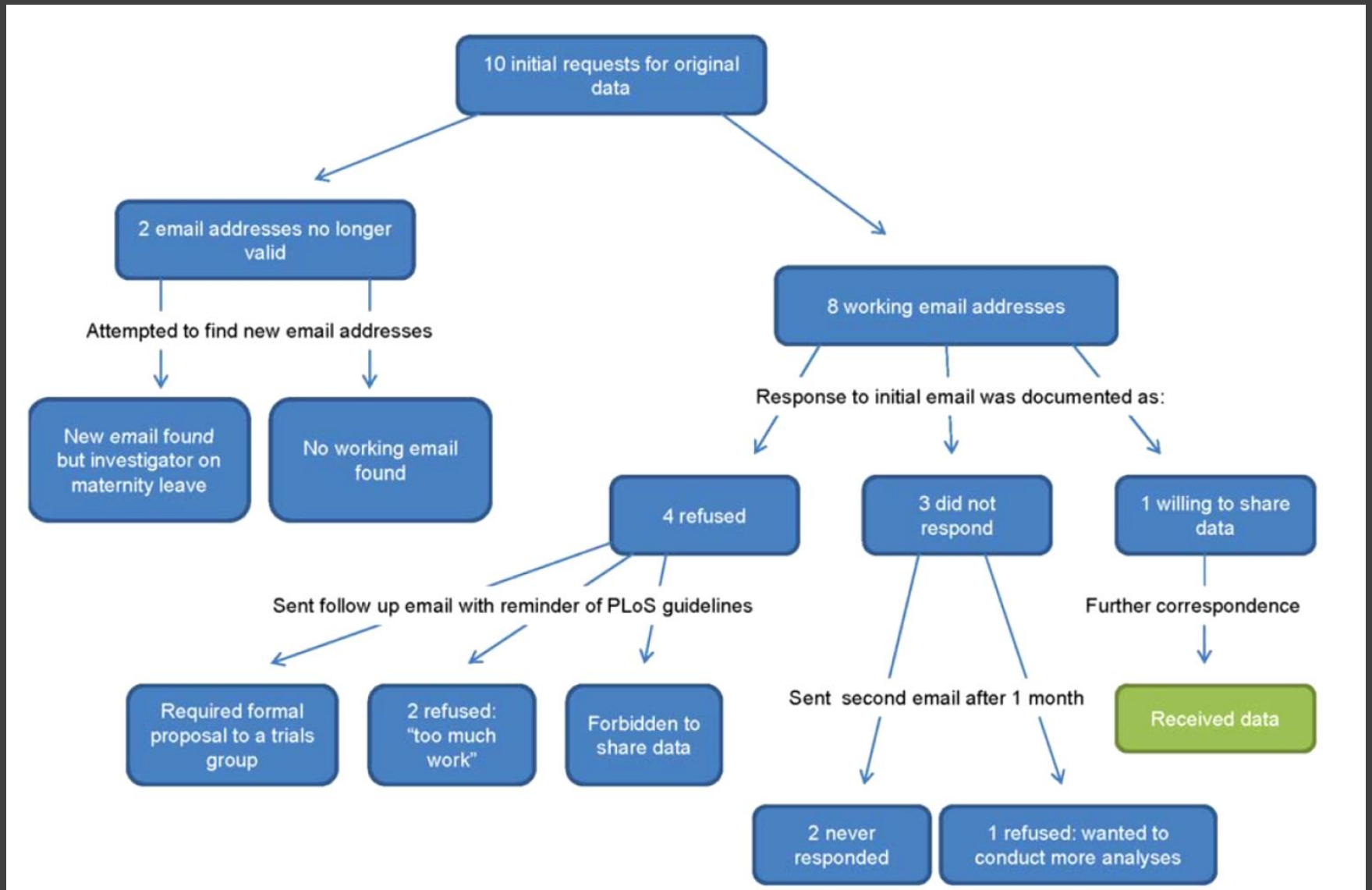
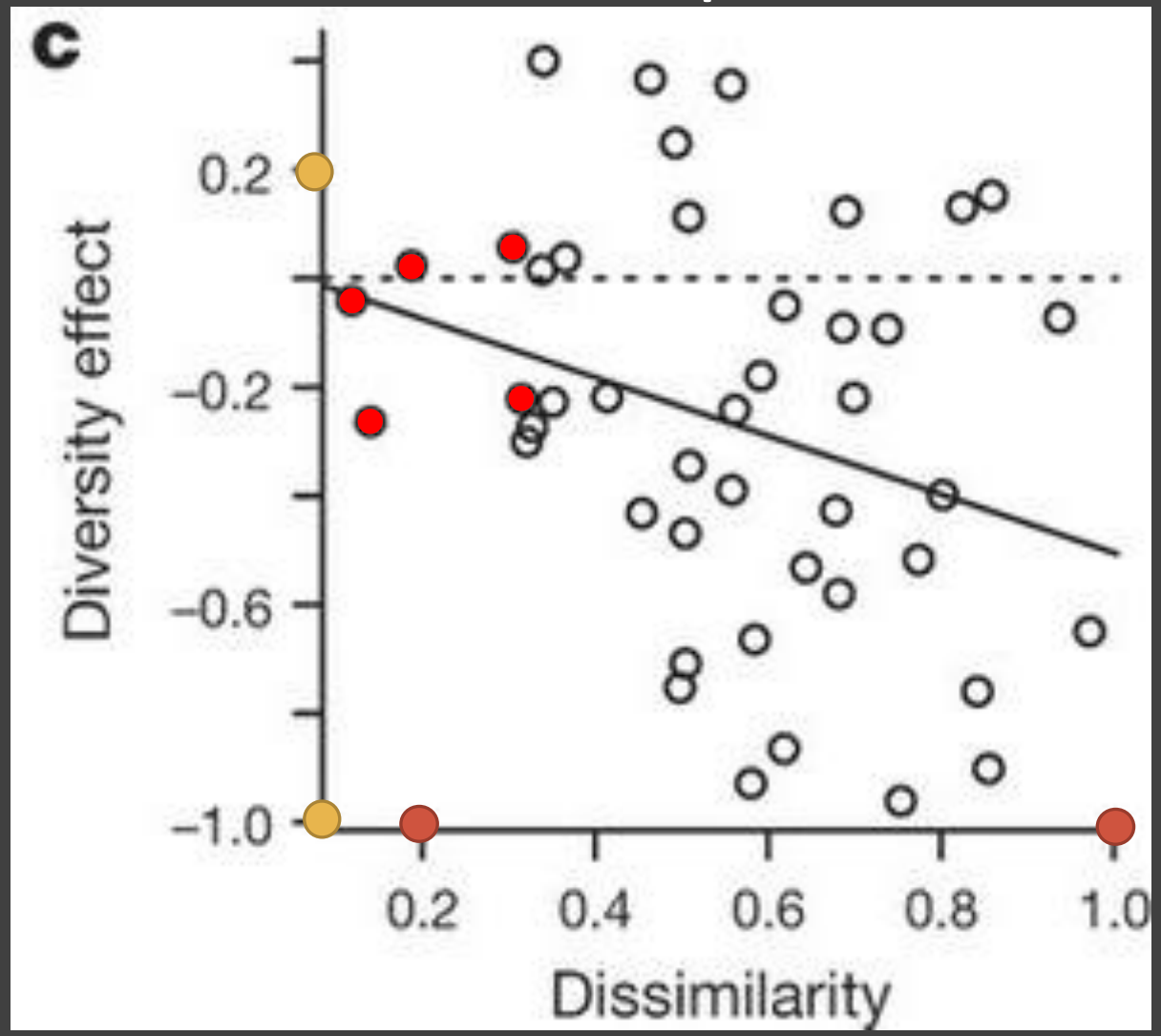


Figure 1. Summary of responses to the 10 initial requests for raw data.
doi:10.1371/journal.pone.0007078.g001

A paper in *Science* 2016

$n = 45$ $R^2 = 0.1$ $p = 0.04$



?

WebPlotDigitizer
Web based tool to extract data from plots, images, and maps

x	y
0.120	-0.040
0.142	-0.261
0.189	0.023
0.307	0.057
0.317	-0.224
...	

→ Linear Regression

C:/science/R-3.5.0/library/reverseR/reverseR - Shiny

http://127.0.0.1:4280 | Open in Browser | Publish

Detection of Significance Reversers in Linear Regression

Data format

The expected input data is a **.csv** spreadsheet with **two** or **three** columns (x-values in first, y-values in second, optional weights in third), with the variable names in the first row (see below):

Time	Fluo	Weights (optional)
4.2	17.22	0.15
5.7	22.31	0.07
...	...	...

Example

(Figure of a 2015 PNAS paper)

Example

Import Data

Data has a Header row

Type of decimal sign

dot (.)

comma (,)

Type of separator

comma (,)

space (\s)

tab (\t)

Statistics | Regression Plot | P-Value Plot | Slope vs. SE Plot | Influence Plot | Leave-Multiple-Out Plot

New Observation Analysis

Message:

(The message 'Error: incorrect number of dimensions' would indicate wrong import parameters)

```
Significant model with p = 0.0406417 has been changed to insignificant by removing one of the followin
g data points (in order of strength):

#41 => 0.08304431
#44 => 0.07207993
#40 => 0.07040441
#9 => 0.06926967
#36 => 0.06551448
#3 => 0.05974426
#4 => 0.05726832
#1 => 0.05400165
#8 => 0.05351361
#11 => 0.05298364
#37 => 0.05187982
```

Summary of full model (all da

```
Call:
lm(formula = DAT[, 2] ~ DAT[, 1], weights = wts)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65061 -0.22042 -0.04951  0.24795  0.62305

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03401    0.15271   0.223   0.8249
DAT[, 1]    -0.54034    0.25580  -2.112   0.0406 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3519 on 42 degrees of freedom
Multiple R-squared:  0.09604,   Adjusted R-squared:  0.07451
F-statistic: 4.462 on 1 and 42 DF,  p-value: 0.04064
```

Statistical Analysis for each Leave-One-Out model. For abbreviations, see below Table.

$$S = 1 - \frac{n_r}{N} = 1 - \frac{11}{45} = 0.76$$

C:/science/R-3.5.0/library/reverseR/reverseR - Shiny

http://127.0.0.1:4280 | Open in Browser

Detection of Significance Revers

Statistics | Regression Plot | P-Value Plot

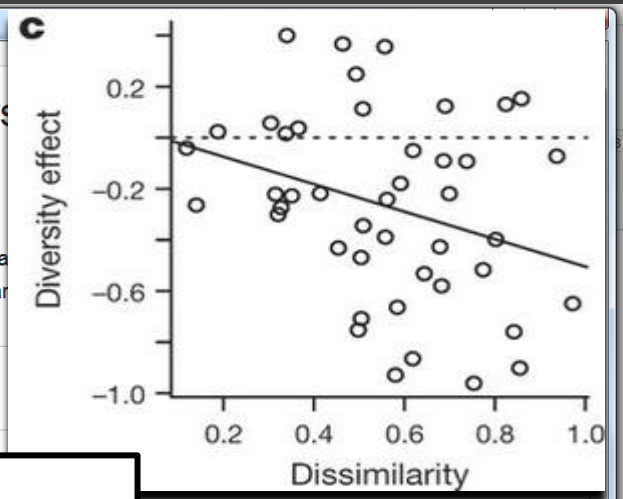
New Observation Analysis

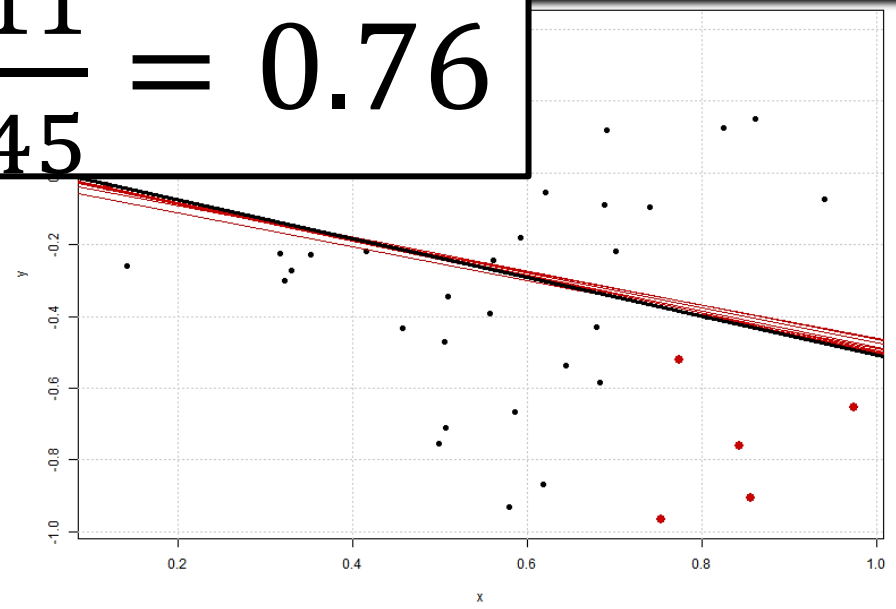
Linear Regression Plot with Influential Data

Significance reversers are depicted with lar

Mouse over data point to see its x/y value.

```
Data under cursor:
No data point.
```





PNAS paper)

Example

Import Data

Data has a Header row

Type of decimal sign

dot (.)

comma (,)

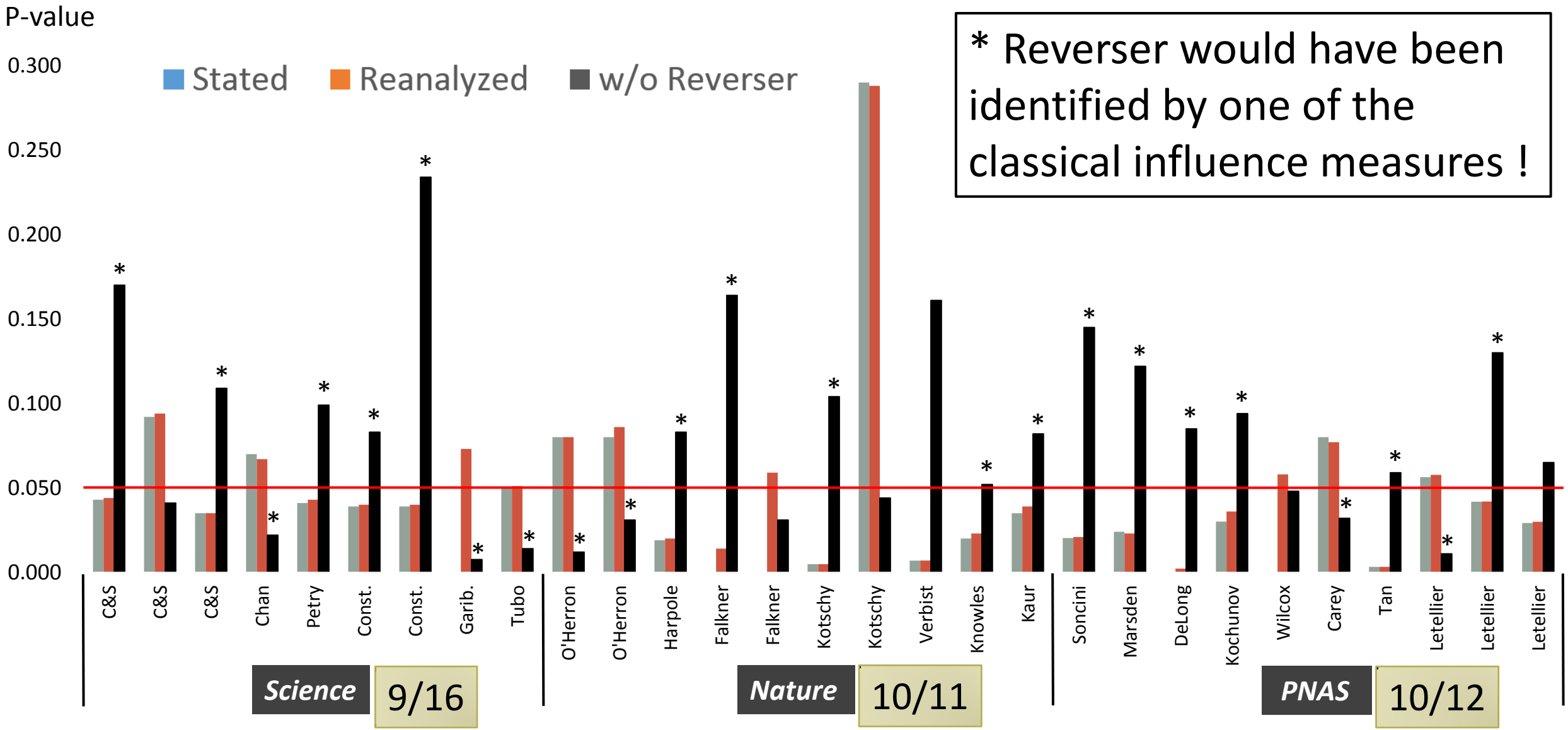
Type of separator

comma (,)

space (\s)

tab (\t)

semicolon (;)


P-value
0.300
0.250
0.200
0.150
0.100
0.050
0.000
Stated
Reanalyzed
w/o Reverser
* Reverser would have been identified by one of the classical influence measures !
C&S
C&S
C&S
Chan
Petry
Const.
Const.
Garib.
Tubo
O'Herron
O'Herron
Harpole
Falkner
Falkner
Kotschy
Kotschy
Verbist
Knowles
Kaur
Soncini
Marsden
DeLong
Kochunov
Wilcox
Carey
Tan
Letellier
Letellier
Letellier
Science
9/16
Nature
10/11
PNAS
10/12

Thanks goes to:

Dr. Stefan Rödiger
Institute of Biotechnology, Brandenburg University of Technology, Cottbus - Senftenberg, Senftenberg, Germany.
Mikhał Burdukiewicz
Department of Genomics, Faculty of Biotechnology, University of Wroclaw, Wrocław, Poland.
Prof. Joel Tellinghuisen
Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, USA.
Prof. Andreas Ziegler
Institute for Medical Biometry and Statistics, Universitätsklinikum Schleswig-Holstein, Lübeck, Germany