How to talk to your bioinformatician?

January Weiner
Core Unit for Bioinformatics, BIH@Charité

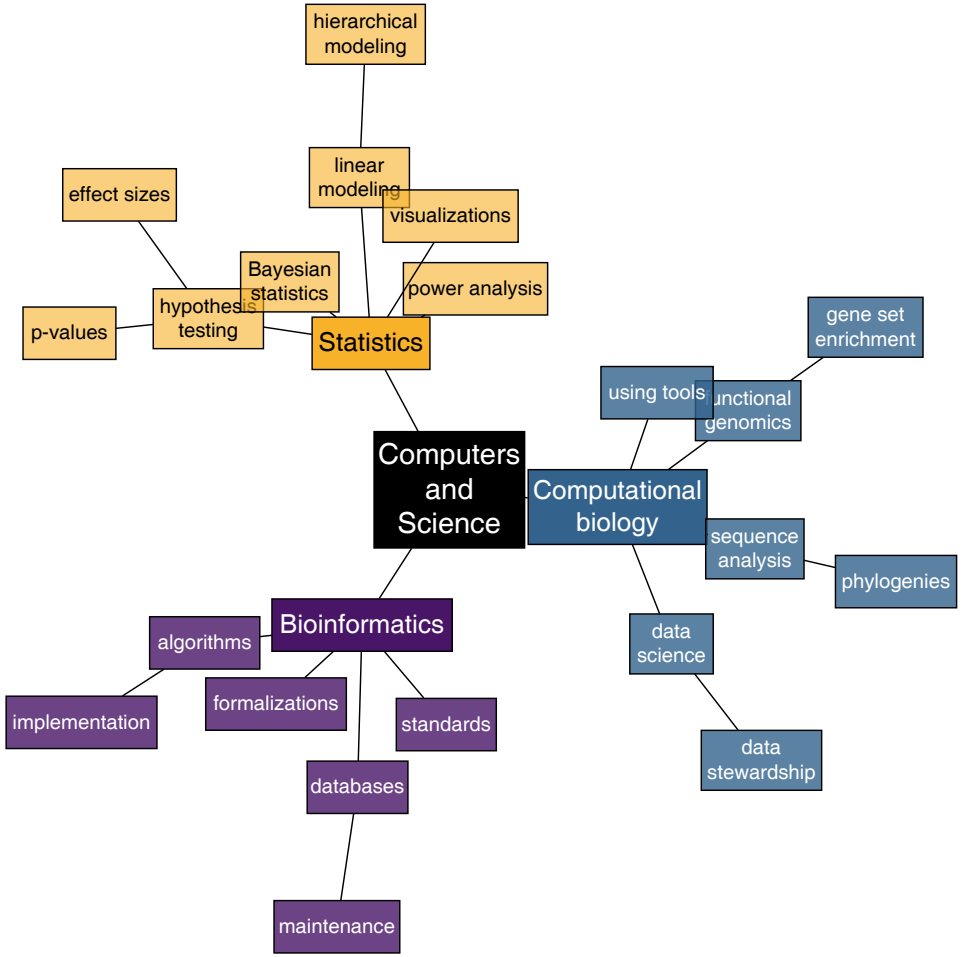
About this presentation

The newest version of this presentation is available at bihealth.github.io/howtotalk.

You can find the sources of the presentation (Qmd file) on github.com/bihealth/howtotalk

I have elaborated parts of this talk into a little brochure, to be found at bihealth.github.io/howtotalk-book.

Who am I to tell you things?



My key advice to you

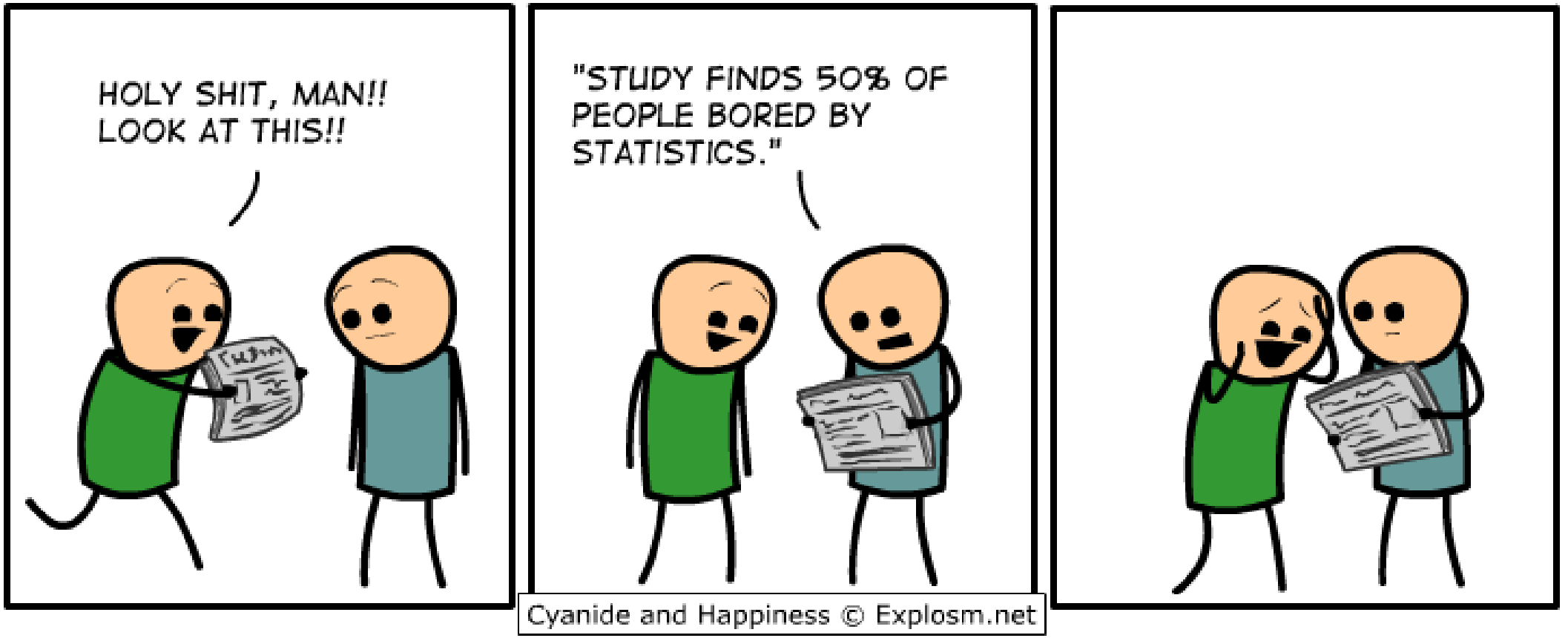
Communication is key

- Keep explaining your project – teach us
- Work iteratively
- Meet frequently

Things bioinformaticians care about

- The biological question
- Statistics
- Experimental design
- Quality control
- Reproducibility
- Consistency

Statistics


HOLY SHIT, MAN!!
LOOK AT THIS!!
"STUDY FINDS 50% OF PEOPLE BORED BY STATISTICS."
Cyanide and Happiness © Explosm.net

What is a p-value?

H_0: The null hypothesis, no effect

H_1: The alternative hypothesis, there is an effect

We run a test, we get a p-value, say 0.03. It is a probability.

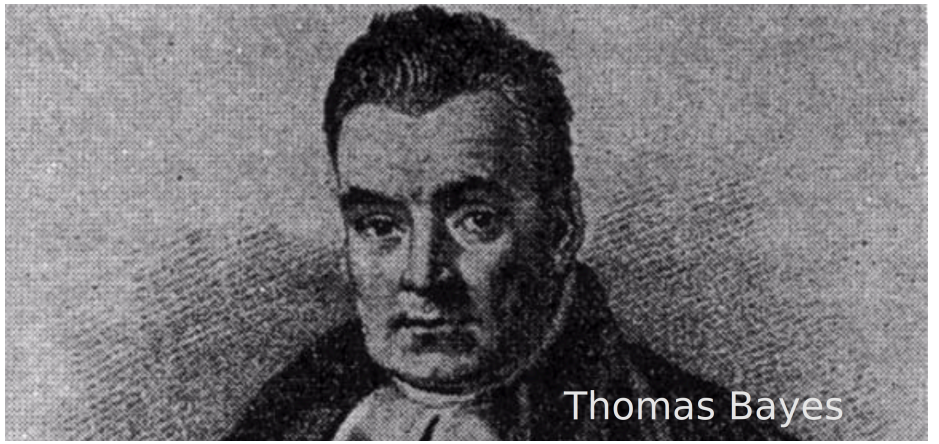
Probability of *what*, exactly?

Raise your hands if you think that the p-value is the...

1. Probability that H_0 is true (probability that there is no difference), given the data
2. Probability that H_1 is true (probability that there is a difference), given the data
3. Probability that the data is random
4. Probability that the observations are due to random chance
5. Probability of getting the same data by random chance

- **Probability of observing an effect at least as extreme given that H_0 is true**

Our intuition is bayesian, not frequentist

Frequentist Statistics	Bayesian Statistics
	
1. Probability is defined as the long-run frequency of events	1. Probability represents a degree of belief or certainty about an event
2. Parameters (like the “true value”) are fixed but unknown quantities.	2. Parameters are treated as random variables with their own probability distributions.
3. Asking about the probability of a hypothesis does not make sense	3. Asking about the probability of a hypothesis is the main goal

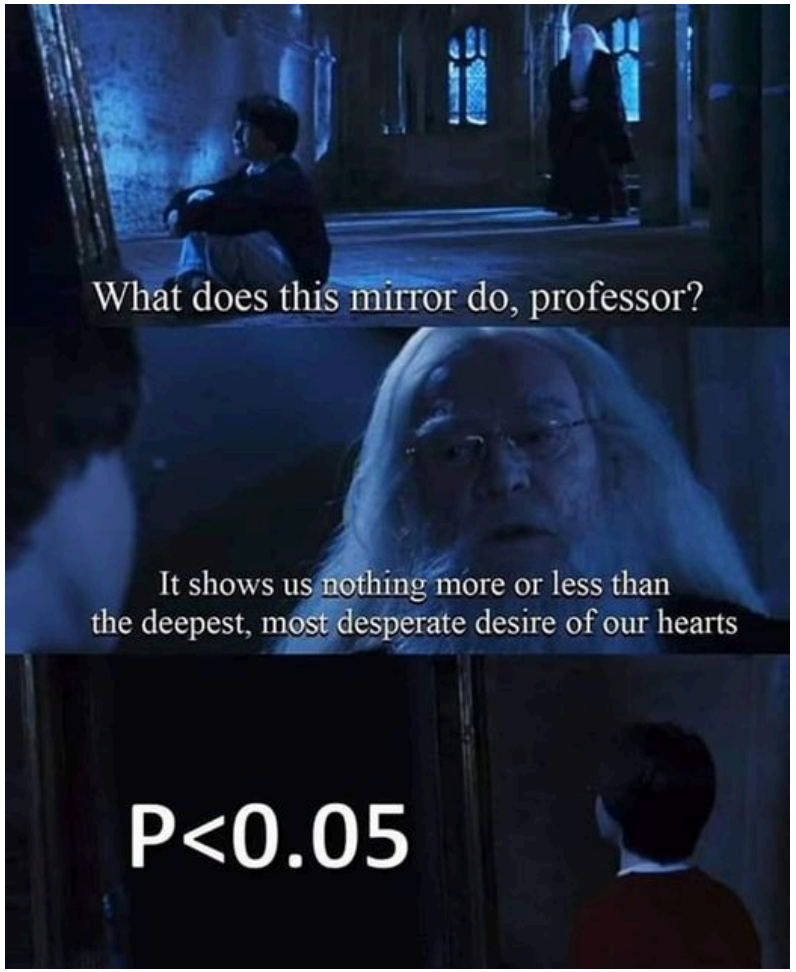
Why is that important?

P-values are part of scientific language

- *Always* use effect sizes
- Never rely on p-values alone

Know their limits:

- they control only type I errors (false positives)
- they **do not** control type II errors (false negatives)

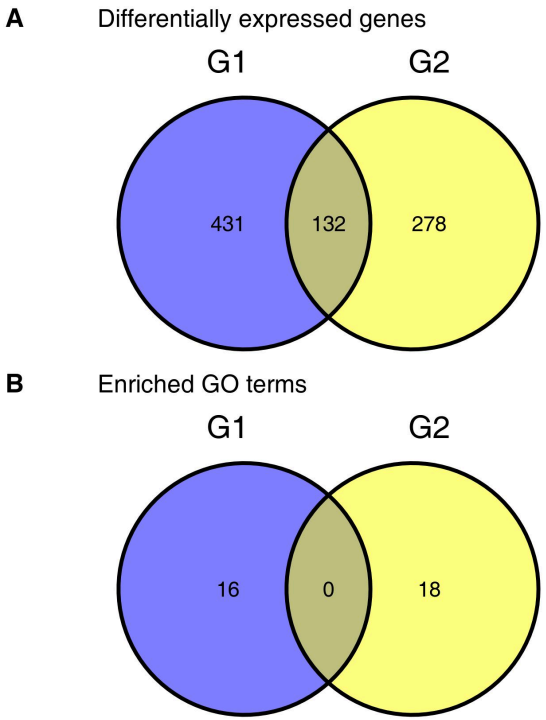


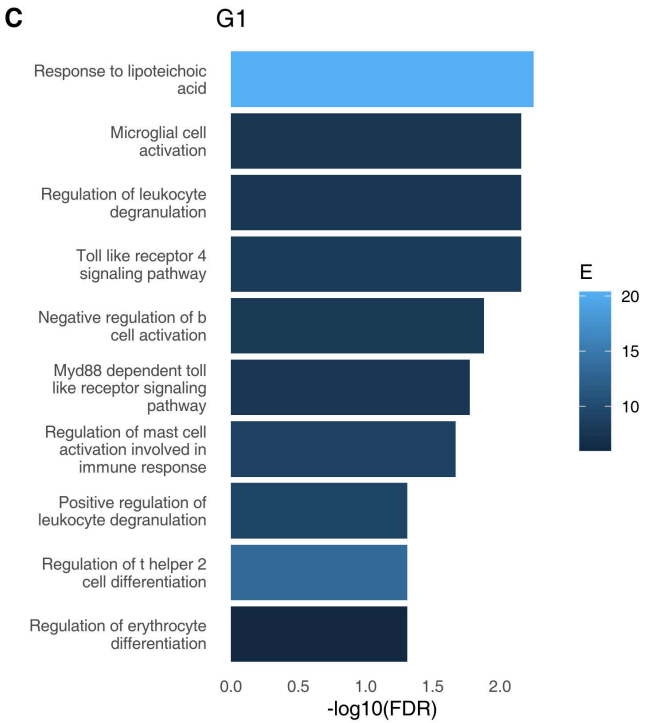
Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. The American

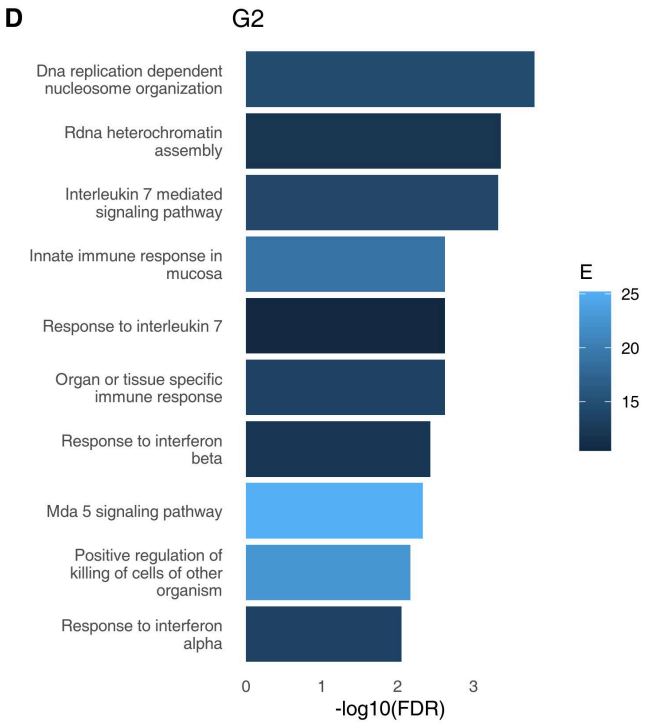
How Venn diagrams can fool scientists

COVID-19 study, both COVID-19 patients and non-COVID-19 patients are compared in two groups of people, *G1* and *G2*.

We wanted to know whether the influence of COVID-19 is different in these two groups.



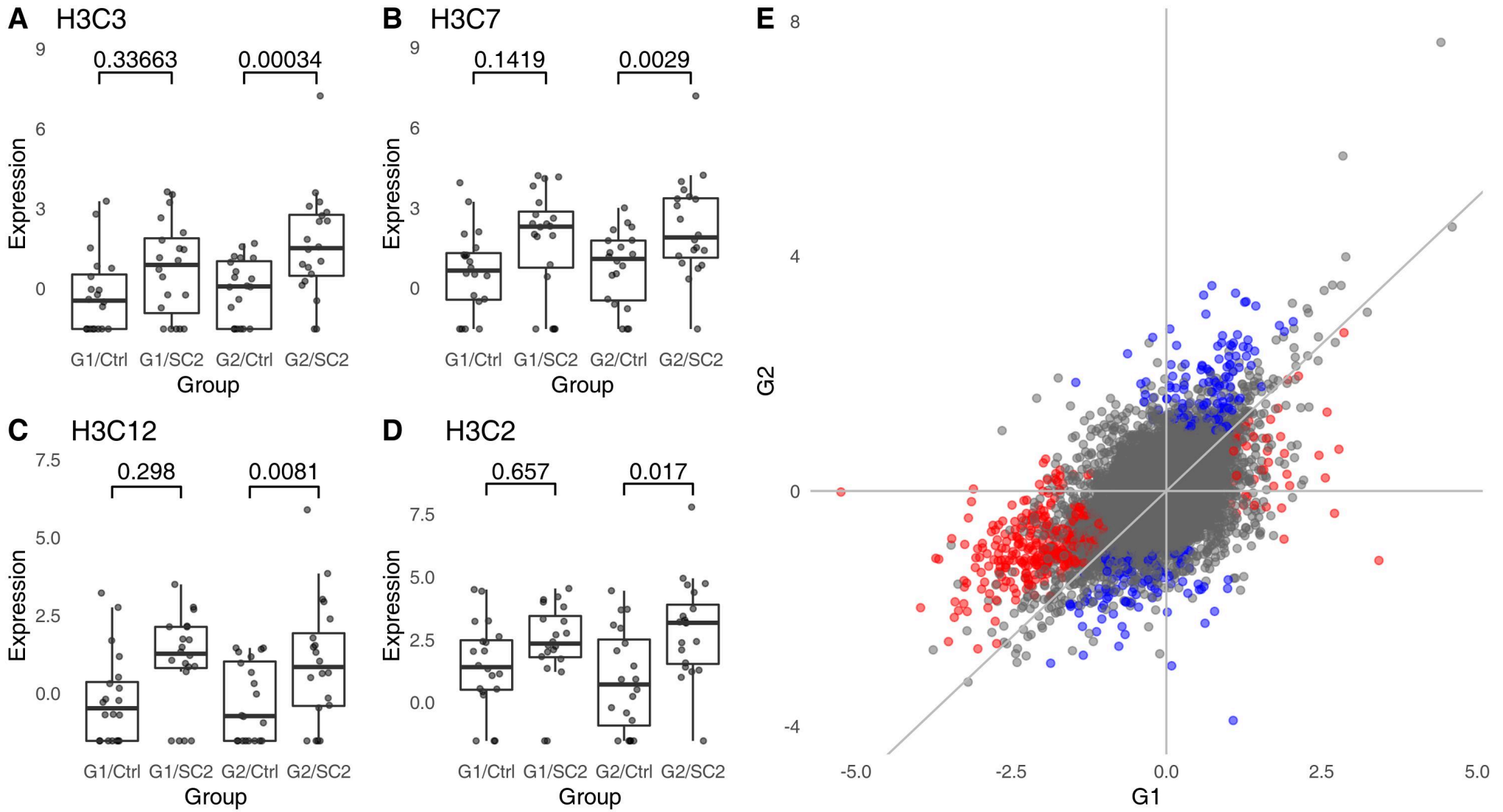




Venn diagrams may indicate erroneous statistical reasoning in transcriptomics. Weiner, Obermayer and Beule,

The results are artifacts!

Groups G1 and G2 were randomly drawn from the same population. They were not different at all.



What happens is, we are comparing significance with non-significance

> The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant

(*Andrew Gelman and Howard Stern*)

If a gene is significant in one comparison, and not significant in another, that does not mean that there is a difference between the two groups.

It simply means that we *failed* to detect the difference in one of the comparisons, but that is actually quite likely to happen!

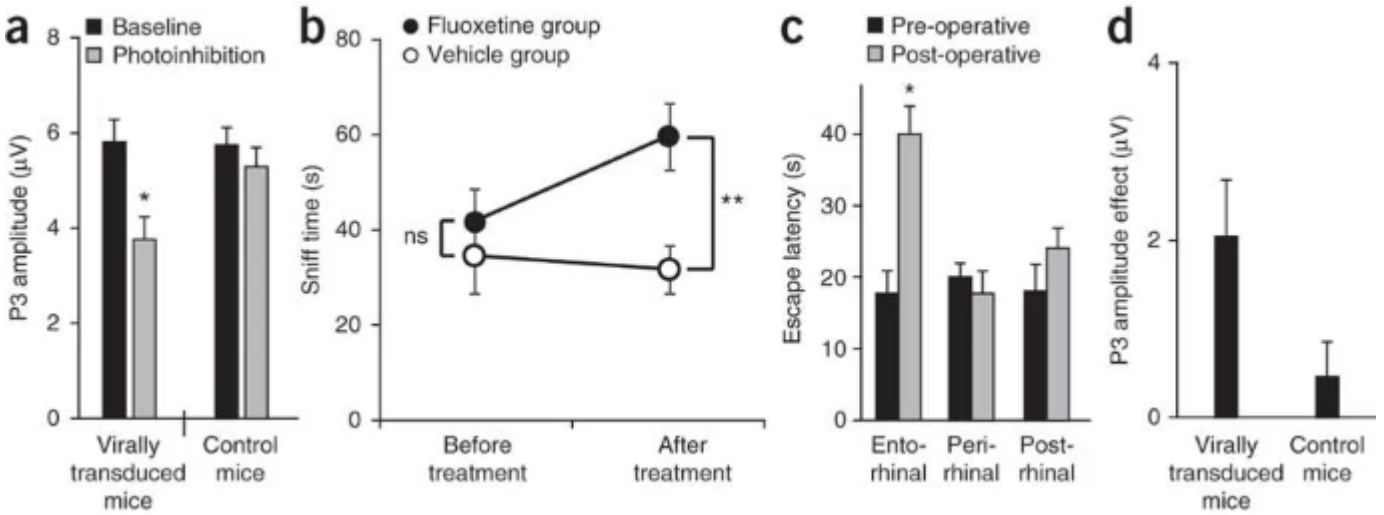
Therefore:

Don’t say “there is no difference”. Say “we did not detect a difference”.

Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant.

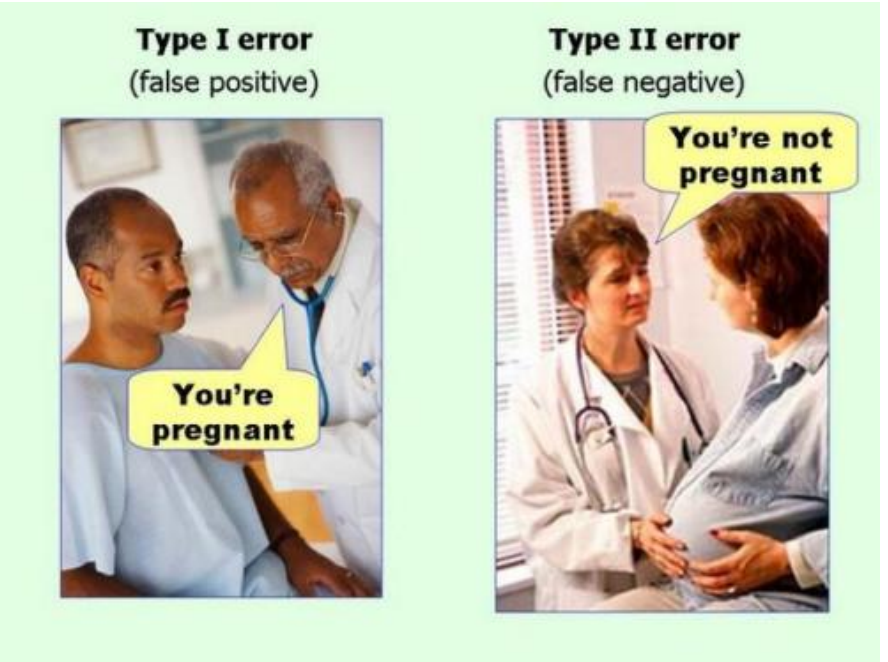
The error is widespread

Nieuwenhuis et al. found that half of the scientists who could have commited this error, did in fact commit this error.



Going beyond p-values

- Estimation rather than testing (e.g. confidence intervals rather than p-values)
- Considering effect sizes
- Power analysis – estimating type II error rates (false negatives)
- Sign / magnitude errors
- Bayesian statistics
- Correcting for multiple testing



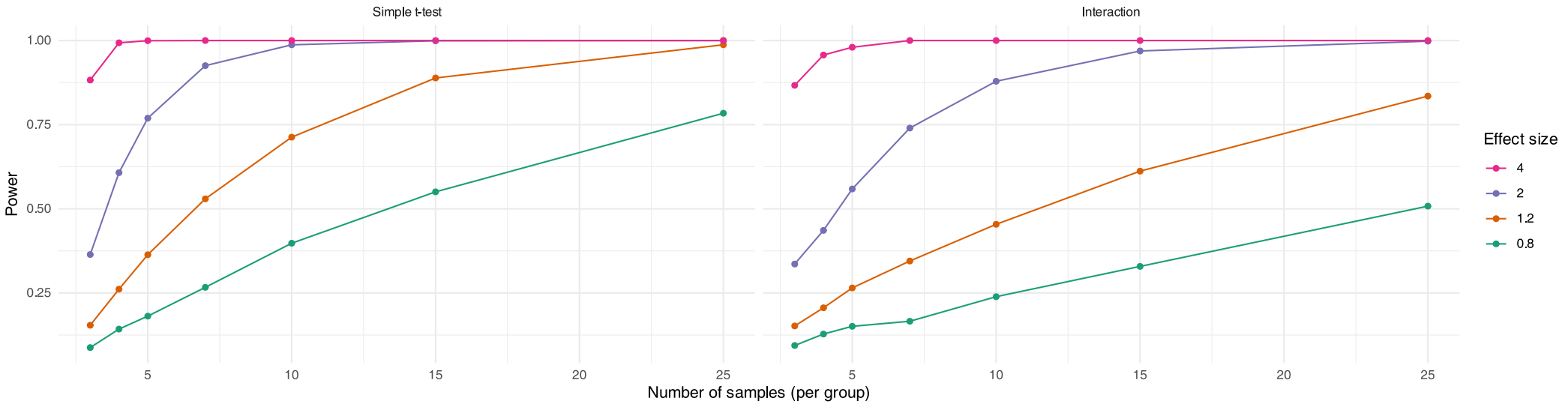
Experimental design

How many samples are sufficient?

Simple comparison between two groups

- Two strains (WT and KO)
- Treatment + control
- Does treatment have a different effect on the KO strain than on the WT strain?

2x2 design, test for interaction term



How many samples are sufficient?

That is not even the worse thing.

Simple calculations show that assuming

- your power is 80% (really great!)
- $p - value$ cutoff is 0.05
- 90% of the H_0 are true (i.e., 10% of the time the differences are real)

then 36% of your “significant” results **are false positives**[1]!

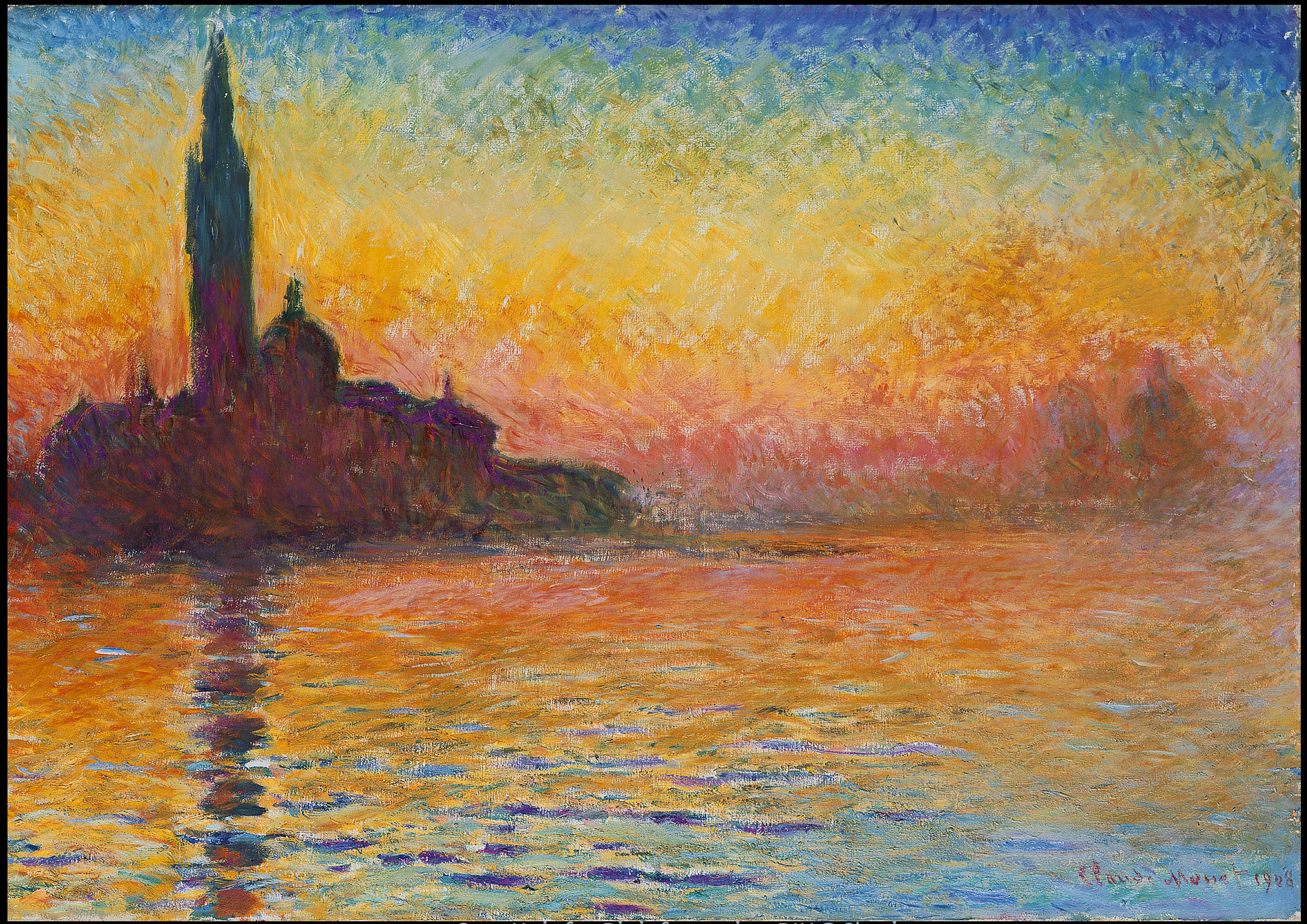
(Plus, you failed to detect 20% of the real differences)

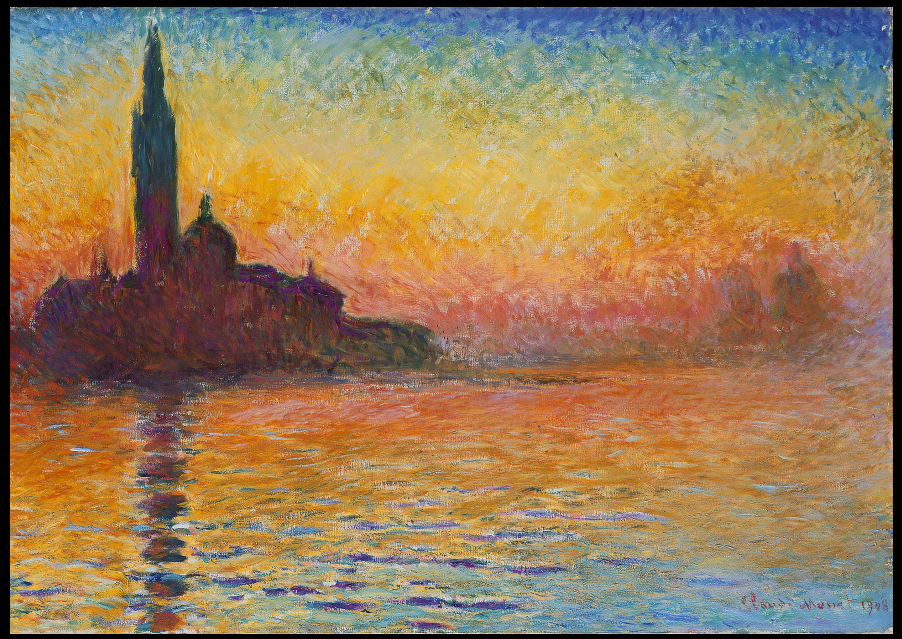
Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society open science. 2014 Nov 19;1(3):140216.

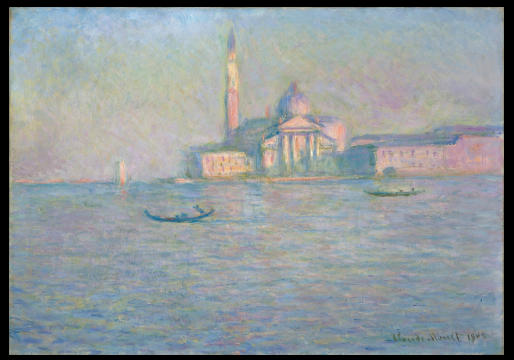
Bottom line

- Talk to your statistician early
- Strive to keep your study design simple
- Use existing data sets and simulations to test your design
- Validate your results with independent methods

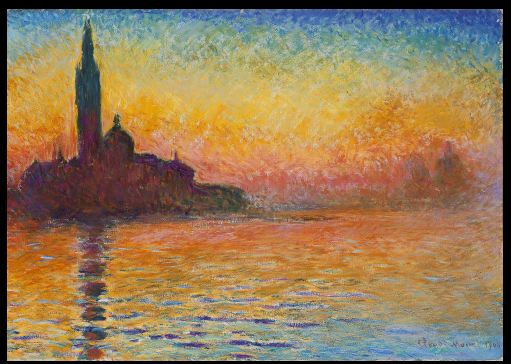
High throughput data


Claude Monet 1908







Explorative vs hypothesis testing

Explorative analysis

Pro:

- No need to define a-priori hypotheses
- Something unexpected and new can be found
- Can be used to generate hypotheses

Con:

- Requires multiple testing correction
- Requires proper validation
- Can’t do it as the last step

Hypothesis-driven analysis

Pro:

- Clear questions
- Clear answers
- More statistical power
- Better story, better paper

Con:

- Requires more planning (and thinking!)
- Can make you miss something unexpected
- If you reject the hypothesis, tough luck

The bottom line

Do

- Formulate clear questions
- Manage your expectations
- Evaluate existing data
- Validate your results

Don’t

- expect miracles
- “let’s just see what we can find”
- try to save money
- make too complex designs

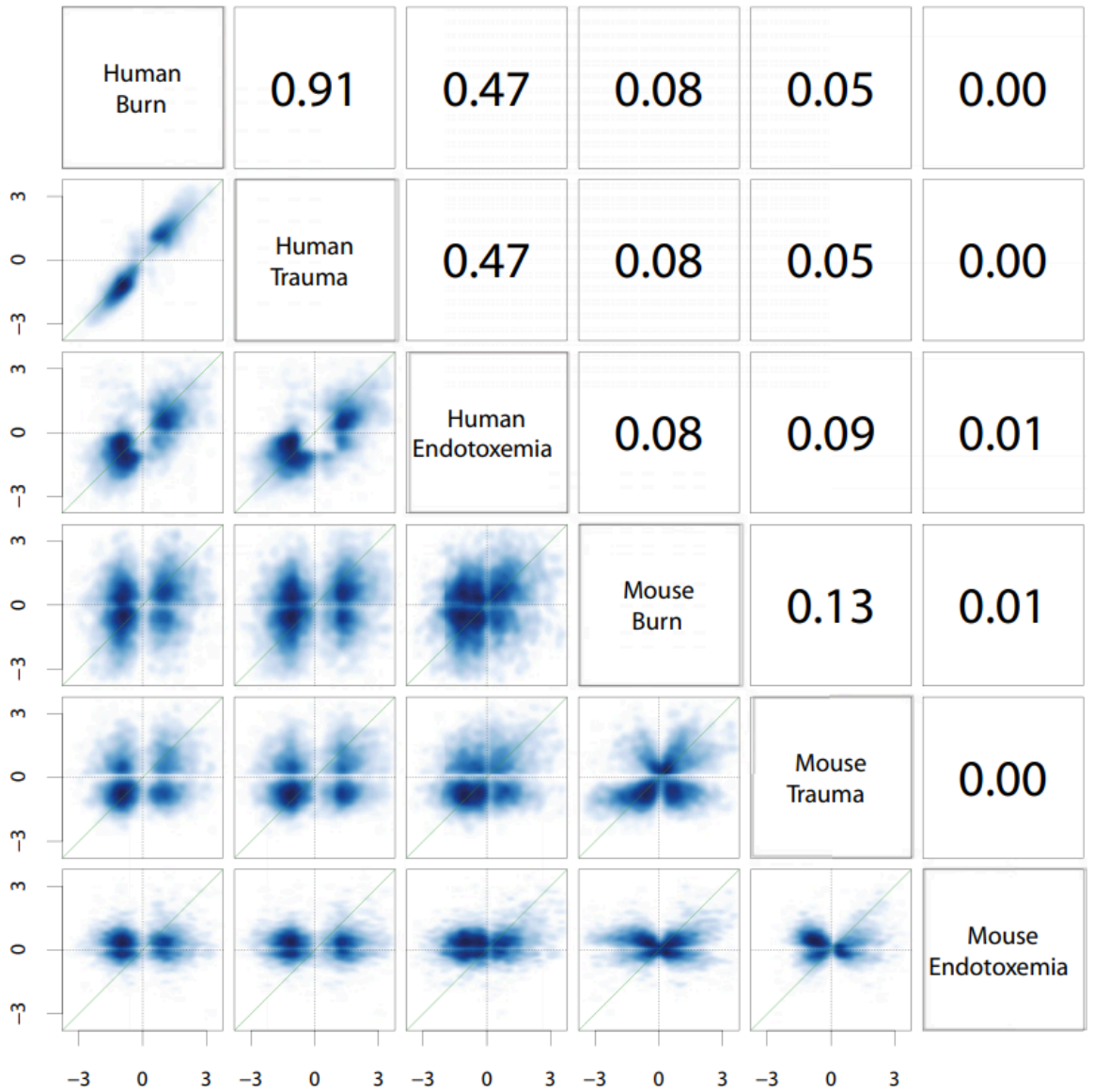
Reproducibility

Tale of two papers

Genomic responses in mouse models poorly mimic human inflammatory diseases

Junhee Seok[a,1], H. Shaw Warren[b,1], Alex G. Cuenca[c,1], Michael N. Mindrinos[a], Henry V. Baker[c], Weihong Xu[a], Daniel R. Richards[d], Grace P. McDonald-Smith[e], Hong Gao[a], Laura Hennessy[f], Celeste C. Finnerty[g], Cecilia M. Lópe
Shari Honari[f], Ernest E. Moore[h], Joseph P. Minei[i], Joseph Cuschieri[j], Paul E. Bankey[k], Jeffrey L. Johnson[h], Jason Spe

Tale of two papers



Tale of two papers

Genomic responses in mouse models poorly mimic human inflammatory diseases

Junhee Seok[a,1], H. Shaw Warren[b,1], Alex G. Cuenca[c,1], Michael N. Mindrinos[a], Henry V. Baker[c], Weihong Xu[a], Daniel R. Richards[d], Grace P. McDonald-Smith[e], Hong Gao[a], Laura Hennessy[f], Celeste C. Finnerty[g], Cecilia M. Lópe
Shari Honari[f], Ernest E. Moore[h], Joseph P. Minei[i], Joseph Cuschieri[j], Paul E. Bankey[k], Jeffrey L. Johnson[h], Jason Spe

Genomic responses in mouse models greatly mimic human inflammatory diseases

Keizo Takao[a,b] and Tsuyoshi Miyakawa[a,b,c,1]

[a]Section of Behavior Patterns, Center for Genetic Analysis of Behavior, National Institute for Physiological Sciences, Okazaki, Aichi 444-8585, Japan; [b]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan; and [c]Division of Syst
Medical Science, Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi 470-1192, Japan

Tale of two papers

Genomic responses in mouse models poorly mimic human inflammatory diseases

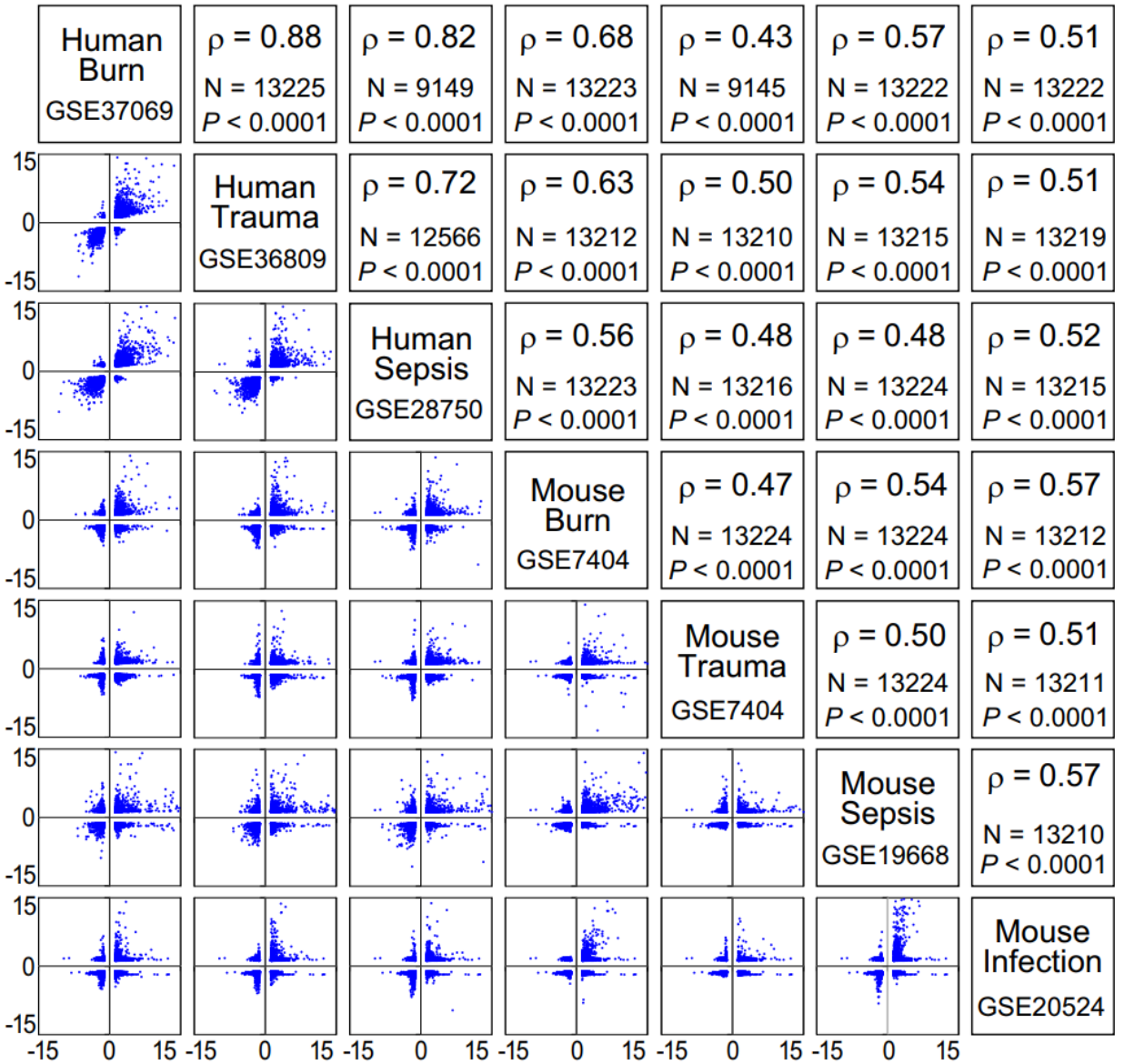
Junhee Seok[a,1], H. Shaw Warren[b,1], Alex G. Cuenca[c,1], Michael N. Mindrinos[a], Henry V. Baker[c], Weihong Xu[a], Daniel R. Richards[d], Grace P. McDonald-Smith[e], Hong Gao[a], Laura Hennessy[f], Celeste C. Finnerty[g], Cecilia M. Lópe
Shari Honari[f], Ernest E. Moore[h], Joseph P. Minei[i], Joseph Cuschieri[j], Paul E. Bankey[k], Jeffrey L. Johnson[h], Jason Spe

Genomic responses in mouse models greatly mimic human inflammatory diseases

Keizo Takao[a,b] and Tsuyoshi Miyakawa[a,b,c,1]

[a]Section of Behavior Patterns, Center for Genetic Analysis of Behavior, National Institute for Physiological Sciences, Okazaki, Aichi 444-8585, Japan; [b]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan; and [c]Division of Syst
Medical Science, Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi 470-1192, Japan

Tale of two papers



Lessons learned

- *A lot* depends on how you analyze your data
- This in turn depends on the questions you ask
- The average “Methods” section is not sufficient for reproducible science!

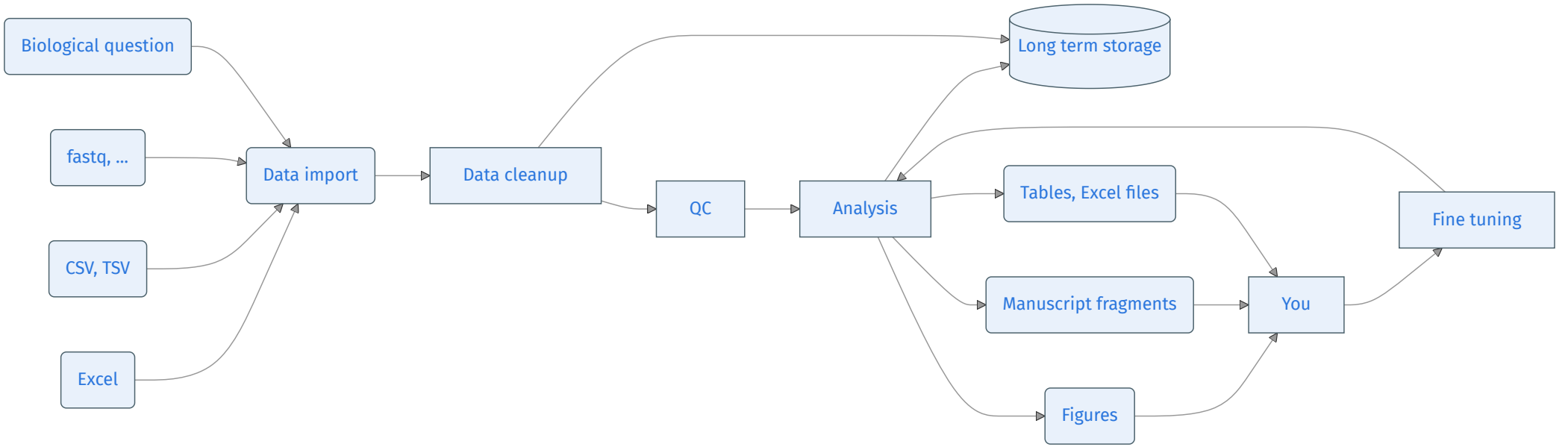
Attempt to replicate 53 high-impact cancer biology papers:

> ” Second, **none of the 193 experiments** were described in sufficient detail in the original paper to enable us to design protocols to repeat the experiments, so we had to seek clarifications from the original authors.” (Errington et al., 2021)

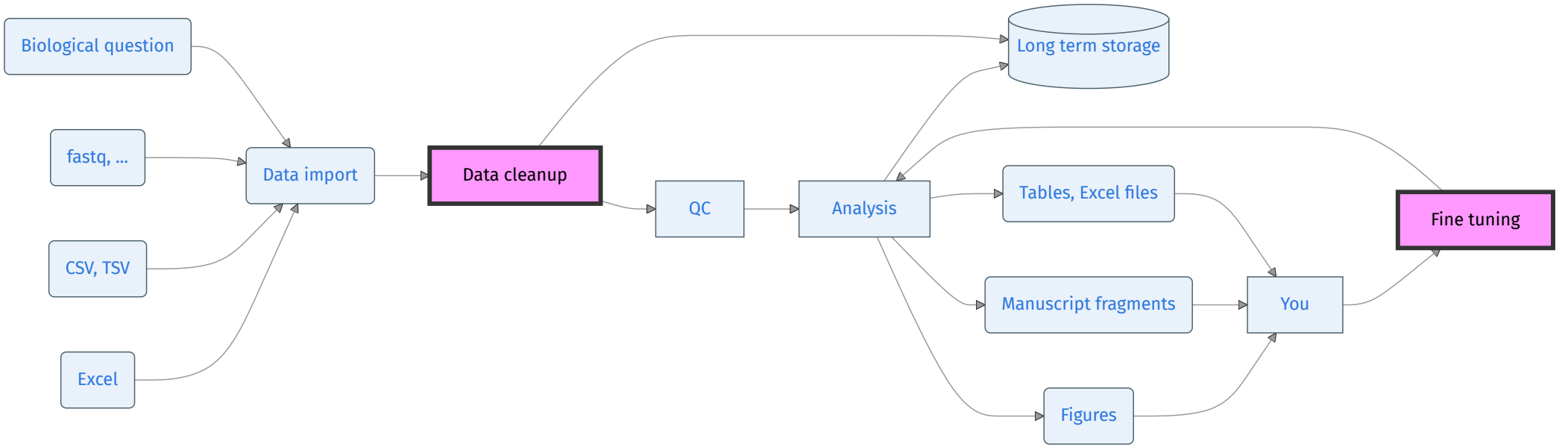
Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. Investigating the replicability of preclinical cancer biology. Elife. 2021 Dec 10;10:e71601.

Managing data

How we work


Biological question
fastq, ...
CSV, TSV
Excel
Data import
Data cleanup
Long term storage
QC
Analysis
Tables, Excel files
Manuscript fragments
Figures
You
Fine tuning

How we work



In the diagram above, two things take usually a lot of hands-on time:

- Understanding and cleaning the data
- Fine-tuning the analysis results

Excel and gene names

Home › Genome Biology › Article

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Gene name errors: Lessons not learned

Mandhri Abeysooriya, Megan Soria, Mary Sravya Kasu, Mark Ziemann

Version 2 Published: July 30, 2021 • https://doi.org/10.1371/journal.pcbi.1008984

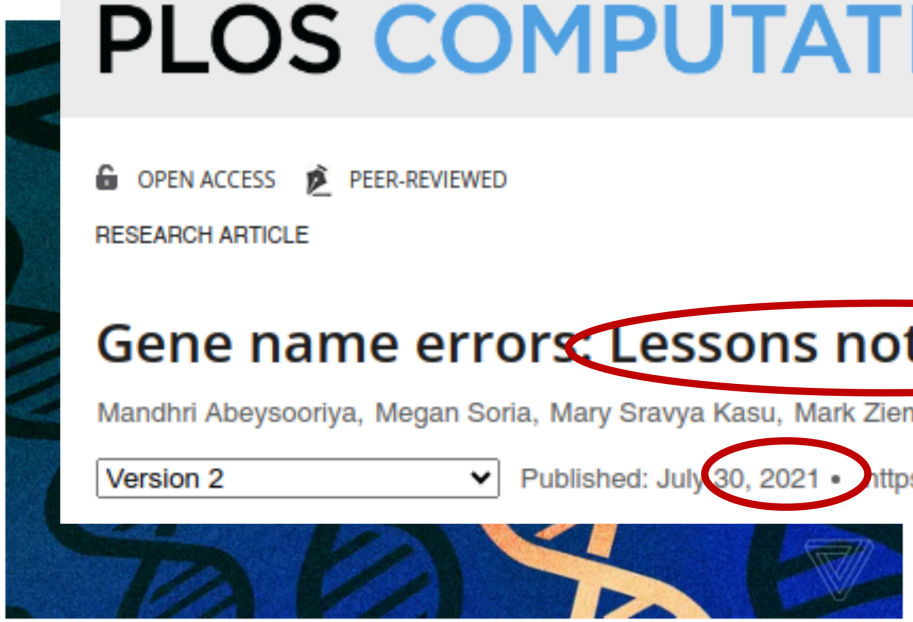


Illustration by Alex Castro / The Verge

If you buy something from a Verge link, Vox Media may earn a commission. ***See our ethics statement.***

Ziemann, Mark, Yotam Eren, and Assam El-Osta. "Gene name errors are widespread in the scientific literature." Genome biology 17 (2016): 1-3; Abeysooriya M, Soria M, Kasu MS, Ziemann M. Gene name errors: Lessons not learned.

Is Excel suitable for science?

- How do you record changes?
- How do you prevent automatic changes?
- In short – how do you ensure reproducibility?

Bottom line

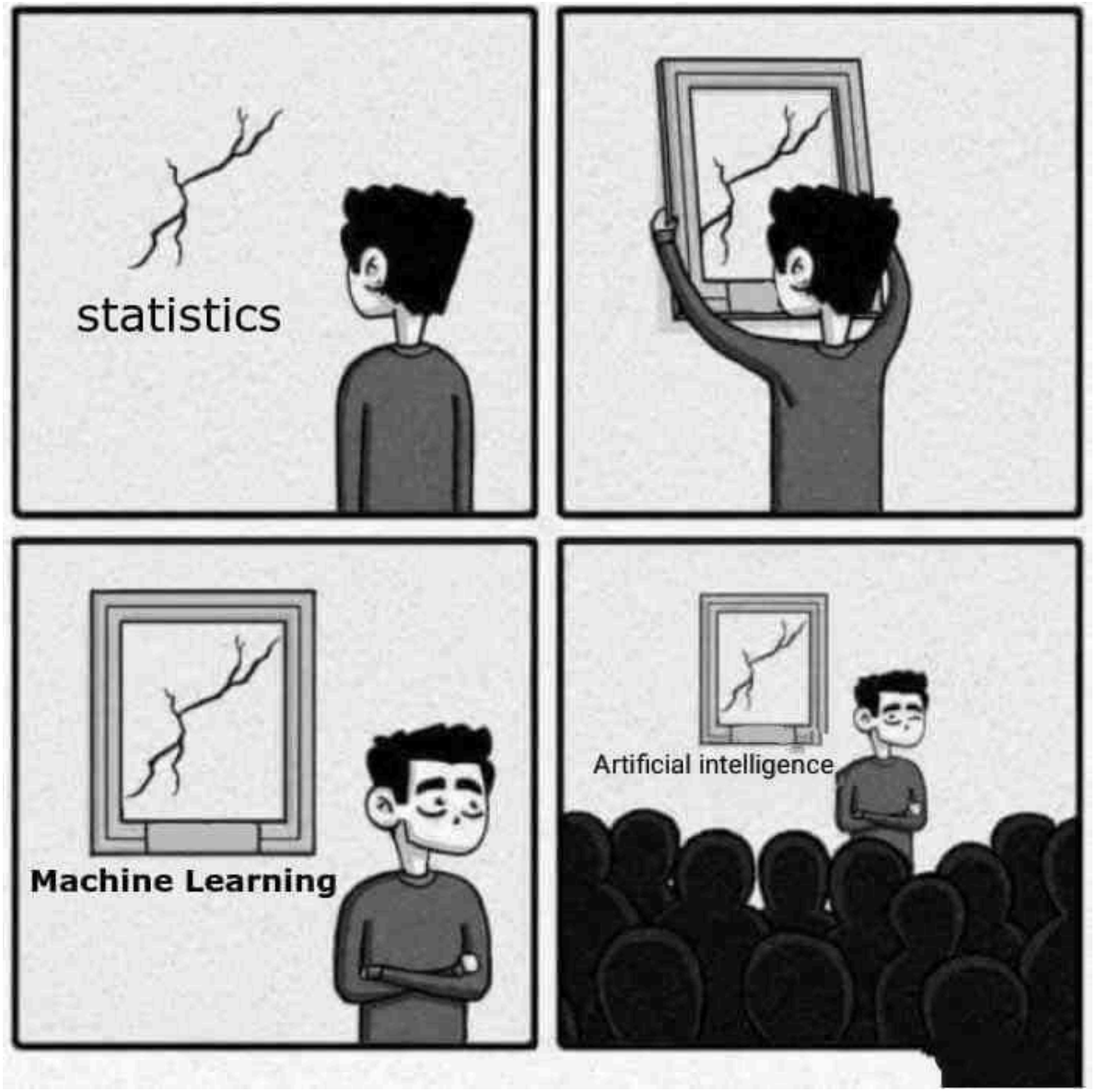
- Learn how to work with Excel
- Learn more suitable tools

Things that you might want to learn

- Statistics
- Coding (likely R or Python)
- Reproducible workflows with Quarto/Rmarkdown or Jupyter

Even if you are not going to use these tools yourself, gaining an insight into how they work will help you to communicate with your bioinformatician.

Will “AI” change the field?



- New deep learning methods are useful, but hard to use
- Some of them are truly revolutionizing the field
- There is still place for simpler ML algorithms
- Ready to use LLMs (ChatGPT & Co.) have their use, but also limitations

Thank you

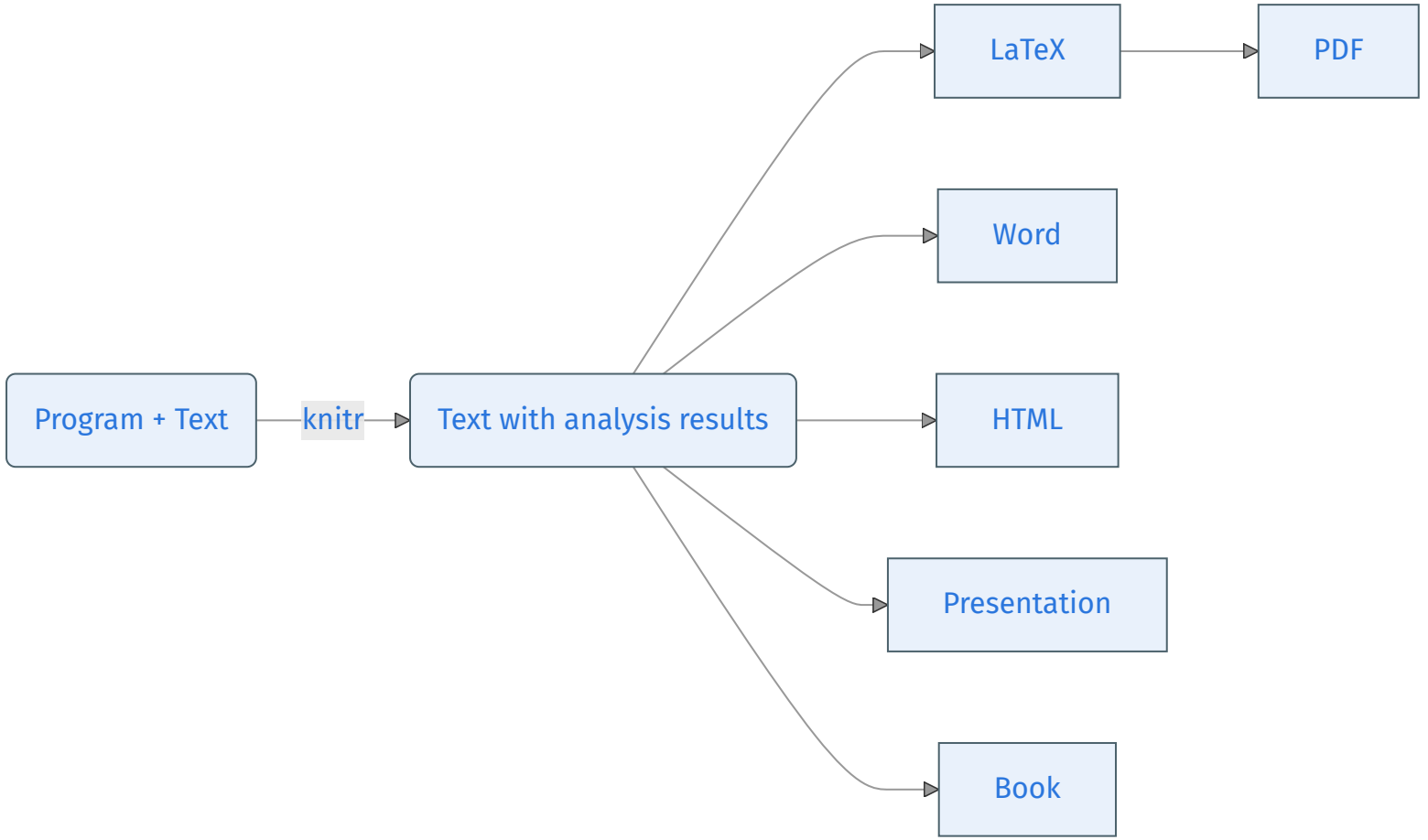
You can find this presentation along its source code at https://github.com/bihealth/howtotalk

I have started writing a little brochure based on this presentation, you can find the current progress (about 10% done) at https://bihealth.github.io/howtotalk-book/

A 5 day R crash course book is available at https://bihealth.github.io/RCrashcourse-book/

Course materials & videos: https://bihealth.github.io/RCrashcourse2023/

Reproducible workflows with Rmarkdown



This can be Rmarkdown, Quarto, Jupyter... the goal is that your code and your text are in one place, and the results of your calculations are entered automatically into the text.

Reproducible workflows with Rmarkdown

In systems such ar R markdown, you can put directly your analysis results in your text. For example, when I write that the p-value is equal to 0.05, I am writing this:

```
In systems such ar R markdown, you can put directly your
analysis results in your text. For example, when I write that the
$p$-value is equal to `r p`, I am writing this:
```

The p-value above is not entered manually (as 0.05), but is the result of a statistical computation. If the data changes, if your analysis changes, the p-value above will automatically change as well.

Identifiers

- Never use “pure numerical” identifiers (`1`, `2`, `3`, ...)
- Never remove columns with “unneeded” identifiers or you can get “involuntary anonymization”
- If you identifier is, say, `S1`, then always refer to that sample as `S1`, not `Smp. 1` or `1` or `Sample 1`
- Better use a unique prefix for a study / cohort / experiment, like `RCDB2024_S1`, `RCDB2024_S2`, `RCDB2025_S1`, `RCDB2025_S2`
- Composite identifiers are fine, as long as you use them consistently: `WT_treatment_1`, `KO_control_2` and not `WT_treatment_1`, `KO-2-control`, `WT_ctrl2`

How (not to) work with Excel

Avoid manually modifying Excel files

- Manual changes cannot be tracked automatically
- You have to record every change you make
- Otherwise, this is not reproducible science!

How (not to) work with Excel

Never use formatting for data

Never encode information as formatting, always use explicit columns

Color / font size / font style cannot be read automatically

	A	B
1	Sample ID (red=treated)	
2	Sample 1	
3	Sample 2	
4	Sample 3	
5	Sample 4	
6	Sample 5	
7	Sample 6	
8	Sample 7	
9		
10		

How (not to) work with Excel

Don’t combine values and comments

Make a separate column for comments

Otherwise the values might be lost[1]

No:

	A	B
1	Sample	Measurement
2	Sample 1	10
3	Sample 2	20 (morning)
4	Sample 3	30
5	Sample 4	40
6	Sample 5	99 (out of range)
7		

Yes:

	A	B	C
1	Sample	Measurement	Comment
2	Sample 1	10	
3	Sample 2	20	morning
4	Sample 3	30	
5	Sample 4	40	
6	Sample 5	99	out of range
7			

How (not to) work with Excel

Don’t put meta-information into column names

Make a separate excel sheet for column meta information

	A	B	C
1	**Sample - standard identifier from the Redcap db**	**Measurement (ug/ml; using the standard plate reader)**	
2	Sample 1	11	
3	Sample 2	13	
4	Sample 3	28	
5	Sample 4	1.5	
6	Sample 5	32	
7			

The statistical testing roulette

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

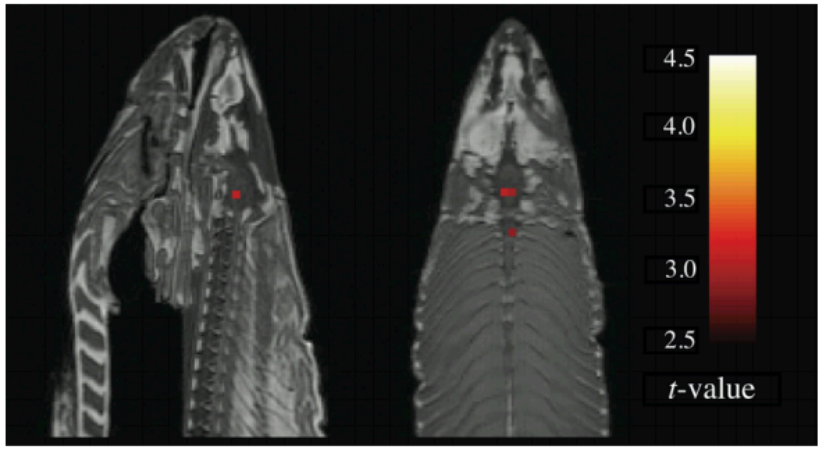
With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was

GLM RESULTS

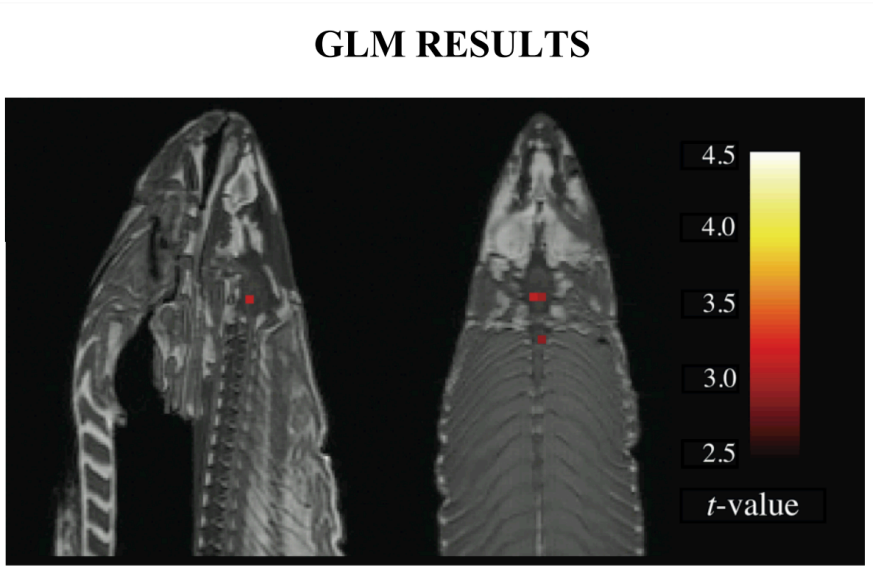


A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

The statistical testing roulette

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

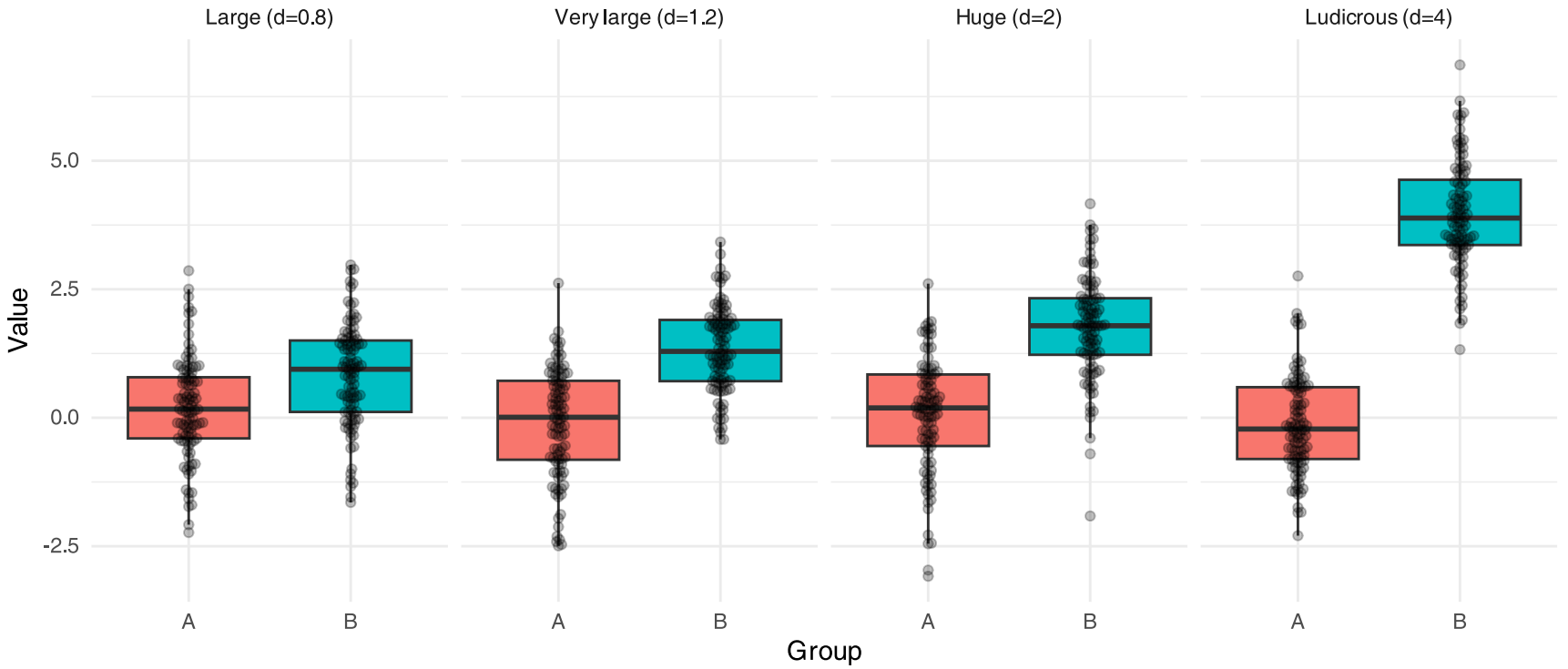
Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.



Bennett CM, Miller MB, Wolford GL. Neural correlates of interspecies perspective taking in the post-mortem

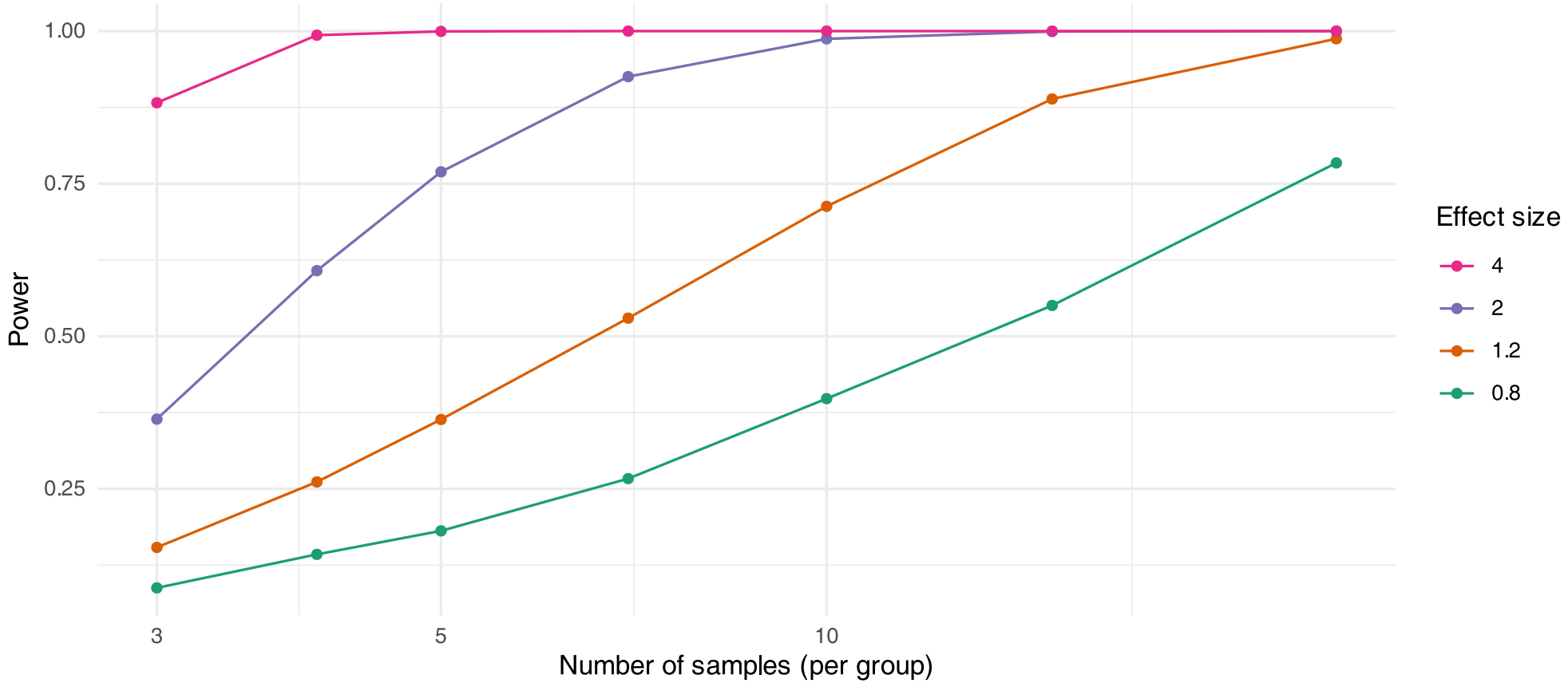
How many samples are sufficient?

Say, we want to compare two groups with a standard $t - test$, nothing fancy. Our ability to detect the differences (the statistical *power*) depends on the sample size and the effect size[1].



How many samples are sufficient?

The y axis on this plot shows how the power of the test – meaning how often, assuming that the groups really differ by d on average, you will be able to detect the difference using a t-test.



Power: power of 50% means that on average, you will be able to see statistical significance in 50% of the experiments *assuming there is a difference!*

How many samples are sufficient?

What about the following setup:

- We have 2 strains (WT and KO)
- We have treatment + control
- We want to know whether the treatment has a different effect on the KO strain than on the WT strain

This is a 2x2 design, and we need to consider the interaction term.

Excel and gene names

- Excel converts some words to dates automatically
- Gene names like `MARCH1` or `SEPT9` are converted to dates
- In most cases[1], you can't switch off this behavior

How to give us (meta-)data

Part of the communication is passing on the data.

1. Make sure the data is **complete** (batches? replicates?)
2. Identifiers should be unique and non-numeric (`ID1` rather than `1`)
3. Use a separate sheet to describe the meaning of columns
4. Explain abbreviations
5. Make the data machine-friendly
6. Disclose precisely all methods (like models, kit labels etc, request them from service providers!)

What you should demand from your bioinformaticians

- Methods used
- Scripts / pipelines used
- Full processed data
- All results as tables (Excel, CSV etc)

Even if you don't know what to do with all that *now*, you might be needing it in the future!

How (not to) work with Excel

(for your reference)

- Avoid manually changing Excel files
- Never use formatting for data
- Don’t combine values and comments
- Don’t put meta-information into column names
- One sheet = one table
- Header = one line
- Do not use merged cells
- Use consistent file names
- Avoid spaces in file and column names (use underscores)

Some more tips and summaries

Things we don’t like

- Cleaning up data
- Data dredging
- P-hacking
- Post-hoc hypotheses
- Excel
- Manual changes like changing fonts in figures
- Non-reproducible science

Things we love

- Clear questions
- A priori hypotheses
- Challenging statistics
- Creating new tools
- R and Rmarkdown, or
- Python and Jupyter
- Reproducible workflows
- Well organized data