Responsible Research

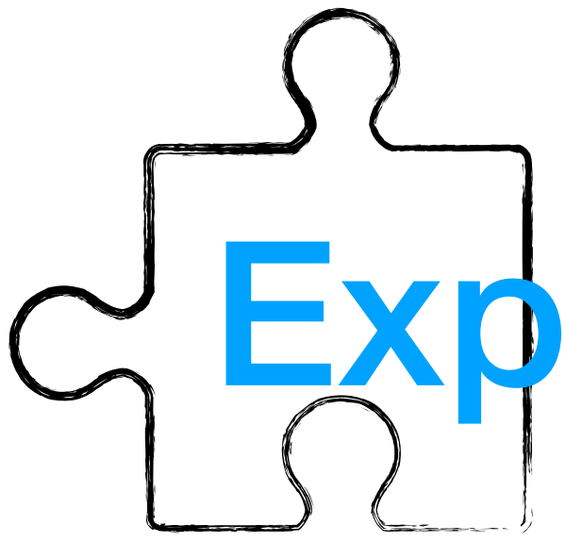
making experiments better

Tobias Straub

tobias.straub@lmu.de

Biomedizinisches Centrum, LMU München

Why am I doing this?



Experiment

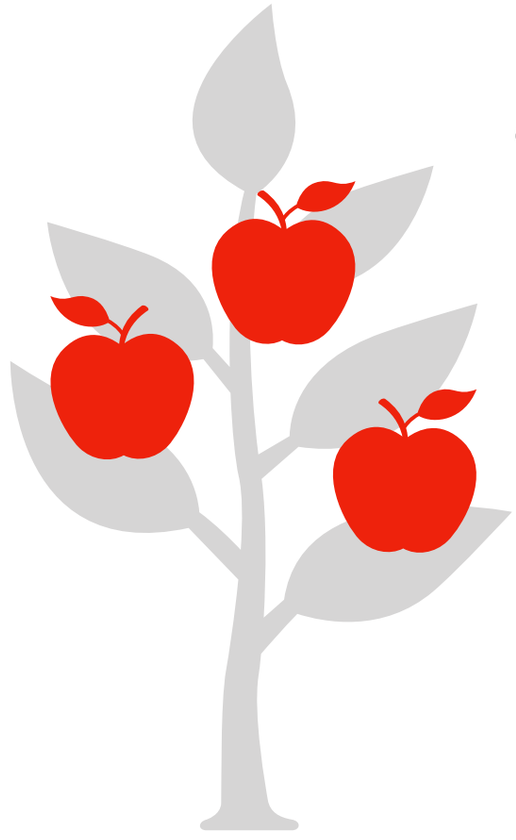
- Purpose: do sth, measure/observe sth, then **predict**!

- Types of experiments
  - Descriptive versus **controlled**
  - Explorative versus **confirmatory**
  - Obscure versus **transparent**
  - Anecdotal versus **representative**
  - Context: Isolated versus **ensemble**
- Types of outcomes
  - **Convincing** versus unconvincing
  - **Reproducible** versus irreproducible
  - Career-changing breakthrough versus failed

- Matters: Data management, documentation, visualisation, interpretation

Taking measurements

descriptive versus controlled

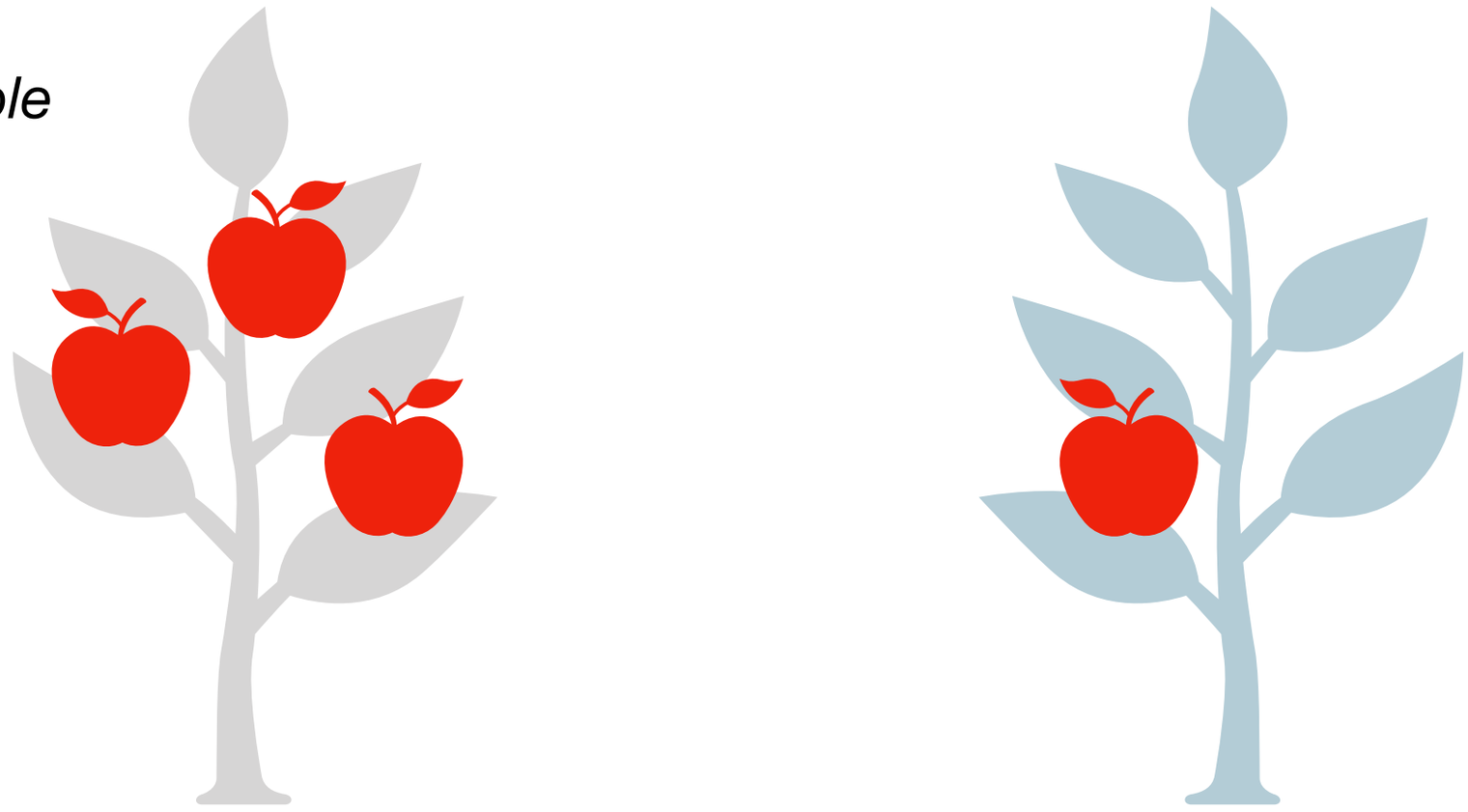


Number of apples
endpoint

descriptive versus **controlled**

Difference in number of apples
The Effect

Number of apples
endpoint
dependent variable

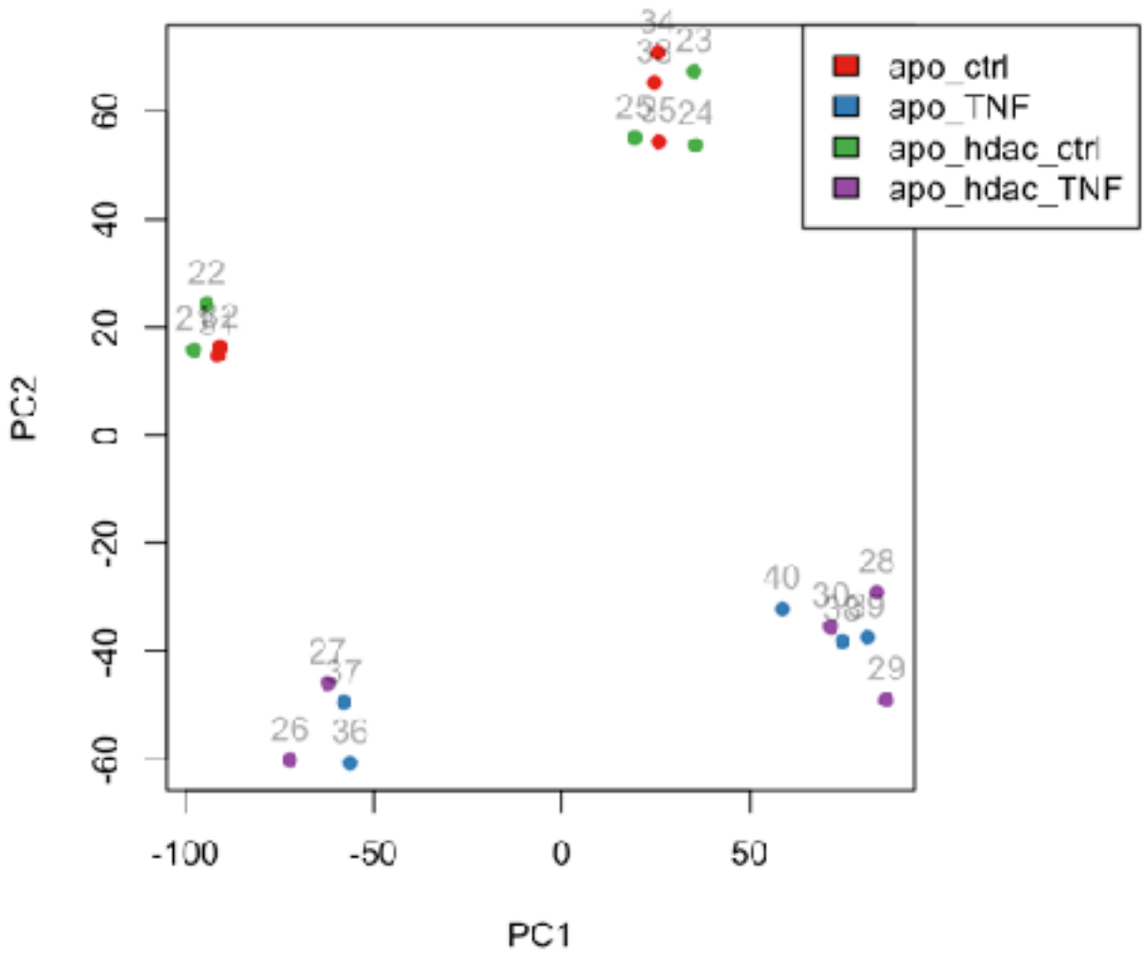


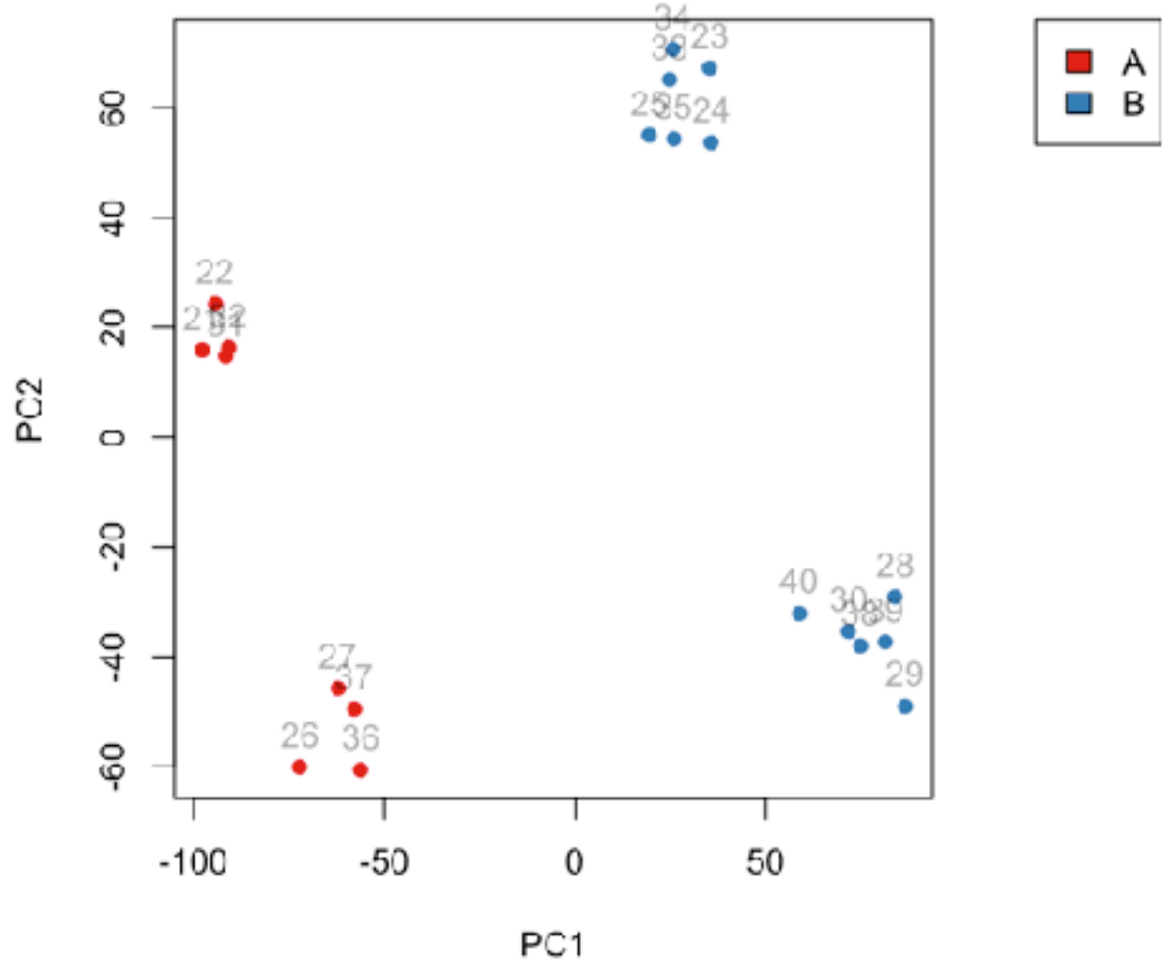
Mutant *Wildtype*

Genotype
independent variable

systematic influences on measurement are eliminated
relate the effect observed to the treatment, causal inference

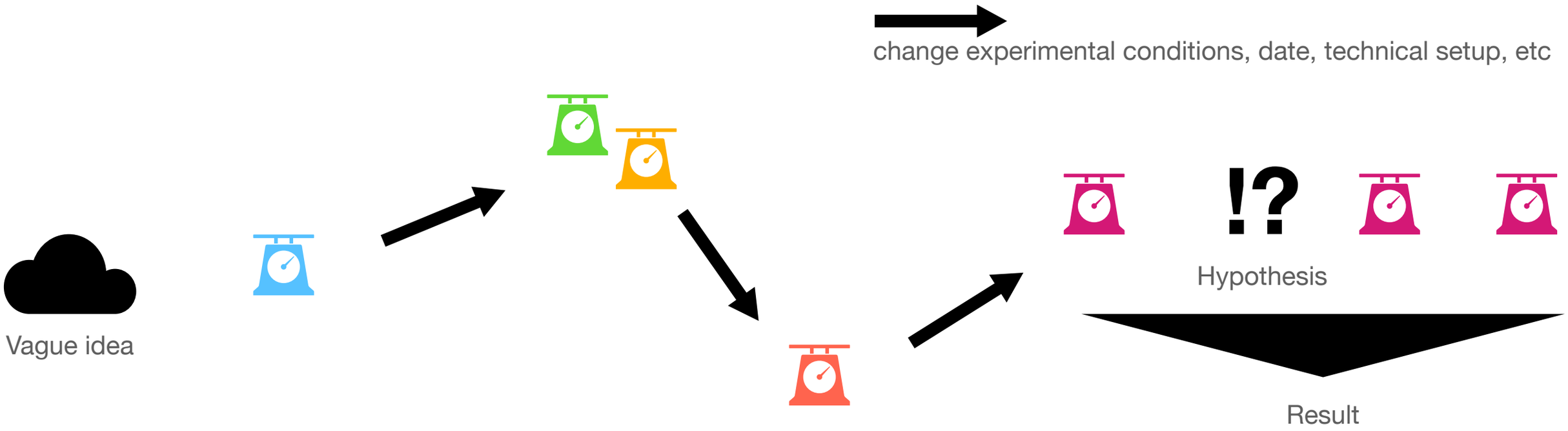
Batch effects





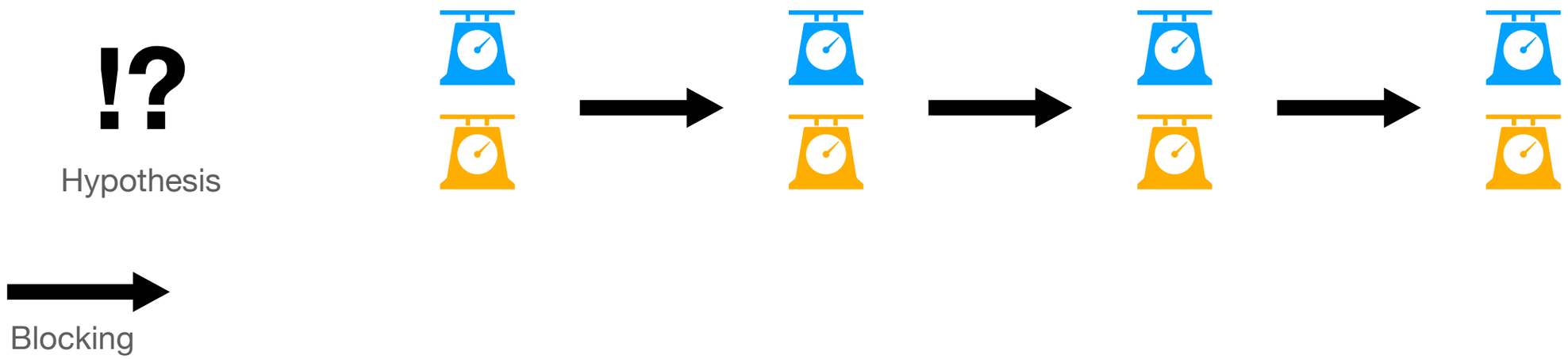
Planing measurements

Exploratory Research



- Tends to be highly **irreproducible**
- Statistical evaluation highly problematic (**sampling bias, power**)
- Data organisation and documentation potentially **compromised** (lack of structure, retrospective collection & assembly)

Confirmatory Research



- **Statistical** evaluation possible
- Standardised measurements/data acquisition
- Estimation of robustness (**reproducibility**) possible
- Requires **Experimental Design**
- Prior definition of Raw Data and Metadata. **Open Science** practices facilitated.

Open Science

State-of-the-art
Documentation and Data Management

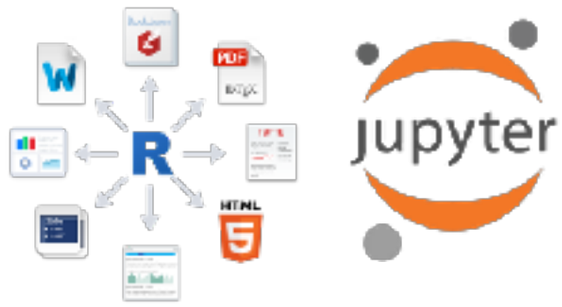
digital documentation

Raw data

Analysis result

Metadata

Data processing

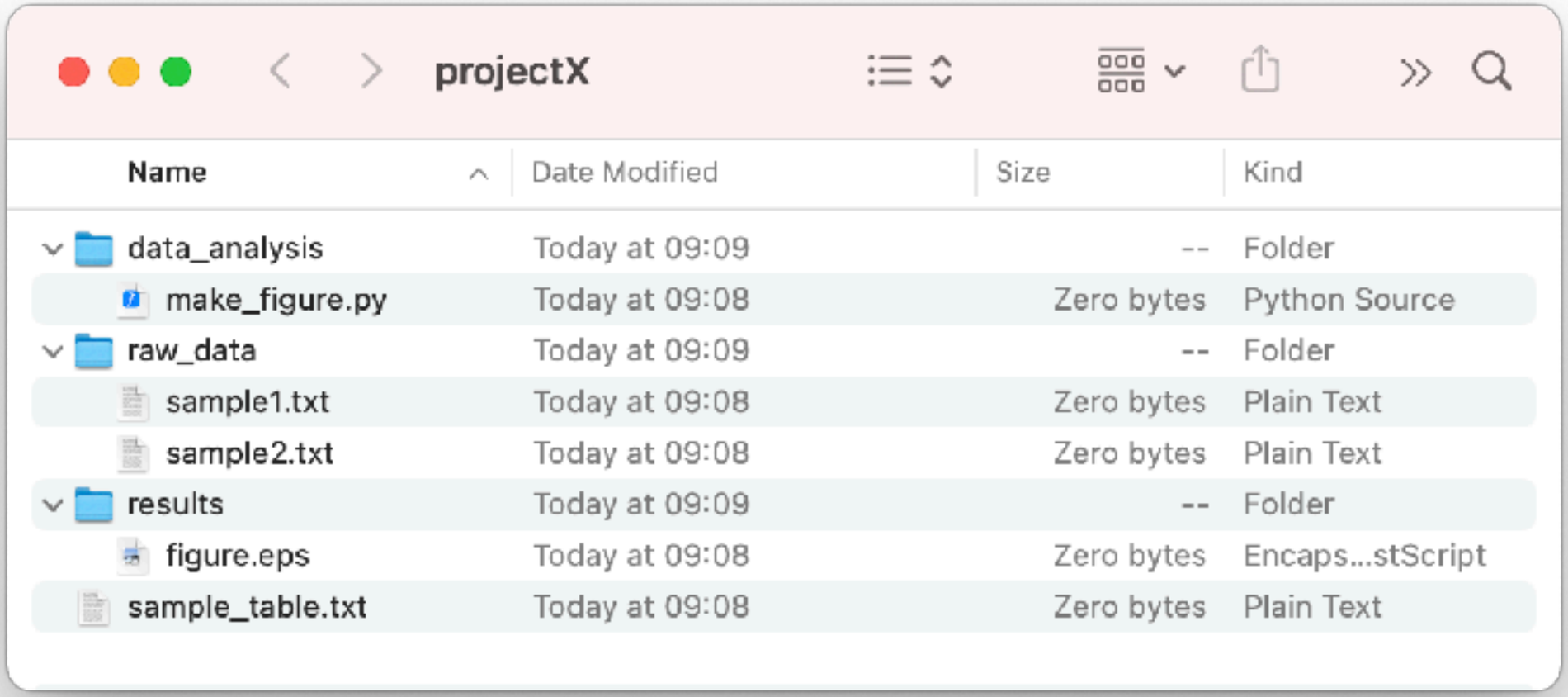


Materials and Methods
Experimental Setup
eLab Journal

Open science practices (in short)

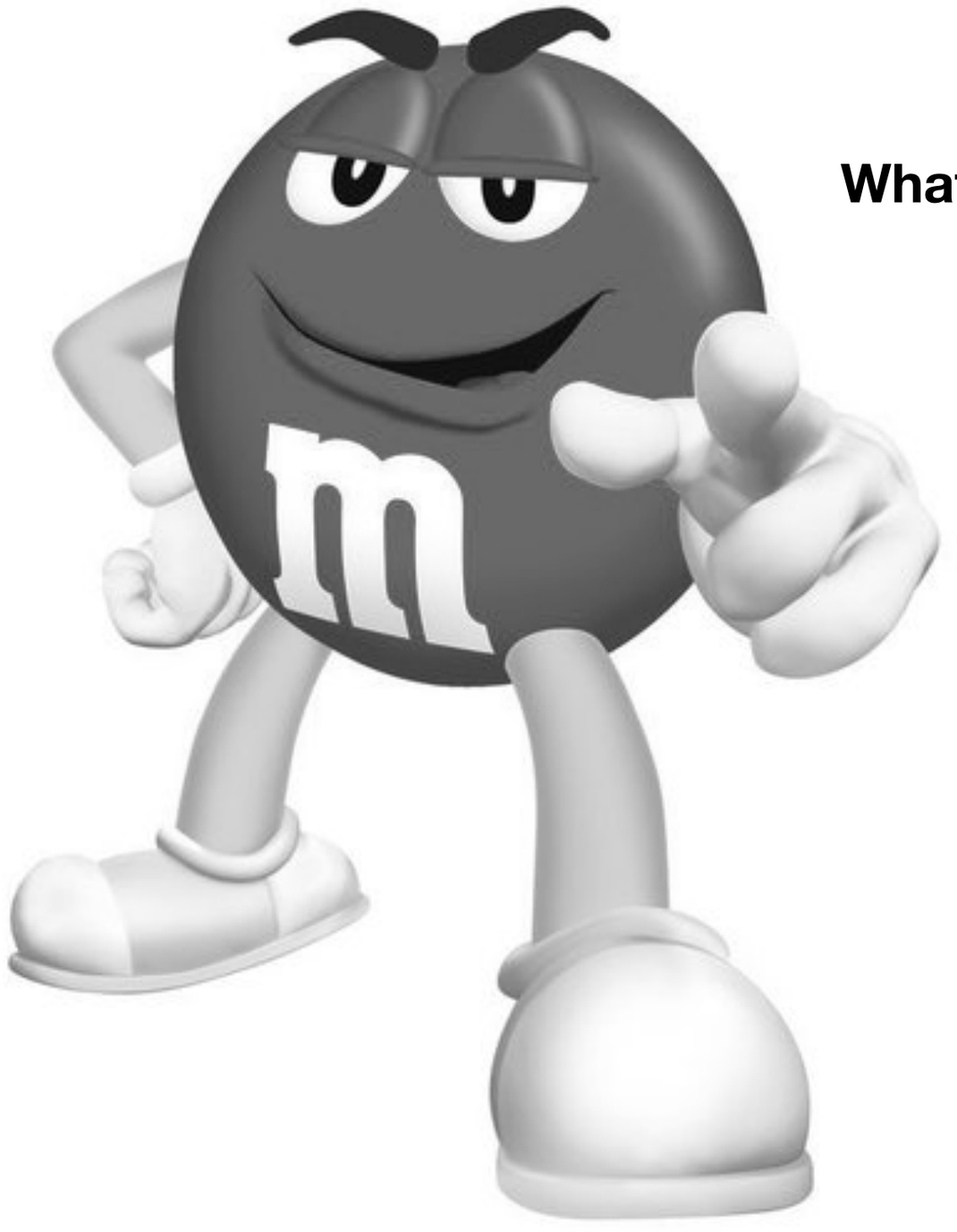
- All files centrally stored in universal format accessible/findable for group/peers/world (FAIR)
- Safely backed up
- Avoid redundancy
- Never edit the raw data files
- All relevant information provided in meta data table.
- Data analyses and environment fully documented to ensure analytical reproducibility

Data organisation



Sampling & Representativity

N is SMALL



What's my colours?

Any expectation?

Experiment

Population

N = usually infinite

Sample

N = 3


LARGE
330g
DAUERHAFT MEHR ZUM TEILEN
m&m's

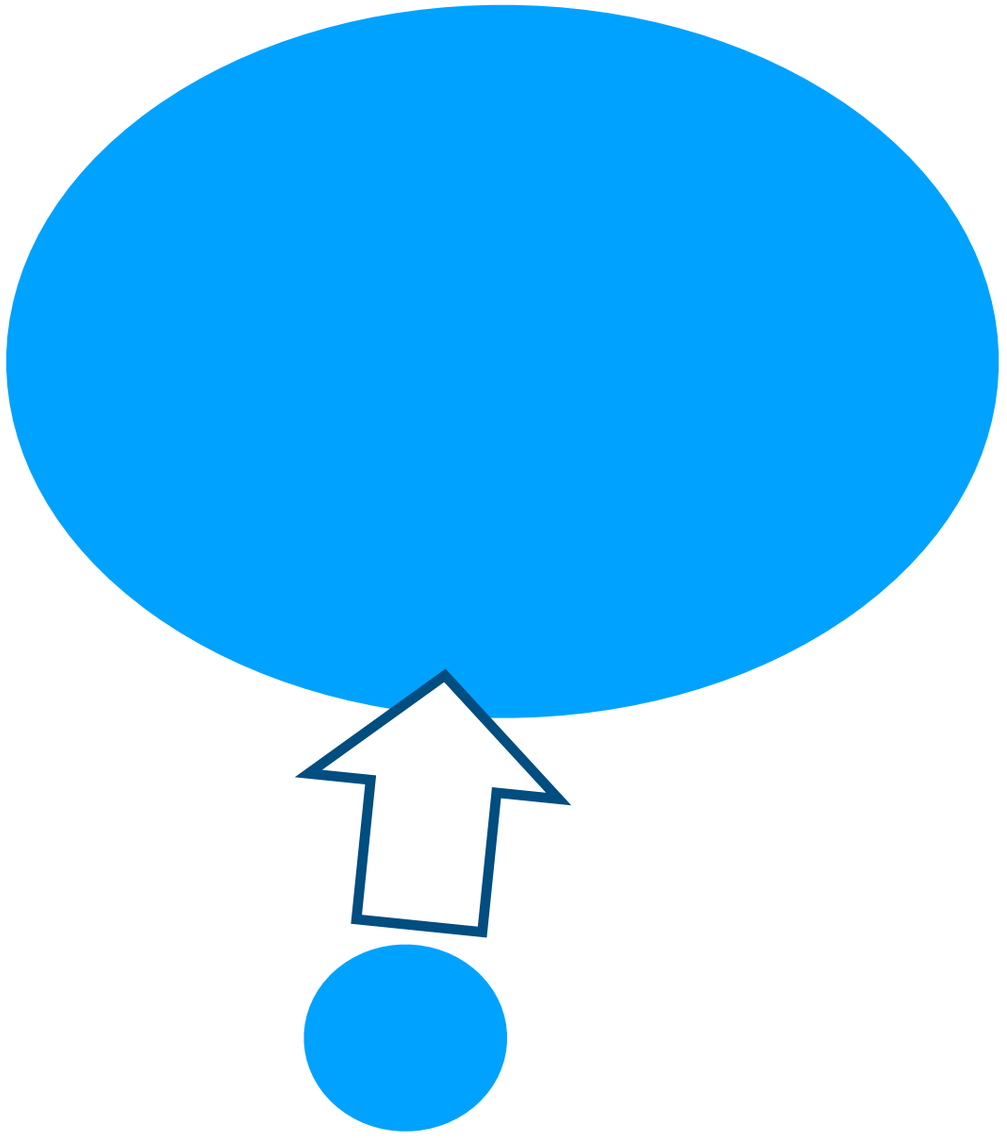

LARGE
330g
m&m's


LARGE
330g
DAUERHAFT MEHR ZUM TEILEN
PEANUT
m&m's


LARGE
330g
DAUERHAFT MEHR ZUM TEILEN
PEANUT
m&m's

Inference

Estimate the colours of the population from the colours of the sample



sweets

Irreproducible anecdote

All M&Ms
are yellow!

Reporting Bias

“My M&Ms are yellow”

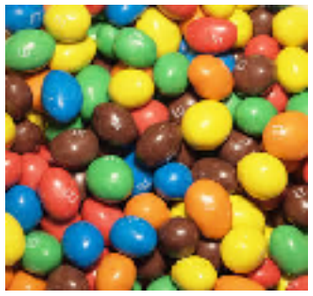
does not equal

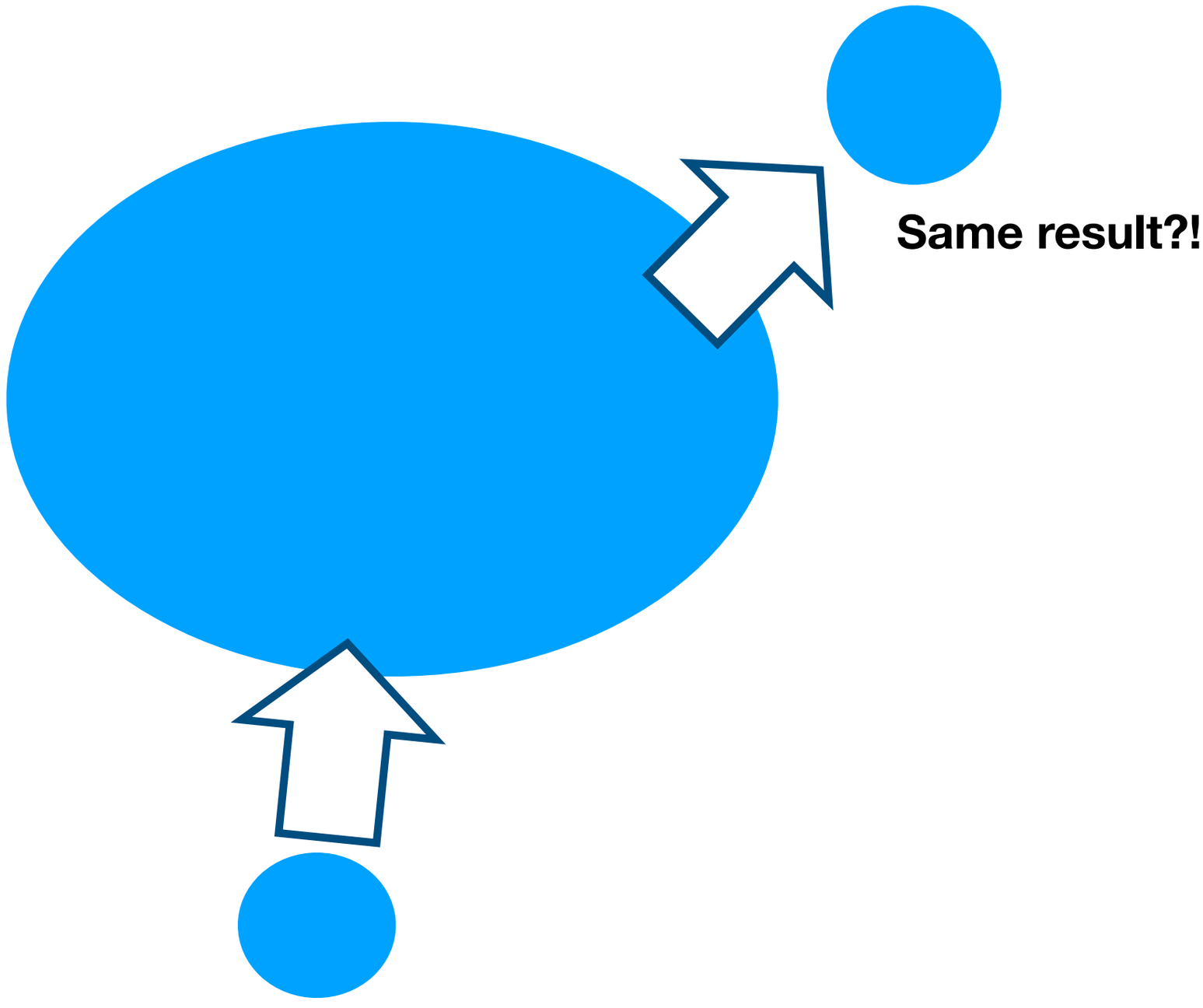
“All M&Ms are yellow”

There is no stronger force in science but the love of a scientist for her/his own results

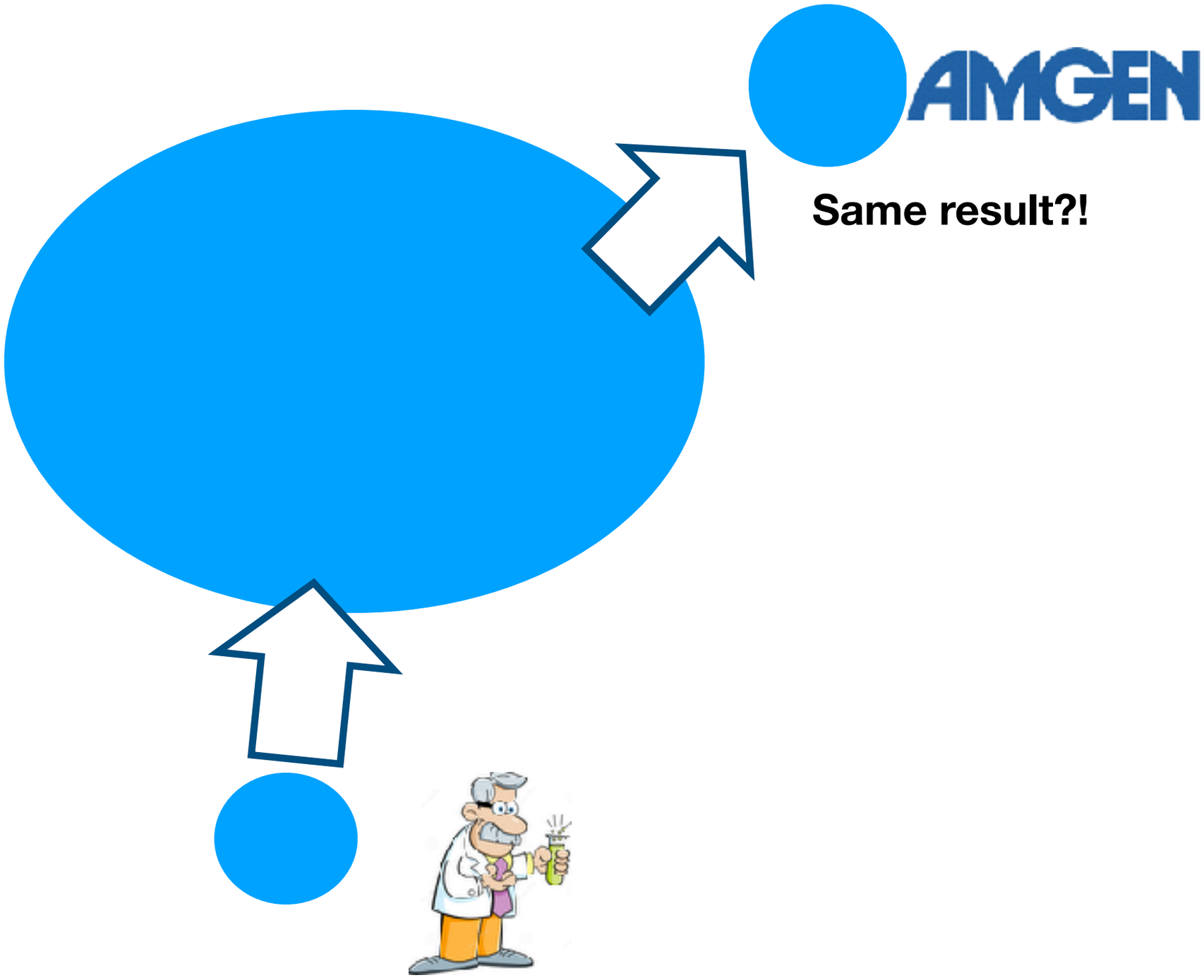
Reproducibility

Direct, Conceptual, Analytical …

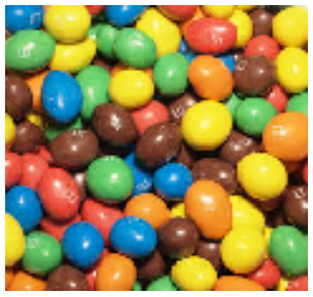


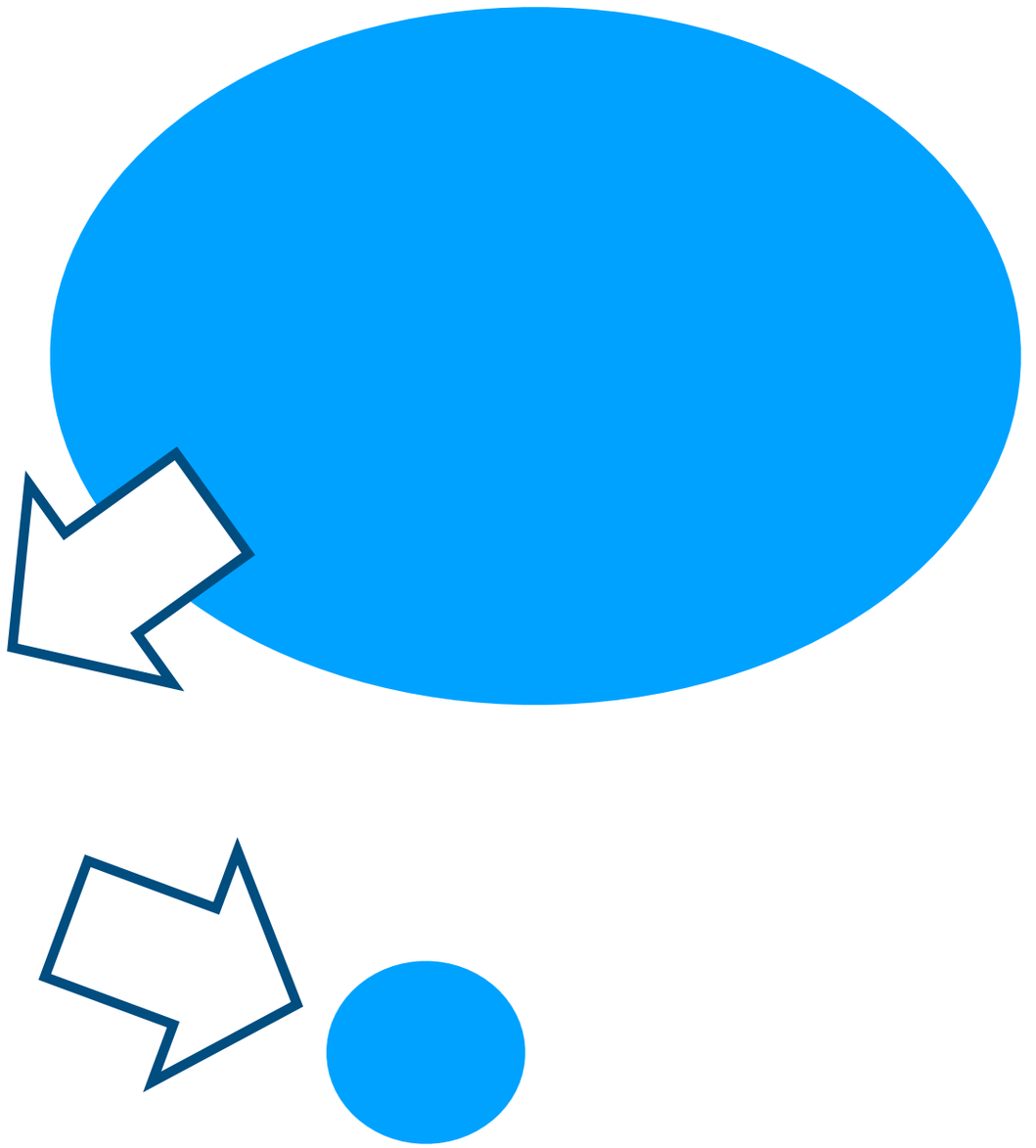

Same result?!

Roadblock to Translation

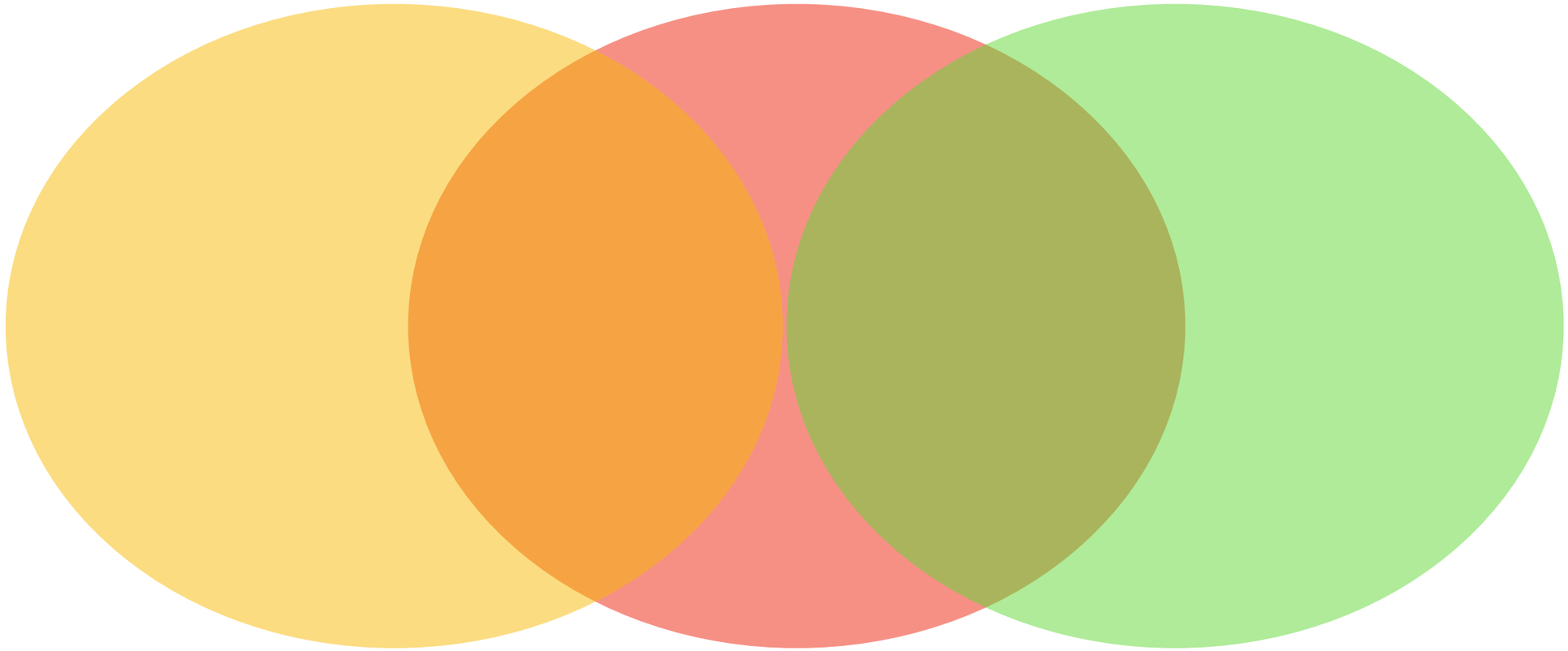


Sampling space = population?





What defines the population?



130-140 mM NaCl

140-150 mM NaCl

Exactly 140 mM NaCl

Reproduction of the original results using the same protocol/reagents/tools

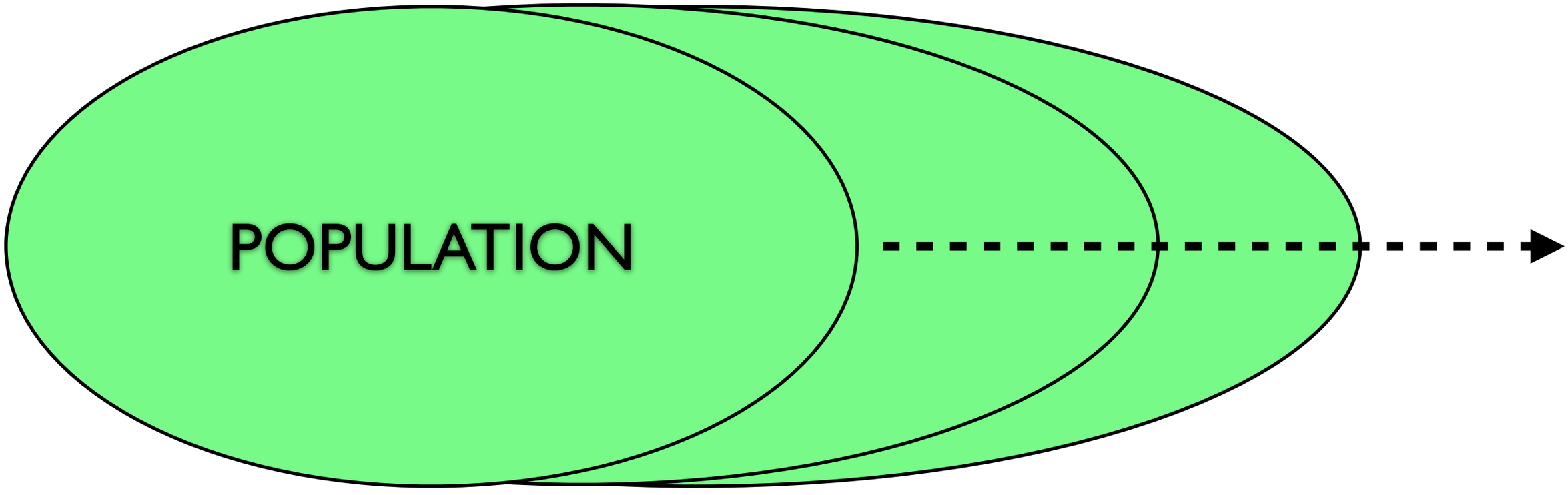
by the same person

by a different person in the lab

by a different person outside the lab

Reproduction using different reagents/ tools but the same protocol by a different person outside the lab

Reproduction just based on text description

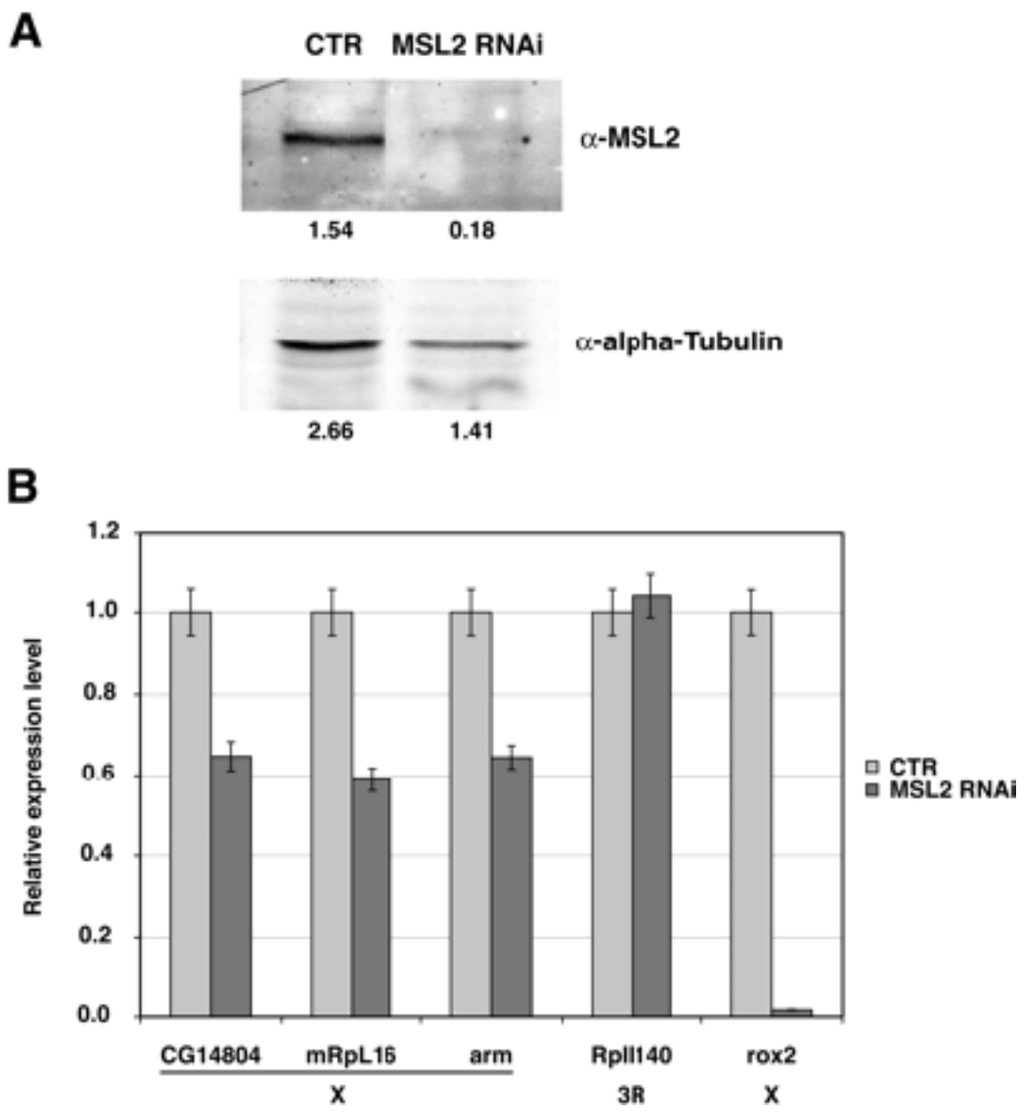


Parameters to consider

- Environmental conditions (animal house, food, health state, mood, season, time of day …)
- Cell type
- Reagents (serum!, antibodies!!!, …)
- Technical (detection limits, …)
- Genomic background, secondary (off-target) mutations
- …

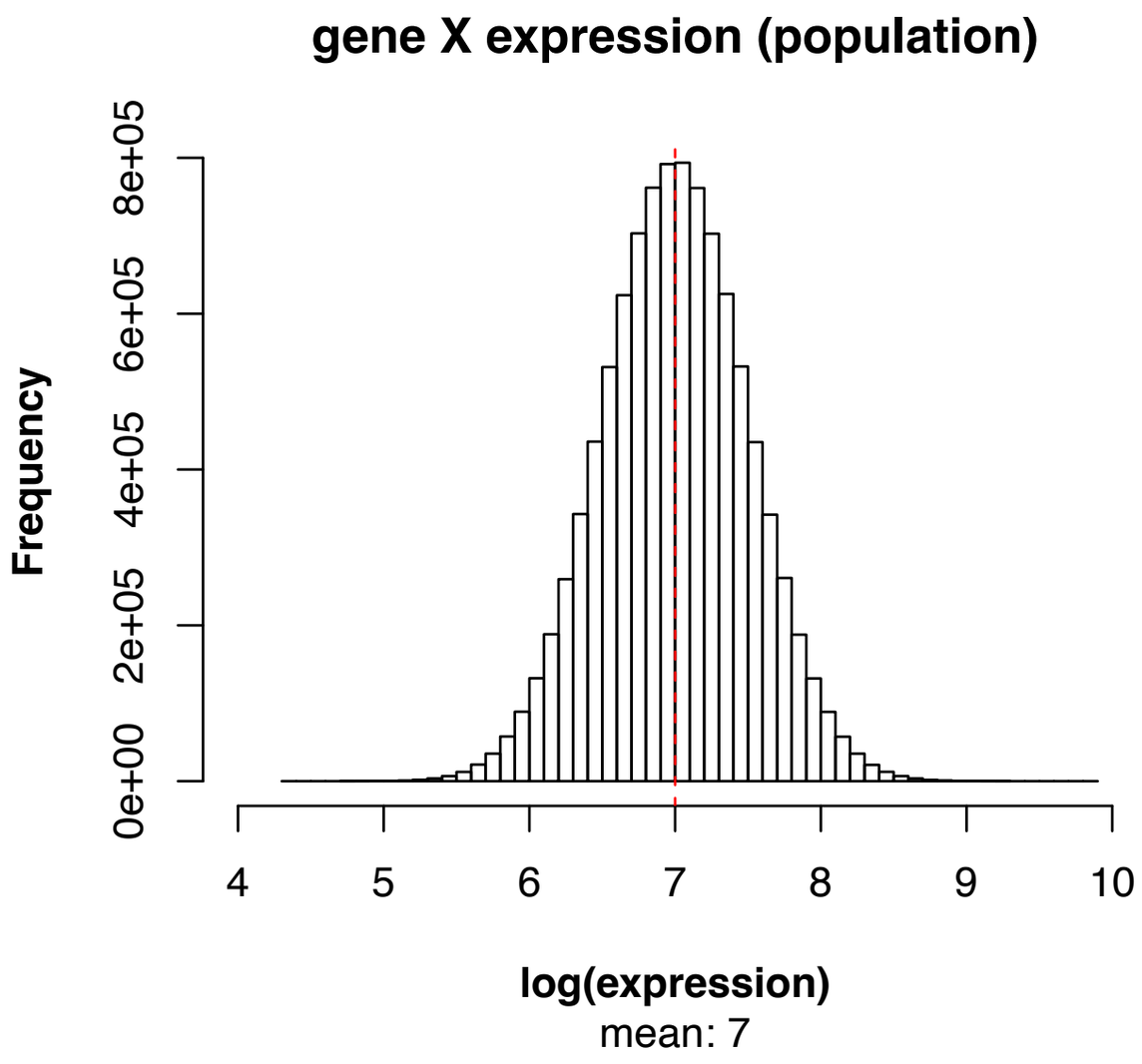
Certainty

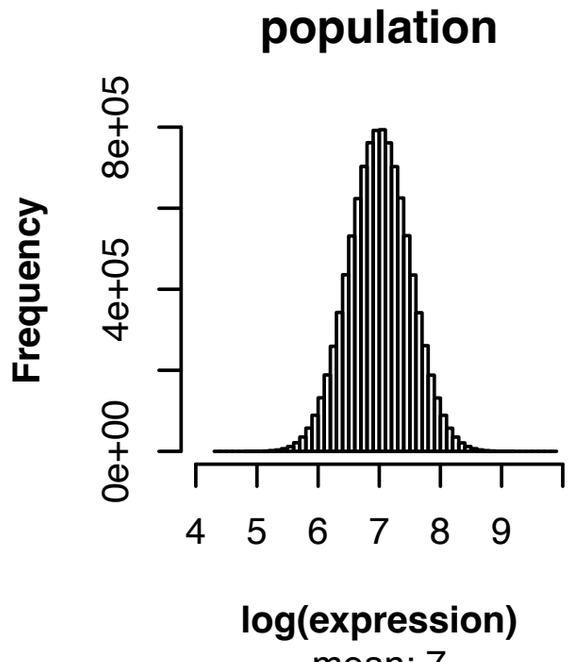
Is This True?

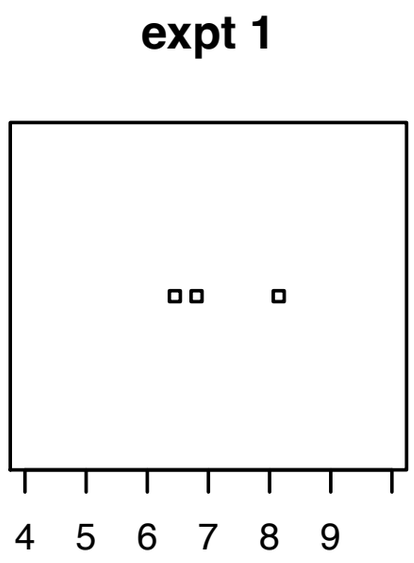


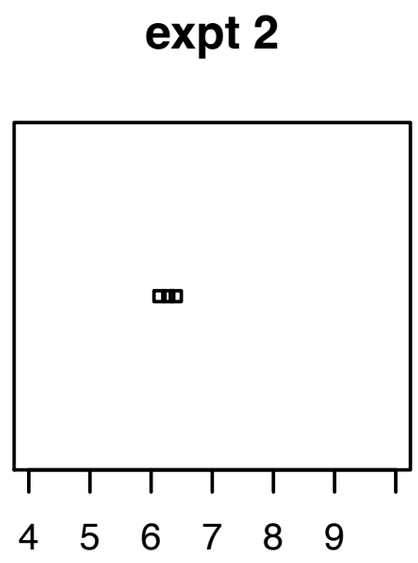
Straub T et al, Genes Dev. 2005

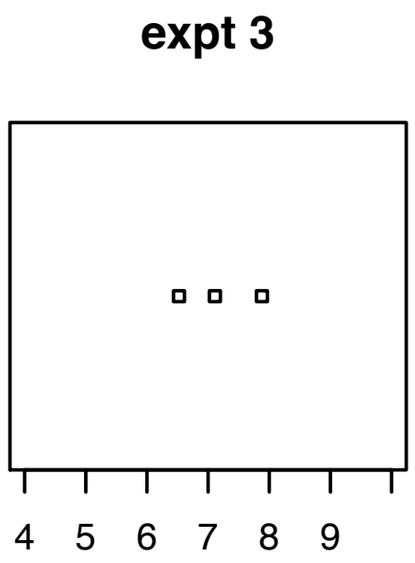
Objective Inference

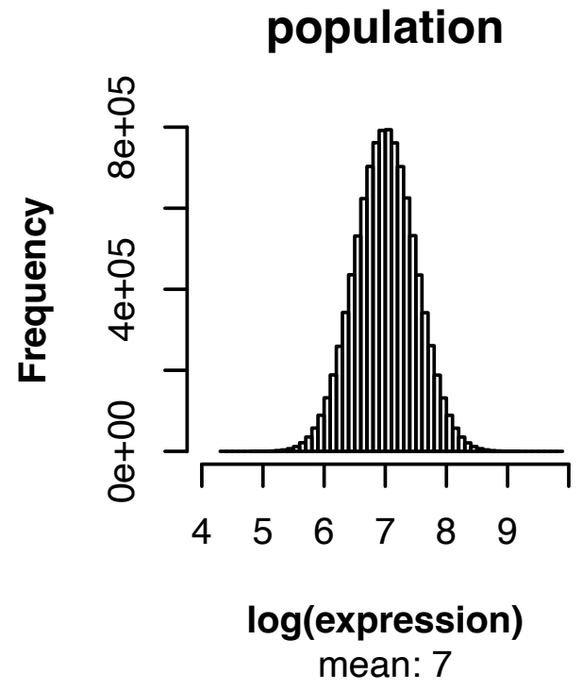

gene X expression (population)
Frequency
0e+00
2e+05
4e+05
6e+05
8e+05
4
5
6
7
8
9
10
log(expression)
mean: 7

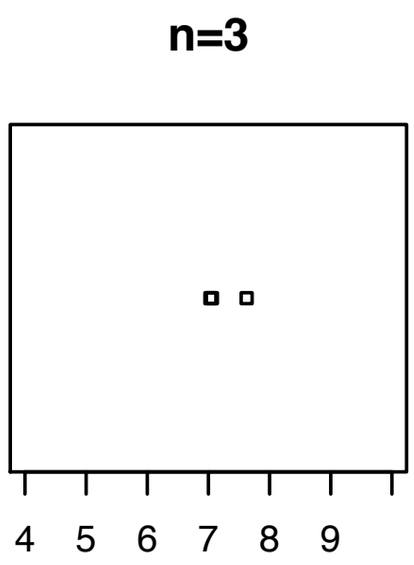

population
Frequency
0e+00
4e+05
8e+05
4
5
6
7
8
9
log(expression)
mean: 7

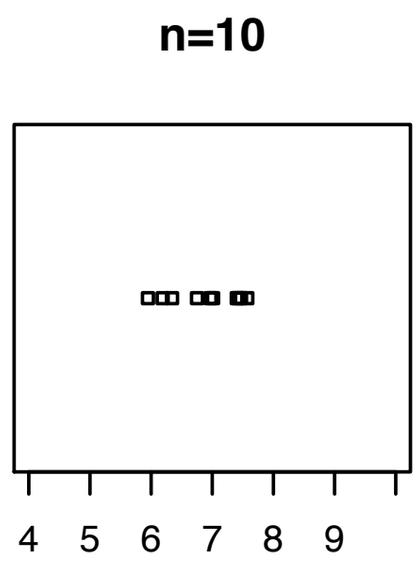

expt 1
4
5
6
7
8
9
log(expression)
mean: 7.135

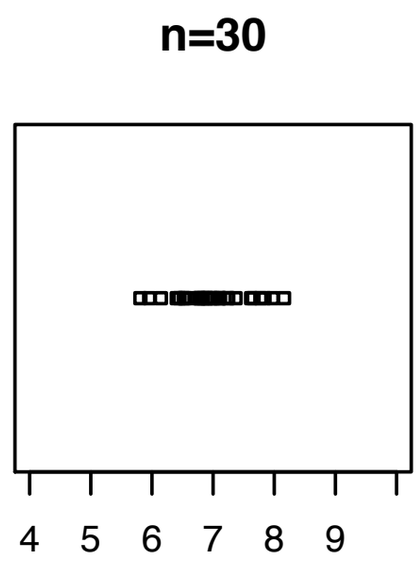

expt 2
4
5
6
7
8
9
log(expression)
mean: 6.274


expt 3
4
5
6
7
8
9
log(expression)
mean: 7.172


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7


n=3
4 5 6 7 8 9
log(expression)
mean: 7.238


n=10
4 5 6 7 8 9
log(expression)
mean: 6.908


n=30
4 5 6 7 8 9
log(expression)
mean: 6.972

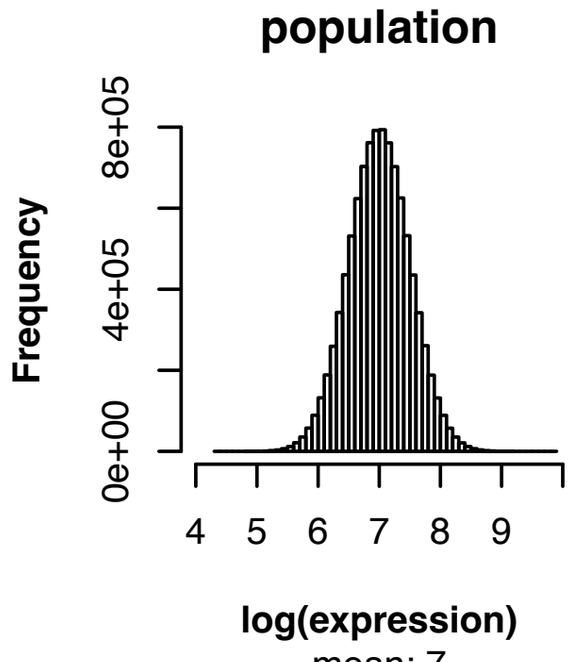
Standard Error (of the mean)

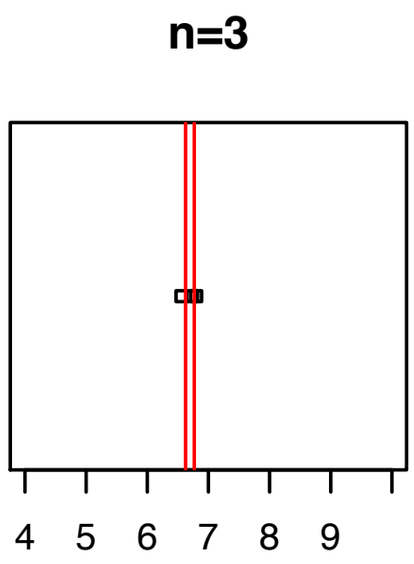
- The standard error of the mean (SEM) is the standard deviation of the sample mean estimate of a population mean.

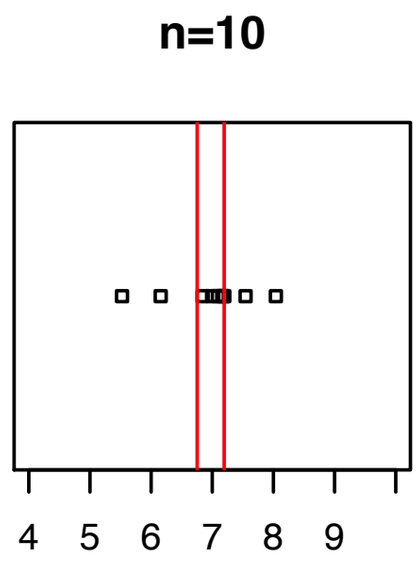
  SEM = standard deviation/square root(n)

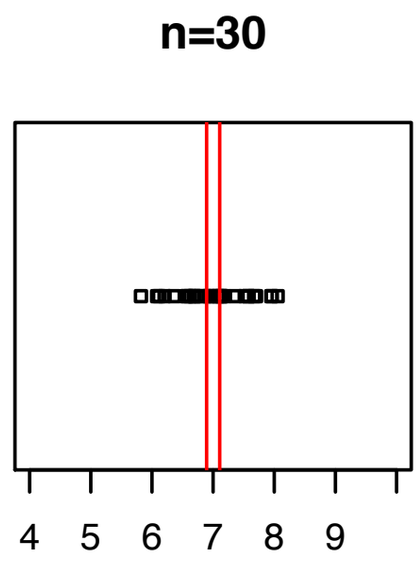
- a small SEM indicates that the sample mean is likely to be quite close to the true population mean

- a large SEM indicates that the sample mean is likely to be far from the true population mean


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7

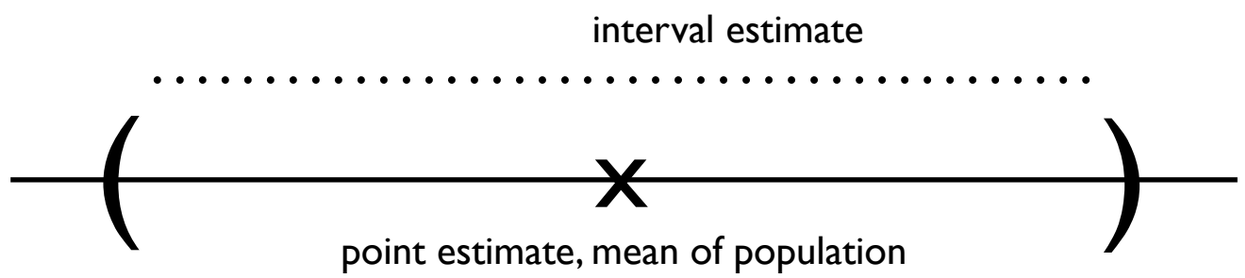

n=3
4 5 6 7 8 9
log(expression)
mean: 6.698


n=10
4 5 6 7 8 9
log(expression)
mean: 6.974

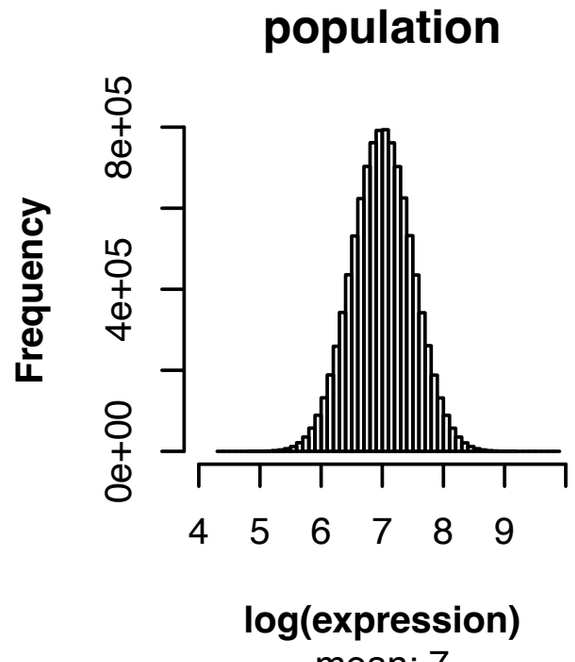

n=30
4 5 6 7 8 9
log(expression)
mean: 7.001

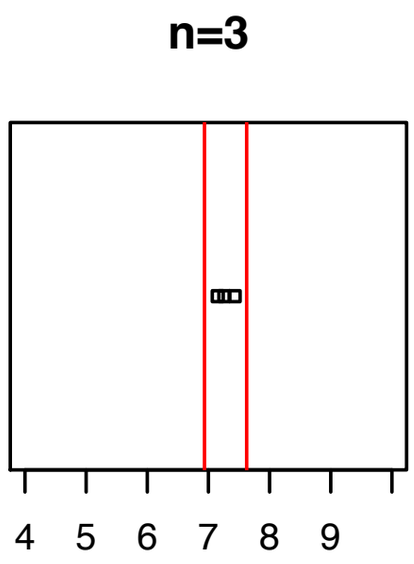
Confidence Intervals

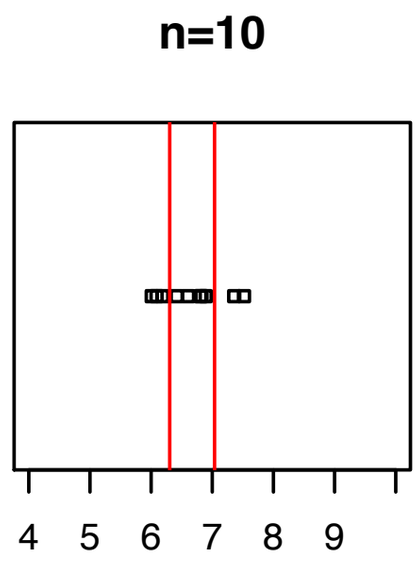
- 95%-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of 95%.

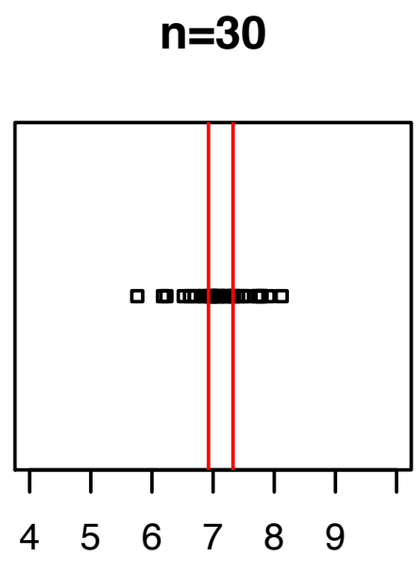


- (1– α)-confidence interval: An estimated interval which contains the „true value“ of a quantity with a probability of (1–α).
  1–α = confidence level, α = error probability


population
Frequency
0e+00
4e+05
8e+05
4 5 6 7 8 9
log(expression)
mean: 7


n=3
4 5 6 7 8 9
log(expression)
mean: 7.28


n=10
4 5 6 7 8 9
log(expression)
mean: 6.671


n=30
4 5 6 7 8 9
log(expression)
mean: 7.125

Practical example

Someone asks: “how many dead cells are in your culture?”

You use a hemocytometer to determine the viability of cells stained with trypan blue. You count 94 unstained cells and 6 stained.

How can the data be reported?

95% CI=0.02-0.13

Example taken from “Intuitive Biostatistics,” H. Motulsky

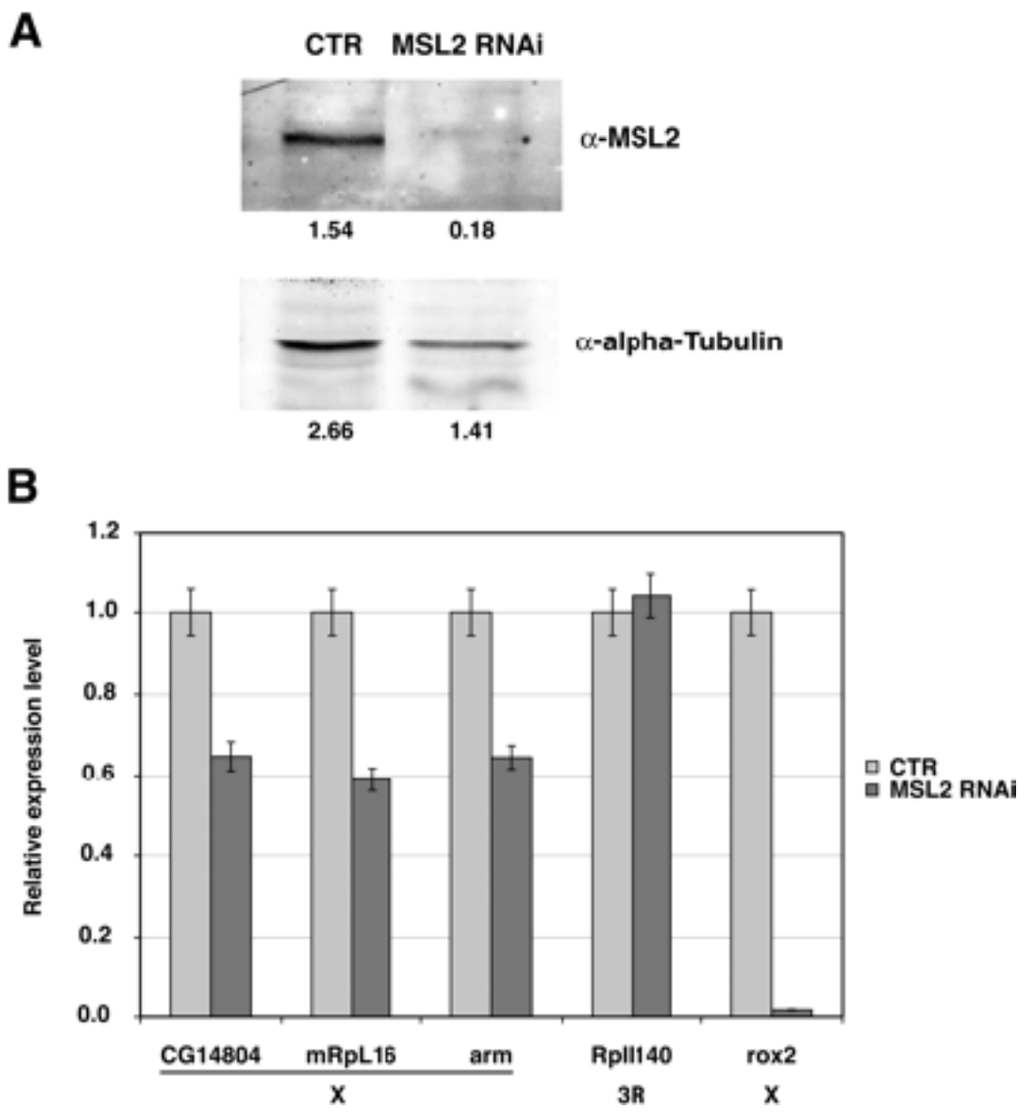
Prerequisites for inference

- the sample has to be representative
- how is representativity achieved?
  - large sample number (N)
  - independent sampling/
    random recruitment of samples
    (BIOLOGICAL replicates)

Unbiased Sampling

- Discard results only if:
  - objective **exclusion criteria** have been defined beforehand. Use positive/negative **controls**.
  - made transparent.
- **Randomisation** measures and blinding help fight the bias
- In “gonzo” designs, selection of sample objects defines the population.

Truth is one, Paths are many @lakens



Straub T et al, Genes Dev. 2005

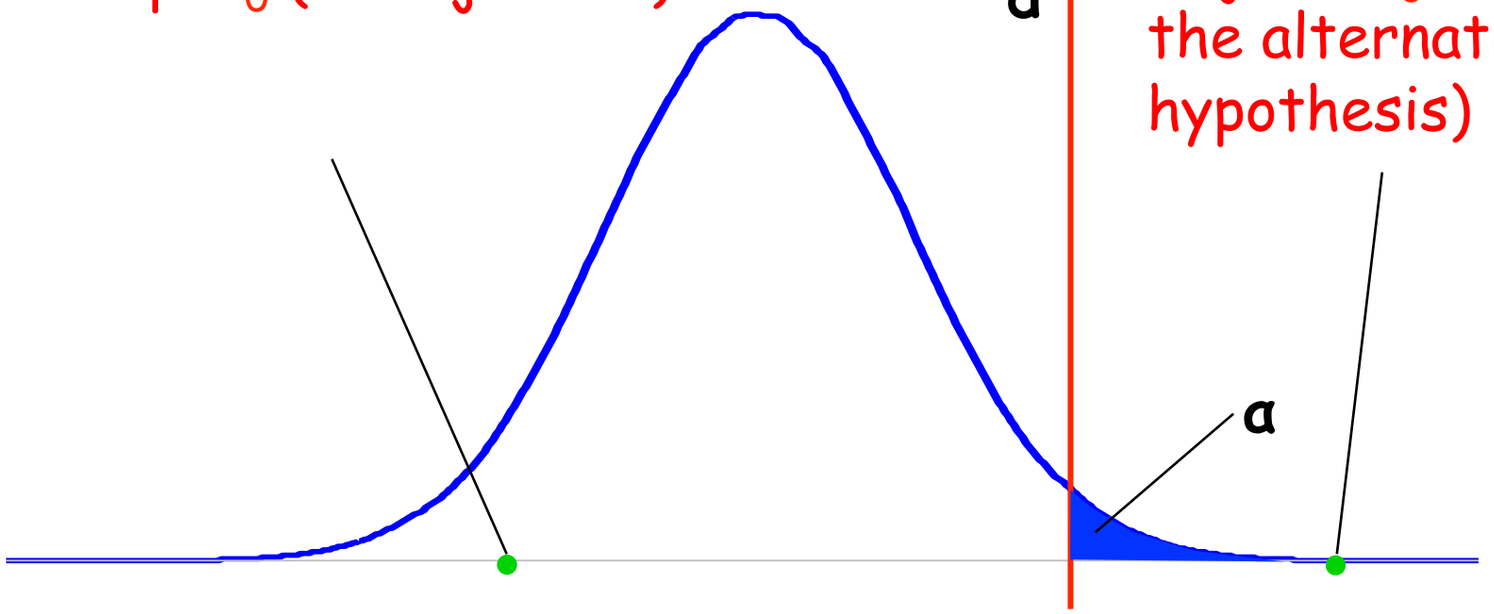
Truth is one, Paths are many @lakens

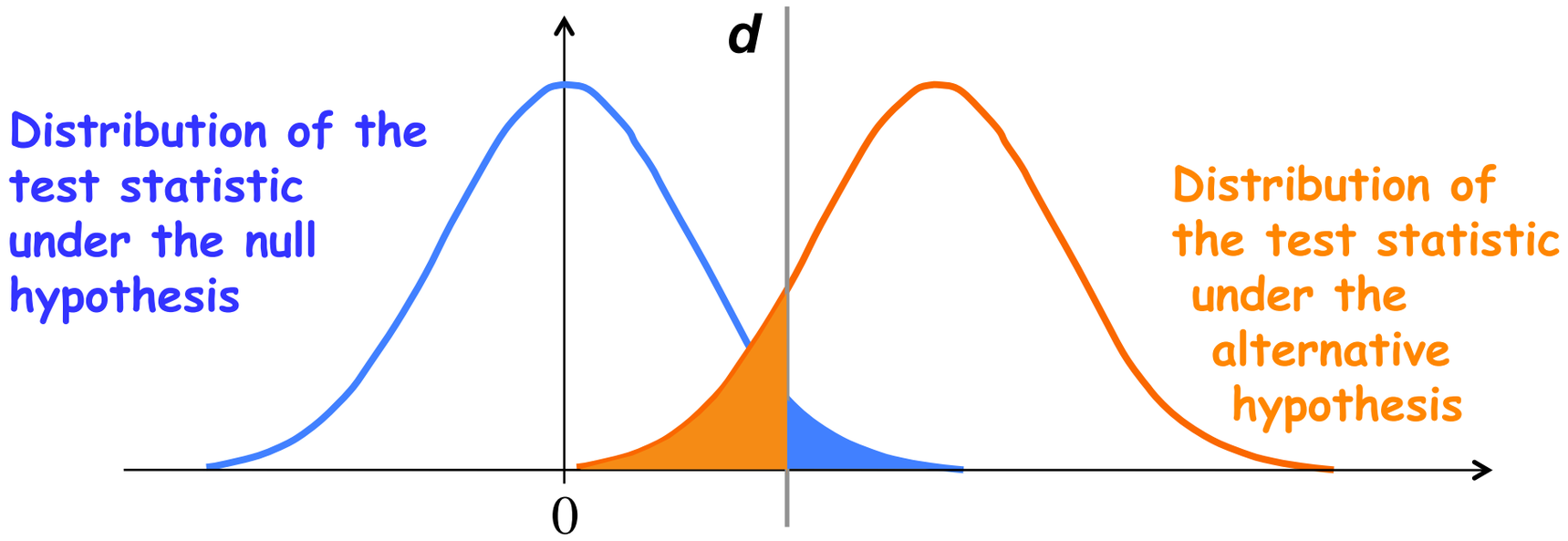
- Frequentist (Neyman Pearson):
  - evaluation of ongoing data collection. Says nothing concrete about the current test
- Likelihood
  - Compare the likelihood of different hypotheses given the data
- Bayesian Statistics
  - Update prior beliefs in light of new data

Keep H_0 (no rejection)

d

Reject H_0 (assume the alternative hypothesis)





	Accept null hypothesis	Reject null hypothesis
null hypothesis is TRUE	correct decision	Type I Error "False Positive"
alternative hypothesis is TRUE	Type II Error "False Negative"	correct decision

Statistical Testing

- Statistical tests are **inferential** procedures.

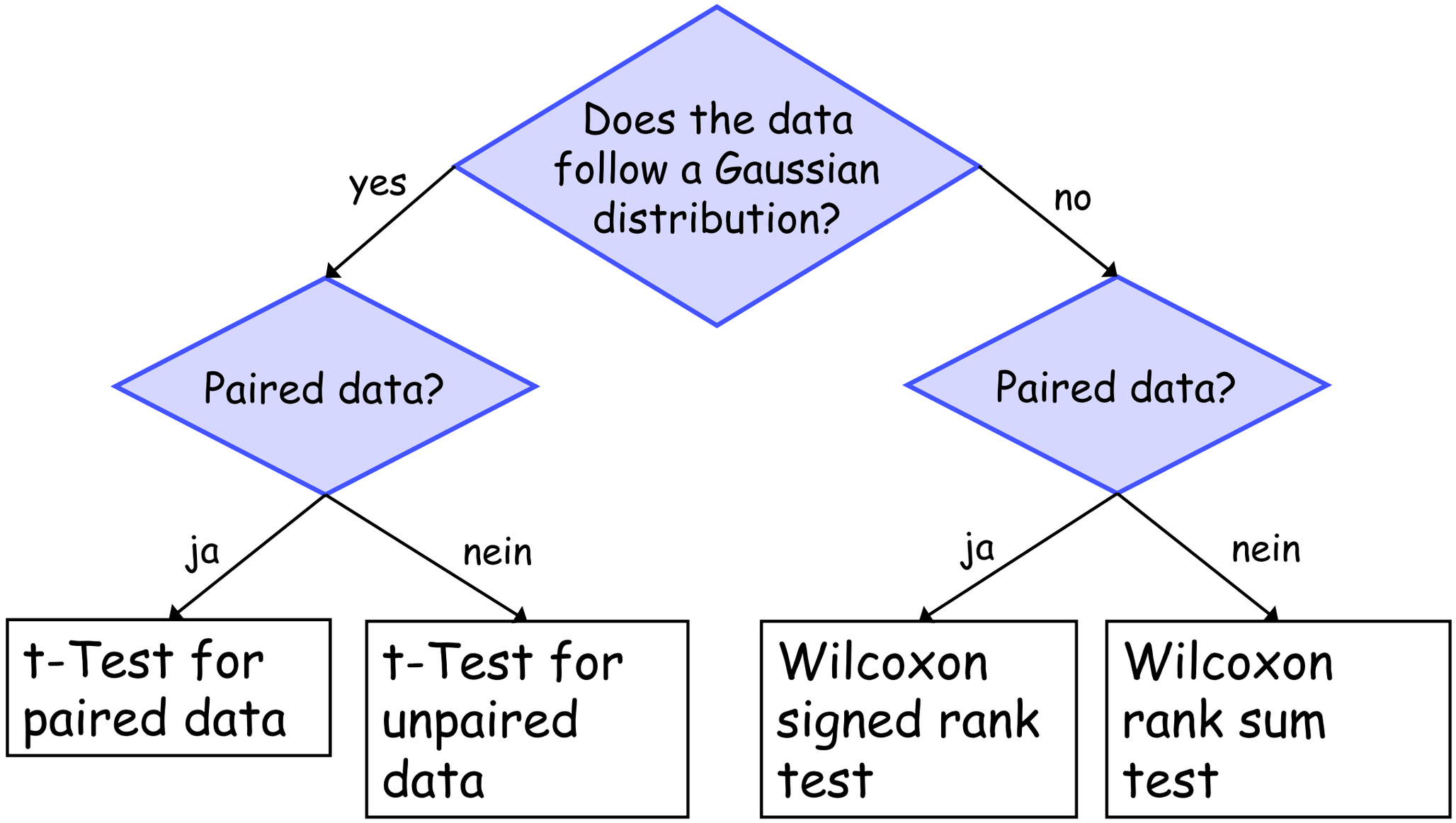
- Statistical tests only address the robustness of effects if applied over independent **biological replicates**.
  (watch out X-omics!)

- No testing without education otherwise you are wrong.
  Understand the **prerequisites** for statistical testing procedure.

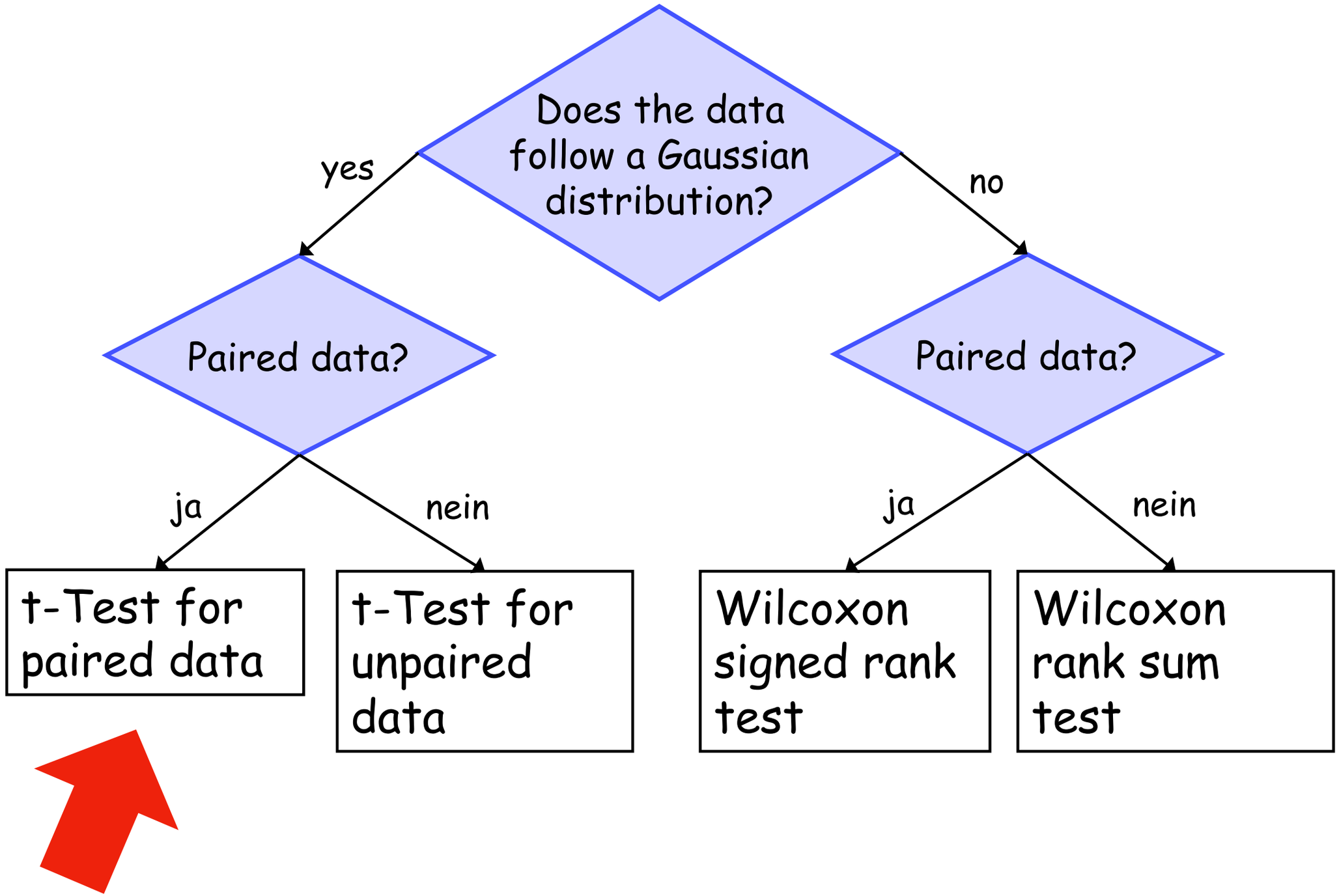
- Understand the meaning of **statistical errors** and the concept of **multiple testing**.

- Understand the meaning of **statistical power**.

Question: Are group 1 and group 2 identical with respect to the distribution of the endpoint?

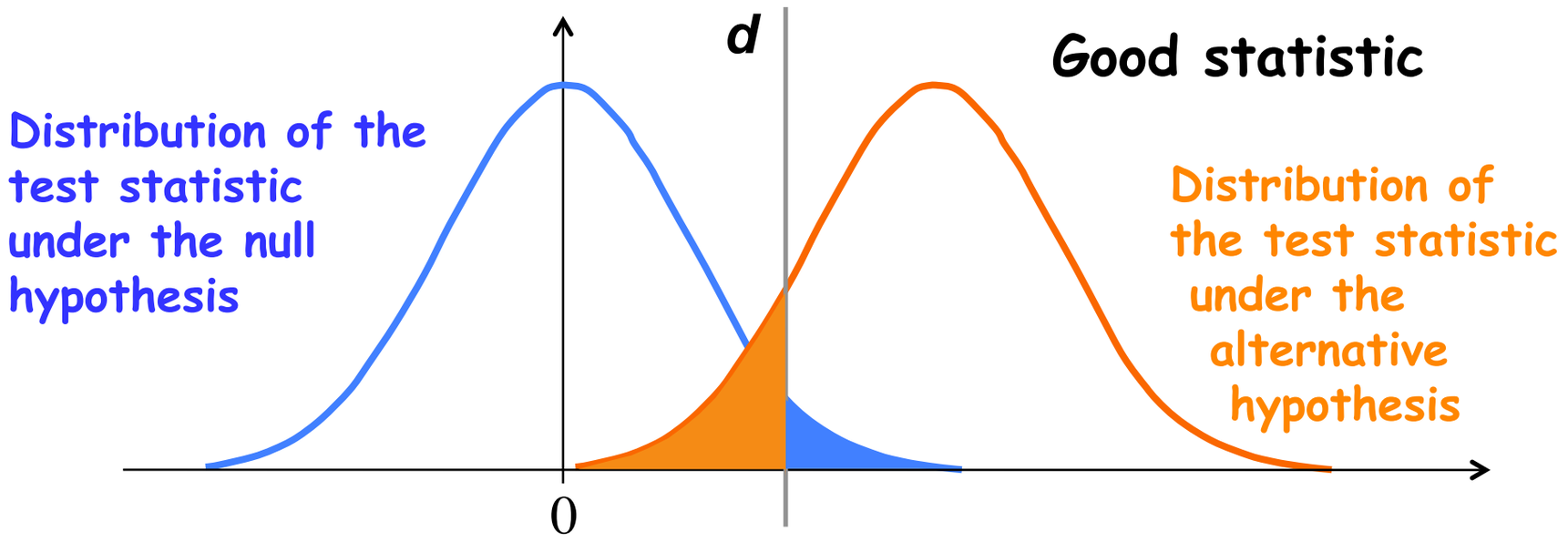


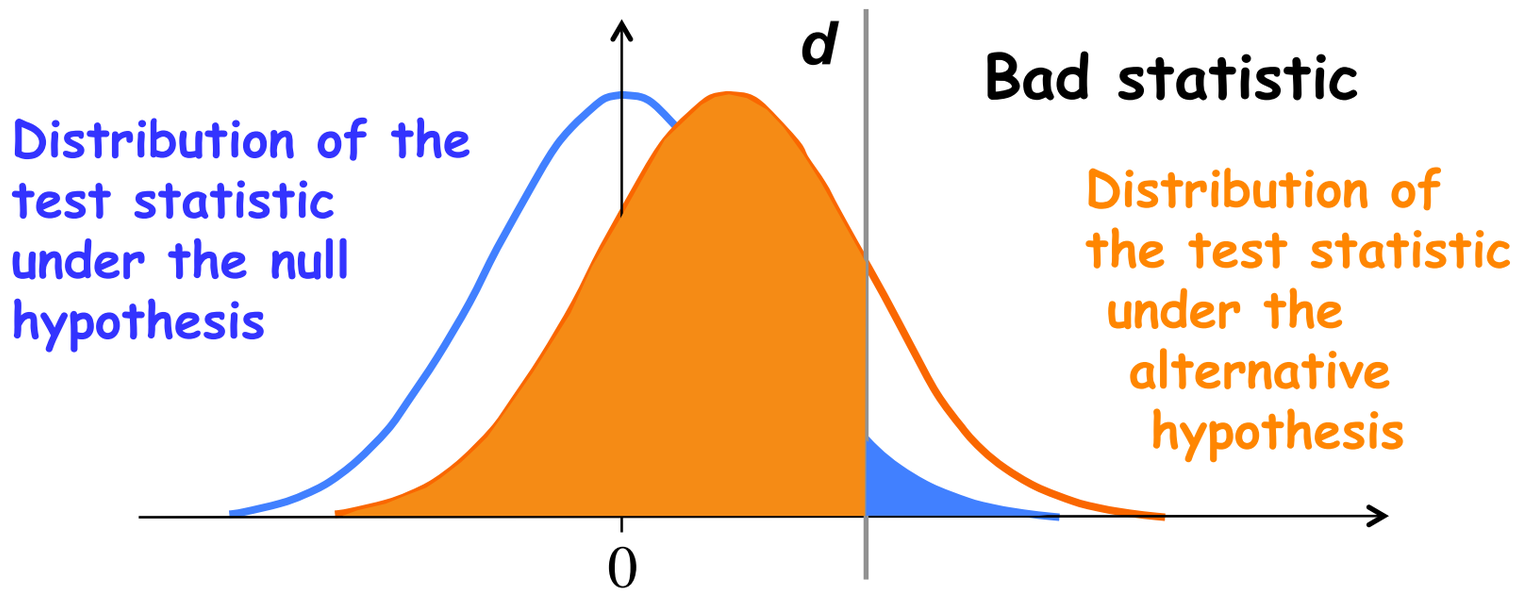
Question: Are group 1 and group 2 identical with respect to the distribution of the endpoint?



Statistical Power

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2


Good statistic
d
Distribution of the test statistic under the null hypothesis
Distribution of the test statistic under the alternative hypothesis
0


Bad statistic
d
Distribution of the test statistic under the null hypothesis
Distribution of the test statistic under the alternative hypothesis
0

- Depending on the magnitude and variability of the effect (effect size) being investigated an experiment requires an **adequate sample size** (N) for reliable evaluation by statistical testing procedures.

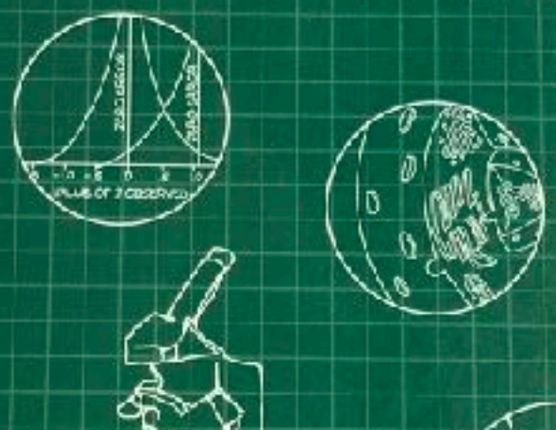
- Very often N is too small, study **underpowered**, true effects not detected.

- If N is very large, small, potentially irrelevant effects become significant. (i.e. **overpowered** study)

Experimental Design

Experimental Design for Biologists

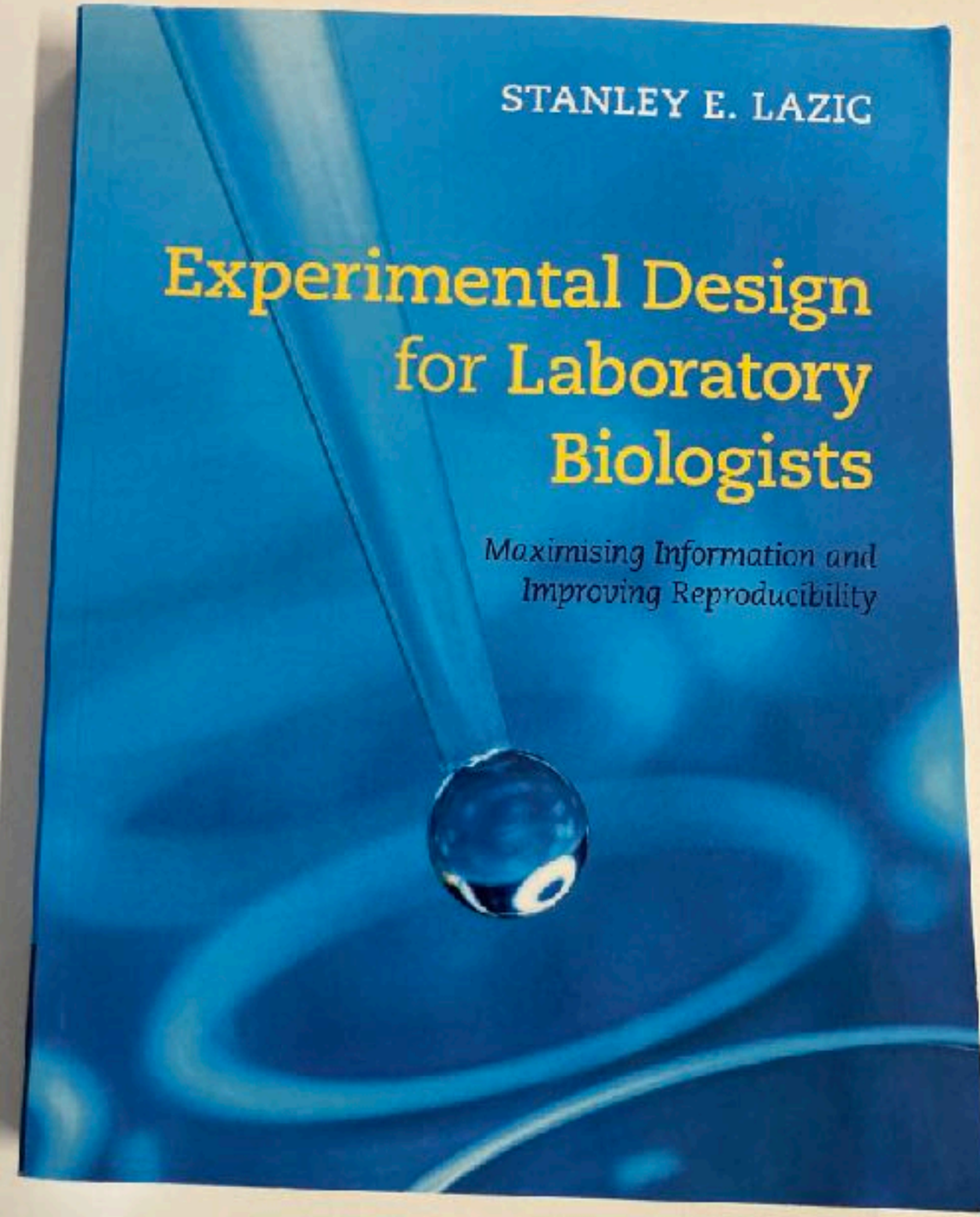
Second Edition



What are the functions of proteins?

At what sequences does EcoRI cut DNA?
BamHI
EcoRI
AatII
HindIII
XbaI

David J. Glass

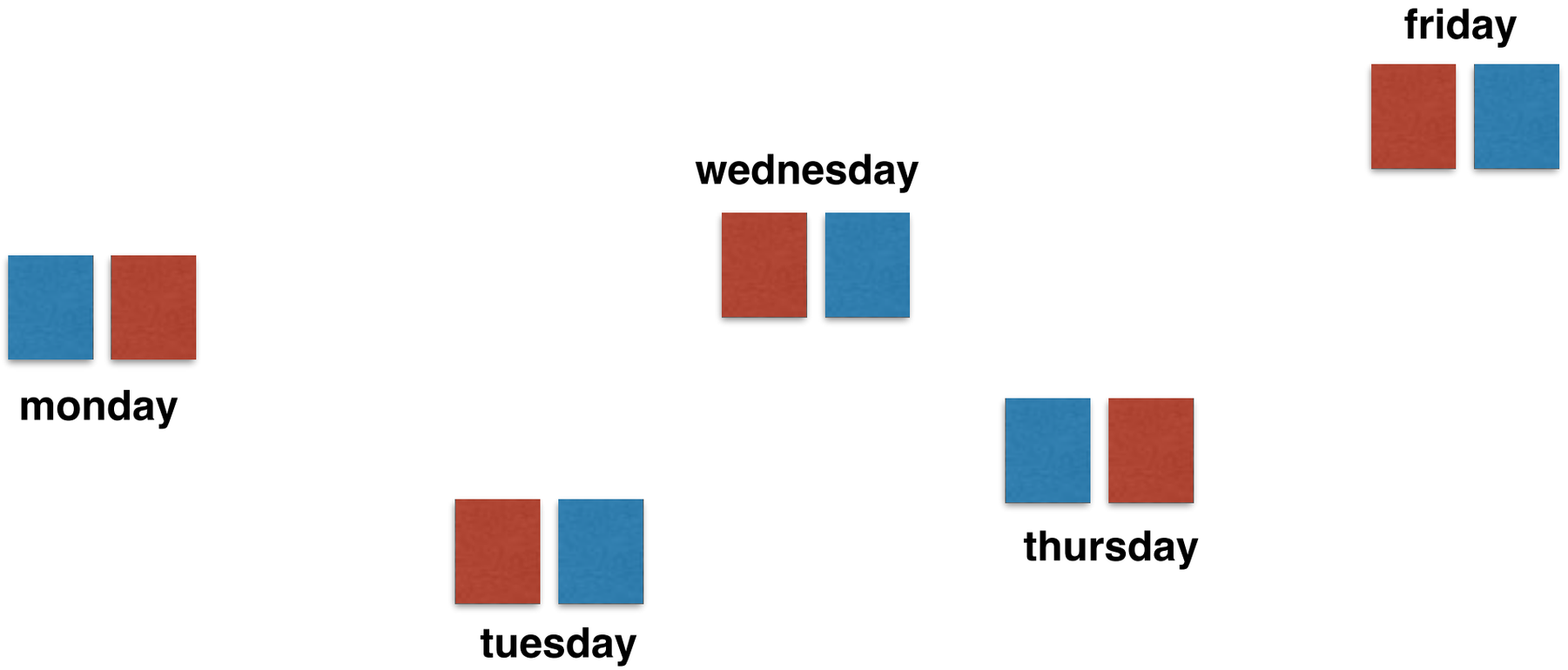


- removes systematic and reduces random errors (**blocking** / **replication**)

- eliminates biases (**randomisation** / **blinding**)

- Maximises sensitivity for the biological effect, (i.e. reveals interesting phenomena that are blurred in gonzo designs)

- can help to **increase statistical power** and/or reduce N

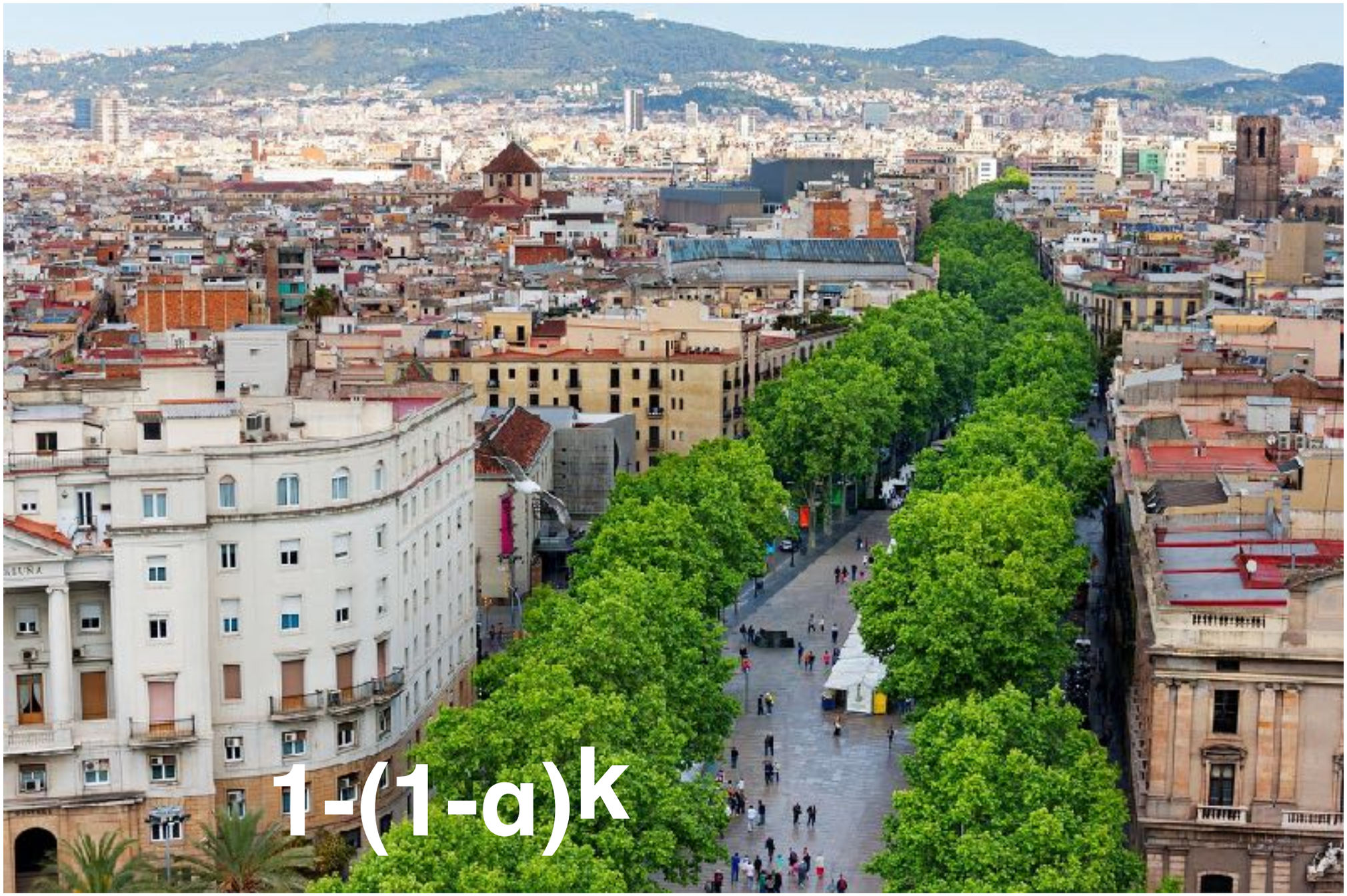


- randomised block design, only 2 factor levels (control, treatment)

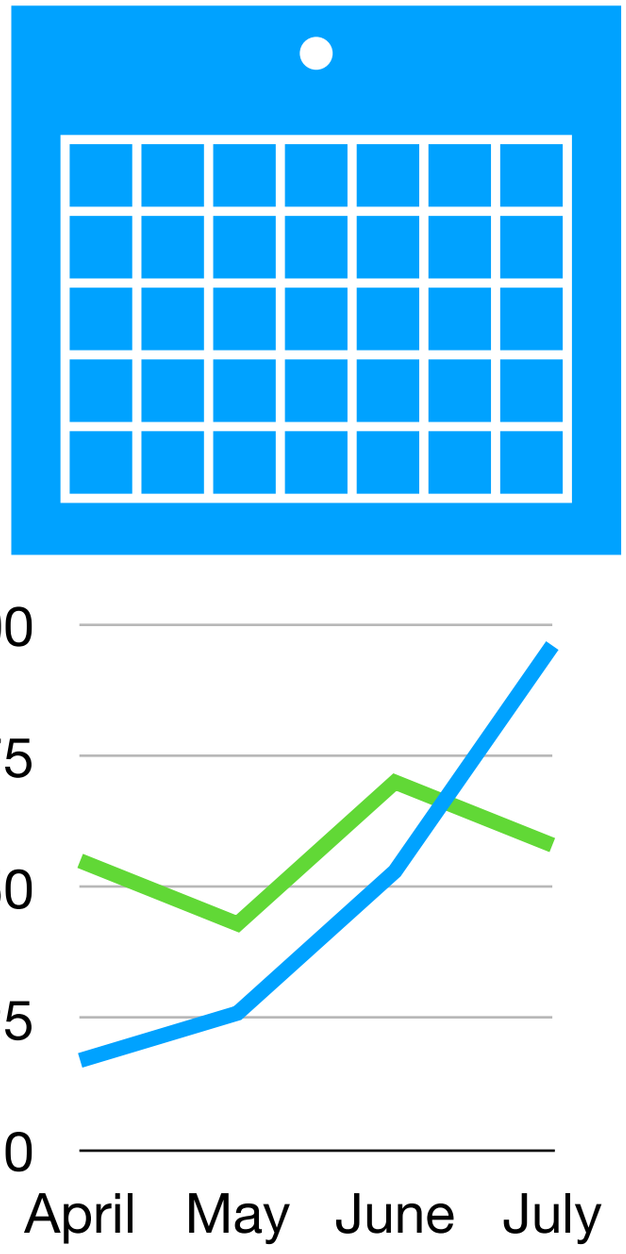
- suited to control for day-to-day fluctuations which are very common. Also change reagents, batches of cells etc. between the blocks as well. Every block a new batch, every block new reagents.

- paired t-test (very powerful!)

Multiple Testing


1-(1-a)k

Many endpoints, more than one independent variable



- In practice, number of endpoints/factors and number of replicates are typically inversely correlated (economic considerations)
- For rather small effects such experiments are typically underpowered (on the level of endpoint)
- Often inconclusive
- But perfect for narrowing the search space
- Figure 1.

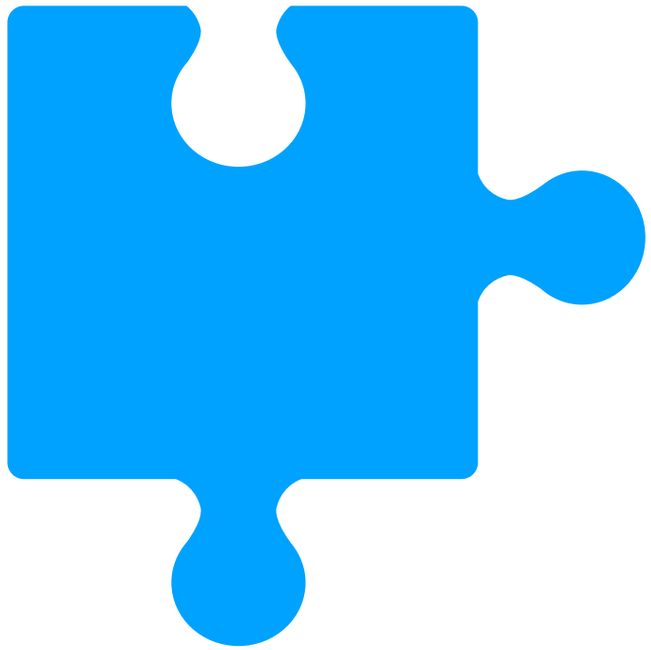
Experiments in the context of a project

How to build the story for publication

Single Experiment

medical phase III/IV trial

Strong Hypothesis



RCT
N >> 100

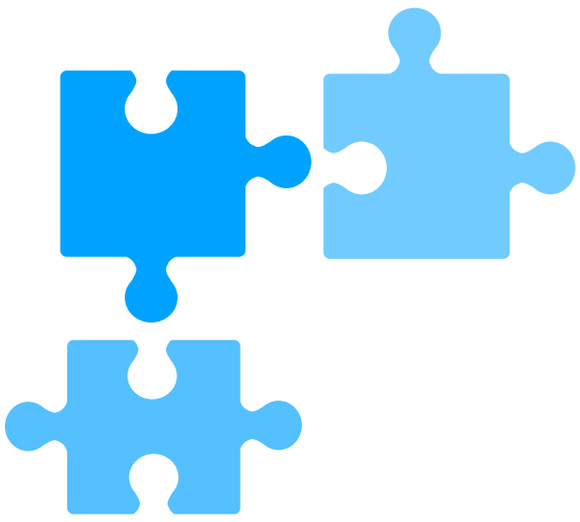
Strong Conclusion

Strong Paper

Preregistration

Triangulation

Vague idea
weak Hypothesis



Ensemble experiments
Of varying evidence strengths

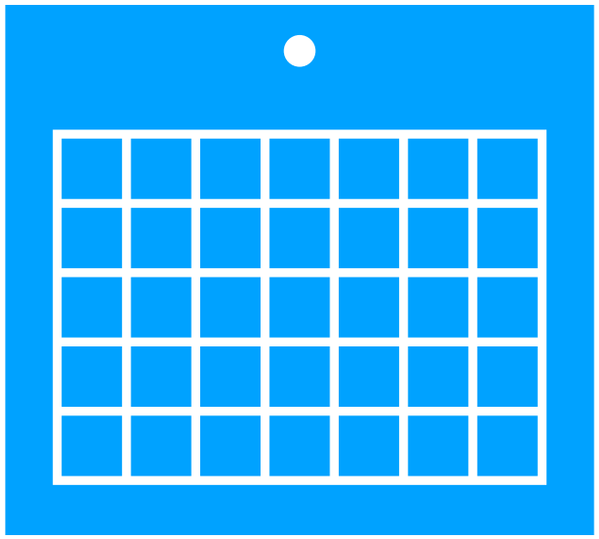
Conclusion

Paper

Screening

?

Vague idea
weak Hypothesis



High throughput
Omics
N < 3

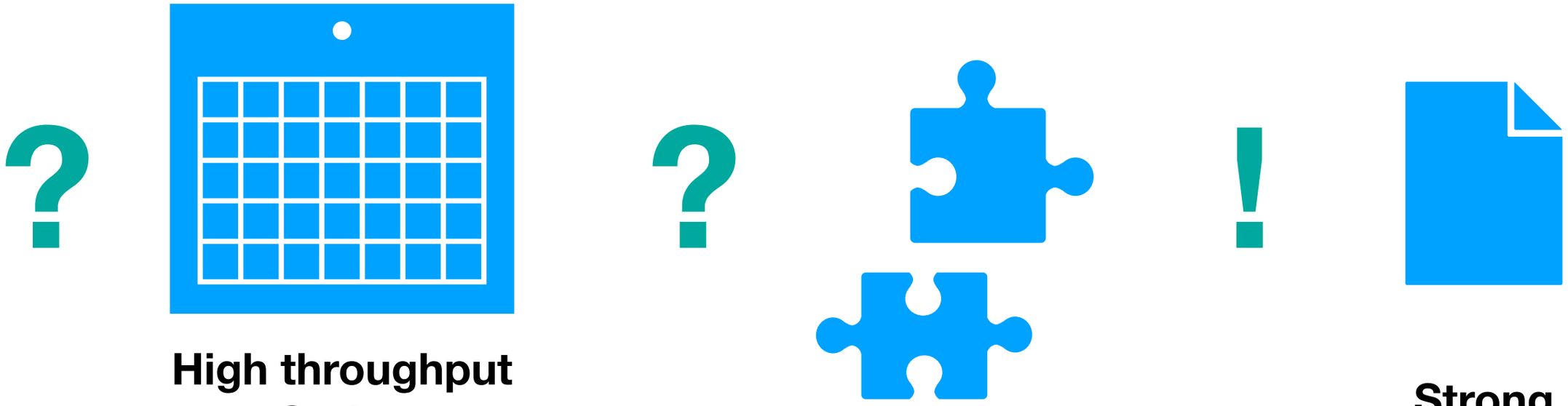
!

Inconclusive

Paper

Screening+

Ensemble experiments
Of varying evidence strengths



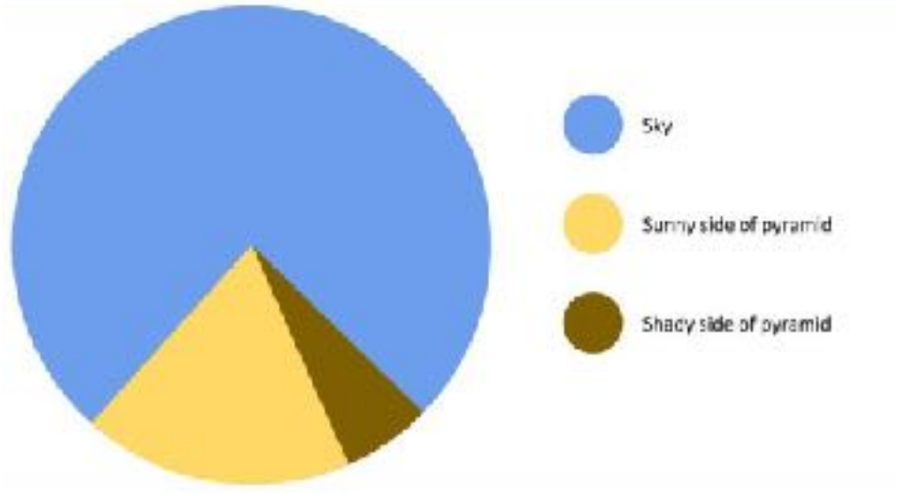
High throughput
Omics
N < 3

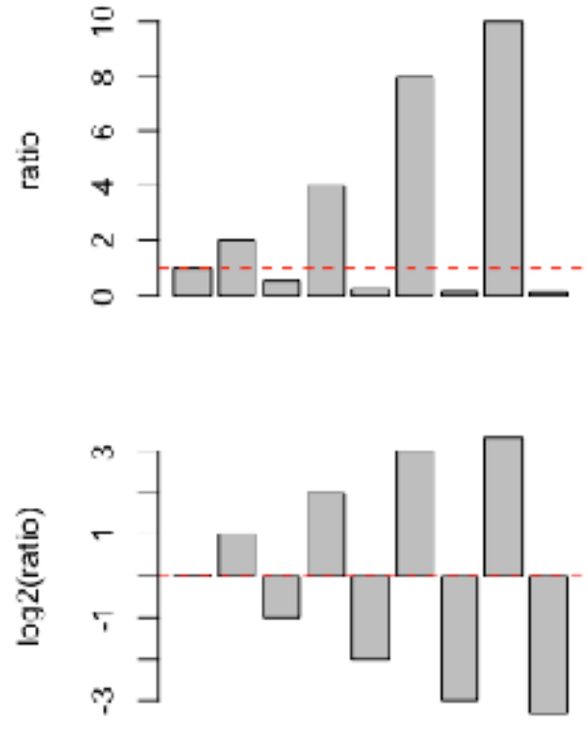
Vague idea
weak Hypothesis

Strong Hypothesis

Strong
Paper

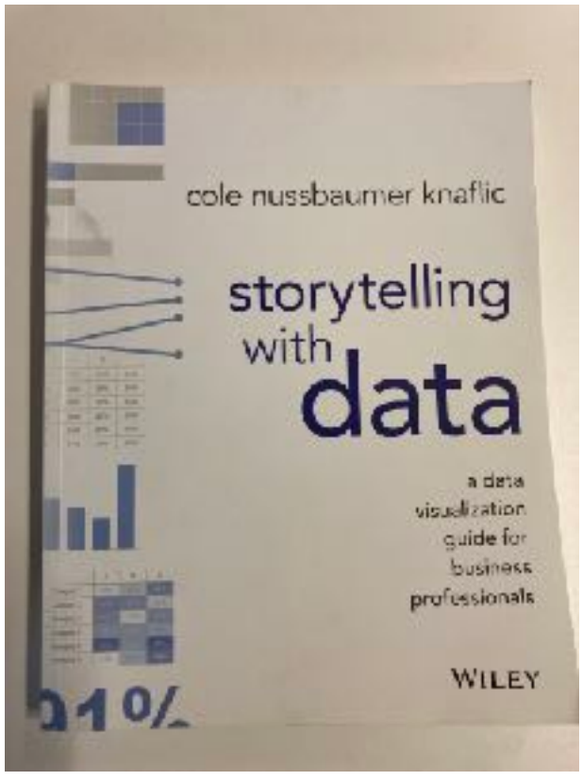
Conclusion

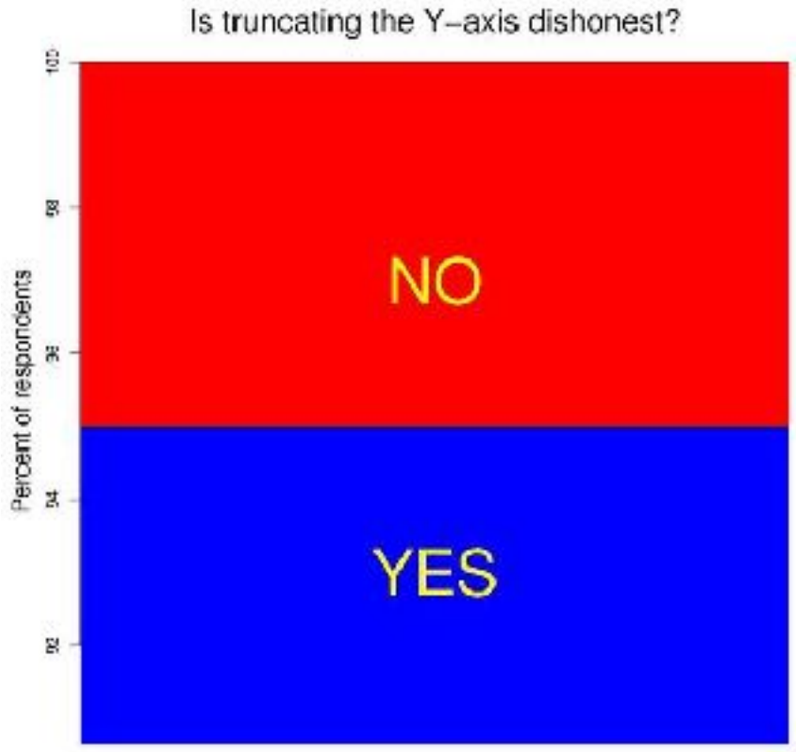




Visualisation

Another day





Publication

Journals now have check lists

- How was the **sample size** chosen to ensure adequate **power** to detect a pre-specified effect size?
- For animal studies, include a statement about sample size estimate even if no statistical methods were used.
- Describe **inclusion/exclusion criteria** if samples or animals were excluded from the analysis. Were the criteria pre-established?
- Were any steps taken to minimise the effects of subjective bias when allocating animals/samples to treatment (e.g. **randomisation** procedure)? If yes, please describe.
  For animal studies, include a statement about randomisation even if no randomisation was used.
- Were any steps taken to minimise the effects of subjective bias during group allocation or/and when assessing results (e.g. **blinding** of the investigator)? If yes please describe.
  For animal studies, include a statement about blinding even if no blinding was done
- For every figure, are statistical tests justified as appropriate?
  Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it. Is there an **estimate of variation within each group** of data?
  Is the variance similar between the groups that are being statistically compared?

Making
better Experiments

- Think ahead, define strategy
- Be representative
- Aim for conclusion: Triangulation and/or confirmatory experiments

- Don’t overdo screening, KISS towards the end.
- Experimental Design
- Open Science practices
  (Transparency, Storage, Analytical Reproducibility)
- Less time, less money, more reliable/sustainable results.

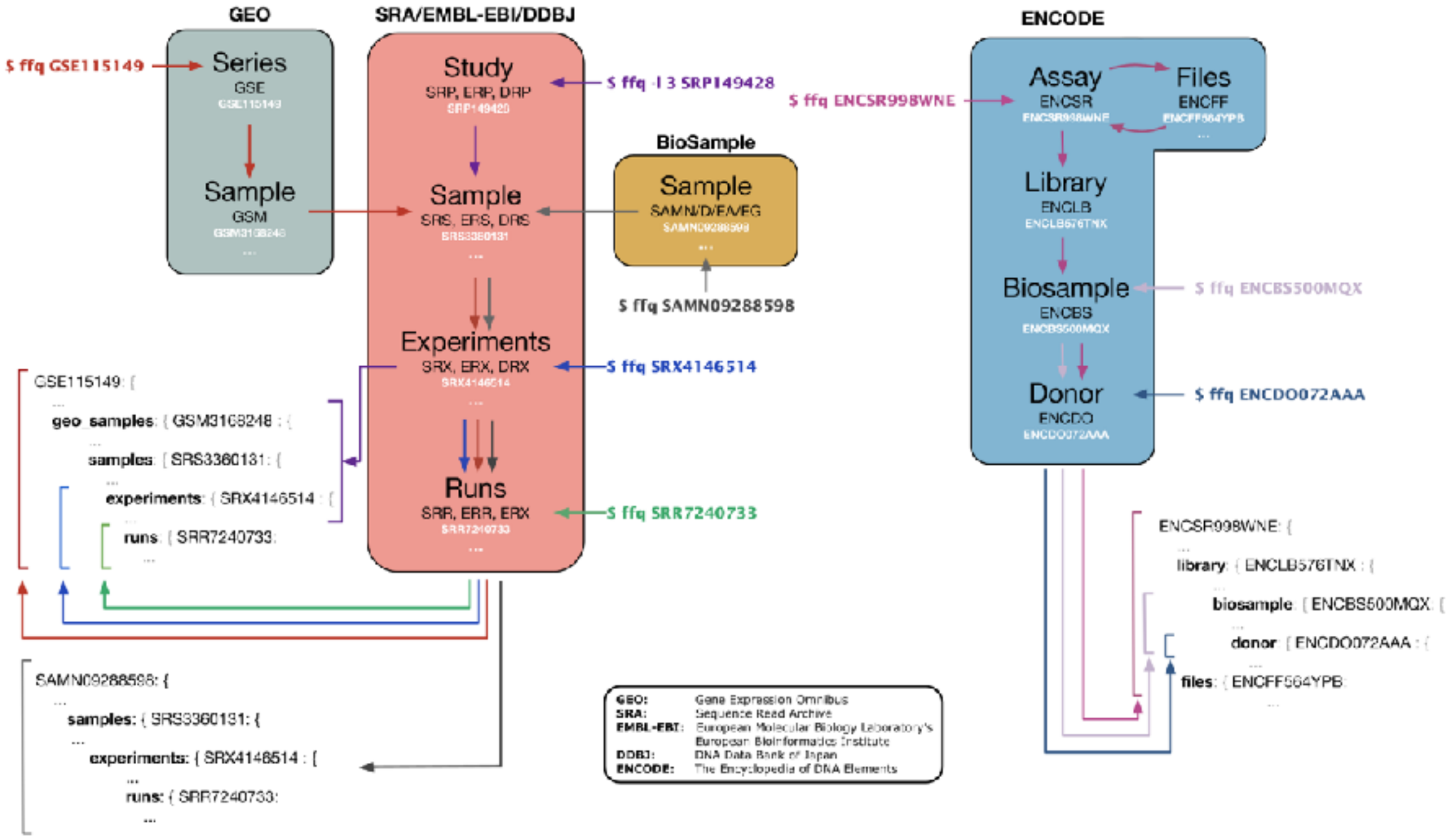
Analysis of Public Transcriptomic Data Sets

Tobias Straub

tobias.straub@lmu.de

Biomedizinisches Centrum, LMU München

Where the data @



https://github.com/pachterlab/ffq

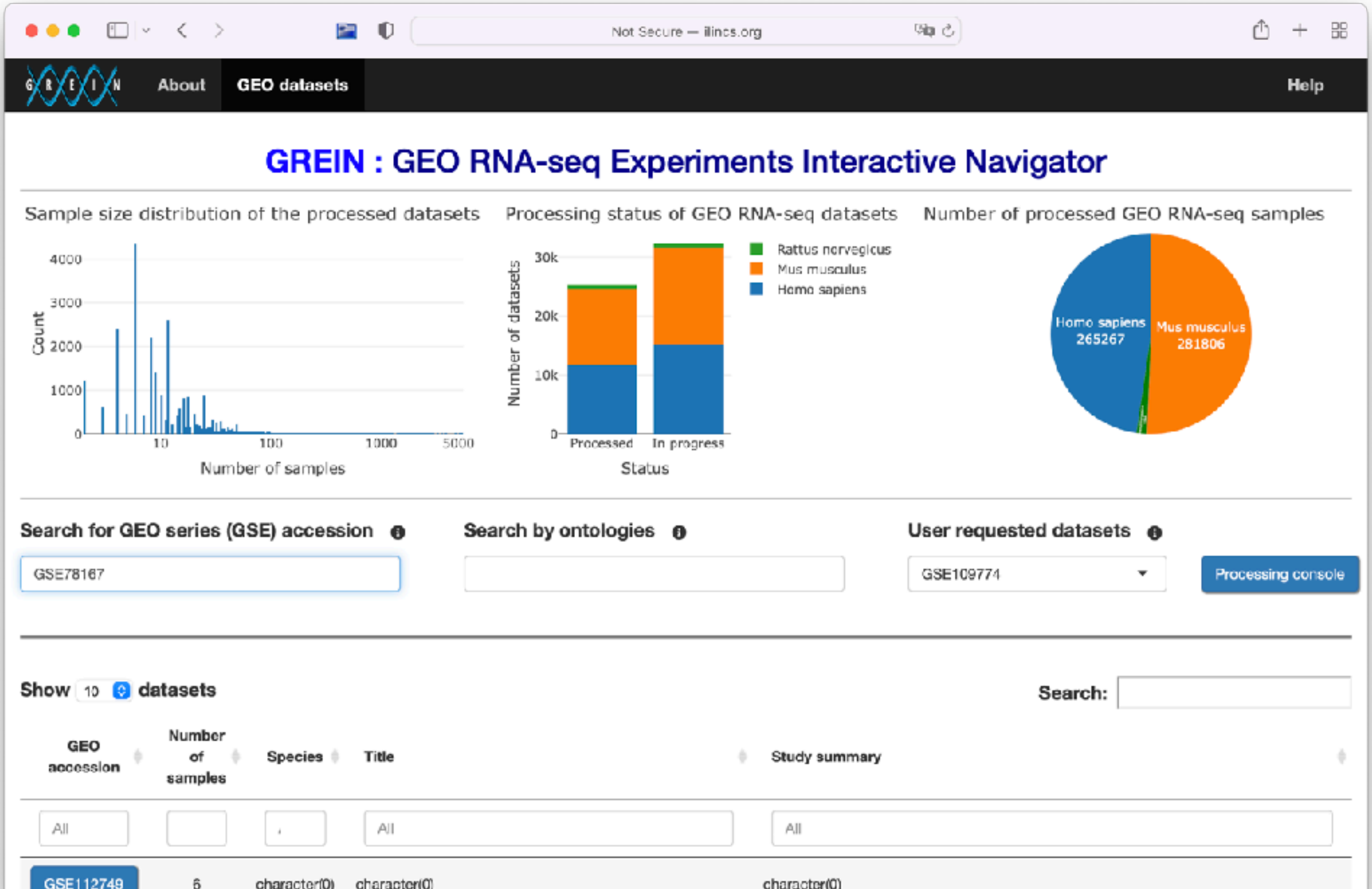
Series GSE78167 — Query DataSets for GSE78167

Status	Public on Aug 19, 2016
Title	An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response [RNA-seq]
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Estrogen receptor α (ERα) is an important biomarker of breast cancer severity and a common therapeutic target. Recent studies have demonstrated that in addition to its role in promoting proliferation, ERα also protects tumors against metastatic transformation. Current therapeutics antagonize ERα and interfere with both beneficial and detrimental signaling pathways stimulated by ERα. The goal of this study is to uncover the dynamics of coding and non-coding RNA (microRNA) expression in response to estrogen stimulation and identify potential therapeutic targets that more specifically inhibit ERα-stimulated growth and survival pathways without interfering with its protective features. To achieve this, we exposed MCF7 cells (an estrogen receptor positive model cell line for breast cancer) to estrogen and prepared a time course of paired mRNA and miRNA sequencing libraries at ten time points throughout the first 24 hours of the response to estrogen. From these data, we identified three primary expression trends—transient, induced, and repressed—that were each enriched for genes with distinct cellular functions. Integrative analysis of paired mRNA and microRNA temporal expression profiles identified miR-503 as the strongest candidate master regulator of the estrogen response, in part through suppression of ZNF217—an oncogene that is frequently amplified in cancer. We confirmed experimentally that miR-503 directly targets ZNF217 and that over-expression of miR-503 suppresses breast cancer cell proliferation. Overall, these data indicate that miR-503 acts as a potent estrogen-induced tumor suppressor microRNA that opposes cellular proliferation and has promise as a therapeutic for breast cancer. More generally, our work provides a systems-level framework for identifying functional interactions that shape the temporal dynamics of gene expression.
Overall design	Quantification of mRNAs in MCF7 cells responding to estrogen following a period of estrogen starvation. Three independent biological replicates (30 samples: 3 replicates x 10 time points) of MCF7 cells were exposed to 10nM Estradiol for 0, 1, 2, 3, 4, 5, 6, 8, 12 , or 24 hours, and total RNA was extracted from the samples. Total RNA was used to generate paired RNA and miRNA sequencing. RNA libraries were prepared using an Illumina TruSeq stranded mRNA library preparation kit.
Contributor(s)	Baran-Gale J, Sethupathy P, Purvis J
Citation(s)	Baran-Gale J, Purvis JE, Sethupathy P. An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response in MCF-7 breast cancer cells. *RNA* 2016 Oct;22(10):1592-603. PMID: 27539763
Submission date	Feb 22, 2016
Last update date	May 15, 2019
Contact name	Jeanette Baran-Gale
E-mail(s)	jtbaran@email.unc.edu
Organization name	University of North Carolina at Chapel Hill
Lab	Sethupathy & Purvis Labs
Street address	120 Mason Farm Rd.
City	Chapel Hill

- The easy way - using convenience tools
  - Convenient
  - Fixed genome/annotation/tools
  - Potentially compromised by bad metadata
  - Reproducibility questionable (documentation)
  - Black box
  - Accessibility/availability nor guaranteed

- The hard way - running analyses from scratch
  - Steep learning curve (but huge reward)
  - Flexible
  - Deposition flaws can be fixed
  - Reproducible, re-usable, time-efficient
  - Publication quality figures

http://www.ilincs.org/apps/grein/



Not Secure — ilincs.org

processed. Please see the following table.

Show 10 datasets

Search:

GEO accession	Number of samples	Species	Title	Study summary
All	All	All	All	All
GSE78167	30	Homo sapiens	An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response [RNA-seq]	Estrogen receptor α (ERα) is an important biomarker of breast cancer severity and a common therapeutic target. Recent studies have demonstrated that in addition to its role in promoting proliferation, ERα also protects tumors against metastatic transformation. Current therapeutics antagonize ERα and interfere with both beneficial and detrimental signaling pathways stimulated by ERα. The goal of this study is to uncover the dynamics of coding and non-coding RNA (microRNA) expression in response to estrogen stimulation and identify potential therapeutic targets that more specifically inhibit ERα-stimulated growth and survival pathways without interfering with its protective features. To achieve this, we exposed MCF7 cells (an estrogen receptor positive model cell line for breast cancer) to estrogen and prepared a time course of paired mRNA and miRNA sequencing libraries at ten time points throughout the first 24 hours of the response to estrogen. From these data, we identified three primary expression trends—transient, induced, and repressed—that were each enriched for genes with distinct cellular functions. Integrative analysis of paired mRNA and microRNA temporal expression profiles identified miR-503 as the strongest candidate master regulator of the estrogen response, in part through suppression of ZNF217—an oncogene that is frequently amplified in cancer. We confirmed experimentally that miR-503 directly targets ZNF217 and that over-expression of miR-503 suppresses breast cancer cell proliferation. Overall, these data indicate that miR-503 acts as a potent estrogen-induced tumor suppressor microRNA that opposes cellular proliferation and has promise as a therapeutic for breast cancer. More generally, our work provides a systems-level framework for identifying functional interactions that shape the temporal dynamics of gene expression.

Showing 1 to 1 of 1 datasets

Previous 1 Next

Selected study

GSE78167

Description | Metadata | Counts table | QC report | Visualization

Study link	GSE78167
No. of GEO samples	30
No. of SRA runs	30
Species	Homo sapiens
Title	An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen response [RNA-seq]
Summary	Estrogen receptor α (ERα) is an important biomarker of breast cancer severity and a common therapeutic target. Recent studies have demonstrated that in addition to its role in promoting proliferation, ERα also protects tumors against metastatic transformation. Current therapeutics antagonize ERα and interfere with both beneficial and detrimental signaling pathways stimulated by ERα. The goal of this study is to uncover the dynamics of coding and non-coding RNA (microRNA) expression in response to estrogen stimulation and identify potential therapeutic targets that more specifically inhibit ERα-stimulated growth and survival pathways without interfering with its protective features. To achieve this, we exposed MCF7 cells (an estrogen receptor positive model cell line for breast cancer) to estrogen and prepared a time course of paired mRNA and miRNA sequencing libraries at ten time points throughout the first 24 hours of the response to estrogen. From these data, we identified three primary expression trends—transient, induced, and repressed—that were each enriched for genes with distinct cellular functions. Integrative analysis of paired mRNA and microRNA temporal expression profiles identified miR-503 as the strongest candidate master regulator of the estrogen response, in part through suppression of ZNF217—an oncogene that is frequently amplified in cancer. We confirmed experimentally that miR-503 directly targets ZNF217 and that over-expression of miR-503 suppresses breast cancer cell proliferation. Overall, these data indicate that miR-503 acts as a potent estrogen-induced tumor suppressor microRNA that opposes cellular proliferation and has promise as a therapeutic for breast cancer. More generally, our work provides a systems-level framework for identifying functional interactions that shape the temporal dynamics of gene expression.

Not Secure — ilincs.org

GREIN | About | GEO datasets | Explore dataset | Analyze dataset | Help

Selected study

GSE78167

Data type

Raw | Normalized

Number of samples to show

30

Show counts table

Download data

Gene level | Transcript level

Description | Metadata | Counts table | QC report | Visualization

Select any row to see boxplot of the selected gene below the table.

Show 8 genes

Search:

	Gene symbol	GSM2068643	GSM2068644	GSM2068645	GSM2068646	GSM2068647	GSM2
ENSG00000000003	TSPAN6	799	568	806	727	610	
ENSG00000000005	TNMD	0	0	0	0	0	
ENSG00000000419	DPM1	2230	1579	2423	2246	2134	
ENSG00000000457	SCYL3	816	410	533	521	595	
ENSG00000000460	C1orf112	375	205	345	291	456	
ENSG00000000938	FGR	1	0	35	28	6	
ENSG00000000971	CFH	1	0	1	0	0	
ENSG00000001036	FUCA2	2716	1832	2888	2567	2909	

Showing 1 to 8 of 27,990 genes

Previous 1 2 3 4 5 … 3499 Next

GREIN | GEO RNA-seq Experiments Interactive Navigator

BD2K-LINCS DATA COORDINATION AND INTEGRATION CENTER

Contact

Not Secure — ilincs.org

GREIN | About | GEO datasets | Explore dataset | Analyze dataset | Help

Selected study

GSE78167

Description | Metadata | Counts table | QC report | Visualization

Download QC report

MultiQC v1.2

- General Stats
- Salmon
- FastQC
  - Sequence Quality Histograms
  - Per Sequence Quality Scores
  - Per Base Sequence Content
  - Per Sequence GC Content
  - Per Base N Content
  - Sequence Length Distribution
  - Sequence Duplication Levels
  - Overrepresented sequences
  - Adapter Content

MultiQC

- /GSE78167/fastqc
- /GSE78167/salmon

Welcome! Not sure where to start? don't show again

Watch a tutorial video (6:06)

General Statistics

Copy table | Configure Columns | Plot | Showing 92/92 rows and 5/7 columns.

Sample Name	% Aligned	M Aligned	%
SRR3182145_GSM2068643_transcripts_quant	92.0%	33.6	
SRR3182145_pass_1			66
SRR3182145_pass_2			64
SRR3182146_GSM2068644_transcripts_quant	90.5%	22.7	
SRR3182146_pass_1			62
SRR3182146_pass_2			60

Toolbox

Not Secure — ilincs.org

GREIN | About | GEO datasets | Explore dataset | Analyze dataset | Help

Selected study

GSE78167

Select a property

treated with

Subset data

No Yes

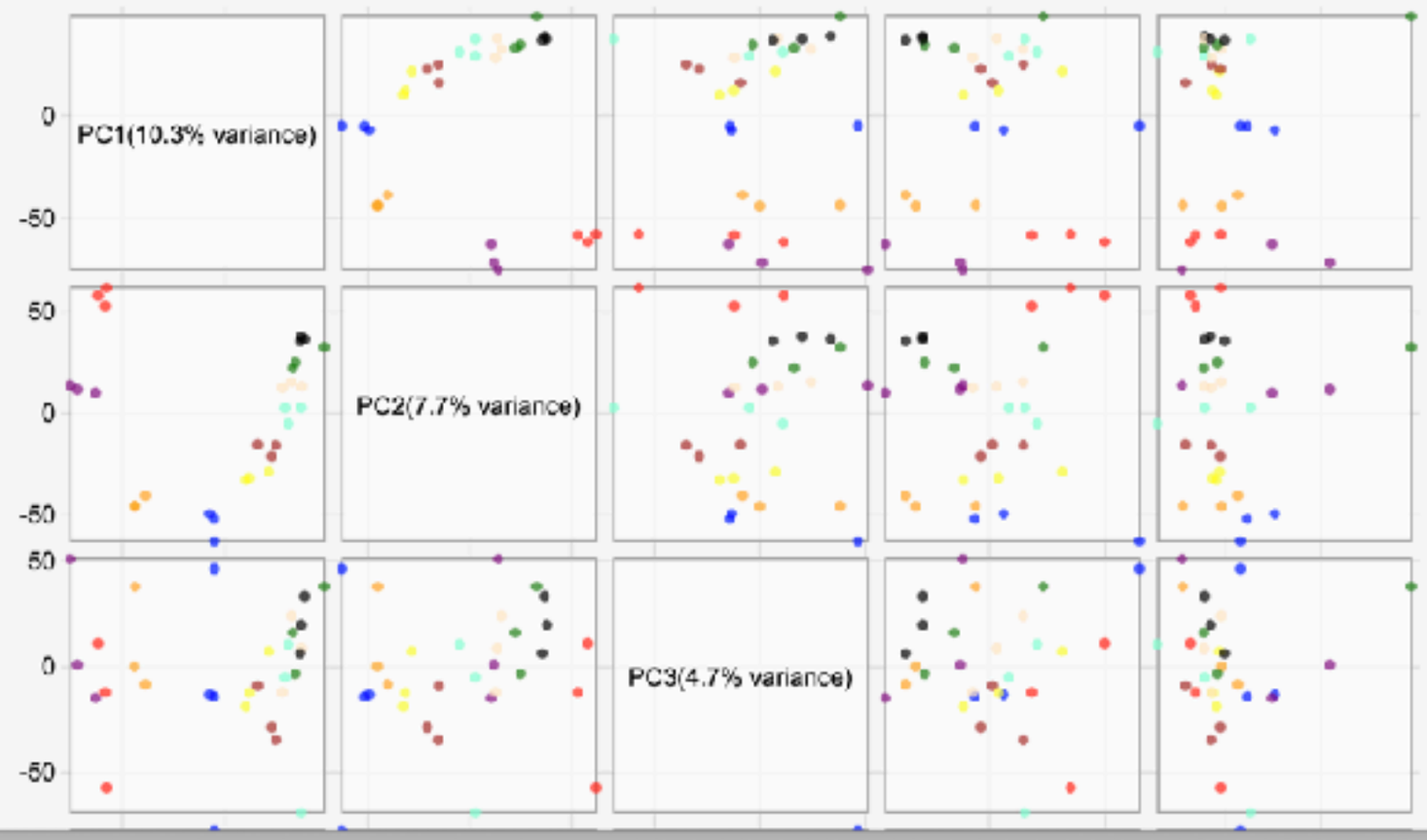
Plot PCA

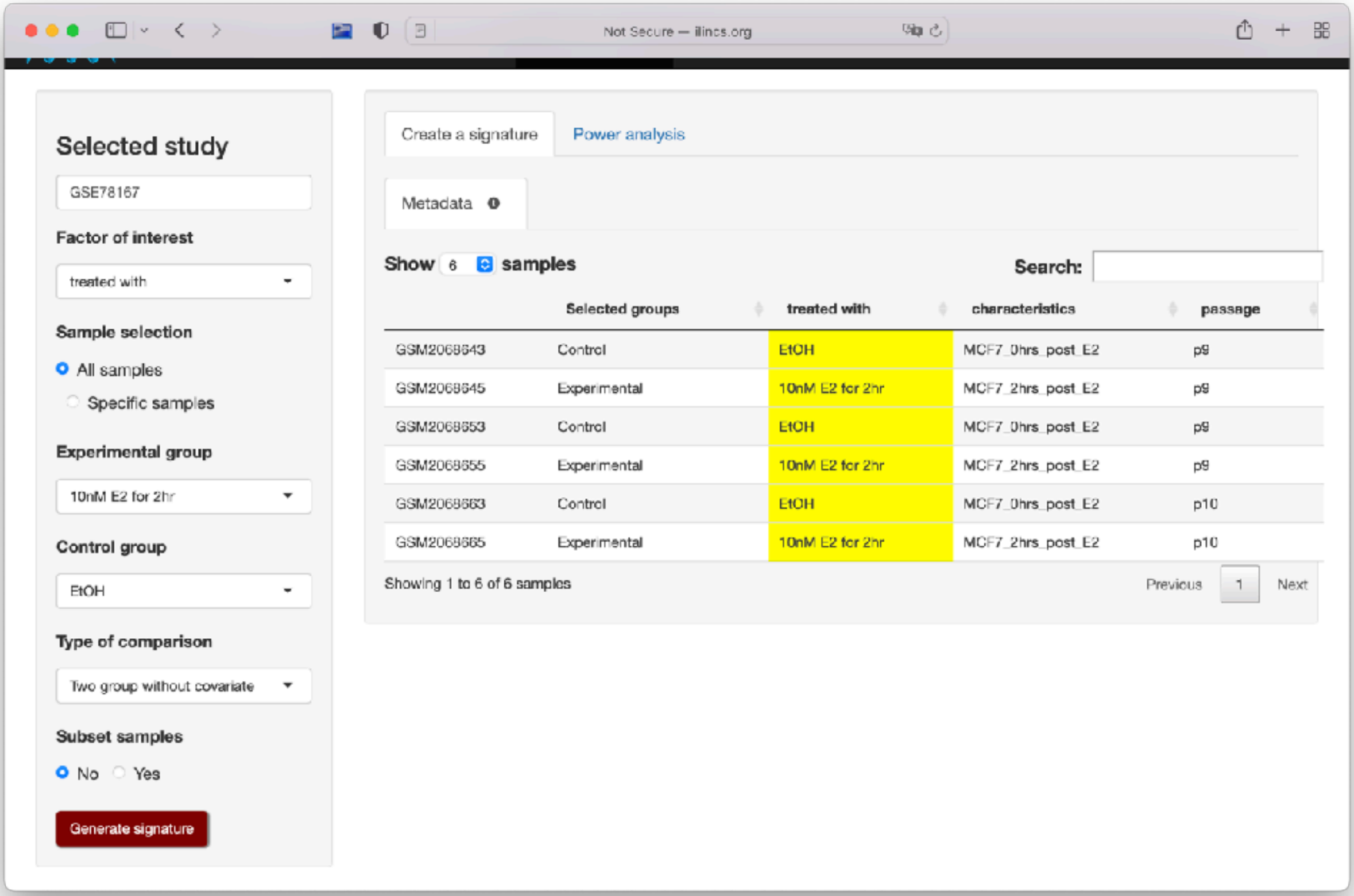
Description | Metadata | Counts table | QC report | Visualization

Correlation plot | Density plot | Heatmap | PCA & t-SNE

2D PCA plot | 3D PCA plot | 2D t-SNE plot | 3D t-SNE plot

Download image




Not Secure — ilincs.org
Selected study
GSE78167
Factor of interest
treated with
Sample selection
All samples
Specific samples
Experimental group
10nM E2 for 2hr
Control group
EtOH
Type of comparison
Two group without covariate
Subset samples
No
Yes
Generate signature
Create a signature
Power analysis
Metadata
Show 6 samples
Search:
Selected groups
treated with
characteristics
passage
GSM2068543 Control EtOH MCF7_0hrs_post_E2 p9
GSM2068545 Experimental 10nM E2 for 2hr MCF7_2hrs_post_E2 p9
GSM2068553 Control EtOH MCF7_0hrs_post_E2 p9
GSM2068555 Experimental 10nM E2 for 2hr MCF7_2hrs_post_E2 p9
GSM2068563 Control EtOH MCF7_0hrs_post_E2 p10
GSM2068565 Experimental 10nM E2 for 2hr MCF7_2hrs_post_E2 p10
Showing 1 to 6 of 6 samples
Previous
1
Next

Not Secure — ilincs.org

GREIN | About | GEO datasets | Explore dataset | Analyze dataset | Help

Selected study

GSE78167

Signature visualization

Download signature

Upload signature to iLINCS

Upload all genes

Upload

Create a signature | Power analysis

Metadata | Signature

Show 10 genes

Search:

Ensembl_ID	Gene_symbol	logFC	logCPM	PValue	FDR
All	All	All	All	All	All
ENSG00000148344	PTGES	4.933	5.122	5.24e-195	1.087e-190
ENSG00000165731	RET	3.176	7.235	2.615e-172	2.711e-168
ENSG00000175591	P2RY2	3.451	7.259	2.268e-165	1.568e-161
ENSG00000099337	KCNK6	3.203	7.657	5.967e-165	3.093e-161
ENSG00000198959	TGM2	5.485	4.008	8.701e-155	3.609e-151
ENSG00000108352	RAPGEFL1	3.123	6.394	1.133e-152	3.917e-149
ENSG00000164626	KCNK5	3.687	6.402	1.114e-146	3.299e-143
ENSG00000188176	SMTNL2	4.908	4.658	2.755e-144	6.402e-141
ENSG00000108375	RNF43	-3.506	5.701	2.778e-144	6.402e-141
ENSG00000115738	ID2	-3.447	6.34	1.667e-143	3.457e-140

1. logFC=Log fold change
2. logCPM=Log counts per million
3. FDR=False discovery rate

Not Secure — ilincs.org

GREIN About GEO datasets Explore dataset Analyze dataset Help

Type of heatmap

- (•) Heatmap across all the samples
- ( ) Heatmap across the comparison samples

Select properties

- [ ] characteristics
- [ ] passage
- [x] treated with

Grouping genes and samples

Pearson correlation

Number of top DE genes

20

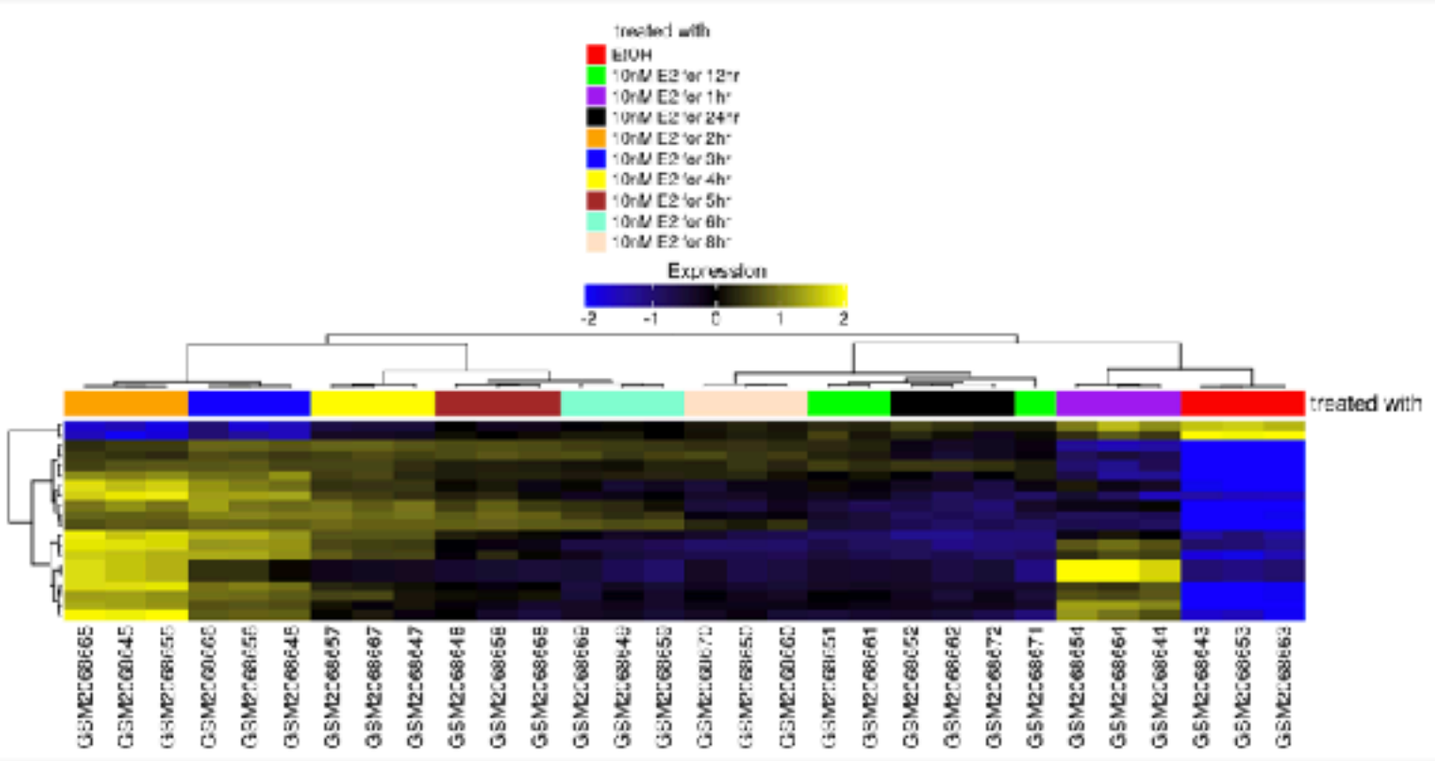
Draw heatmap

Download (•) png ( ) csv

Static heatmap | Interactive heatmap | MA plot | Gene detectability plot

Heatmap of top DE genes (ranked by adjusted p-values)

Fit in screen | Scrollable



Close

Not Secure — ilincs.org

GREIN About GEO datasets Explore dataset Analyze dataset Help

False discovery rate

0.05

Log2(Fold change)

-8.9 -1 1 8.9

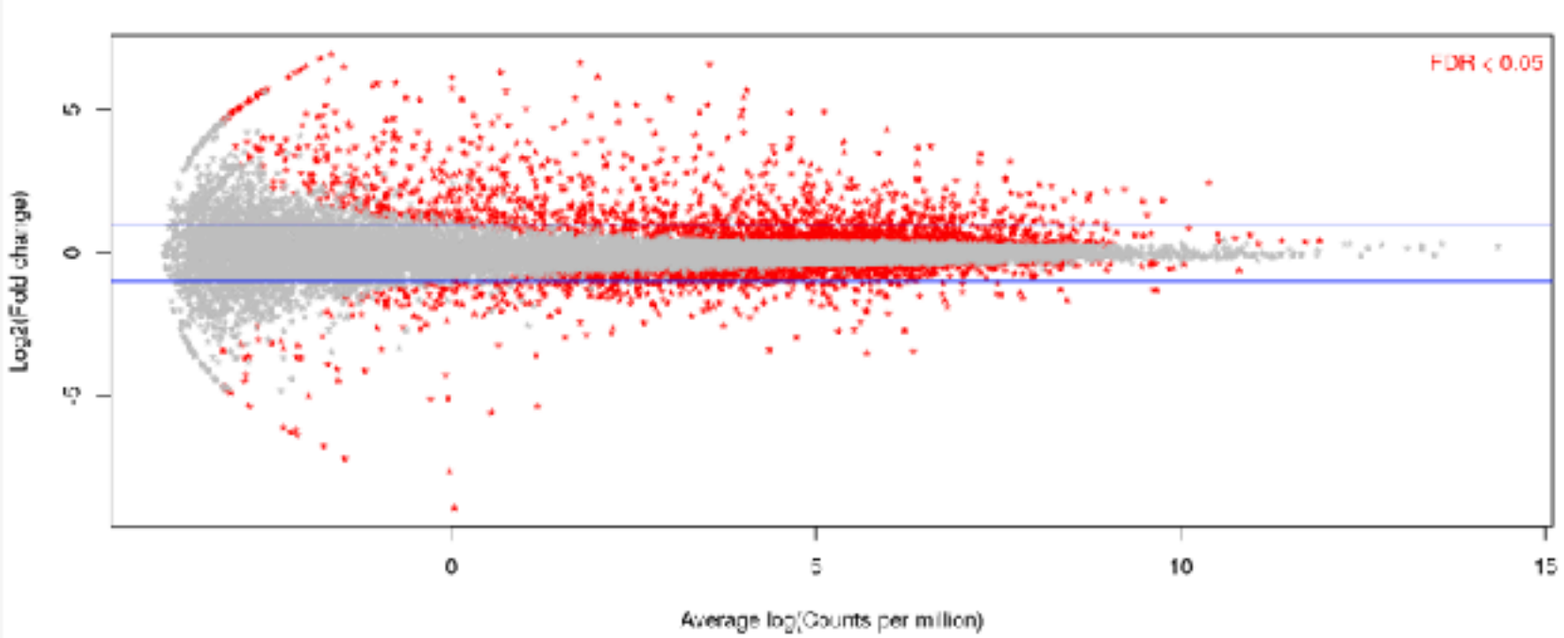
8.9 7.3 5.7 4.1 2.6 0.3 0.7 2.3 3.9 5.5 8.2

Draw MA plot

Static heatmap | Interactive heatmap | MA plot | Gene detectability plot

Log-fold change (M) vs. Log-average expression (A) plot

Download image



Click on any point or select any area to see the values of the genes. To reset the plot, click 'Draw MA plot'

Not Secure — ilincs.org

GREIN | DCIC integrative LINCS genomics data portal | iLINCS

iLINCS Signatures Datasets Genes iLINCS Paper new

Search for signatures / Upload a signature / Uploaded Signature

Uploaded Signature

Signature analysis

Modify the list of selected genes »

Other analyses with selected genes »

Signature Info

Session ID: Sat_May_28_09_46_50_2022_580356

File name: GSE78167_signatureData_13:43:43_2022-05-28_10124_up.txt

Genes not Found: C6ORF132, C14ORF132, FCMR, C1ORF226, C10ORF2, C2ORF54, DISP3, C15ORF59, C1ORF109, C1ORF111, PHF24, C8ORF46, C4ORF19, C11... More

Found **9471** out of **9873** submitted entries.

Complete signature (9471) | **Selected genes (100)**

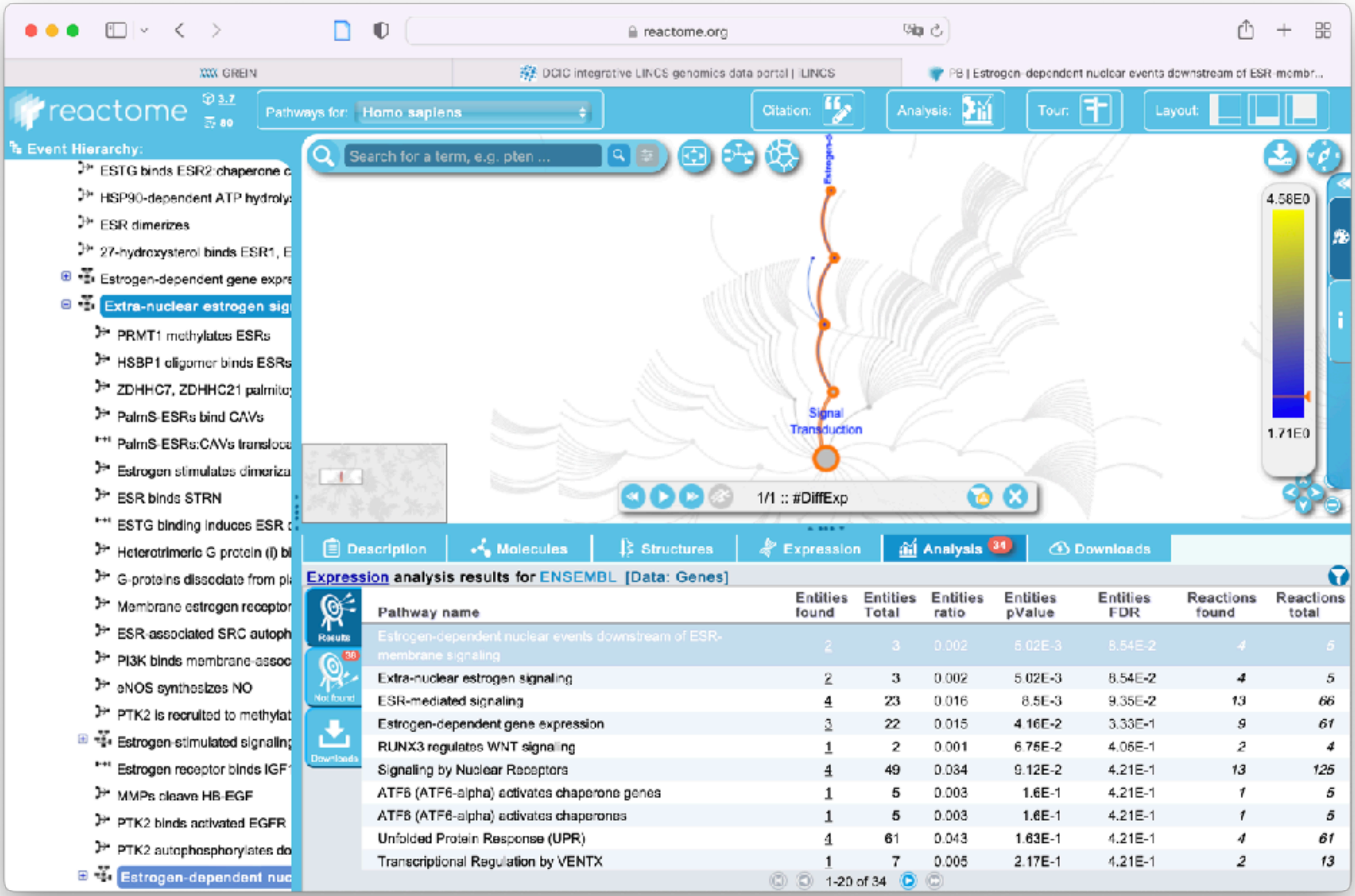
Download

Signature Analysis Tools | Signature Data | Connected Signatures | Connected Perturbations

Pathway Analysis

Enrichr | DAVID | ToppFun | Reactome

Network Analysis



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total
Estrogen-dependent nuclear events downstream of ESR-membrane signaling	2	3	0.002	5.02E-3	8.54E-2	4	5
Extra-nuclear estrogen signaling	2	3	0.002	5.02E-3	8.54E-2	4	5
ESR-mediated signaling	4	23	0.016	8.5E-3	9.35E-2	13	66
Estrogen-dependent gene expression	3	22	0.015	4.16E-2	3.33E-1	9	61
RUNX3 regulates WNT signaling	1	2	0.001	6.75E-2	4.05E-1	2	4
Signaling by Nuclear Receptors	4	49	0.034	9.12E-2	4.21E-1	13	125
ATF6 (ATF6-alpha) activates chaperone genes	1	5	0.003	1.6E-1	4.21E-1	1	5
ATF6 (ATF6-alpha) activates chaperones	1	5	0.003	1.6E-1	4.21E-1	1	5
Unfolded Protein Response (UPR)	4	61	0.043	1.63E-1	4.21E-1	4	61
Transcriptional Regulation by VENTX	1	7	0.005	2.17E-1	4.21E-1	2	13

1-20 of 34

Not Secure — ilincs.org

GREIN | DCIC integrative LINCS genomics data portal | iLINCS

iLINCS Signatures Datasets Genes iLINCS Paper new

Search for signatures / Upload a signature / Uploaded Signature / Pathway Analysis

SPIA Functional Pathway Analysis

KEGG pathway name	KEGG pathway ID	Genes in Pathway	DE Genes in Pathway	ORA pval	Topology Score	Top pval	SPIA pval	SPIA padj	Status	KEGG link
									All	
Intestinal immune network for IgA production	hsa04672	5	2	0.0024	-0.26	0.818	0.0144	0.3998	Inhibited	KEGG
Osteoclast differentiation	hsa04380	52	3	0.0195	-13.9124	0.075	0.0215	0.3998	Inhibited	KEGG
Amoebiasis	hsa05146	14	1	0.2017	-4.904	0.03	0.037	0.3998	Inhibited	KEGG
cGMP-PKG signaling pathway	hsa04022	67	2	0.2697	19.8674	0.028	0.0472	0.3998	Activated	KEGG
Hepatitis B	hsa05161	56	1	0.5951	20.3431	0.014	0.0482	0.3998	Activated	KEGG
Taste transduction	hsa04742	4	1	0.0523	3.567	0.153	0.0539	0.3998	Activated	KEGG
RNA degradation	hsa03018	5	1	0.0772	0	NA	0.0772	0.3998	Inhibited	KEGG
Sphingolipid signaling pathway	hsa04071	40	2	0.1331	15.4392	0.112	0.0776	0.3998	Activated	KEGG
Axon guidance	hsa04360	69	3	0.0971	16.8184	0.155	0.0782	0.3998	Activated	KEGG
T cell receptor signaling pathway	hsa04660	41	1	0.4838	20.93	0.041	0.0976	0.4489	Activated	KEGG

5 10 50

First « 1 2 3 4 5 » Last

Page 1 of 5, of 43 entries

Download data

iLINCS Paper: Connecting omics signatures of diseases, drugs, and mechanisms of

LINCS Tools

Contact

The hard but rewarding way

```
@BS-DSU-ELLAC_0:4:1:6214:930
NNTTCCTCGGCTGGTAGCTTAAATAATAGAGCTTAA
+
##***(*)+(@@@@@1211557755.,55575775@
@BS-DSU-ELLAC_0:4:1:8611:931
NNACAAATGAGCGTGAGCTTCTCGCCATCTTATGGG
+
####################################
@BS-DSU-ELLAC_0:4:1:9150:948
NNTTTTATATTTCCTAATTACATATGTACAAAAGTT
```

Raw data

+ Metadata

+ Genomic data

UNIX/LINUX

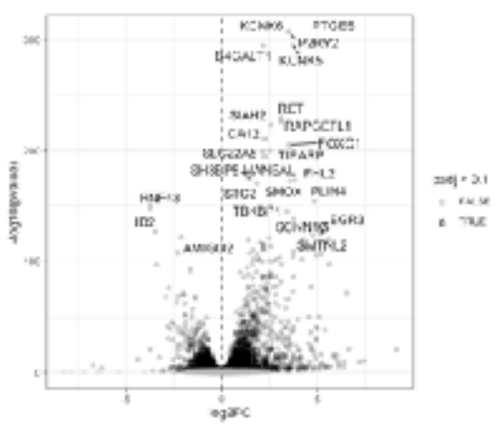
Pre-processing (aka the engineering part)

Processed data

`SummarizedExperiment`

Data analysis (the scientific part)

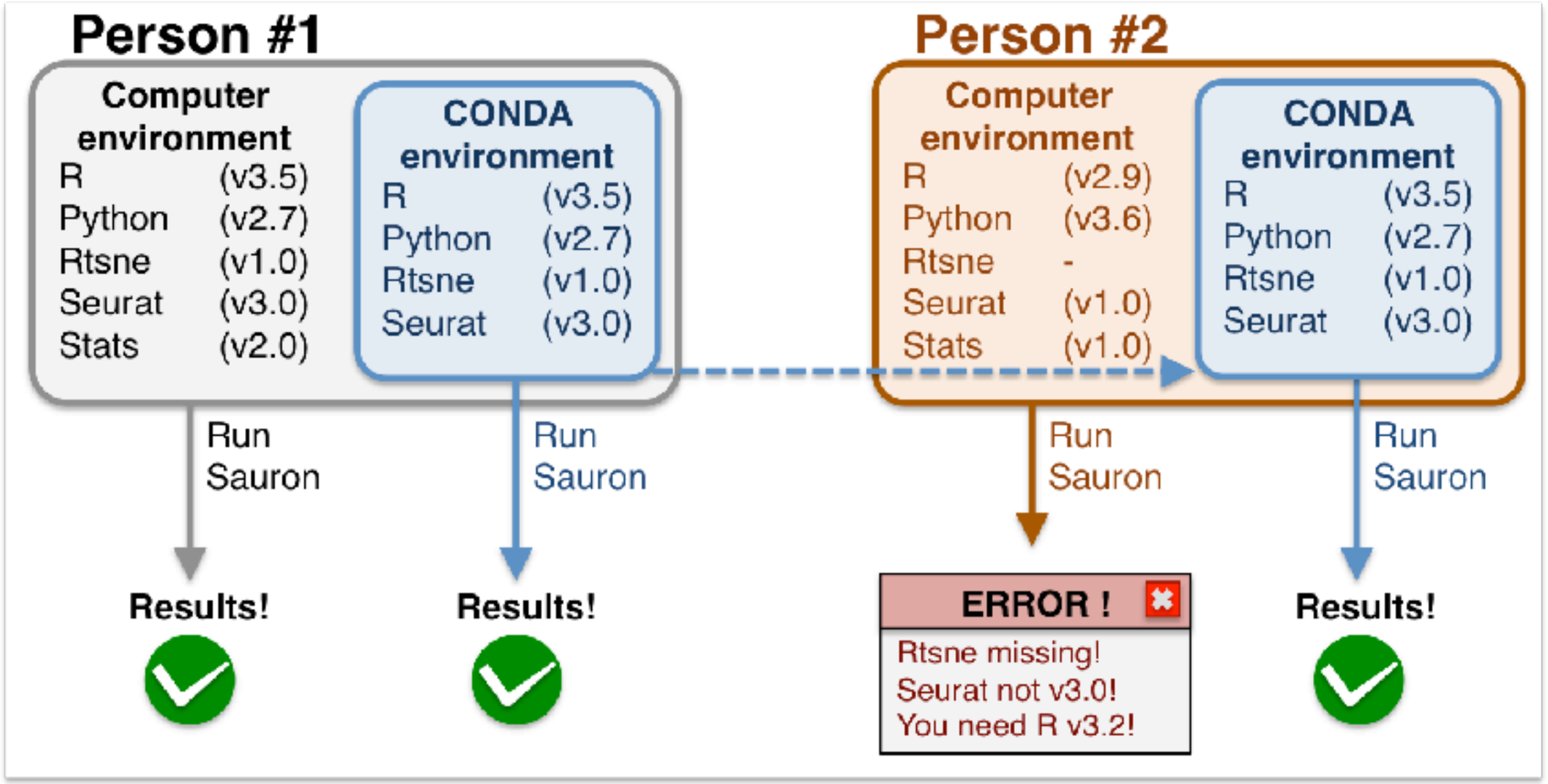
R/python



Requirements

- Desire, commitment
- Computer hardware / good Network connection
- Skills:
  - Google, copy, paste skills
  - (Very basic) Unix
  - Software mangagement with conda
  - Snakemake
  - R/bioconductor

Conda


Person #1
Computer environment
R (v3.5)
Python (v2.7)
Rtsne (v1.0)
Seurat (v3.0)
Stats (v2.0)
CONDA environment
R (v3.5)
Python (v2.7)
Rtsne (v1.0)
Seurat (v3.0)
Run Sauron
Run Sauron
Results!
Results!
Person #2
Computer environment
R (v2.9)
Python (v3.6)
Rtsne -
Seurat (v1.0)
Stats (v1.0)
CONDA environment
R (v3.5)
Python (v2.7)
Rtsne (v1.0)
Seurat (v3.0)
Run Sauron
Run Sauron
ERROR !
Rtsne missing!
Seurat not v3.0!
You need R v3.2!
Results!

Snakemake

Tell Snakemake what files you want to be created

```
rule:
  input: "A.txt", "B.txt", "C.txt"
```

Produce the files you want to have from some intermediate result

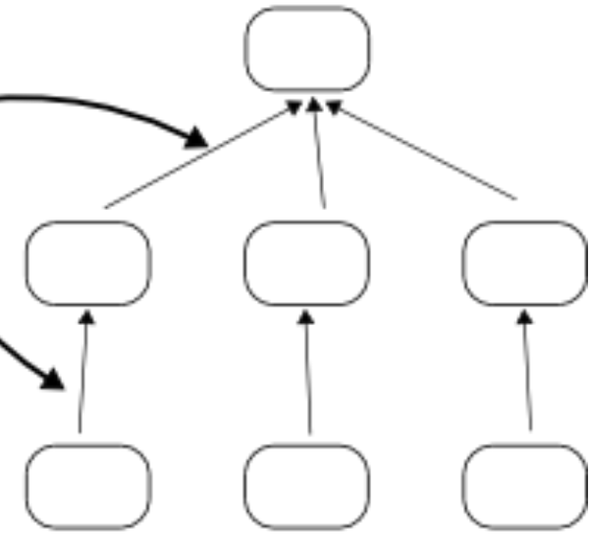
```
rule:
  input: "{sample}.inter"
  output: "{sample}.txt"
  shell: "somecommand {input} {output}"
```

Create a needed intermediate result

```
rule:
  input:  "{sample}.in"
  output: "{sample}.inter"
  run:
    somepythoncode()
```

Use wildcards to write general rules for all samples

Snakemake determines the dependencies for you



Library preparation kit with 50 ng of depleted RNA; Strand orientated RNA-Seq

Library strategy	RNA-Seq
Library source	transcriptomic
Library selection	cDNA
Instrument model	Illumina HiSeq 2500

Description	growth in M63 minimal media plus 0.4% glucose plus 50 µM NiCl2
Data processing	Basecalls performed using HCS 2.0.5 and RTA 1.17.20 RNA-seq reads were aligned to the W3110 genome using CASAVA 1.8.2 Gene expression (based on known genes) was determined using Cufflinks 2.0.2, only PF reads were retained Genome_build: W3110 (NC_007779) Supplementary_files_format_and_content: excel file with coverage and FPKM measurements

Submission date	Dec 18, 2015
Last update date	May 15, 2019
Contact name	Agnès Rodrigue
E-mail(s)	agnes.rodrigue@insa-lyon.fr
Organization name	INSA Lyon CNRS
Lab	MAP UMR 5240
Street address	10 rue Dubois
City	Villeurbanne
ZIP/Postal code	69621
Country	France

Platform ID	GPL18133
Series (1)	GSE76167 Mechanisms of nickel toxicity in bacteria

Relations

BioSample	SAMN04351201
SRA	SRX1492019

Supplementary data files not provided

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

NCBI SRA Run Selector

Accession: PRJNA312017 Search

Filters List

1 Bases
2 Bytes
3 Passage
4 source_name
5 treated_with

Common Fields

BioProject	PRJNA312017
Consent	PUBLIC
Assay Type	RNA-Seq
AvgSpotLen	98
Cell Line	MCF7
Center Name	GEO
DATASTORE filetype	PASTIS, SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1

Select

Run	Bytes	Bases	Download	Cloud DataDelivery	Computing
Total	30	55.59 Gb	54.48G	Metadata or Accession List	
Selected	0	0	0	Metadata or Accession List or RMT Cart	Deliver Data	Galaxy

Found 30 Items

Search within results Clear

	Run	BioSample	Bases	Bytes	Experiment	GEO_Accession	Passage	Sample Name	source_name	treated_with
1	SRR3132145	SAMN04506919	365G	2.18 Gb	SRX1556372	GSM2068642	p9	GSM2068643	MCF7_0hrs_post_E2	EtOH
2	SRR3132146	SAMN04506920	252G	1.53 Gb	SRX1556373	GSM2068644	p9	GSM2068644	MCF7_1Hrs post E2	10nM E2 for 1hr
3	SRR3132147	SAMN04506921	392G	2.35 Gb	SRX1556374	GSM2068645	p9	GSM2068645	MCF7_2hrs_post_E2	10nM E2 for 2hr
4	SRR3132148	SAMN04506922	301G	1.80 Gb	SRX1556375	GSM2068646	p9	GSM2068646	MCF7_3hrs_post_E2	10nM E2 for 3hr
5	SRR3132149	SAMN04506923	314G	1.88 Gb	SRX1556376	GSM2068647	p9	GSM2068647	MCF7_4hrs_post_E2	10nM E2 for 4hr
6	SRR3132150	SAMN04506924	360G	2.16 Gb	SRX1556377	GSM2068648	p9	GSM2068648	MCF7_5hrs_post_E2	10nM E2 for 5hr
7	SRR3132151	SAMN04506925	300G	1.81 Gb	SRX1556378	GSM2068649	p9	GSM2068649	MCF7_6hrs_post_E2	10nM E2 for 6hr
8	SRR3132152	SAMN04506926	264G	1.57 Gb	SRX1556379	GSM2068650	p9	GSM2068650	MCF7_8hrs_post_E2	10nM E2 for 8hr
9	SRR3132153	SAMN04506927	330G	2.00 Gb	SRX1556380	GSM2068651	p9	GSM2068651	MCF7_12hrs_post_E2	10nM E2 for 12hr
10	SRR3132154	SAMN04506928	333G	1.98 Gb	SRX1556381	GSM2068652	p9	GSM2068652	MCF7_24hrs_post_E2	10nM E2 for 24hr
11	SRR3132155	SAMN04506929	319G	1.94 Gb	SRX1556382	GSM2068653	p9	GSM2068653	MCF7_0hrs_post_E2	EtOH
12	SRR3132156	SAMN04506930	293G	1.75 Gb	SRX1556383	GSM2068654	p9	GSM2068654	MCF7_1hrs_post_E2	10nM E2 for 1hr
13	SRR3132157	SAMN04506931	341G	2.06 Gb	SRX1556384	GSM2068655	p9	GSM2068655	MCF7_2hrs_post_E2	10nM E2 for 2hr
14	SRR3132158	SAMN04506873	309G	1.87 Gb	SRX1556385	GSM2068656	p9	GSM2068656	MCF7_3hrs_post_E2	10nM E2 for 3hr
15	SRR3132159	SAMN04506874	379G	2.25 Gb	SRX1556386	GSM2068657	p9	GSM2068657	MCF7_4hrs_post_E2	10nM E2 for 4hr
16	SRR3132160	SAMN04506875	349G	2.11 Gb	SRX1556387	GSM2068658	p9	GSM2068658	MCF7_5hrs_post_E2	10nM E2 for 5hr
17	SRR3132161	SAMN04506876	275G	1.66 Gb	SRX1556388	GSM2068659	p9	GSM2068659	MCF7_6hrs_post_E2	10nM E2 for 6hr
18	SRR3132162	SAMN04506877	280G	1.69 Gb	SRX1556389	GSM2068660	p9	GSM2068660	MCF7_8hrs_post_E2	10nM E2 for 8hr

Project — ~/Desktop/mountjwork/project/tobias/GSE78167

Project

- GSE78167
  - .snakemake
  - GSM2068643_out
  - GSM2068644_out
  - GSM2068645_out
  - GSM2068646_out
  - GSM2068647_out
  - GSM2068648_out
  - GSM2068649_out
  - GSM2068650_out
  - GSM2068651_out
  - GSM2068652_out
  - GSM2068653_out
  - GSM2068654_out
  - GSM2068655_out
  - GSM2068656_out
  - GSM2068657_out
  - GSM2068658_out
  - GSM2068659_out
  - GSM2068660_out
  - GSM2068661_out
  - GSM2068662_out
  - GSM2068663_out
  - GSM2068664_out
  - GSM2068665_out
  - GSM2068666_out
  - GSM2068667_out
  - GSM2068668_out
  - GSM2068669_out
  - GSM2068670_out
  - GSM2068671_out
  - GSM2068672_out
  - logs
  - scripts
    - toSe.R
  - slurm
  - config.yaml
  - run.sh
  - se.rds
  - Snakefile
  - sra_kallisto_paired.yaml
  - SraRunTable.txt

Snakefile

```
#
configfile: "config.yaml"
#
import pandas as pd
#
sample_table = pd.read_table(config["runtable"], sep=",")
samples = list(sample_table['Sample Name'].unique())
#
# everything that need network connection for
localrules: prefetch, get_index_files, create_se
#
rule all:
    input:
        "se.rds"
#
rule create_se:
    input:
        expand("{SRS}_out/{SRS}.abundance.tsv", SRS=samples),
        "genome.gtf.gz"
    output:
        "se.rds"
    script:
        "scripts/toSe.R"
#
rule kallisto:
    input:
        files1=lambda wildcards: expand("raw/{sample}_1.fastq.gz",
            sample=list(sample_table[sample_table['Sample Name']==wildcards.SRS].Run)),
        files2=lambda wildcards: expand("raw/{sample}_2.fastq.gz",
            sample=list(sample_table[sample_table['Sample Name']==wildcards.SRS].Run)),
        gtf="genome.gtf.gz",
        index="kallisto_index/genome"
    output:
        "{SRS}_out/{SRS}.abundance.h5",
        "{SRS}_out/{SRS}.abundance.tsv"
    threads: 32
    params:
        stranded=config['strandedness']
    resources:
        mem_mb=10000
    shell:
        """
        if [ {params.stranded} == "reverse" ]
        then
            kallisto quant --index={input.index} --gtf {input.gtf} --output-dir={wildcards.SRS}_out --rf-stranded --threads={threads} -b 30 {input.files1} {input.files2}
        elif [ {params.stranded} == "forward" ]
        then
            kallisto quant --index={input.index} --gtf {input.gtf} --output-dir={wildcards.SRS}_out --fr-stranded --threads={threads} -b 30 {input.files1} {input.files2}
        else
            kallisto quant --index={input.index} --gtf {input.gtf} --output-dir={wildcards.SRS}_out --threads={threads} -b 30 {input.files1} {input.files2}
        fi
        mv {wildcards.SRS}_out/abundance.h5 {wildcards.SRS}_out/{wildcards.SRS}.abundance.h5
        mv {wildcards.SRS}_out/abundance.tsv {wildcards.SRS}_out/{wildcards.SRS}.abundance.tsv
        """
#
#
rule fastq_dump:
    input:
        "sra/{SRR}.sra"
    output:
        temp("raw/{SRR}_1.fastq.gz"),
        temp("raw/{SRR}_2.fastq.gz")
    shell:
        """
```

Snakefile 15 — CRLF UTF-8 Python GitHub Git (0)

Project — ~/Desktop/mountjwork/project/tobias/GSE78167

Project

- GSE78167
  - .snakemake
  - GSM2068643_out
  - GSM2068644_out
  - GSM2068645_out
  - GSM2068646_out
  - GSM2068647_out
  - GSM2068648_out
  - GSM2068649_out
  - GSM2068650_out
  - GSM2068651_out
  - GSM2068652_out
  - GSM2068653_out
  - GSM2068654_out
  - GSM2068655_out
  - GSM2068656_out
  - GSM2068657_out
  - GSM2068658_out
  - GSM2068659_out
  - GSM2068660_out
  - GSM2068661_out
  - GSM2068662_out
  - GSM2068663_out
  - GSM2068664_out
  - GSM2068665_out
  - GSM2068666_out
  - GSM2068667_out
  - GSM2068668_out
  - GSM2068669_out
  - GSM2068670_out
  - GSM2068671_out
  - GSM2068672_out
  - logs
  - scripts
    - toSe.R
  - slurm
  - config.yaml
  - run.sh
  - se.rds
  - Snakefile
  - sra_kallisto_paired.yaml
  - SraRunTable.txt

sra_kallisto_paired.yaml

```yaml
channels:
  - bioconda
  - conda-forge
dependencies:
  - snakemake-minimal =7.2
  - python
  - pandas
  - sra-tools =2.11
  - kallisto =0.48
  - r >=4.0
  - bioconductor-tximport
  - bioconductor-SummarizedExperiment
  - bioconductor-biomaRt
  - bioconductor-GenomicFeatures

```

sra_kallisto_paired.yaml 1:1 LF UTF-8 YAML GitHub Git (0)

Project — ~/Desktop/mountjwork/project/tobias/GSE78167

Project

- GSE78167
  - .snakemake
  - logs
  - scripts
  - slurm
  - config.yaml
  - run.sh
  - Snakefile
  - sra_kallisto_paired.yaml
  - SraRunTable.txt

sra_kallisto_paired.yaml

```yaml
channels:
  - bioconda
  - conda-forge
dependencies:
  - snakemake-minimal =7.2
  - python
  - pandas
  - sra-tools =2.11
  - kallisto =0.48
  - r >=4.0
  - bioconductor-tximport
  - bioconductor-SummarizedExperiment
  - bioconductor-biomaRt
  - bioconductor-GenomicFeatures

```

```
        then
            kallisto quant --index=kallisto_index/genome --gtf genome.gtf.gz --output-dir=GSM2068659_out --fr-stranded --threads=1 -b 30 raw/SRR3182161_1.fastq.gz raw/SRR3182161_2.fastq.gz
        else
            kallisto quant --index=kallisto_index/genome --gtf genome.gtf.gz --output-dir=GSM2068659_out --threads=1 -b 30 raw/SRR3182161_1.fastq.gz raw/SRR3182161_2.fastq.gz
        fi
        mv GSM2068659_out/abundance.h5 GSM2068659_out/GSM2068659.abundance.h5
        mv GSM2068659_out/abundance.tsv GSM2068659_out/GSM2068659.abundance.tsv


[Sun May 29 15:54:10 2022]
localrule create_se:
    input: GSM2068643_out/GSM2068643.abundance.tsv, GSM2068644_out/GSM2068644.abundance.tsv, GSM2068645_out/GSM2068645.abundance.tsv, GSM2068646_out/GSM2068646.abundance.tsv, GSM2068647_out/GSM
2068647.abundance.tsv, GSM2068648_out/GSM2068648.abundance.tsv, GSM2068649_out/GSM2068649.abundance.tsv, GSM2068650_out/GSM2068650.abundance.tsv, GSM2068651_out/GSM2068651.abundance.tsv, GSM206
8652_out/GSM2068652.abundance.tsv, GSM2068653_out/GSM2068653.abundance.tsv, GSM2068654_out/GSM2068654.abundance.tsv, GSM2068655_out/GSM2068655.abundance.tsv, GSM2068656_out/GSM2068656.abundance
.tsv, GSM2068657_out/GSM2068657.abundance.tsv, GSM2068658_out/GSM2068658.abundance.tsv, GSM2068659_out/GSM2068659.abundance.tsv, GSM2068660_out/GSM2068660.abundance.tsv, GSM2068661_out/GSM20686
61.abundance.tsv, GSM2068662_out/GSM2068662.abundance.tsv, GSM2068663_out/GSM2068663.abundance.tsv, GSM2068664_out/GSM2068664.abundance.tsv, GSM2068665_out/GSM2068665.abundance.tsv, GSM2068666_
out/GSM2068666.abundance.tsv, GSM2068667_out/GSM2068667.abundance.tsv, GSM2068668_out/GSM2068668.abundance.tsv, GSM2068669_out/GSM2068669.abundance.tsv, GSM2068670_out/GSM2068670.abundance.tsv,
 GSM2068671_out/GSM2068671.abundance.tsv, GSM2068672_out/GSM2068672.abundance.tsv, genome.gtf.gz
    output: se.rds
    jobid: 1
    resources: tmpdir=/tmp

[Sun May 29 15:54:10 2022]
localrule all:
    input: se.rds
    jobid: 0
    resources: tmpdir=/tmp

Job stats:
job               count    min threads    max threads
---------------  -------  -------------  -------------
all                   1              1              1
create_se             1              1              1
fastq_dump           30              1              1
get_index_files       1              1              1
kallisto             30              1              1
kallisto_index        1              1              1
prefetch             30              1              1
total                94              1              1

This was a dry-run (flag -n). The order of jobs does not reflect the order of execution.
(sra_kallisto_paired) [tobiasst@master GSE78167]$
```

sra_kallisto_paired.yaml 1/1 LF UTF-8 YAML GitHub Git (0)

Project — ~/Desktop/mountjwork/project/tobias/GSE78167

Project

- GSE78167
  - .snakemake
  - logs
  - scripts
  - slurm
  - config.yaml
  - run.sh
  - Snakefile
  - sra_kallisto_paired.yaml
  - SraRunTable.txt

config.yaml

```yaml
runtable: SraRunTable.txt
transcriptsFTP: http://ftp.ensembl.org/pub/release-106/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz
gtfFTP: http://ftp.ensembl.org/pub/release-106/gtf/homo_sapiens/Homo_sapiens.GRCh38.106.gtf.gz
strandedness: reverse
organism: human
annotationVersion: 106

```

```
kallisto_index        1             1             1
prefetch             30             1             1
total                94             1             1

This was a dry-run (flag -n). The order of jobs does not reflect the order of execution.
(sra_kallisto_paired) [tobiasst@master GSE78167]$
```

config.yaml 1:1 LF UTF-8 YAML GitHub Git (0)

RNA_Seq_reference - RStudio

01_project_overview.Rmd

```
---
title: "Project Overview"
author: "tobiasst"
date: "`r format(Sys.time(), "%d %B, %Y")`"
output:
  html_document:
    df_print: paged
    code_folding: hide
    toc: true
---

```{r libs}
suppressPackageStartupMessages({
  library(SummarizedExperiment)
  library(RColorBrewer)
  library(ggplot2)
  library(ggrepel)
})
```

Series GSE78167

An integrative transcriptomics approach identifies miR-503 as a candidate master regulator of the estrogen respo

Quantification of mRNAs in MCF7 cells responding to estrogen following a period of estrogen starvation. Three in

```{r init_vars}
cluster.dir <- "/Users/tobiasst/Desktop/mount/work/project/tobias/GSE78167/"
```

group variable

```{r}
group_var <- "treated_with"
group_var
```
```

Collect data from clus...
Read distribution
Genes detected
PCA
sample based
PCA
gene based

15:24 Chunk 1: libs — R Markdown

Console Terminal

R 4.2.0 · ~/Library/Mobile Documents/com~apple~CloudDocs/Documents/RNA_Seq_reference/

```
> cluster.dir <- "/Users/tobiasst/Desktop/mount/work/project/tobias/GSE78167/"
> group_var <- "treated_with"
> group_var
[1] "treated_with"
> batch_var <- "Passage"
> batch_var
[1] "Passage"
>
```

Environment | History | Connections | Tutorial

Import | 630 MiB | List

R | Global Environment

Data

se	Large SummarizedExperiment (260..

Values

batch_var	"Passage"
cluster.dir	"/Users/tobiasst/Desktop/mount/wor..
group_var	"treated_with"

Files | Plots | Packages | Help | Viewer

com~apple~CloudDocs › Documents › RNA_Seq_reference

Name	Size	Modified
..		
.Rhistory	15.4 KB	May 29,
01_project_overview.html	2.1 MB	May 28,
01_project_overview.Rmd	3.9 KB	May 28,
02_DGE.html	1 MB	May 6, 2
02_DGE.Rmd	1.8 KB	May 9, 2
03_EnrichR.html	5.3 MB	May 29,
03_EnrichR.Rmd	2.3 KB	May 29,
03_Heatmaps.html	1.1 MB	May 28,
03_Heatmaps.Rmd	2 KB	May 28,
03_MA_Vulcano.html	1.6 MB	May 28,
03_MA_Vulcano.Rmd	2.1 KB	May 28,
03_pathways_2h_alternative.Rmd	8.1 KB	May 29,
03_pathways_2h.html	2.4 MB	May 28,
03_pathways_2h.Rmd	7.2 KB	May 28,
03_pathways_24h.html	3.3 MB	May 28,

PCA

sample based

colored for group

Hide

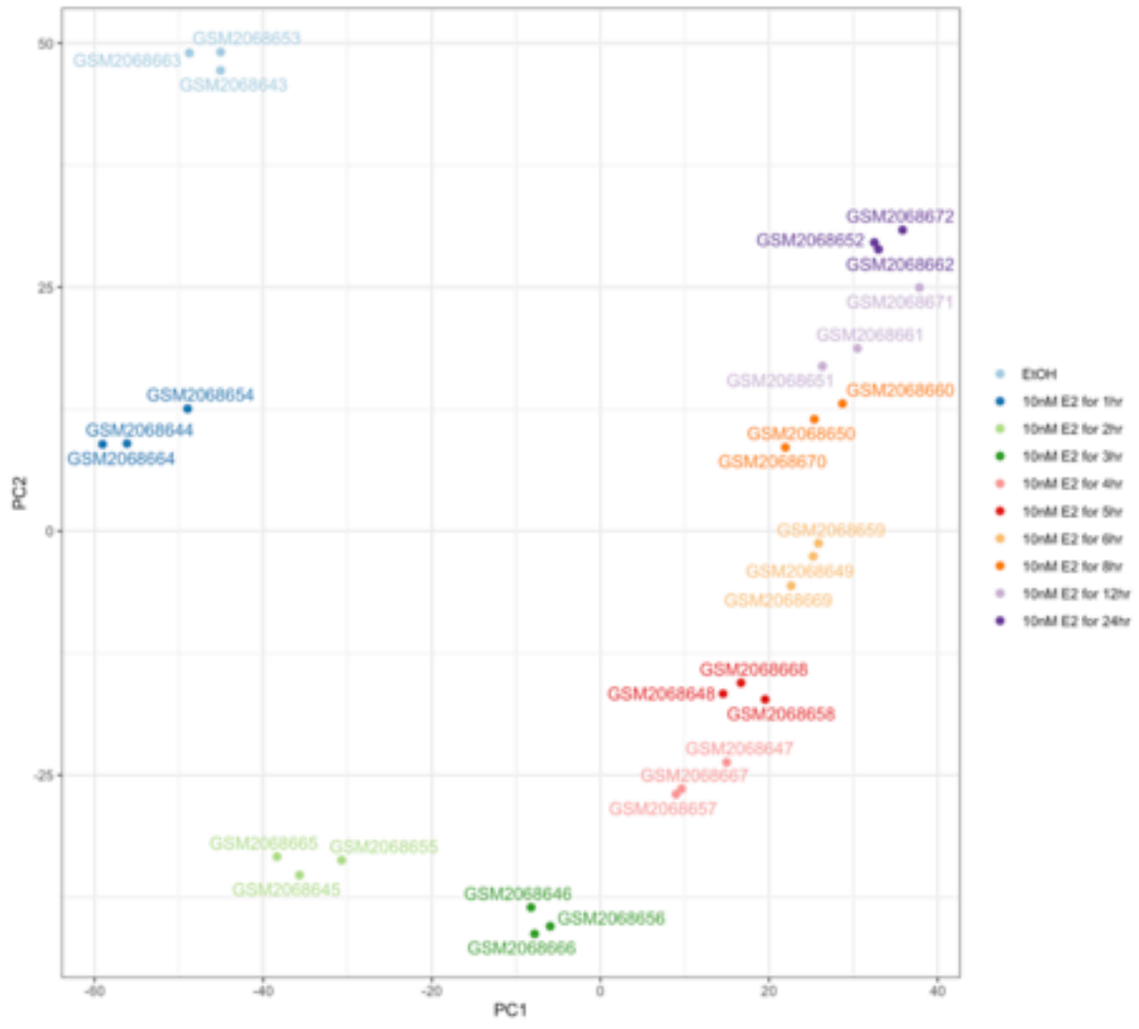
```
se <- se[!is.na(rowData(se)$entrez),]

mat <- log2(assays(se)$tpms+0.01)
rownames(mat) <- rowData(se)$symbol
mat <- mat[apply(mat,1,var)>0,]

pca <- prcomp(t(mat))

ggdf <- data.frame(pca$x[,1:2])
ggdf$class  <- colData(se)[,group_var]
ggdf$sample_id <- colnames(se)

ggplot(ggdf, aes(x=PC1, y=PC2, col=class)) +
  geom_point(size=2) +
  geom_text_repel(aes_string(label = "sample_id"), show.legend = FALSE, max.overla
ps = 10) +
  scale_color_manual(values = class.cols) +
  theme_bw() +  theme(legend.title = element_blank())
```



immediate early

Hide

```
res <- results(dds, contrast=c("treated_with","10nM.E2.for.2hr","EtOH"))
summary(res)
```

```
##
## out of 26098 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 2502, 9.6%
## LFC < 0 (down)     : 2127, 8.2%
## outliers [1]       : 2, 0.0077%
## low counts [2]     : 10120, 39%
## (mean count < 4)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

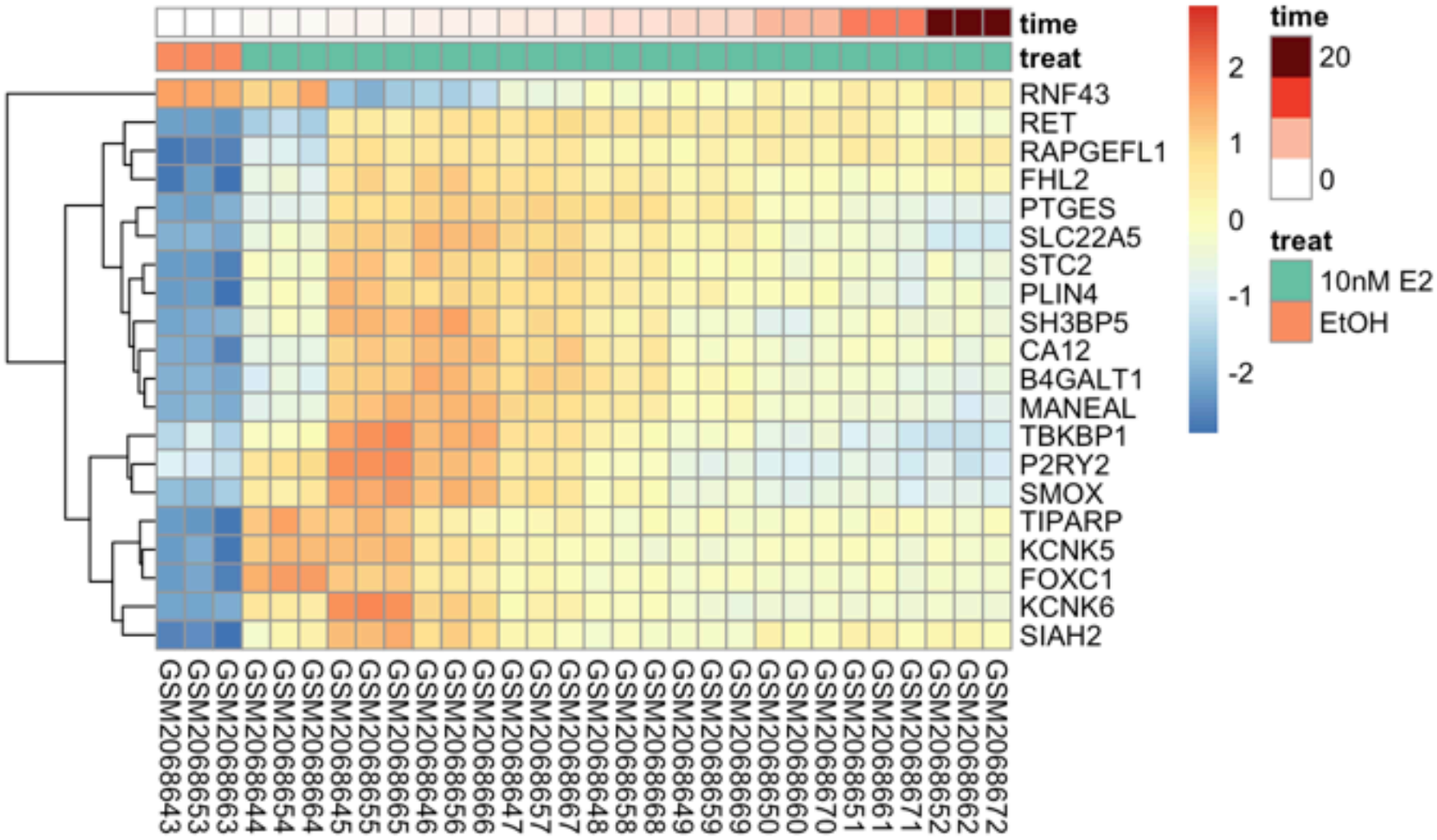
Code

2h

top 20 de-regulated genes

Code

Code

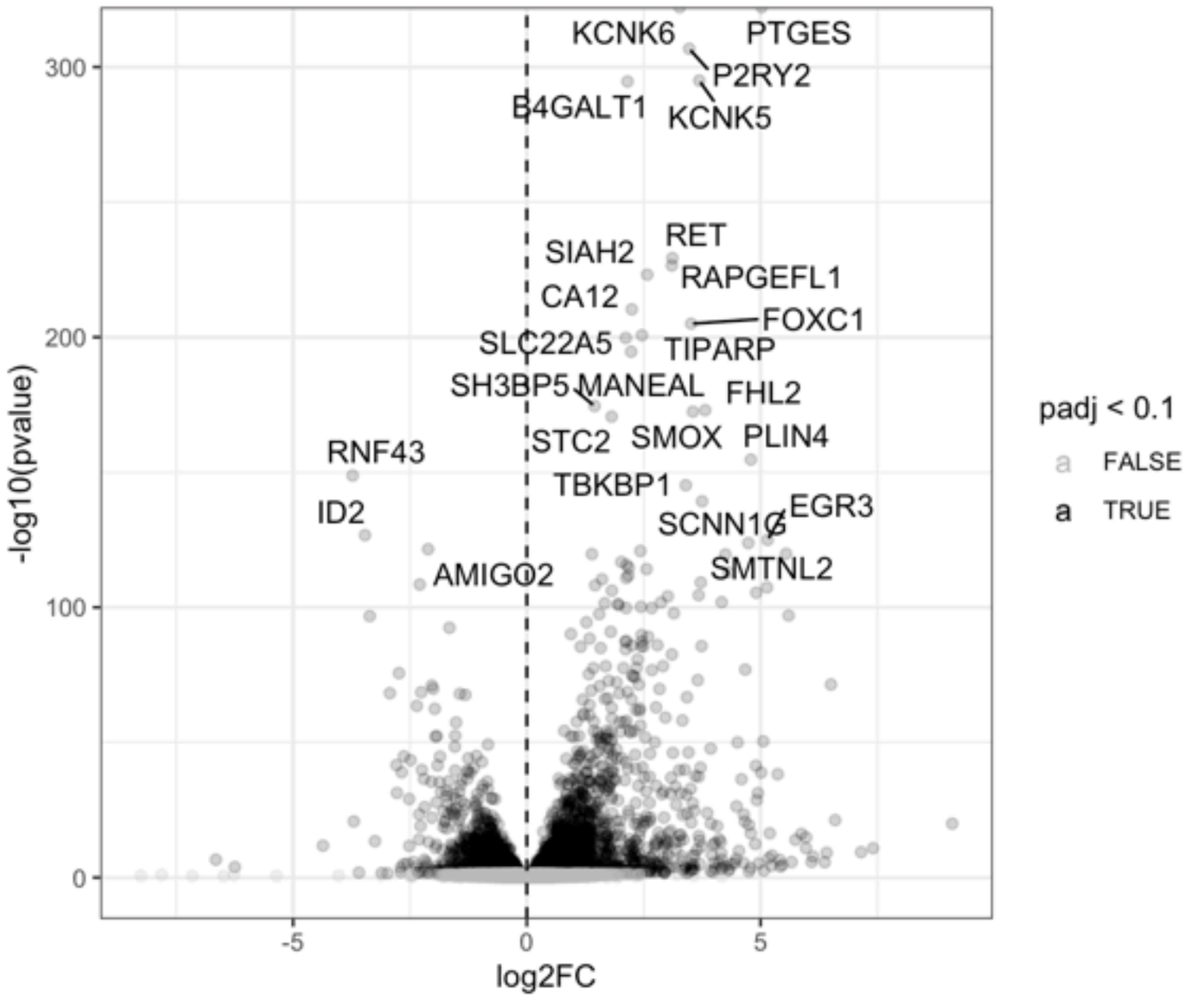


Vulcano

Code

Hide

```
ggplot(ggdf, aes(x=log2FC, y=-log10(pvalue), label=gene, col=padj<0.1)) + geom_point(alpha=0.2) +
  theme_bw() + geom_vline(xintercept = 0, linetype=2) +
  geom_text_repel() +
  scale_color_manual(values=c("grey","black"))
```



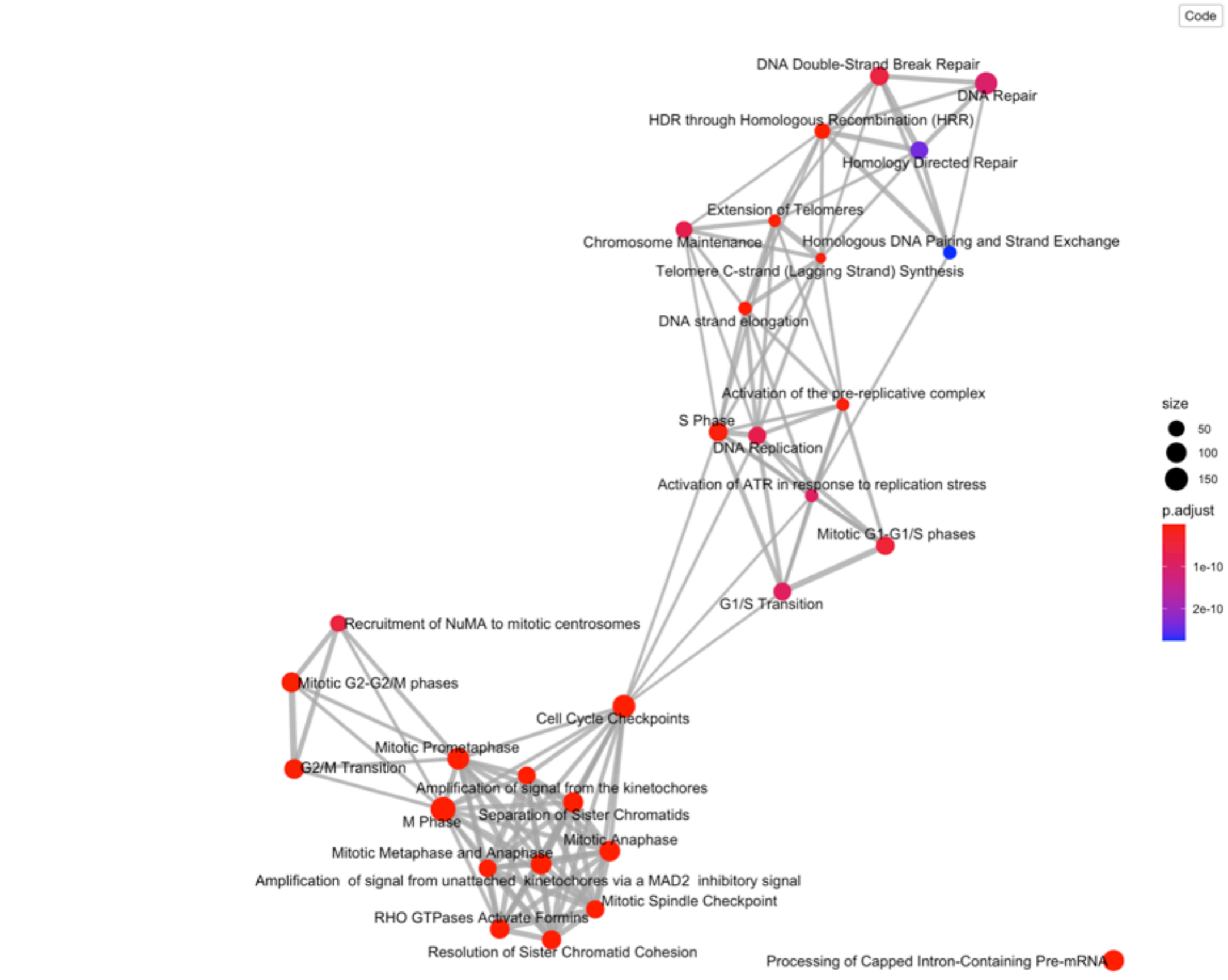
GSEA

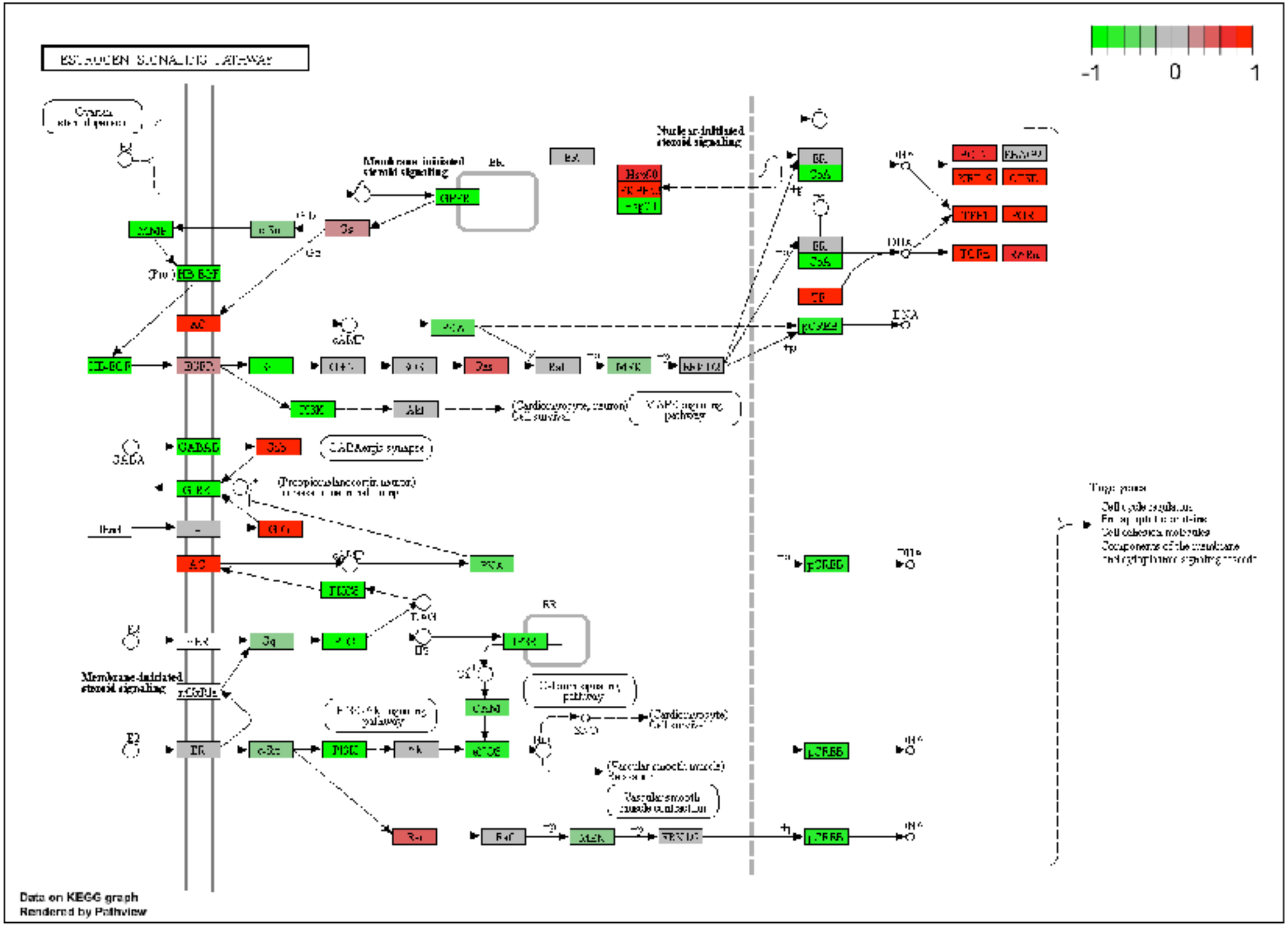
Code

Hallmark gene sets

Code

Pathway	Gene ranks	NES	pval	padj
HALLMARK_ESTROGEN_RESPONSE_EARLY		2.71	1.0e-10	2.5e-09
HALLMARK_ESTROGEN_RESPONSE_LATE		2.52	1.0e-10	2.5e-09
HALLMARK_UNFOLDED_PROTEIN_RESPONSE		1.81	1.5e-05	2.5e-04
HALLMARK_MYC_TARGETS_V2		1.76	1.4e-03	1.0e-02
HALLMARK_HYPOXIA		1.63	1.4e-04	1.7e-03
HALLMARK_WNT_BETA_CATENIN_SIGNALING		-1.27	1.2e-01	2.4e-01
HALLMARK_INTERFERON_GAMMA_RESPONSE		-1.28	1.3e-02	3.9e-02
HALLMARK_INTERFERON_ALPHA_RESPONSE		-1.40	4.3e-02	1.1e-01
HALLMARK_E2F_TARGETS		-1.42	3.5e-03	1.4e-02
HALLMARK_UV_RESPONSE_DN		-1.46	6.4e-03	2.3e-02


Code
DNA Double-Strand Break Repair
DNA Repair
HDR through Homologous Recombination (HRR)
Homology Directed Repair
Extension of Telomeres
Chromosome Maintenance
Homologous DNA Pairing and Strand Exchange
Telomere C-strand (Lagging Strand) Synthesis
DNA strand elongation
Activation of the pre-replicative complex
S Phase
DNA Replication
Activation of ATR in response to replication stress
Mitotic G1-G1/S phases
G1/S Transition
Recruitment of NuMA to mitotic centrosomes
Mitotic G2-G2/M phases
Cell Cycle Checkpoints
Mitotic Prometaphase
G2/M Transition
Amplification of signal from the kinetochores
M Phase
Separation of Sister Chromatids
Mitotic Anaphase
Mitotic Metaphase and Anaphase
Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal
Mitotic Spindle Checkpoint
RHO GTPases Activate Formins
Resolution of Sister Chromatid Cohesion
Processing of Capped Intron-Containing Pre-mRNA
size
50
100
150
p.adjust
1e-10
2e-10


-1
0
1
Data on KEGG graph
Rendered by Pathview

curated gene sets

Code

Pathway	Gene ranks	NES	pval	padj
DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP		3.13	1.0e-10	3.4e-09
DUTERTRE_ESTRADIOL_RESPONSE_6HR_UP		2.98	1.0e-10	3.4e-09
KOBAYASHI_EGFR_SIGNALING_24HR_DN		2.88	1.0e-10	3.4e-09
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN		2.85	1.0e-10	3.4e-09
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER		2.85	1.0e-10	3.4e-09
BHAT_ESR1_TARGETS_VIA_AKT1_DN		-2.62	1.0e-10	3.4e-09
BHAT_ESR1_TARGETS_NOT_VIA_AKT1_DN		-2.77	1.0e-10	3.4e-09
FRASOR_RESPONSE_TO_ESTRADIOL_DN		-2.79	1.0e-10	3.4e-09
DUTERTRE_ESTRADIOL_RESPONSE_6HR_DN		-3.05	1.0e-10	3.4e-09
DUTERTRE_ESTRADIOL_RESPONSE_24HR_DN		-3.10	1.0e-10	3.4e-09