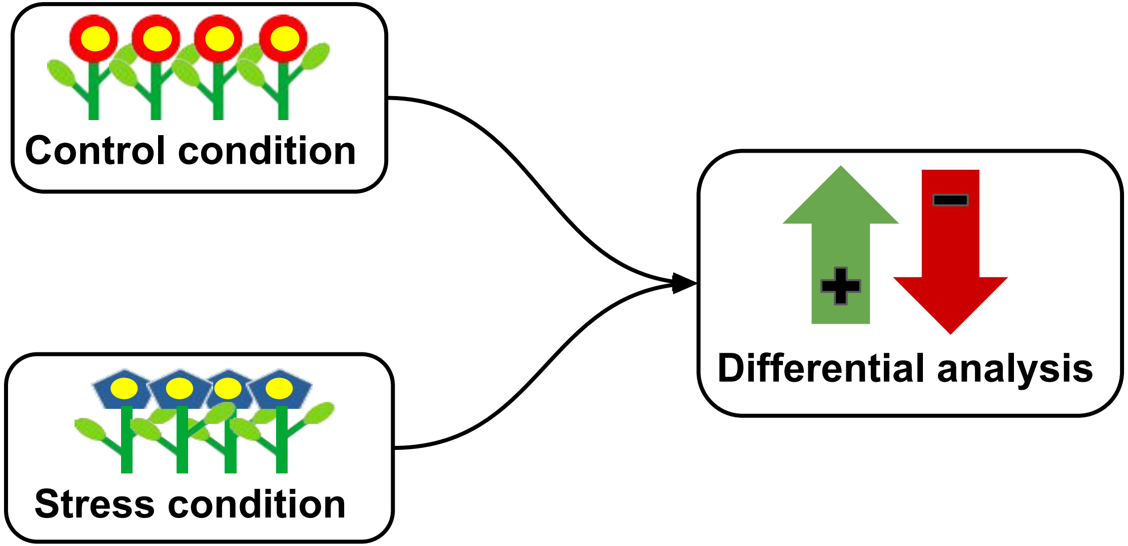
Impact of replicate number on omics analyses

Jeremy FERRARO

25/11/2025

Statistical power



Statistical power ≠ 1

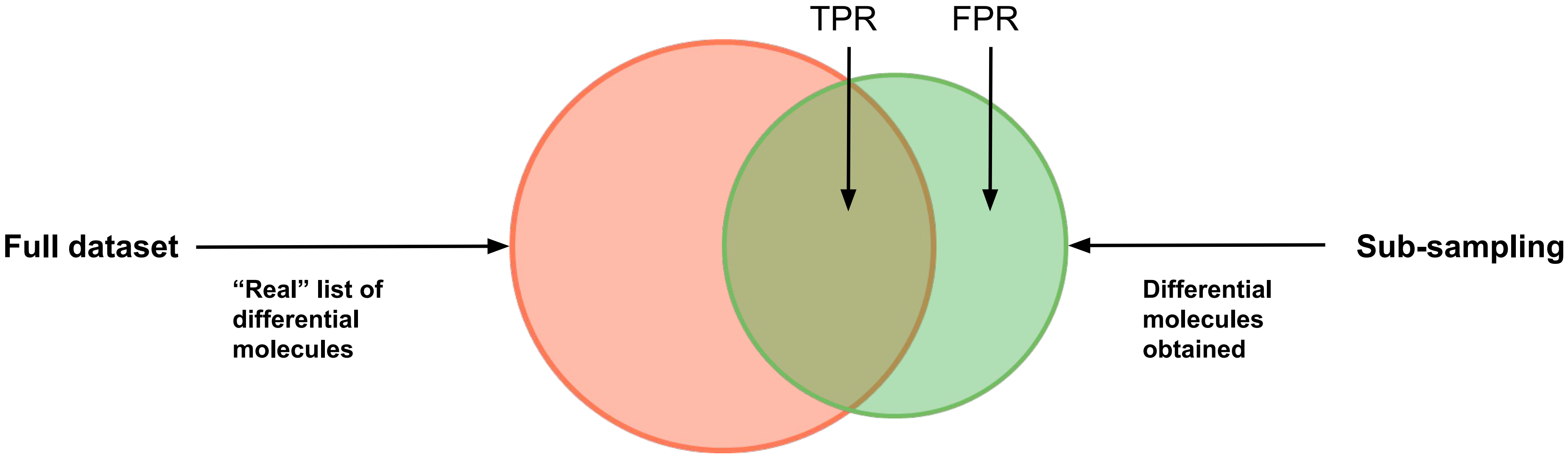
↓

Unable to identified all differential molecules

Estimation of statistical power: TPR and FPR

TPR= True Positive Rate

FPR= False Positive Rate



TPR

FPR

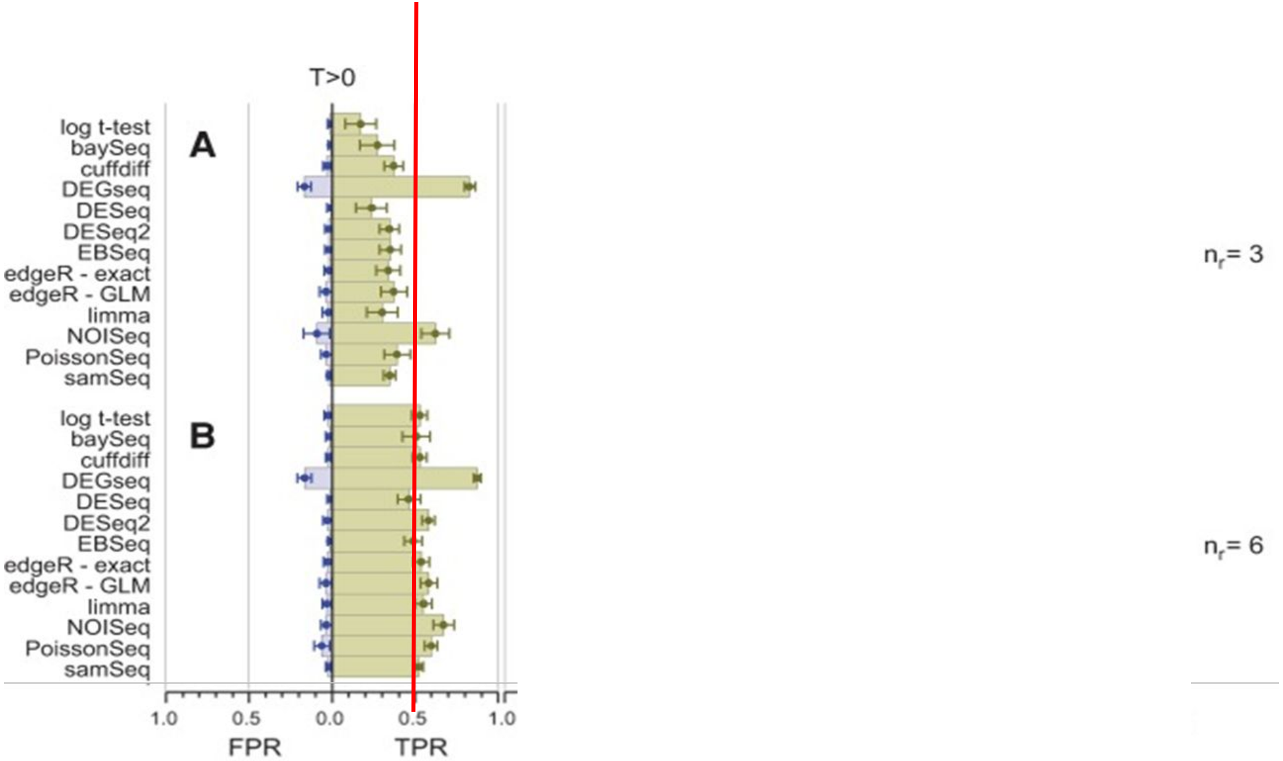
Full dataset

“Real” list of differential molecules

Sub-sampling

Differential molecules obtained

How statistical power is influenced?

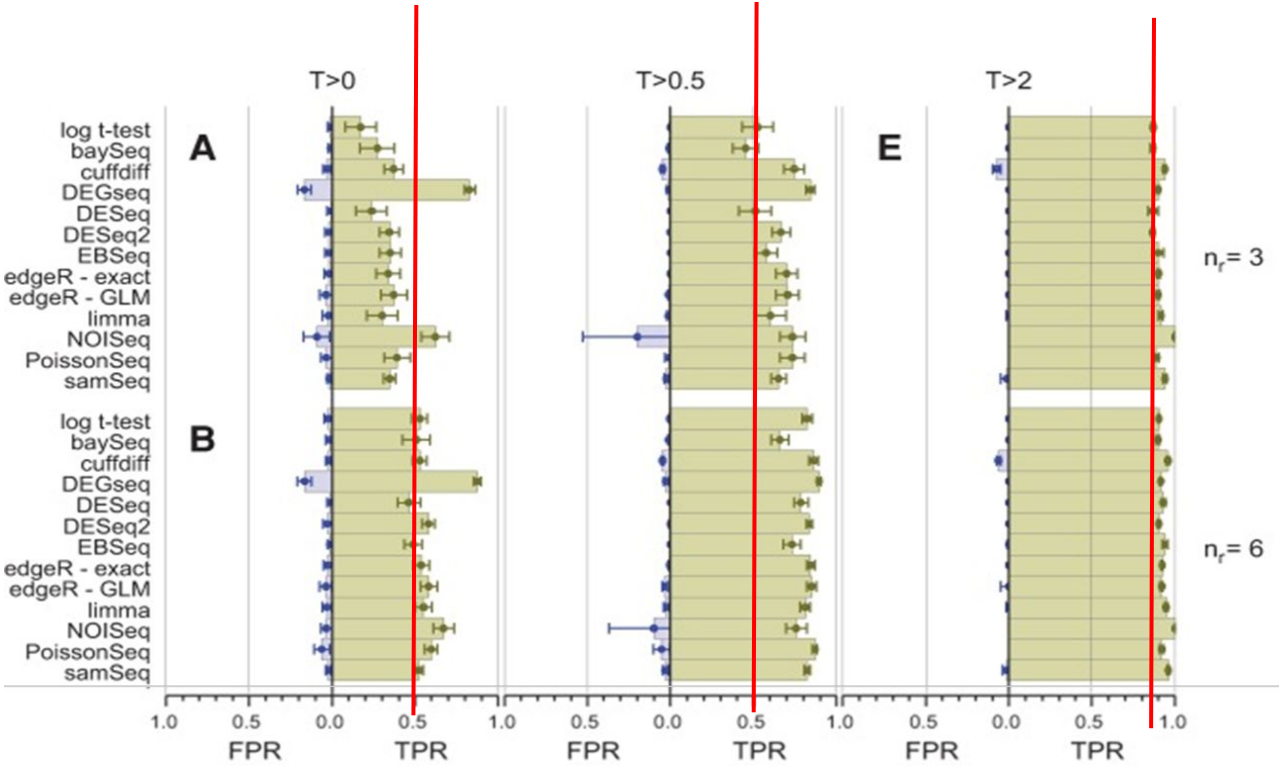


Schurch et al. 2016

RNAseq

48 yeast replicates

Increased power by increasing the number of replicates



Schurch et al. 2016

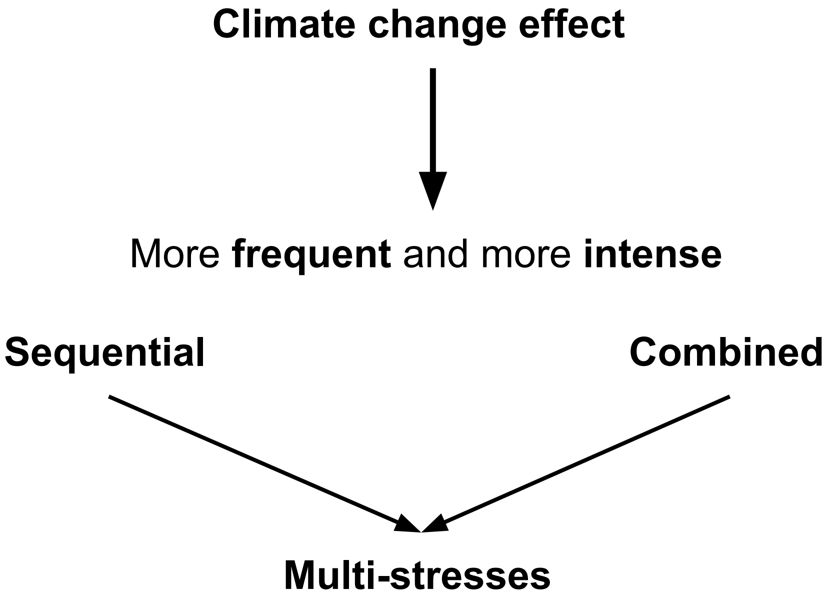
RNAseq

48 yeast replicates

LFC threshold

Number of replicates

Decoding multistress with limited power



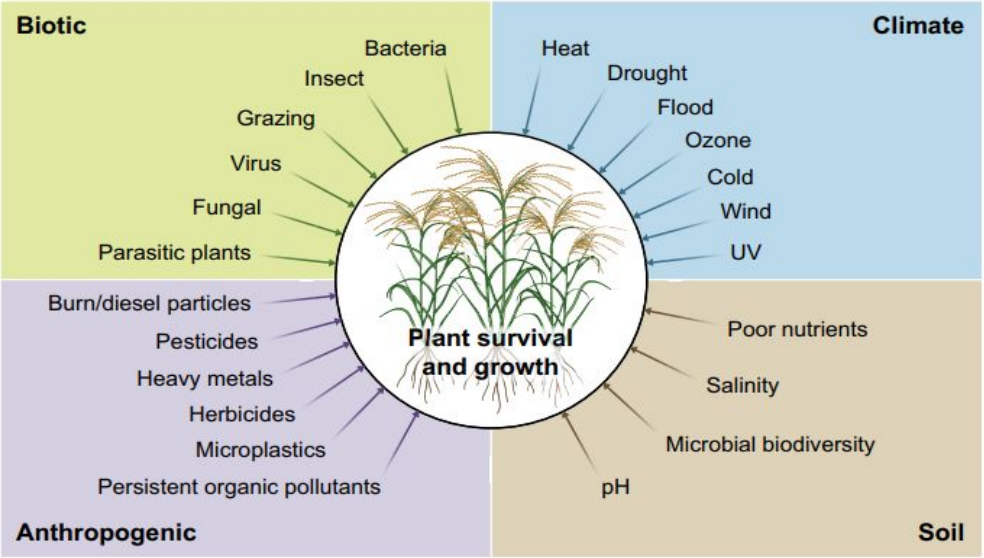
Climate change effect

More **frequent** and more **intense**

Sequential

Combined

Multi-stresses



Zandalinas et al. 2022

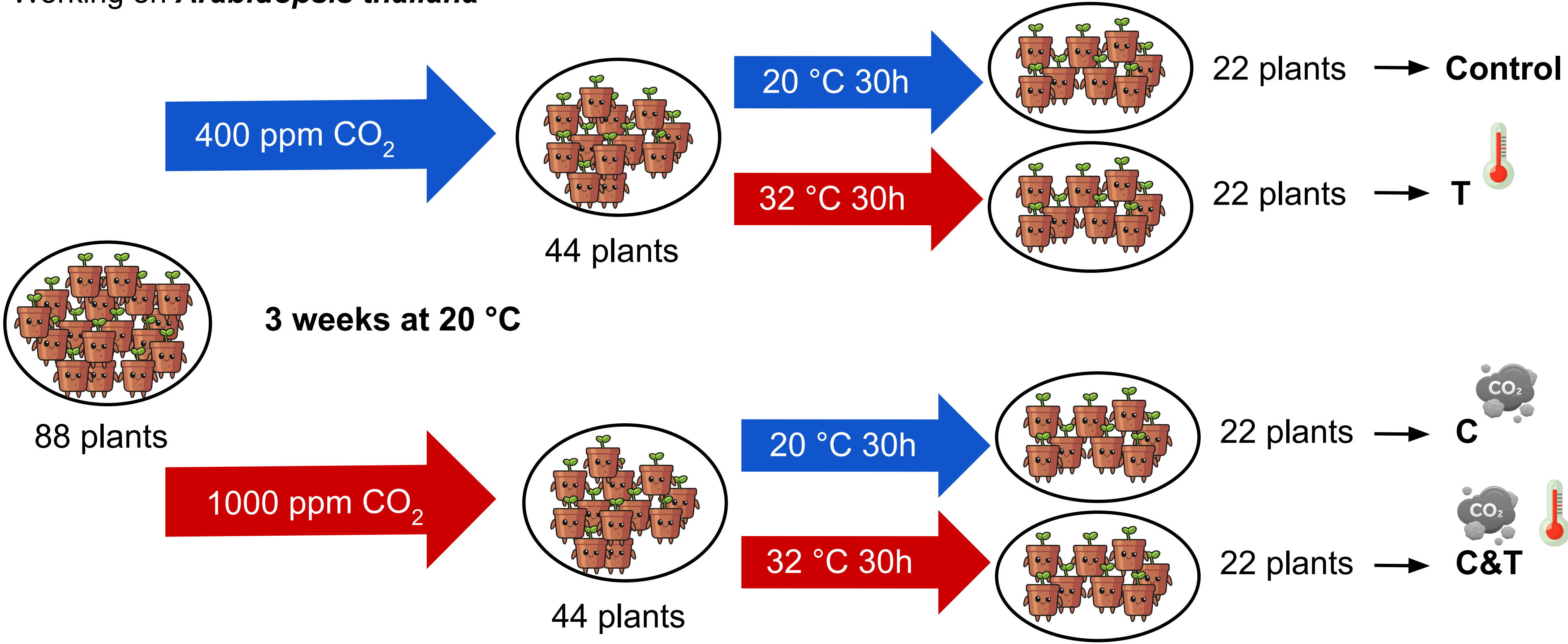
A lot of **comparison** for few replicates

Objective

How does the lack of replication impact interpretations made in multistress studies?

Experimental design: Based the most pessimistic projections for 2100

Working on ***Arabidopsis thaliana***



88 plants

3 weeks at 20 °C

400 ppm CO_2

44 plants

20 °C 30h

22 plants → **Control**

32 °C 30h

22 plants → **T**

1000 ppm CO_2

44 plants

20 °C 30h

22 plants → **C**

32 °C 30h

22 plants → **C&T**

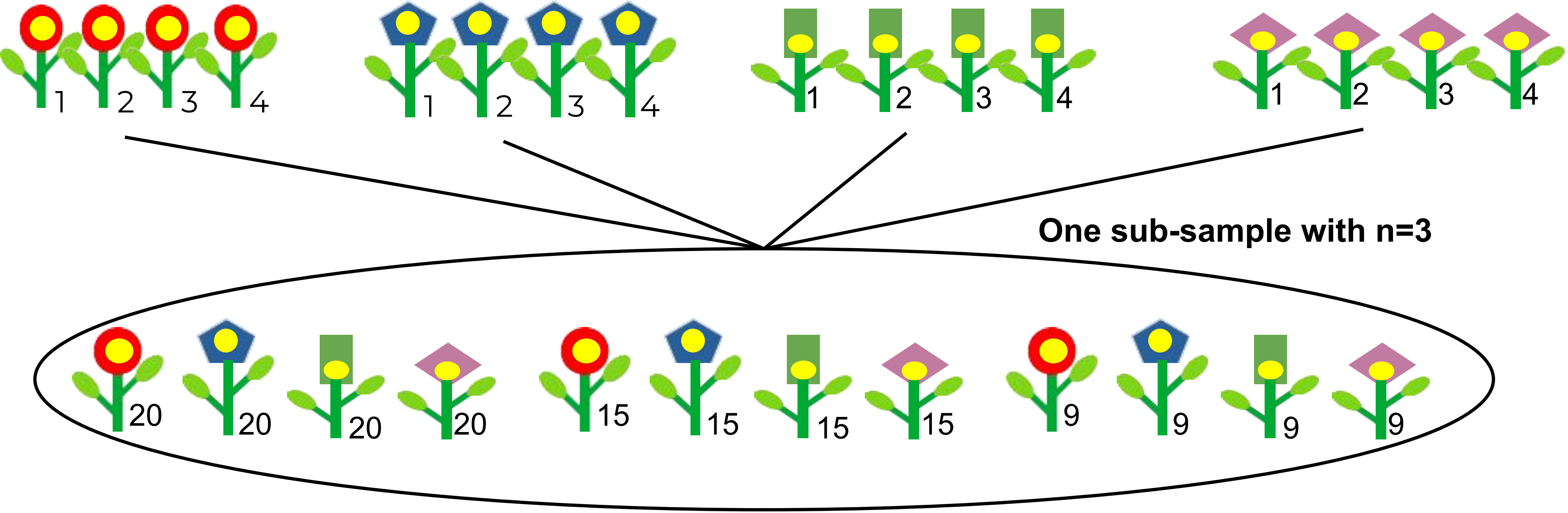
Sub-sampling

Control

Stress CO_2

Stress temperature

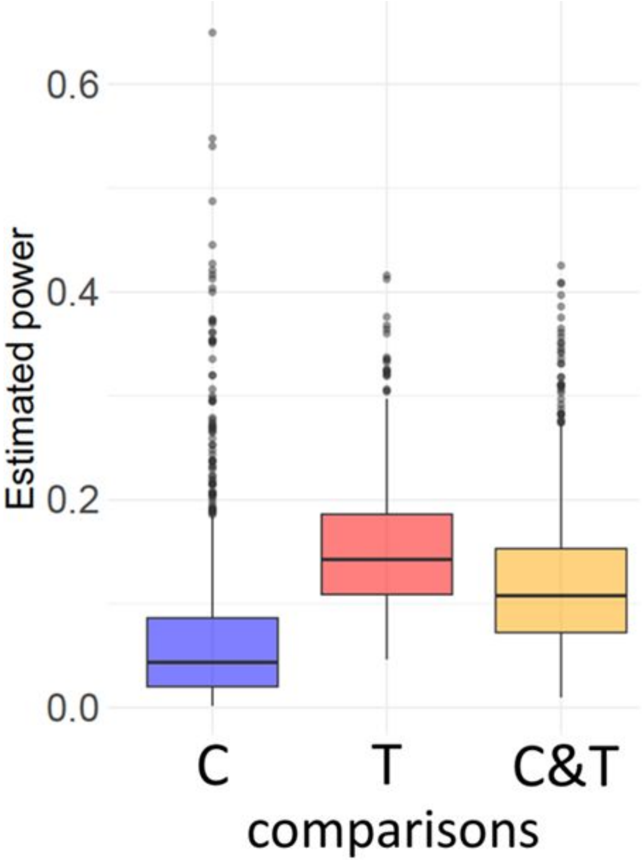
Combined stress



One sub-sample with n=3

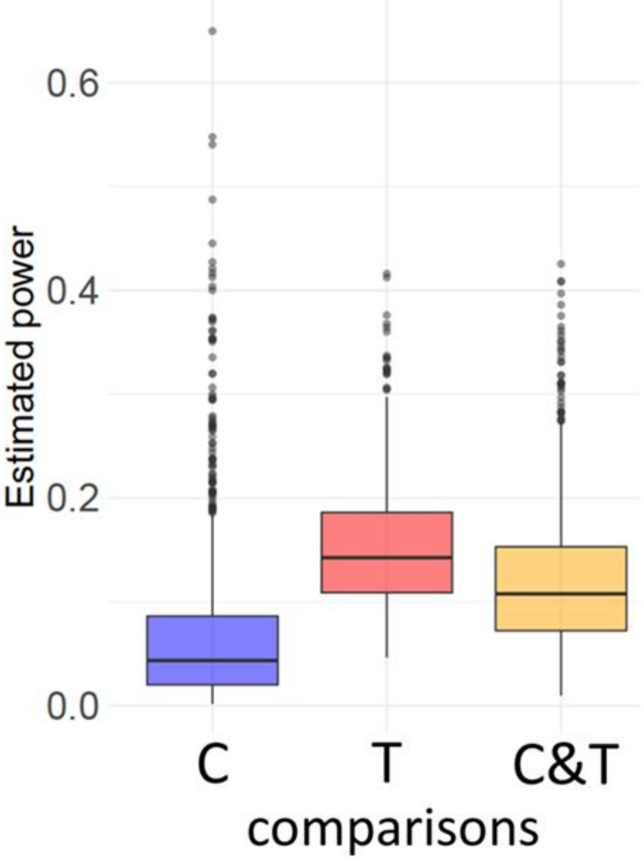
1000 random sub-samples for each size

Statistical power at three replicates

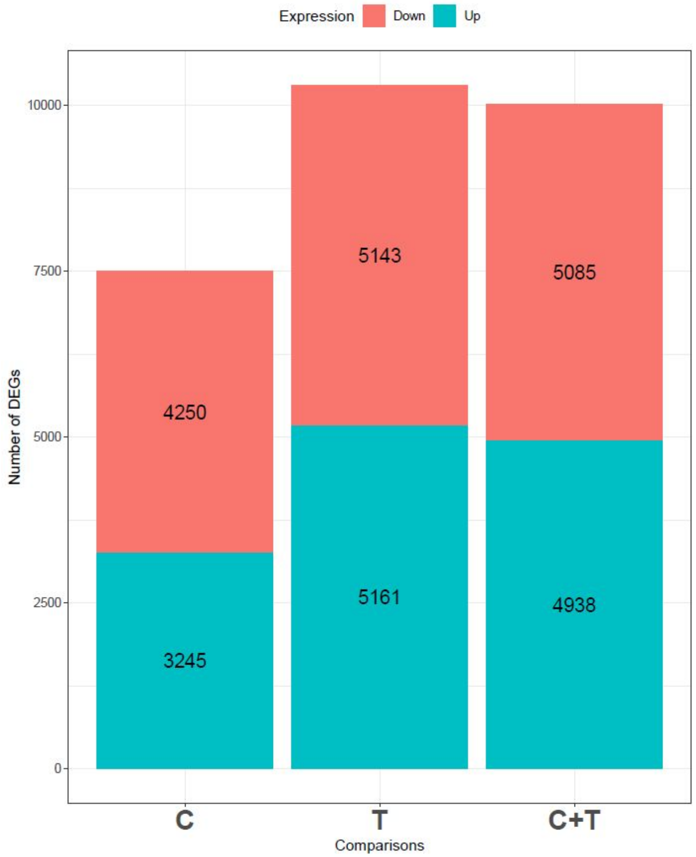


Statistical power **<20%**

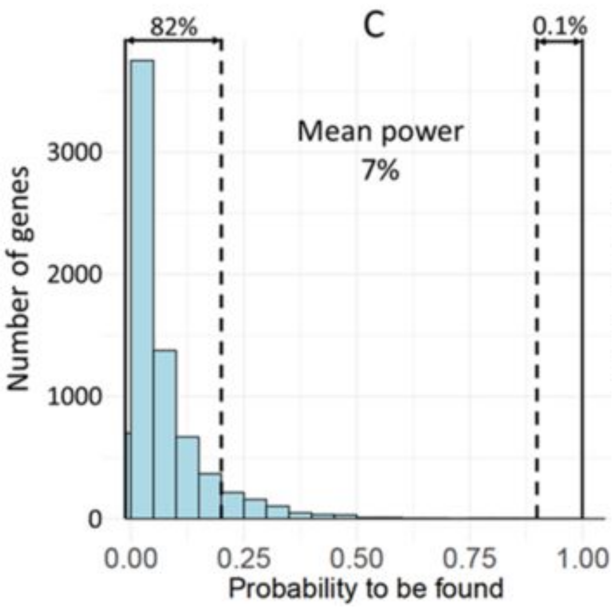
Statistical power depends on gene response



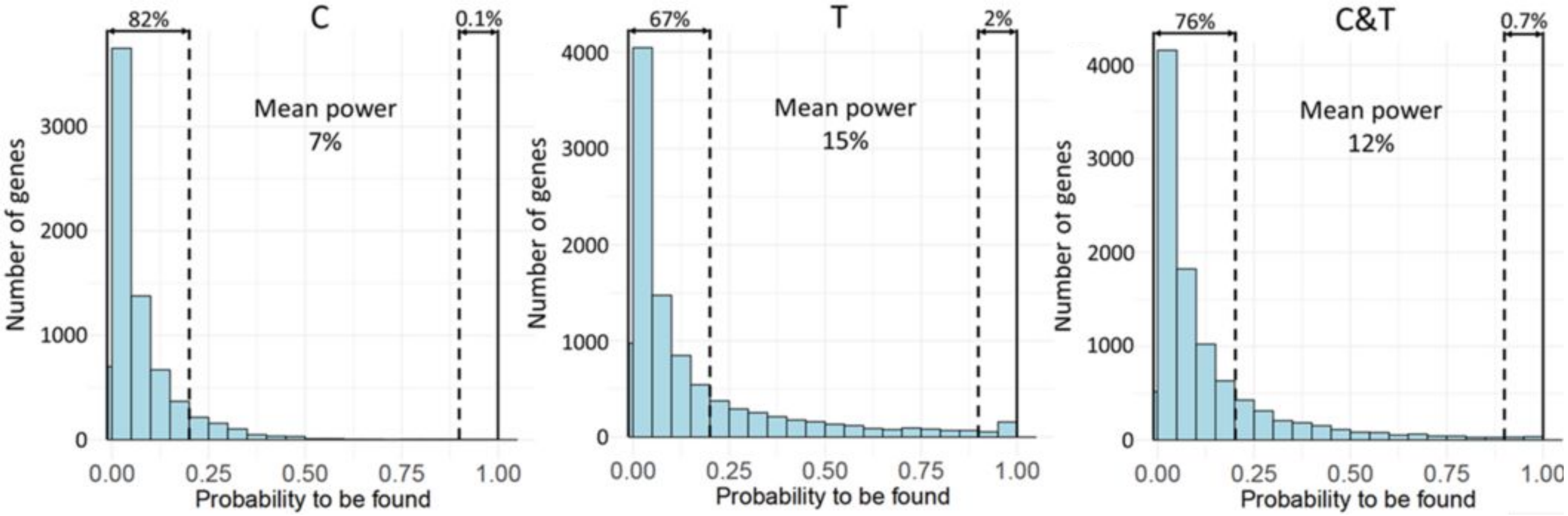
Statistical power **<20%**



Reproducibility at three replicates

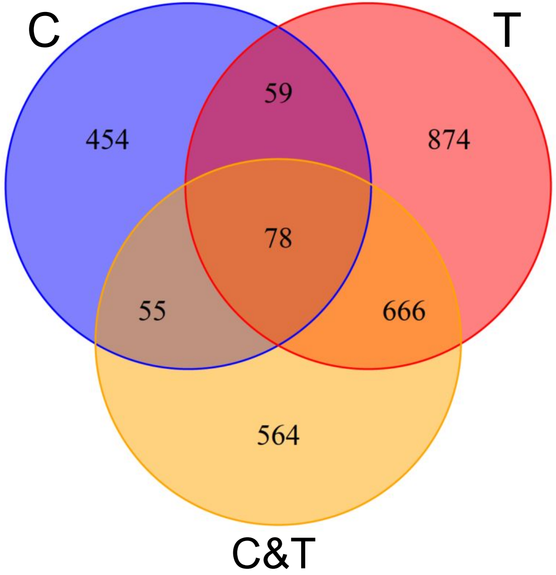


Very poor reproducibility in all conditions



Venn diagrams widely used in omics

427 papers published in Nature Plants between 2020 and July 2024 → **108** studies comparing DEG lists **Venn or UpsetR diagrams**

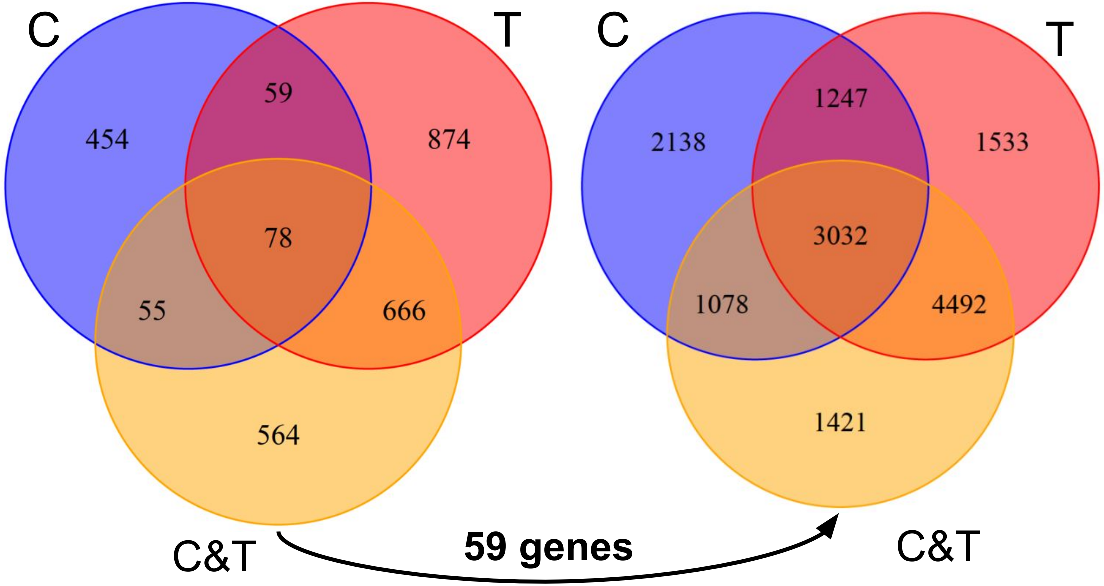


n=3

Considered **non-predictive**

Effect C=0 Effect T=0 Effect C&T=0

Low replication inflates false specificity

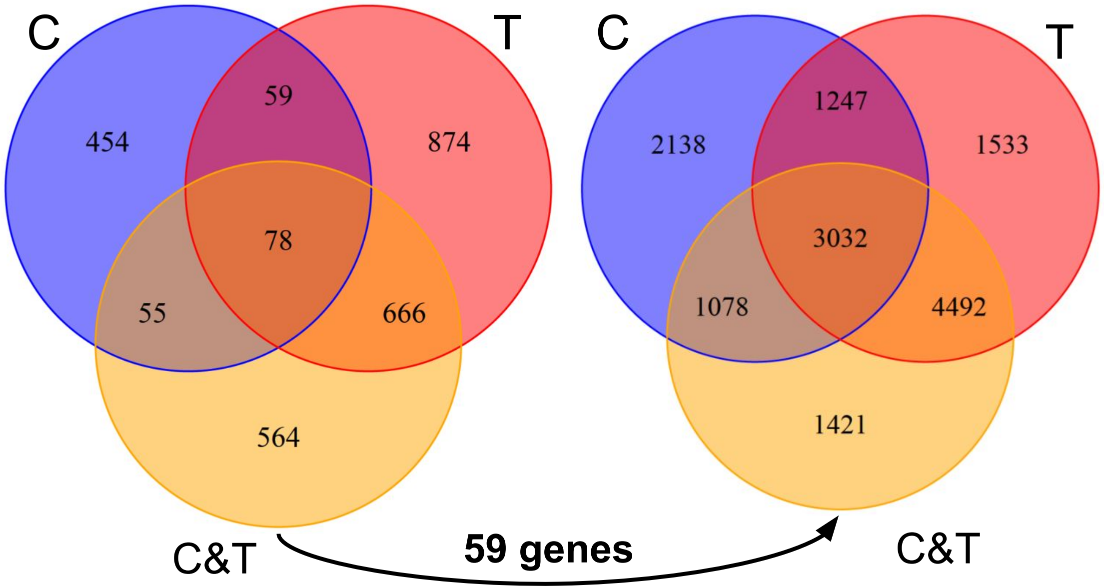


n=3

n=22

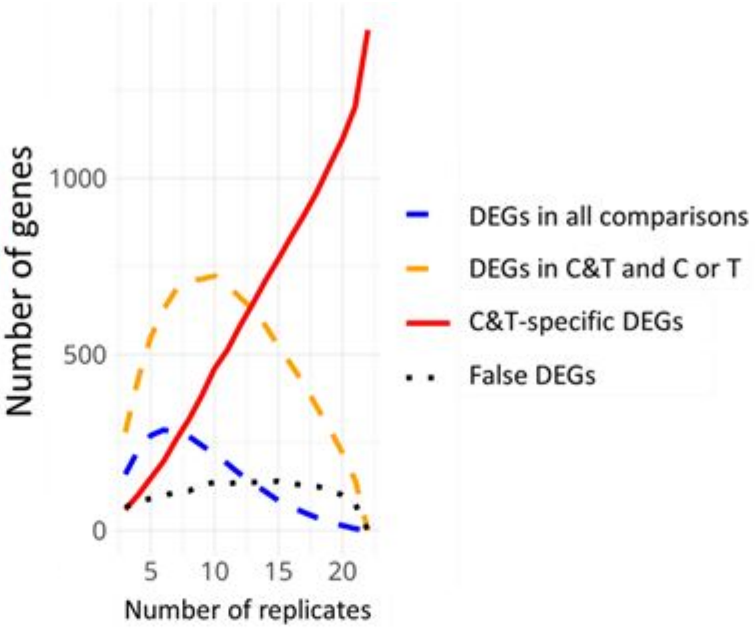
Low replication inflates false specificity

H_0: There is no difference



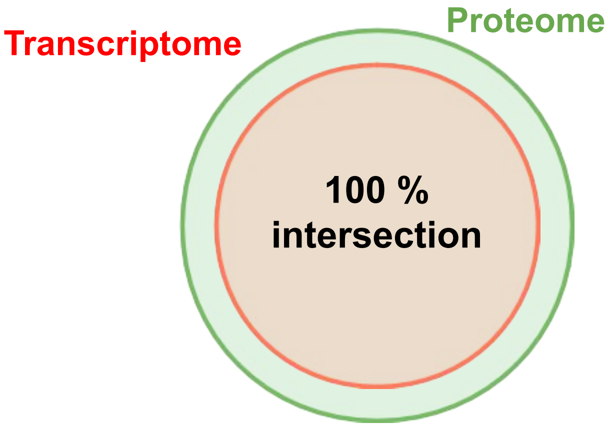
n=3

n=22



Integration analysis

Assumption: Transcript deregulation ⇒ Protein deregulation

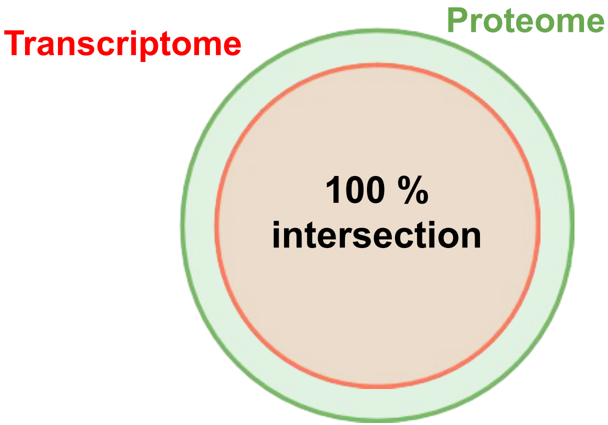


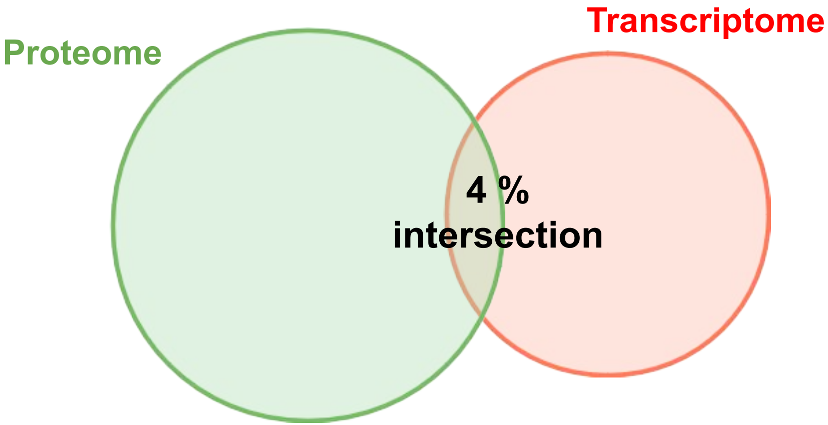
Statistical power = 100%

➡ Consistency between variations

Integration analysis

Assumption: Transcript deregulation ⇒ Protein deregulation





Statistical power = 100%

- Consistency between variations

Statistical power = 20%

- No consistency between variations
- Wrong biological conclusion

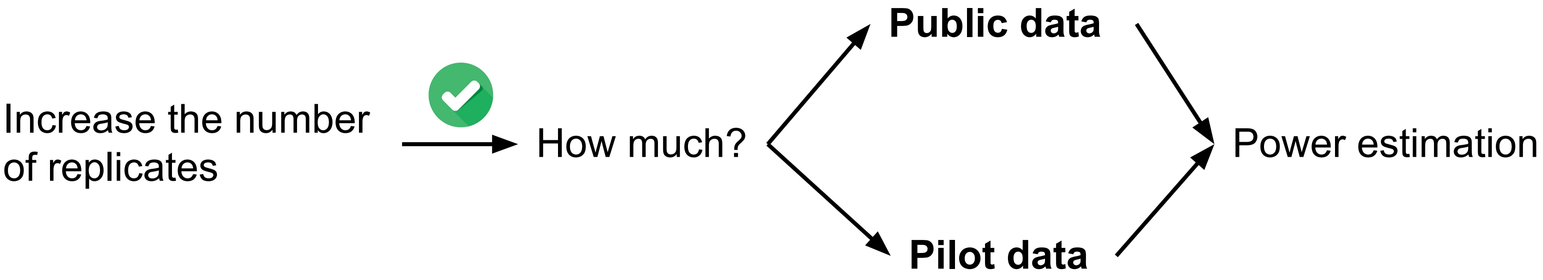
Conclusion

Poor **statistical power** and poor **reproducibility**

Enrichment or LFC **threshold** does **not compensate** for the lack of replicates

Limitations → **Interpreting** our data

Guidelines



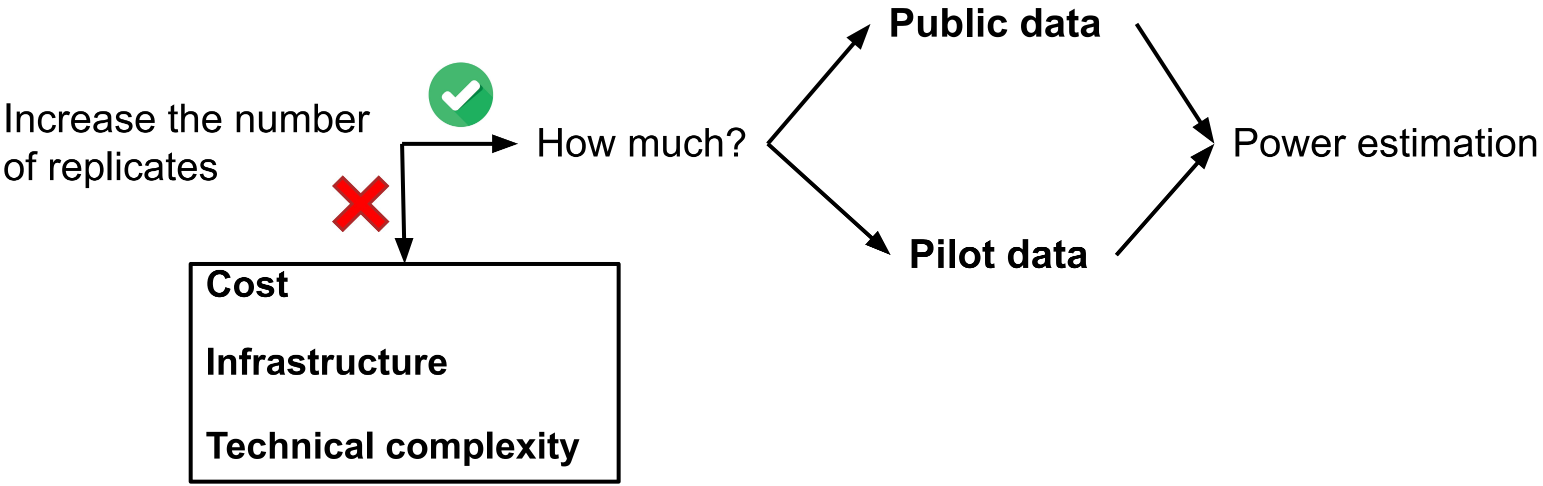
Increase the number of replicates ✓ → How much?

Public data

Pilot data

Power estimation

Guidelines



Increase the number of replicates

✓ How much?

Public data

Pilot data

Power estimation

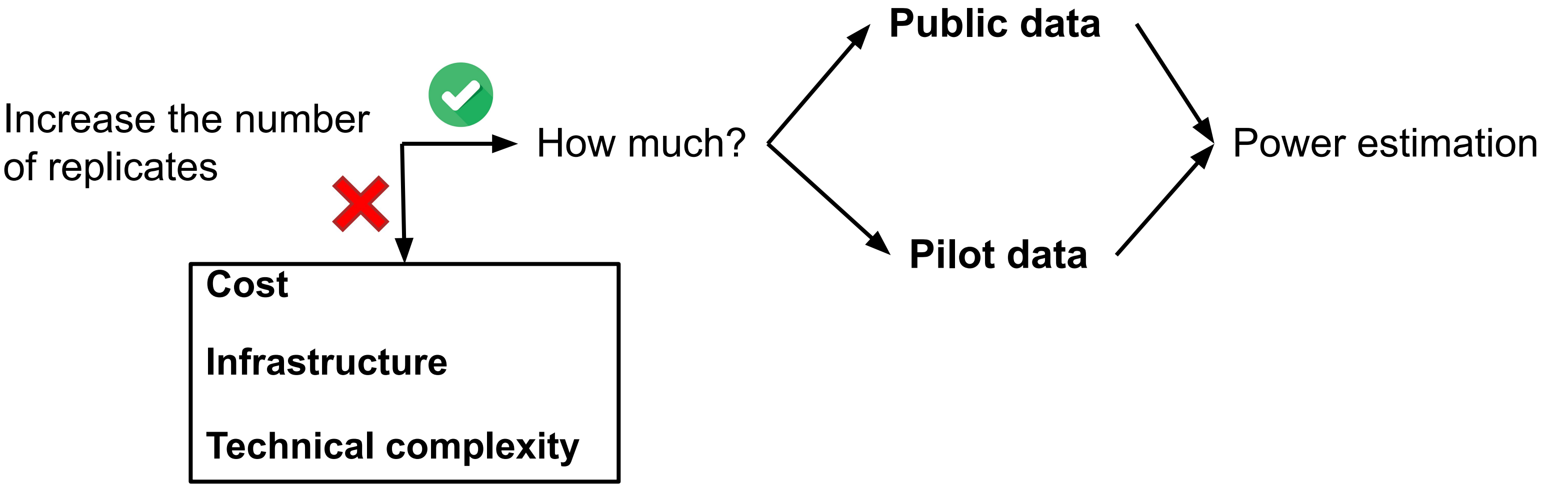
✗

Cost

Infrastructure

Technical complexity

Guidelines



Increase the number of replicates

✓ → How much?

- **Public data** → Power estimation
- **Pilot data** → Power estimation

✗ →

Cost

Infrastructure

Technical complexity

Interaction term → Full factorial design

- **2^2 = 4 conditions**
- **2^5 = 32 conditions**
- **…**

≠ Fractional factorial designs

Identify specific differences → Composite hypothesis testing

UNIVERSITÉ ÉVRY
PARIS-SACLAY

université
PARIS-SACLAY

Université
Paris Cité

ijpb

AgroParisTech

cnrs

MIA
PARIS-SACLAY

LaMME
Laboratoire de
Mathématiques
et Modélisation
d'Évry

ensIIE

INRAE

IPS2
Institute of Plant Sciences
Paris - Saclay

Virginie Noël

Axel de Zelicourt

Michael Hodges

Elodie Gilbault

Olivier Loudet

José Caius

Alexandra Launay-Avon

Stéphanie Pateyron

Christine Paysant Le Roux

Marie-Laure Martin

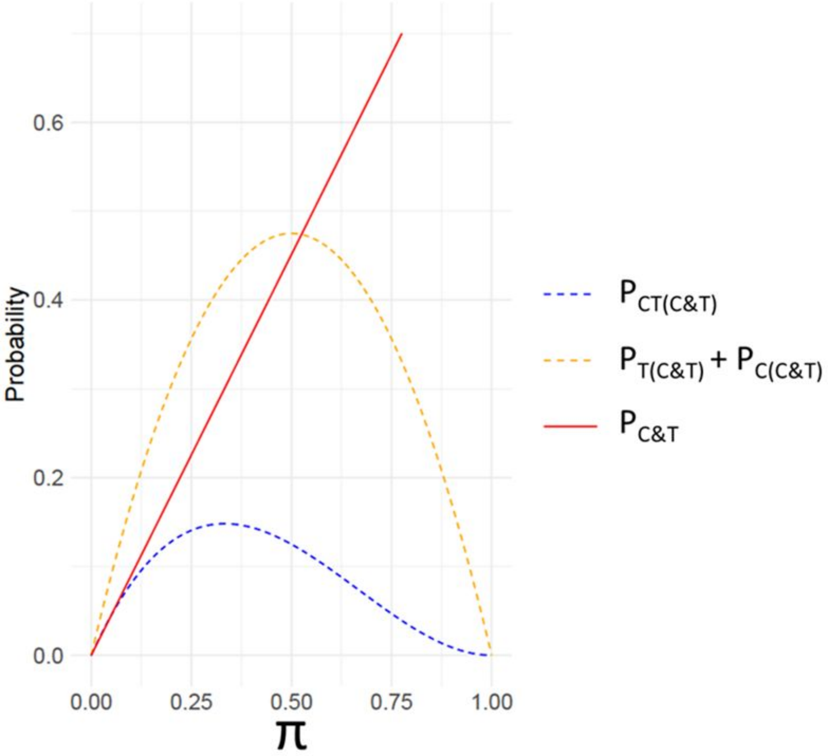
Benoît Castandet

Etienne Delannoy

Guillem Rigaill

Annex

Probabilistic model



power= π false detection rate= α

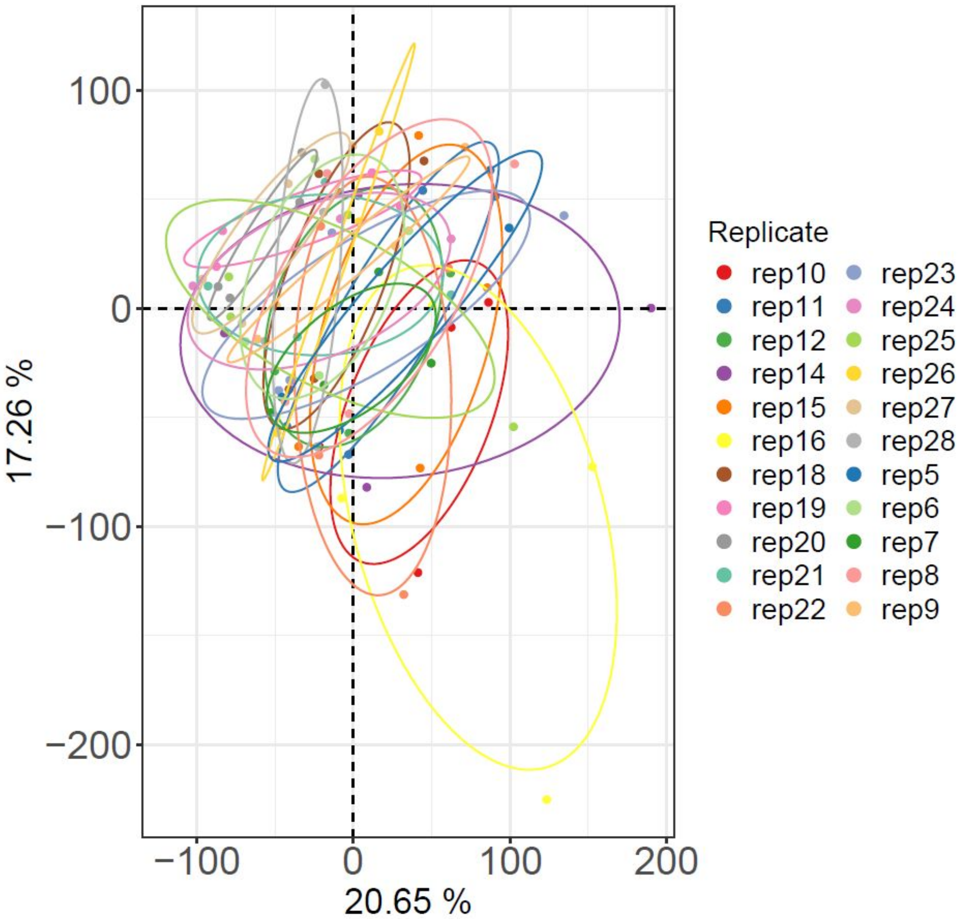
Truly specific to C&T:

C&T (π)

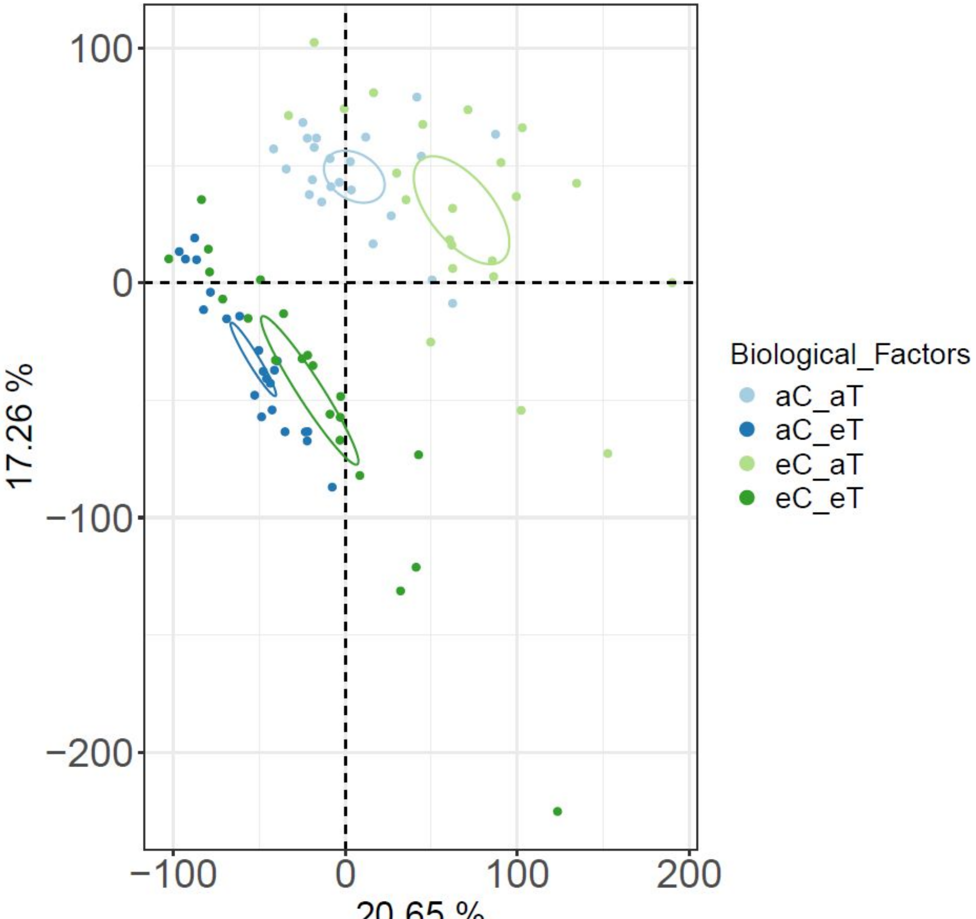
not falsely detecting it in C (1-α) and in T (1-α)

$$P_{(C\&T)}=\pi(1-\alpha)^2=\pi-2\pi\alpha+\pi\alpha^2$$

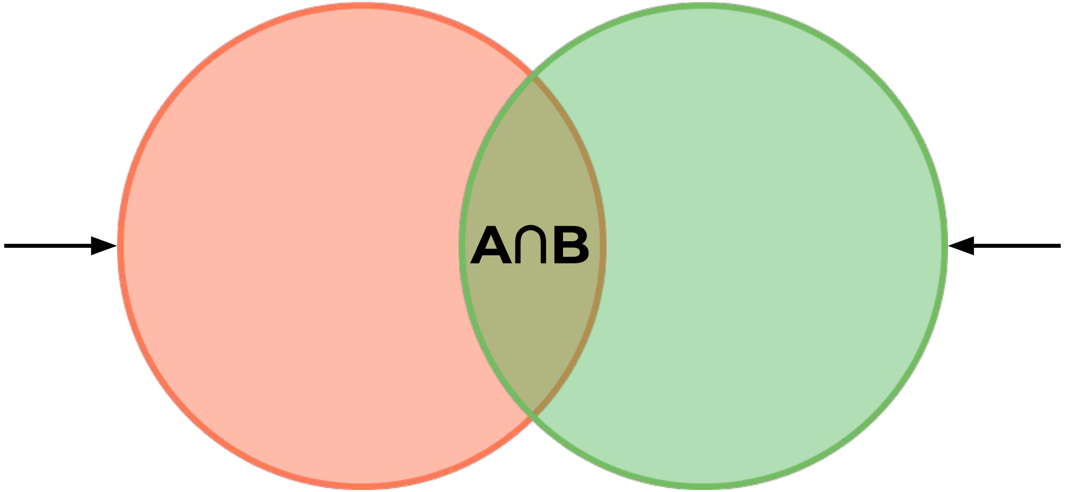
Quality control on normalized counts: Replicates



Quality control on normalized counts: Conditions



Reproducibility: Jaccard index

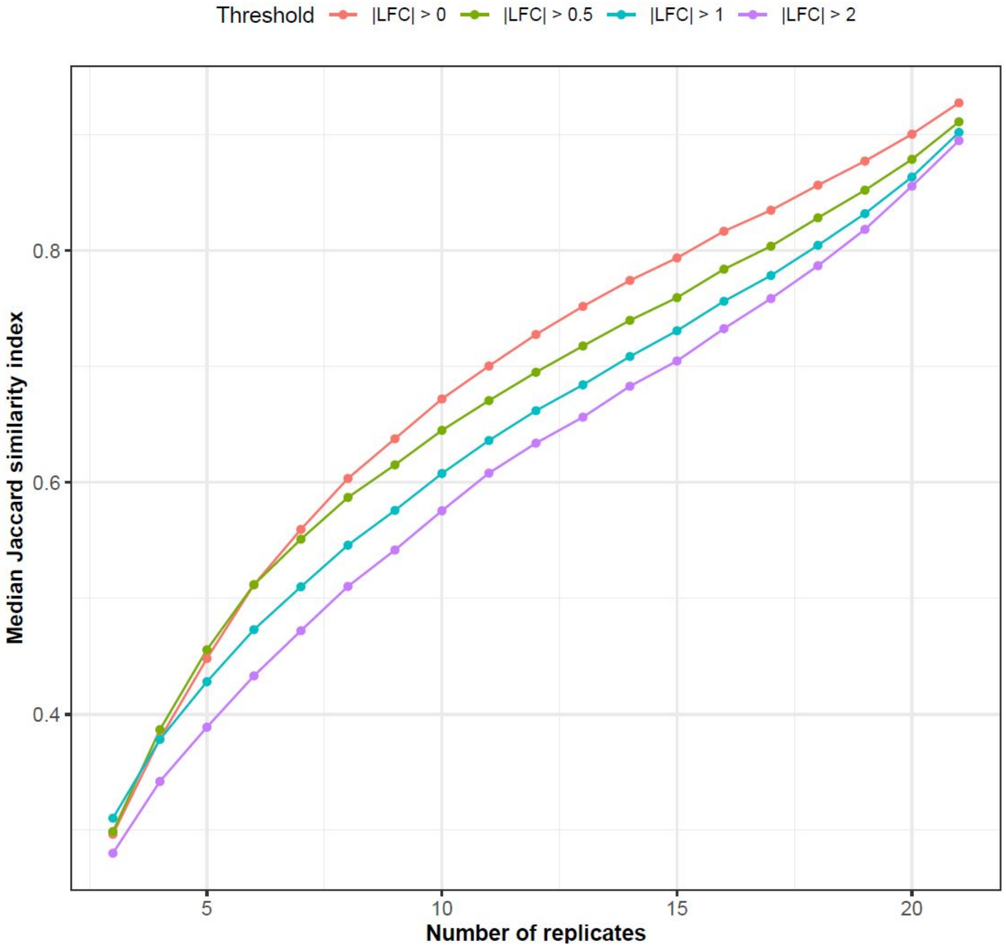


Sub-sample A:
Replicates 1 - 7 - 14

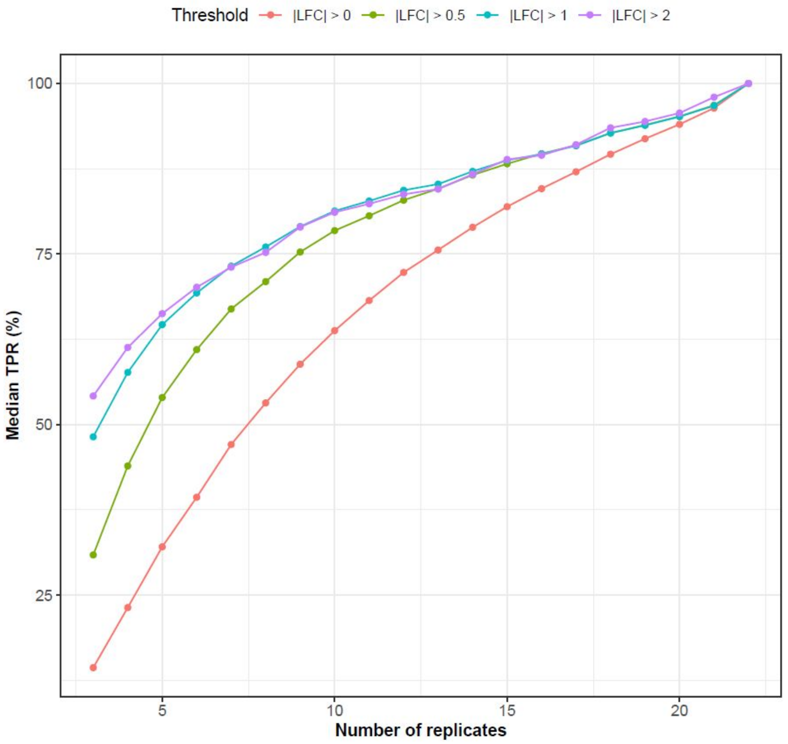
Sub-sample B:
Replicates 8 - 12 - 21

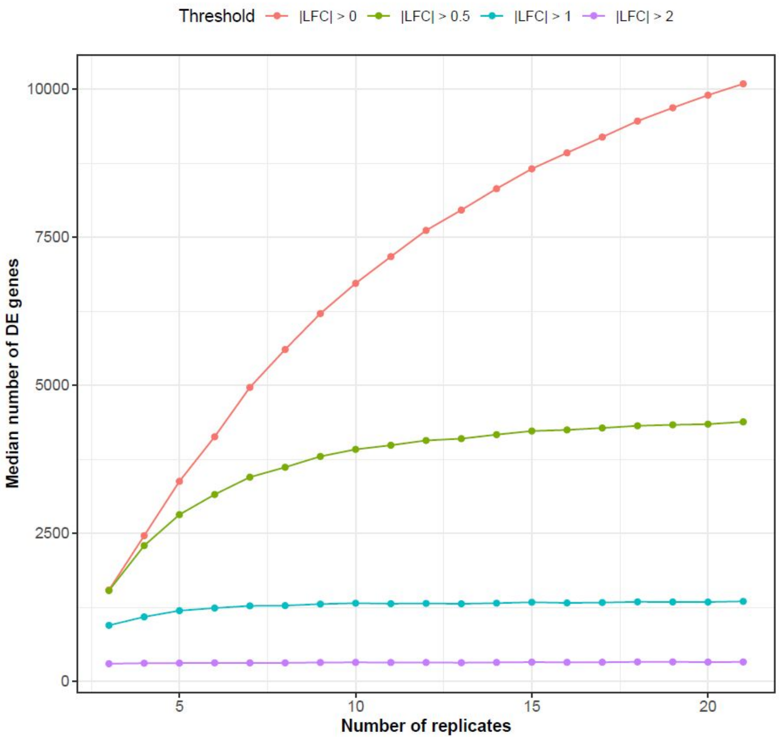
Calculated between pairs of subsamples of the **same size**

Reproducibility: Jaccard index in temperature stress



LFC thresholds in temperature stress: A tempting but misleading idea





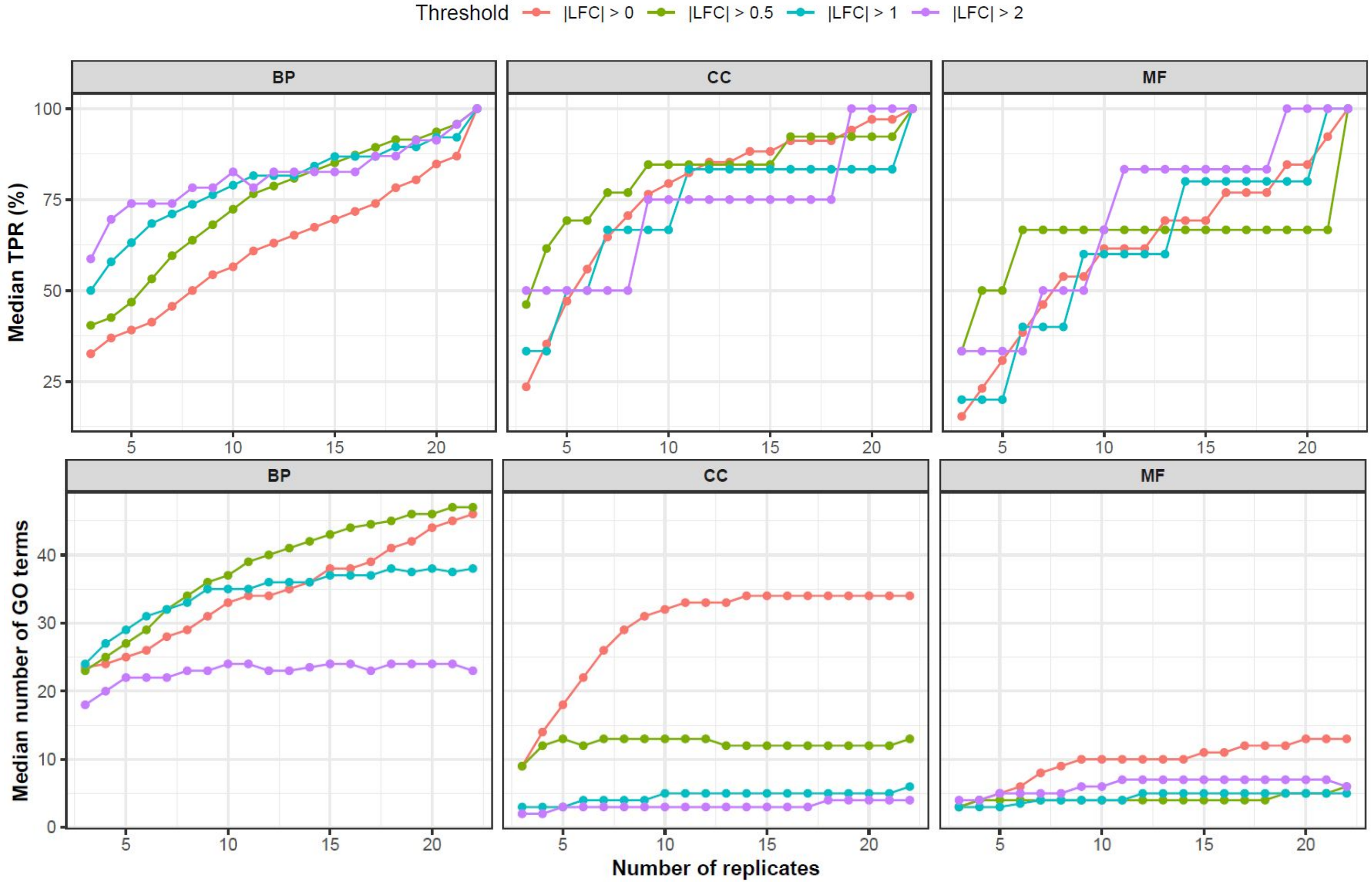
Better power

Gap narrows as the number of replicates increases

Loss of a lot of genes

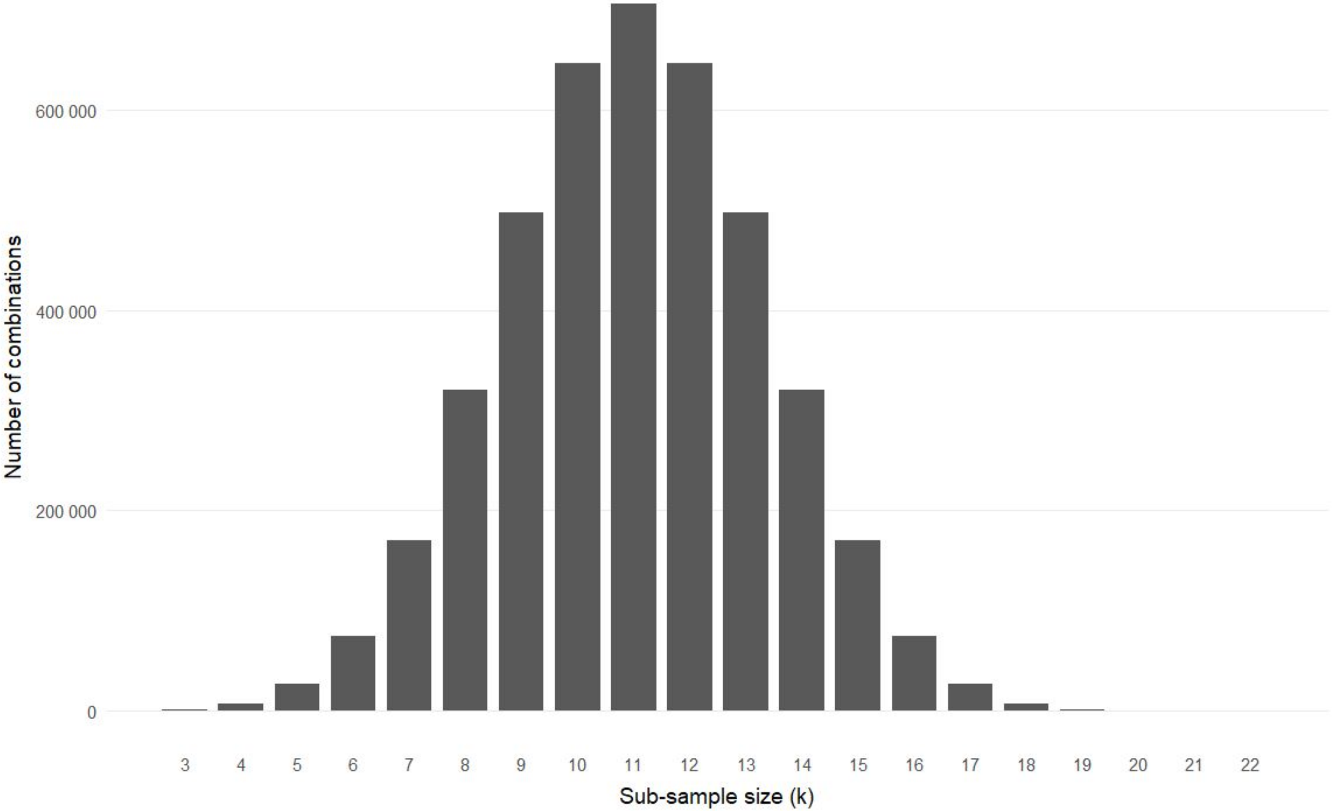
The meaning of taking high LFC → **Biological relevance** ?

Enrichment analysis in temperature stress

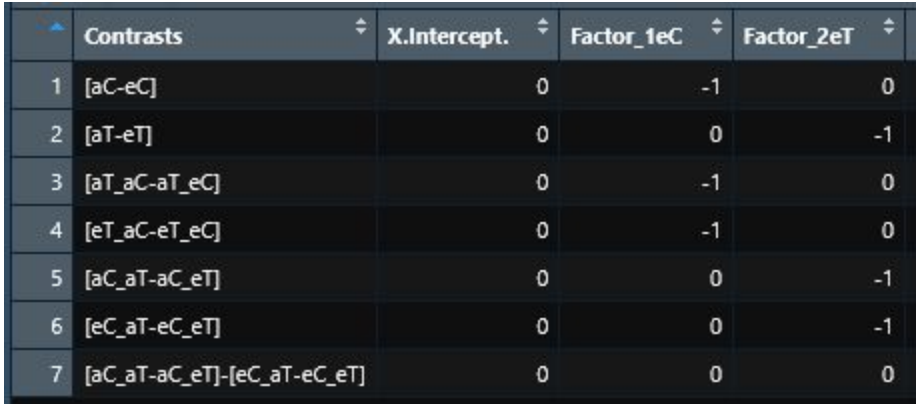


Same logic as before

Number of combinations



Contrast matrix



	Contrasts	X.Intercept.	Factor_1eC	Factor_2eT
1	[aC-eC]	0	-1	0
2	[aT-eT]	0	0	-1
3	[aT_aC-aT_eC]	0	-1	0
4	[eT_aC-eT_eC]	0	-1	0
5	[aC_aT-aC_eT]	0	0	-1
6	[eC_aT-eC_eT]	0	0	-1
7	[aC_aT-aC_eT]-[eC_aT-eC_eT]	0	0	0

[aC_aT-eC_eT]

Design with 4 factors

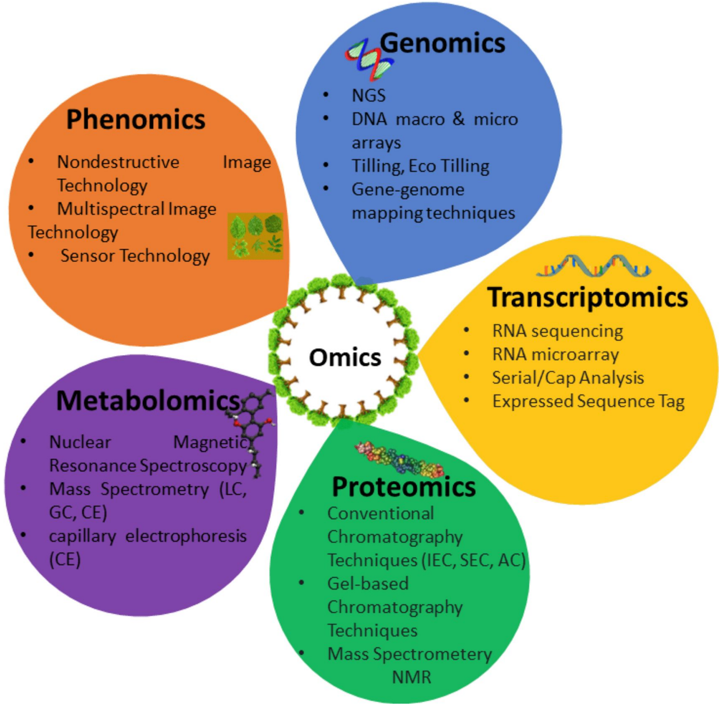
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CO_2	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
Temp	-	-	+	+	-	-	+	+	-	-	+	+	-	-	+	+
Sec	-	-	-	-	+	+	+	+	-	-	-	-	+	+	+	+
UV	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+

Omics?

A set of **high-throughput technologies**

Detects **molecules**

A **systems-level perspective**



Genomics

- NGS
- DNA macro & micro arrays
- Tilling, Eco Tilling
- Gene-genome mapping techniques

Phenomics

- Nondestructive Image Technology
- Multispectral Image Technology
- Sensor Technology

Omics

Transcriptomics

- RNA sequencing
- RNA microarray
- Serial/Cap Analysis
- Expressed Sequence Tag

Metabolomics

- Nuclear Magnetic Resonance Spectroscopy
- Mass Spectrometry (LC, GC, CE)
- capillary electrophoresis (CE)

Proteomics

- Conventional Chromatography Techniques (IEC, SEC, AC)
- Gel-based Chromatography Techniques
- Mass Spectrometery
  NMR

Isewon I, Apata O, Oluwamuyiwa F et al. 2022

Objectives

Effect of the number of replicates
- Statistical power
- Reproducibility

Does it have an effect on **interpretation**?
- LFC threshold
- Enrichment analysis

Does it have an impact on **integration data**?