SCIENCE MEETS LIFE

Design of comparative experiments

ABB lecture series – lecture 1

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

21/02/2019

Veronique Storme

Designing comparative experiments

- Comparative experiments estimate differences in response between treatments
- An experiment has
  - Experimental factors
  - Experimental units
  - A method to assign the experimental factors to the experimental units
  - Responses

Experimental factors

- Usually a classification factor
    - Genotype or variety
    - Treatment
    - Temperature
    - Dosage
    - Time
    - …

Experimental units

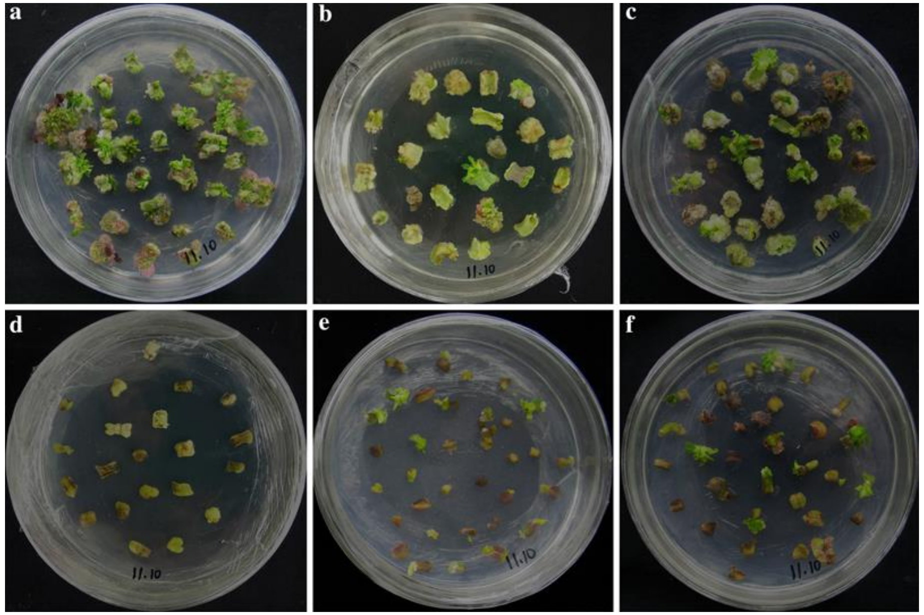
That object to which a treatment or condition is **independently** applied.

- An experimental unit should be as homogeneous as possible
  - Eg. Explants cultured in test tubes containing 15 ml of regeneration medium. There are 2 types of medium (=experimental factor). There are 24 tubes per treatment. All tubes are completely randomly placed in the rack. The number of shoots per explant was recorded 12 weeks after culture initiation. The experimental unit is the test tube.

Experimental units

Subsampling (or pseudoreplication)

- Cotyledons of 1 genotype are arranged in 3 plates with medium A and in 3 plates with medium B. What is the experimental unit now?
- Measurements on different plant parts within a plant
- The measurements are not independent of each other. The statistical analysis must **account for the correlation** between measurements done on the same plate



VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

UNIVERSITEIT GENT

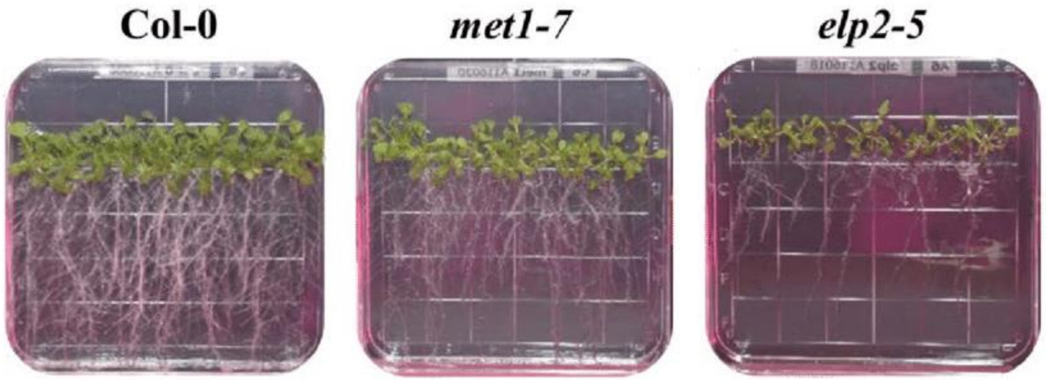
SCIENCE MEETS LIFE

Randomization

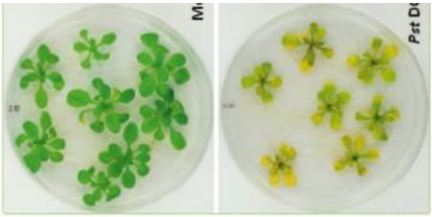
- Randomization is the way of assigning subjects to treatment groups. All experimental units should have an equal chance of being assigned to any of the treatment groups.
- Randomization protects against **confounding**
  - Confounding occurs when the effect of one factor cannot be distinguished from the effect of another factor
- Randomization of measurements: measuring first all treatment 1 units, and the all treatment 2 units will introduce serial dependency or autocorrelation in the data

Confounding examples

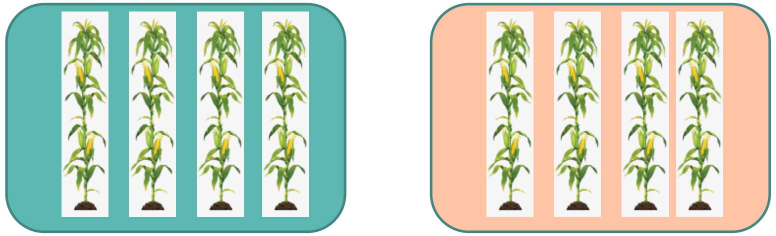
- Comparing different genotypes: all plates should contain all genotypes!



- Confounding of treatment with experimental units
  - Treatment-plate confounding



  - Treatment-block confounding



Confounding examples

- 20 B73 maize plants were grown under normal water conditions and 20 under drought conditions according to a CRD design. Metabolomics data were collected but first all watered plants were analysed and then all drought plants.

Performing a randomization

- Physical randomisation
  - Coin tosses
  - Rolls of a die
  - Tickets in a hat
- Numerical randomisation
  - Using random number generators with computer software
- Example:
  - assigning 16 plants to two treatments
  - Number the plants
  - Sample from 1 to 16 without replacement
  - 16 9 3 14 11 6 4 15 5 2 1 7 12 8 10 13
  - Assign treatment A to plants 16 9 3 14 11 6 4 15
  - Assign treatment B to plants 5 2 1 7 12 8 10 13

Types of outcome

- Continuous outcome
  - Eg. area, height, weight, length, biomass
- Count data
  - Eg. number of chloroplast cells, nr of seeds, nr of ovules, nr of leaves, nr of lateral roots
  - Count rate data: lateral root density
- Binary data:
  - Eg. germinated or not, dead or not, peptide presence
- Ordinal data:
  - Eg. ovule stages, disease stage, bud formation stage

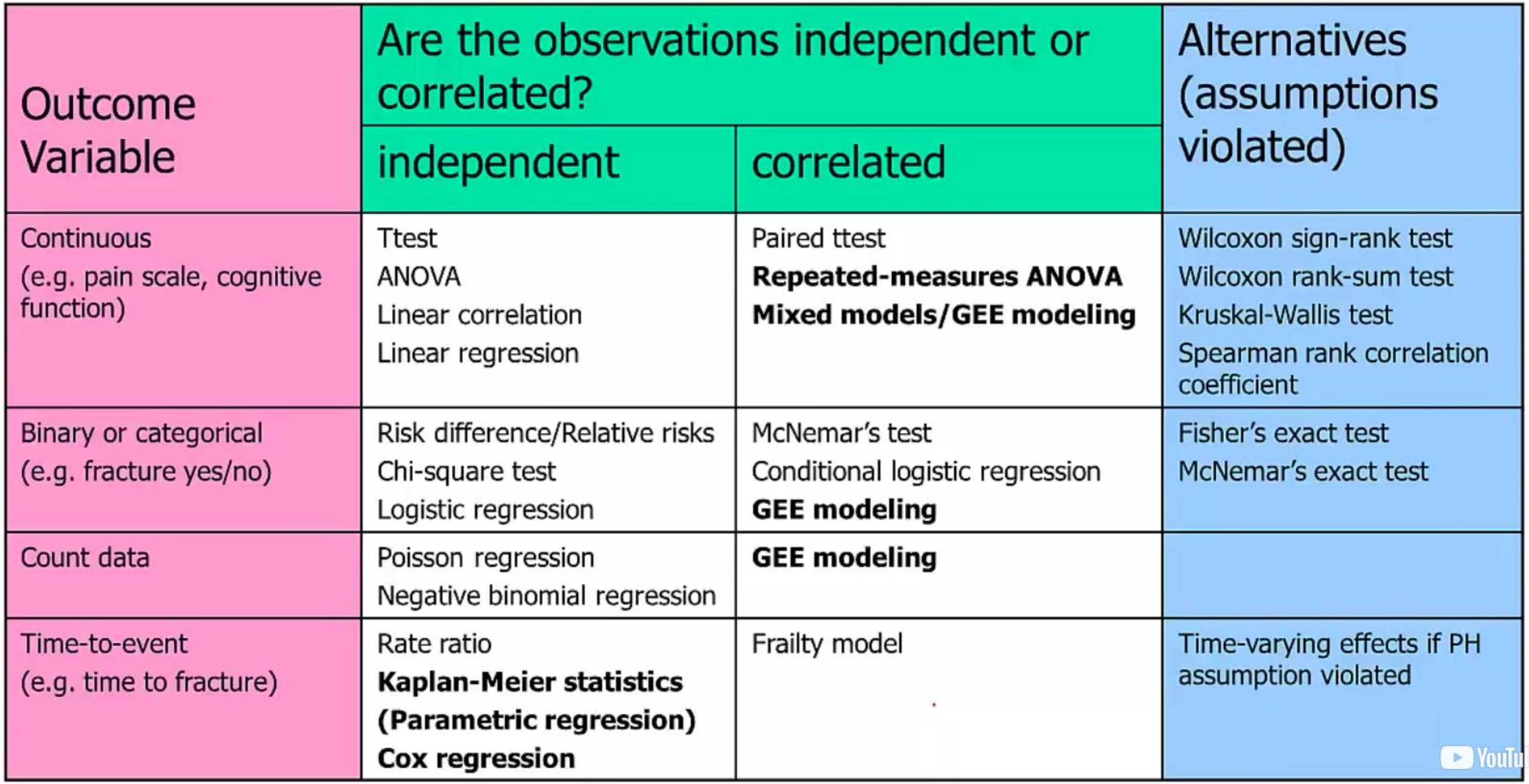
Types of outcome

- Repeated measurements
  - Multiple measurements on the same experimental unit
  - Examples:
    - Leaf area measured on 2 leaves from the same plant
    - Root length measured on seedlings grown on the same plate
    - Yield measured on maize plants from the same block

    Hierarchical or clustered data

    - Leaf area measured over time on the same plant

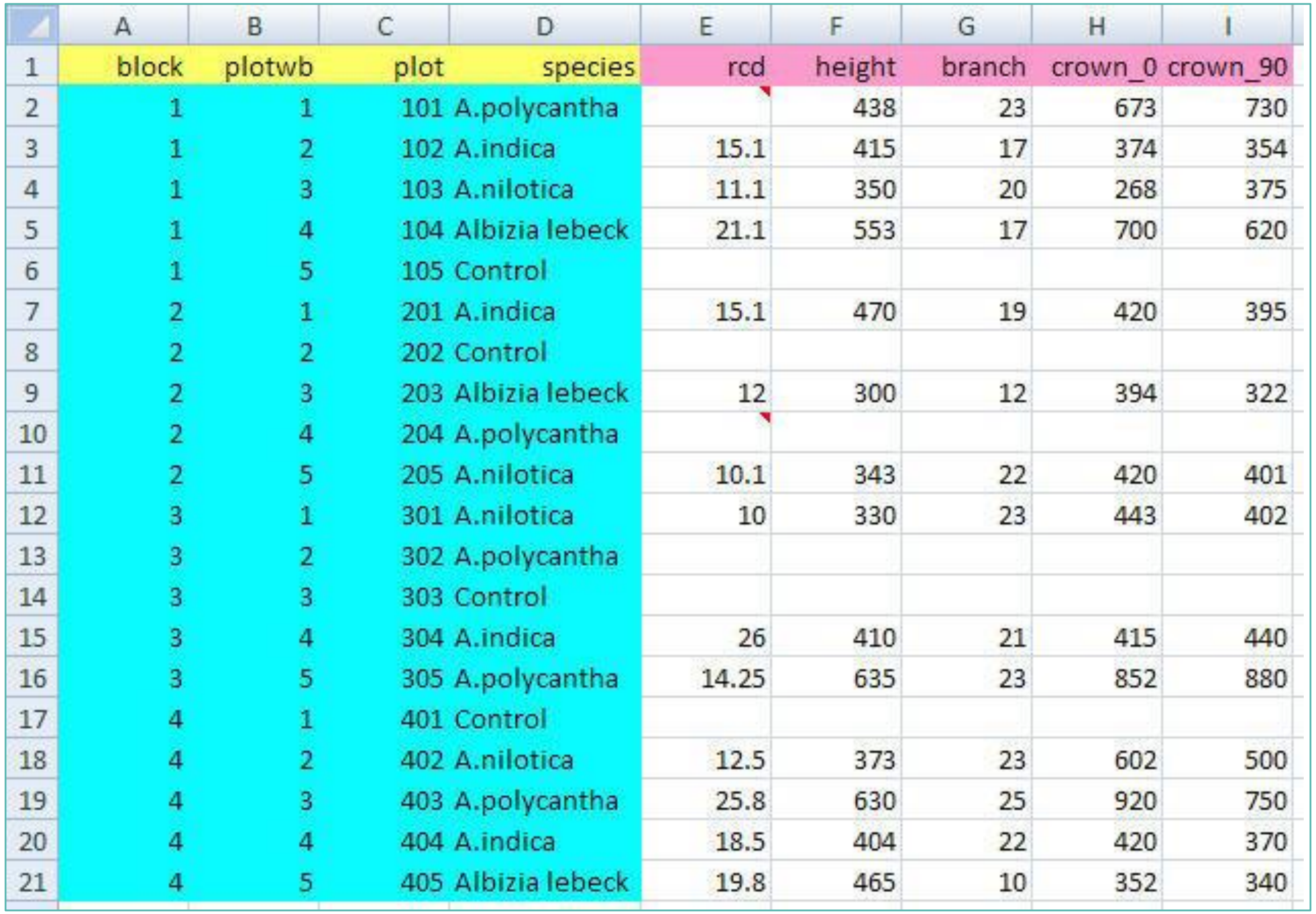
    Longitudinal data
  - Analysed with mixed models;
    - Continuous outcome: linear mixed model
    - Counts, binary, ordinal: generalized mixed models



Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest **Repeated-measures ANOVA** **Mixed models/GEE modeling**	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression **GEE modeling**	Fisher's exact test McNemar's exact test
Count data	Poisson regression Negative binomial regression	**GEE modeling**	
Time-to-event (e.g. time to fracture)	Rate ratio **Kaplan-Meier statistics** **(Parametric regression)** **Cox regression**	Frailty model	Time-varying effects if PH assumption violated

Arrangement of the data

- All data on the same subject should be on one row
- More examples on the psb wiki
- Follow name conventions for the headers
- Use . as decimal separator
- Provide the meta data



	A	B	C	D	E	F	G	H	I
1	block	plotwb	plot	species	rcd	height	branch	crown_0	crown_90
2	1	1	101	A.polycantha		438	23	673	730
3	1	2	102	A.indica	15.1	415	17	374	354
4	1	3	103	A.nilotica	11.1	350	20	268	375
5	1	4	104	Albizia lebeck	21.1	553	17	700	620
6	1	5	105	Control					
7	2	1	201	A.indica	15.1	470	19	420	395
8	2	2	202	Control					
9	2	3	203	Albizia lebeck	12	300	12	394	322
10	2	4	204	A.polycantha					
11	2	5	205	A.nilotica	10.1	343	22	420	401
12	3	1	301	A.nilotica	10	330	23	443	402
13	3	2	302	A.polycantha					
14	3	3	303	Control					
15	3	4	304	A.indica	26	410	21	415	440
16	3	5	305	A.polycantha	14.25	635	23	852	880
17	4	1	401	Control					
18	4	2	402	A.nilotica	12.5	373	23	602	500
19	4	3	403	A.polycantha	25.8	630	25	920	750
20	4	4	404	A.indica	18.5	404	22	420	370
21	4	5	405	Albizia lebeck	19.8	465	10	352	340

A good experimental design must…

- avoid systematic error (=bias) -> randomize
- be precise. Precision depends on
  - the size of the random errors on the responses
  - the number of experimental units
  - the design
- allow estimation of the random (=experimental) error
  - *i.e.* the variability in the outcome among **identically** and **independently** treated experimental units
- have broad validity
  - The experimental units should reflect the population about which we wish to draw inference.
- be sufficiently powerful to detect treatment effects of biological significance -> adequate number of replications

A wise experimenter will consider the analysis when planning an experiment

Experimental error

- describes the variability in the outcome among **identically** and **independently** treated experimental units
- Minimize experimental error by
  - Selecting experimental units which are as homogeneous as possible
  - Taking measurements in as uniform as possible. Be aware of how the following can introduce variability:
    - Person-to-person differences in how measurements are taken
    - Within person differences (how tired were you when collecting the measurements?)
    - Run-to-run differences in how measurements are recorded
    - Differences in equipment: pipettes, microtiter plates,…
    - Differences in supplies (soils, growth media,…)

Experimental error

- Minimize experimental error by
  - Controlling the environment to be as uniform as possible (temperature, humidity, light, …)
  - Taking the same seed stock
  - Harvesting at the same time
  - Studies that take time to complete should keep track of the days/weeks/months on which measurements are collected (temporal, seasonal variability)
  - Studies involving repeated measurements on each experimental unit need to keep track of which measurements came from which unit.

Replication

- *i.e.* assigning several **experimental units** to each treatment group
- Replication allows the researcher to demonstrate **reproducibility**
- Replication guards against an experiment failing (plants died, seed did not germinate,…)
- Replication allows to measure experimental error
- Replication improves the precision of the treatment effect estimate
- How many replications should we use?
    - Perform sample size study

Power and sample size

- Sample size analysis
  - Prior knowledge necessary on treatment effects, error variance, type I error rate and power
- Power analysis
  - Prior knowledge necessary on treatment effects, error variance, type I error rate and sample sizes
- For complex experiments no closed formulas available, simulation studies are necessary.

Generalizability

- The conclusions we draw from an experiment are directly applicable to the experimental units used in that experiment
  - If the units were randomly selected from some population of units, then the conclusions can be applied to that population.
  - If the conclusions are applied to any other group of potential experimental units, then we are extrapolating. This may not be valid.

Climate chamber experiments

- Weiss chamber
  - Thermostatic chamber

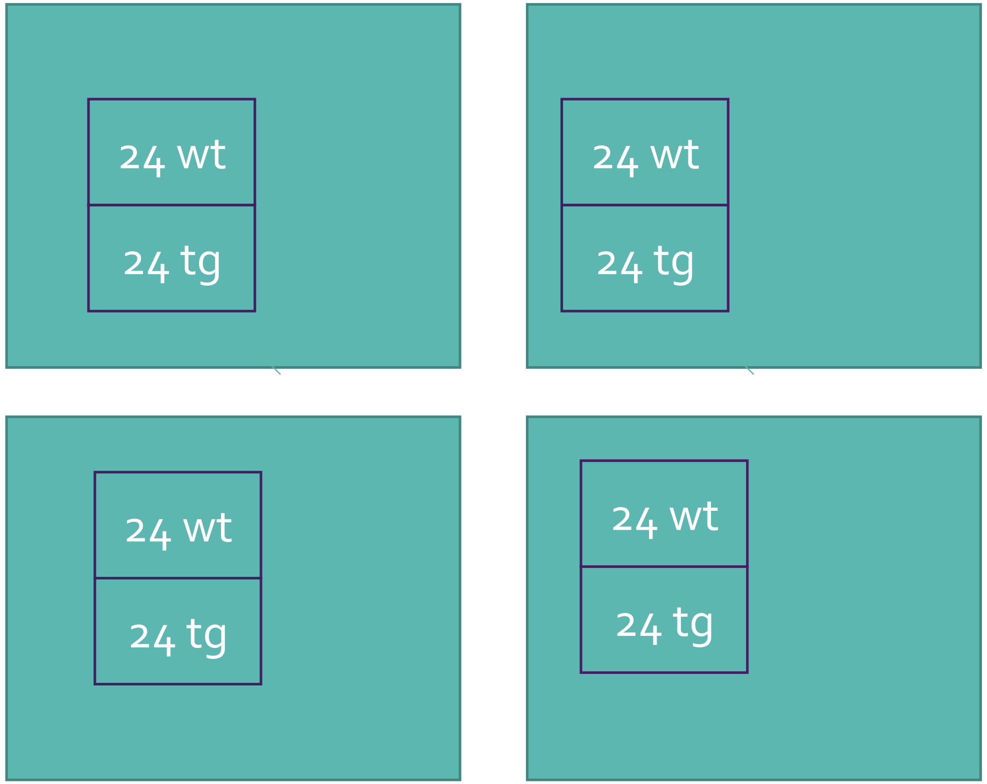
- Lovibond
  - Thermostatic chamber

Climate chamber experiments

- The treatment condition (temperature or light intensity) is randomly applied to the growth chamber, thus **the growth chamber becomes the experimental unit**.
- Pots, petri-dishes, trays, plants included within a growth chamber are considered subsamples and should not be treated as true replicates.
- To provide **valid replications** when the number of growth chambers are limited, the experiments should be repeated by using the same growth chamber with different randomly assigned treatment levels.

Climate chamber example

- 4 growth chambers
- 1tray/growth chamber
- 2 genotypes/tray
  - Col: 24 plants
  - Transgene: 24 plants
- 16 plants/gt/tray selected
- 1 ovary/plant
- Count the ovules/ovary and decide on stage (5 stages)
  - problem of no fixed sample size
  - Problem of complete separation
- Unbalanced data
- Clustered data
- Proportional odds model



Growth chamber experiments

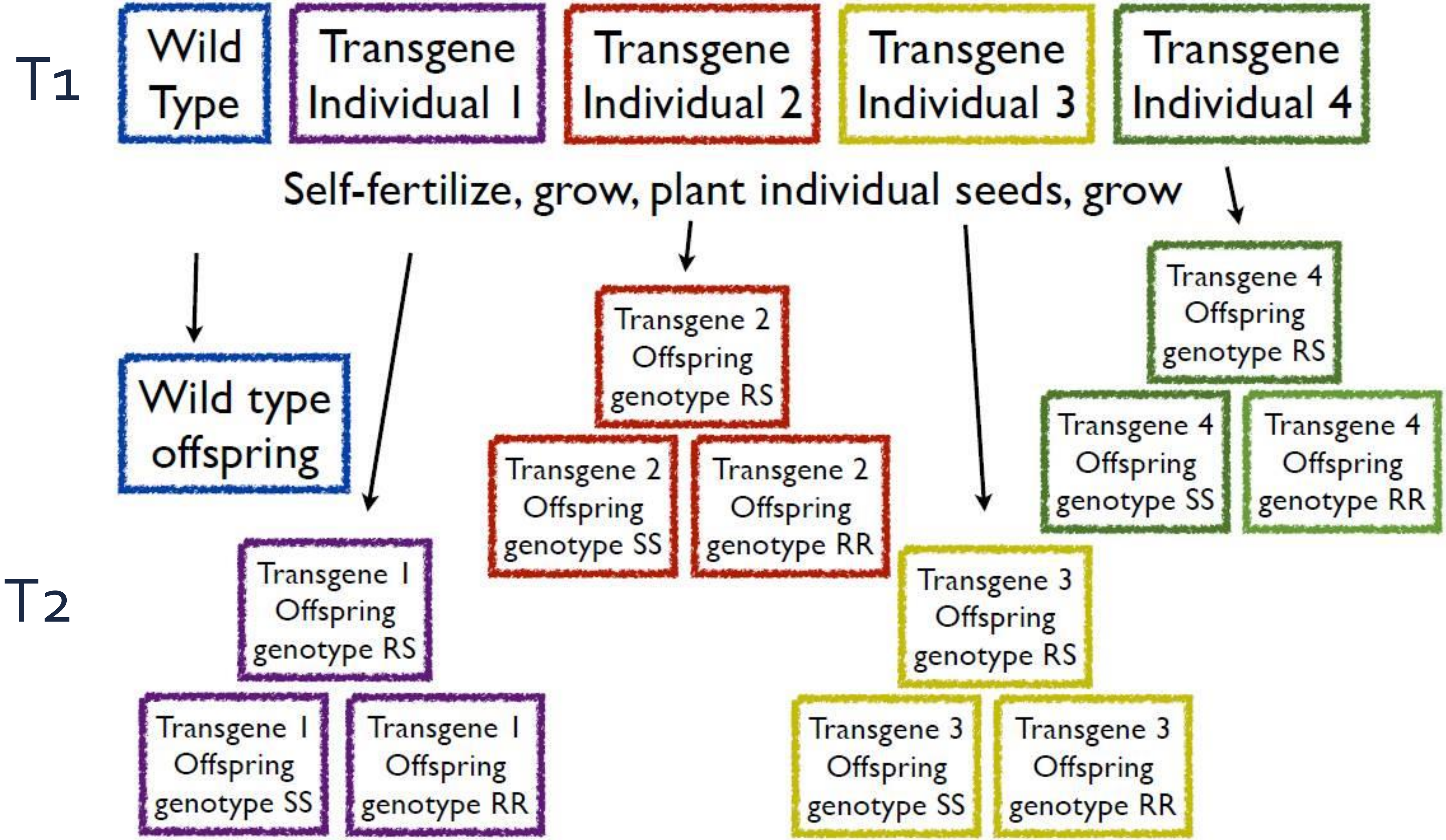
- Commonly used experimental units:
  - Individual pots
  - Trays
  - Plates
- Recommended to measure also air temperature, humidity, moisture variation.
- Use suitable blocking and possibly additional covariates
  - Shelf to shelf variation
  - position on the shelf (under the light, close to the wall, to the door,…)

Greenhouse experiments

- Commonly used experimental units:
  - Individual pots
  - Trays
- Recommended to measure also air temperature, humidity, solar radiation, moisture variation.
- Use suitable blocking and possibly additional covariates.
  - When assigning blocks take into account known directional variation (movement of the sun)
  - ***Blocking is preferred over relocating the plants!***
    - Brien et al. Plant methods 2013

Arabidopsis example

- A gene from a related plant is introduced into the genomes of 4 separate Arabidopsis Col-0 plants.
- Each of these plants is the progenitor of a transgenic line.
- A wild-type Col-0 plant is included in the experiment.
- These 5 plants represent the T1 generation.
- Each T1 plant is grown, self-fertilized and seed is collected.
- 25 seeds from each plant are potted individually, grown and self-fertilized. These plants are the T2 generation. Fruit length is measured from 10 fruit of each T2 plant



T1

Wild Type

Transgene Individual 1

Transgene Individual 2

Transgene Individual 3

Transgene Individual 4

Self-fertilize, grow, plant individual seeds, grow

T2

Wild type offspring

Transgene 1 Offspring genotype RS

Transgene 1 Offspring genotype SS

Transgene 1 Offspring genotype RR

Transgene 2 Offspring genotype RS

Transgene 2 Offspring genotype SS

Transgene 2 Offspring genotype RR

Transgene 3 Offspring genotype RS

Transgene 3 Offspring genotype SS

Transgene 3 Offspring genotype RR

Transgene 4 Offspring genotype RS

Transgene 4 Offspring genotype SS

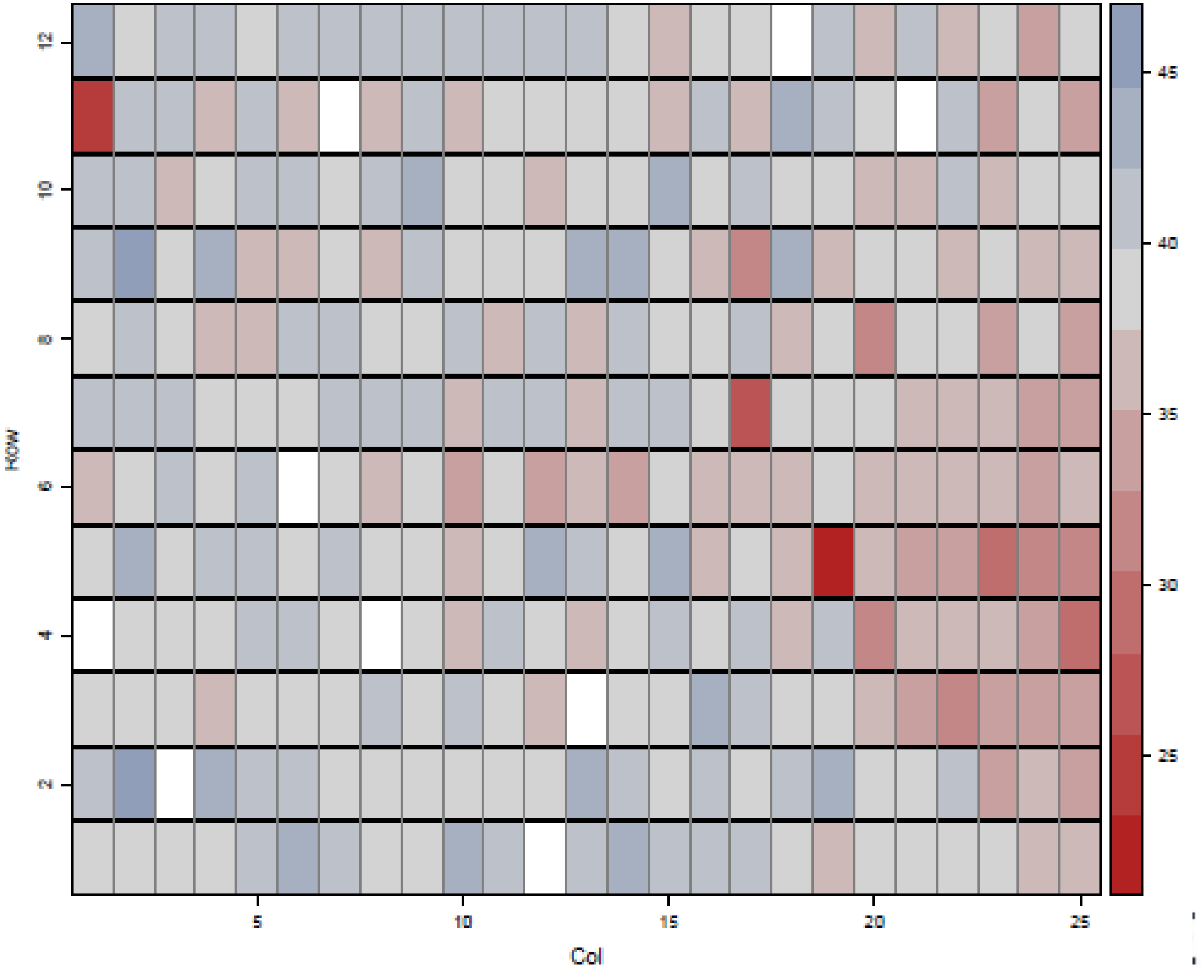
Transgene 4 Offspring genotype RR

Arabidopsis example from the lab

- T2 plants are sown over 5 plates (unknown genotype Aa, aa or AA (characterised under bino).
  - As a result: highly unbalanced over the trays
- 5 lines of T2 plants (1 line is the ctrl line)
- 2 conditions
- Y measured over time (longitudinal data)
- Measure Y on 20 to 30 plants/line*condition

High-throughput phenotyping platforms

- start with a **uniformity trial**
  - All blocks receive the same treatment
  - To investigate heterogeneity on the platform/field
  - Spatially correlated data
    - Spatial analysis is required (possible within the mixed model framework)



Field experiments

- Suitable for conducting
  - Yield trials
  - Genotype screening
  - Compare irrigation, nutrition, pest, disease, weed control,…
- **Multiyear** studies to investigate the interaction between the treatment and the year to remove confounding effects of climatic variation.
- **Multisite** studies in a given year to evaluate the treatment effects across different sites and to investigate the treatment-site interaction to remove confounding effects of soil variation.

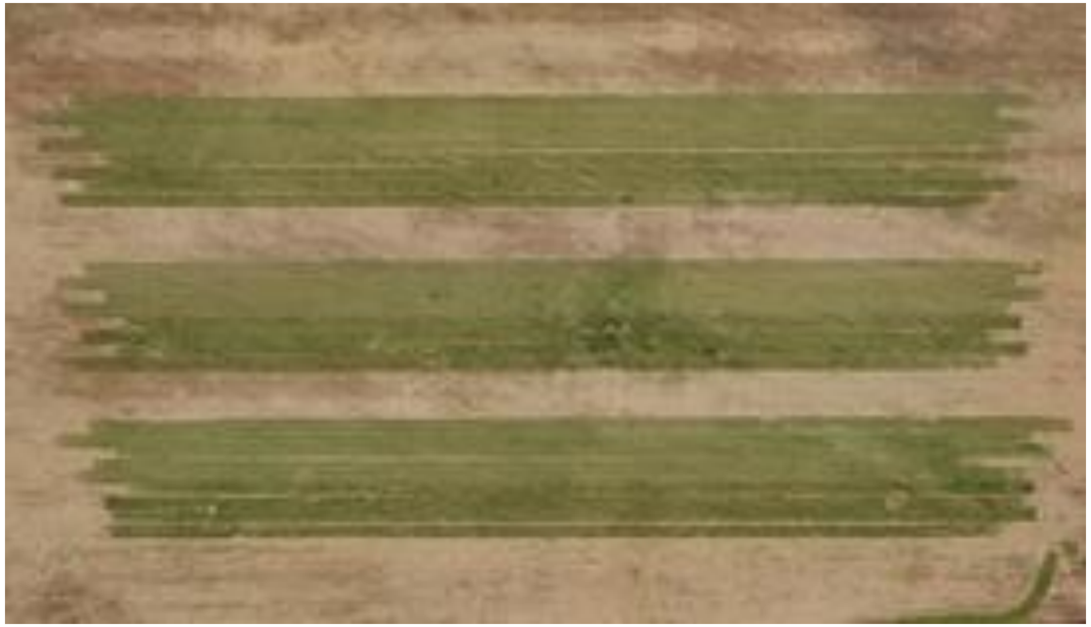
Field experiments

- Highly recommended:
  - Blocking along the field variability gradient
  - Measuring additional continuous covariates

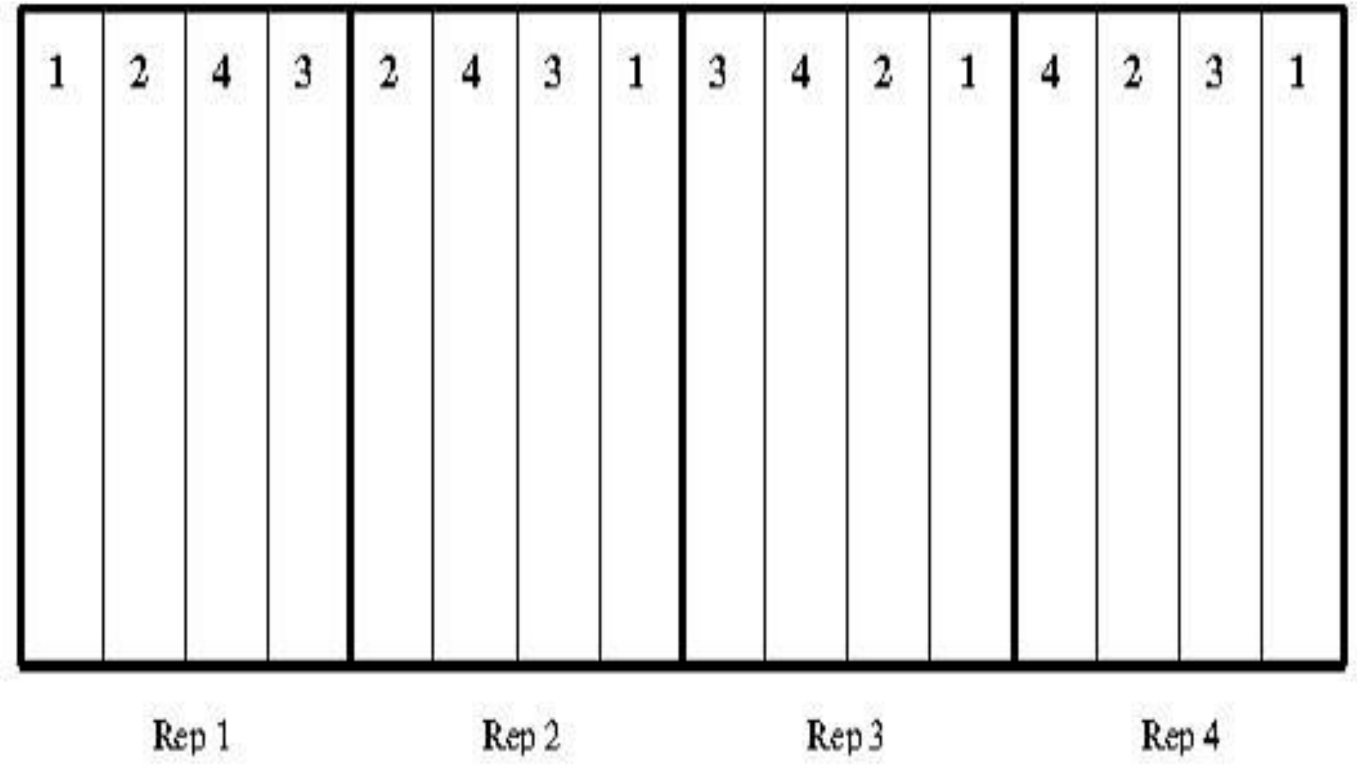
- Common designs
  - Strip-plot design
  - Randomized complete block design (RCB)
  - Split-plot design
  - Factorial design
  - Row-column design

Strip-plot design

- 3 varieties each planted in 8 rows



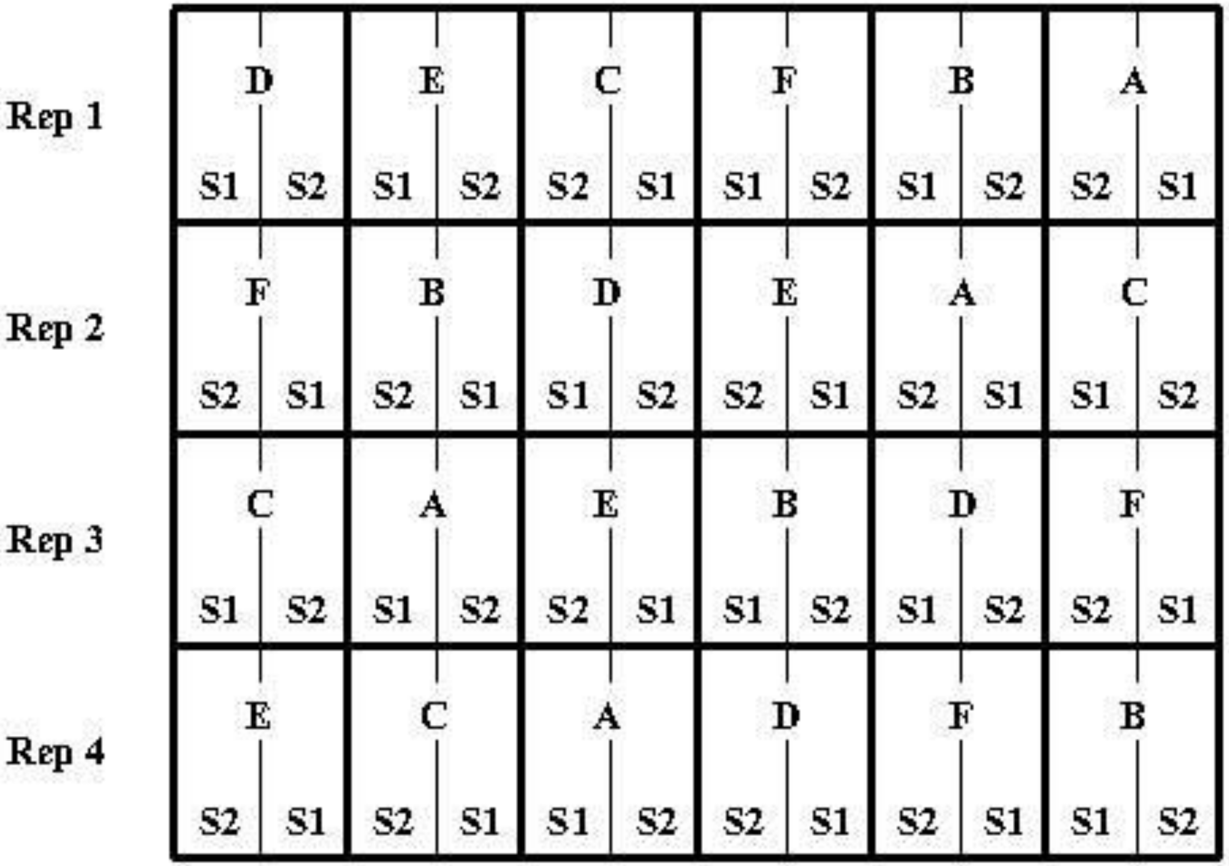
Randomized complete block design



1	2	4	3	2	4	3	1	3	4	2	1	4	2	3	1
Rep 1				Rep 2				Rep 3				Rep 4			

Split-plot design

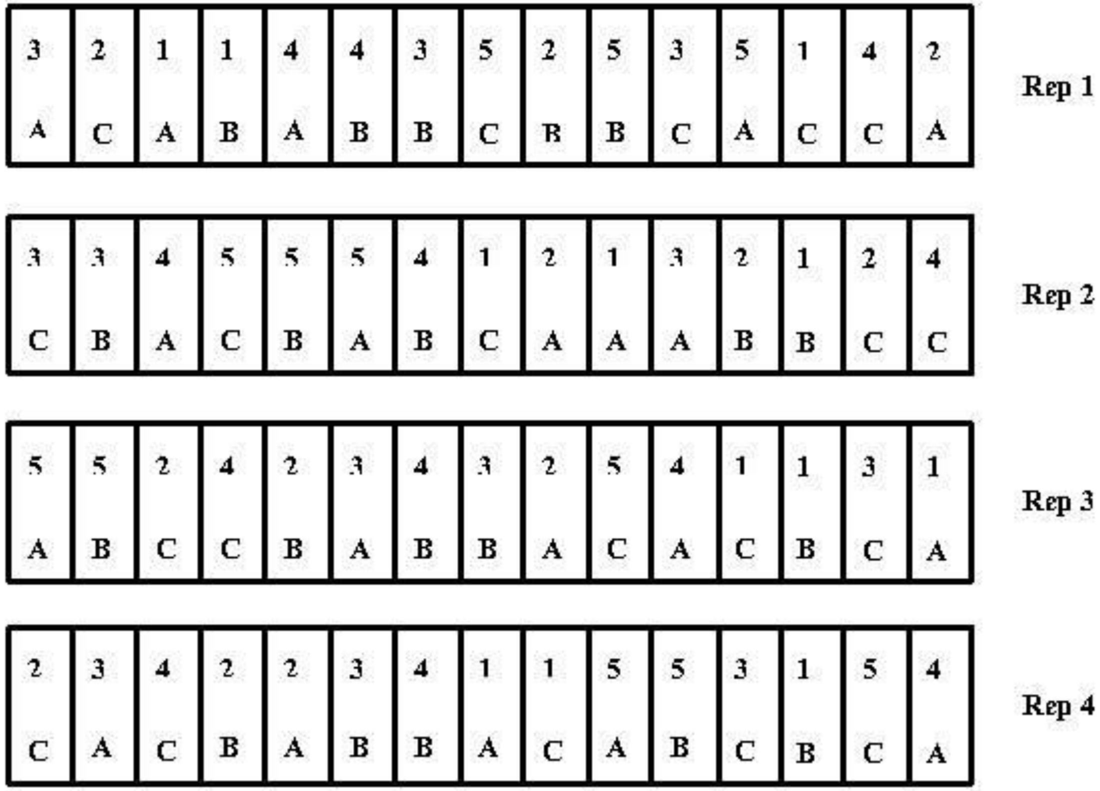
- 2 factors
  - One is hard to change
    - Main plots A--D
  - One is easy to change
    - Sub plots S1 S2



Rep 1	D		E		C		F		B		A	
	S1	S2	S1	S2	S2	S1	S1	S2	S1	S2	S2	S1
Rep 2	F		B		D		E		A		C	
	S2	S1	S2	S1	S1	S2	S2	S1	S2	S1	S1	S2
Rep 3	C		A		E		B		D		F	
	S1	S2	S1	S2	S2	S1	S1	S2	S1	S2	S2	S1
Rep 4	E		C		A		D		F		B	
	S2	S1	S2	S1	S1	S2	S2	S1	S2	S1	S1	S2

Factorial design

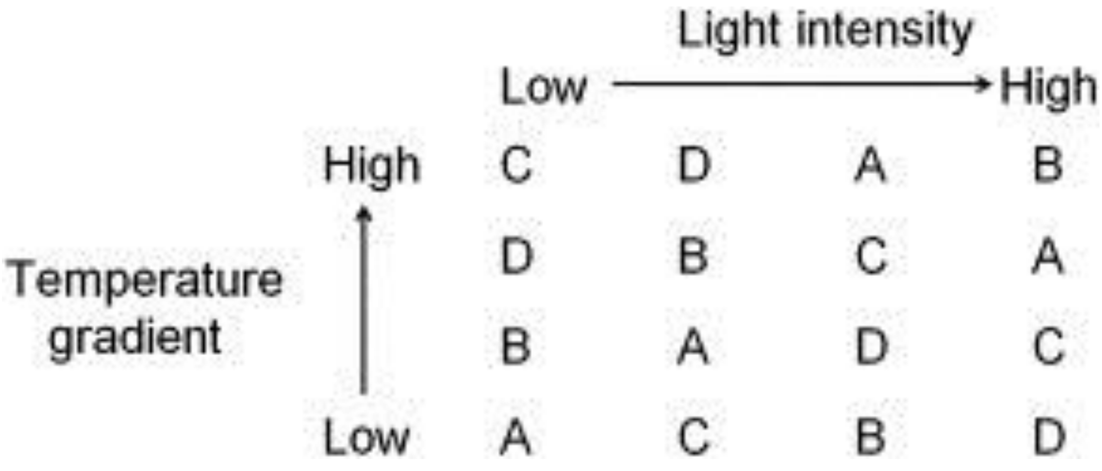
- Eg 5 canola varieties and 3 tillage systems



3	2	1	1	4	4	3	5	2	5	3	5	1	4	2	Rep 1
A	C	A	B	A	B	B	C	B	B	C	A	C	C	A	
3	3	4	5	5	5	4	1	2	1	3	2	1	2	4	Rep 2
C	B	A	C	B	A	B	C	A	A	A	B	B	C	C	
5	5	2	4	2	3	4	3	2	5	4	1	1	3	1	Rep 3
A	B	C	C	B	A	B	B	A	C	A	C	B	C	A	
2	3	4	2	2	3	4	1	1	5	5	3	1	5	4	Rep 4
C	A	C	B	A	B	B	A	C	A	B	C	B	C	A	

Row-column design

- Latin square
  - Complete row-column design
  - When a gradient in 2 directions is expected



Light intensity

Low → High

Temperature gradient

High	C	D	A	B
	D	B	C	A
	B	A	D	C
Low	A	C	B	D

Row-column design

- Incomplete row-columns
  - 7 treatments arranged in 4 rows and 7 columns

1	**3**	**5**	**7**	**2**	**4**	**6**
2	1	7	5	3	6	4
3	6	2	4	7	5	1
4	5	6	3	1	2	7

- General row-column setting
  - v treatments arranged in a p x q array

Incomplete block design

- Occurs when not all treatments are allocated in every block

BIB with 5 treatments, 10 blocks of size k=3
Balanced:

- Each treatment is replicated 6 times
- each pair of treatments appears 3 times within a block

IB designs are used for GWAS applications
Dedicated software is used to plan the design

- SAS proc optex
- cycDesignN

block	treatment				
block	A	B	C	D	E
1		x	x		x
2	x			x	x
3		x		x	x
4			x	x	x
5	x	x		x	
6	x	x			x
7	x	x	x		
8		x	x	x	
9	x		x		x
10	x		x	x	

Statistical analysis

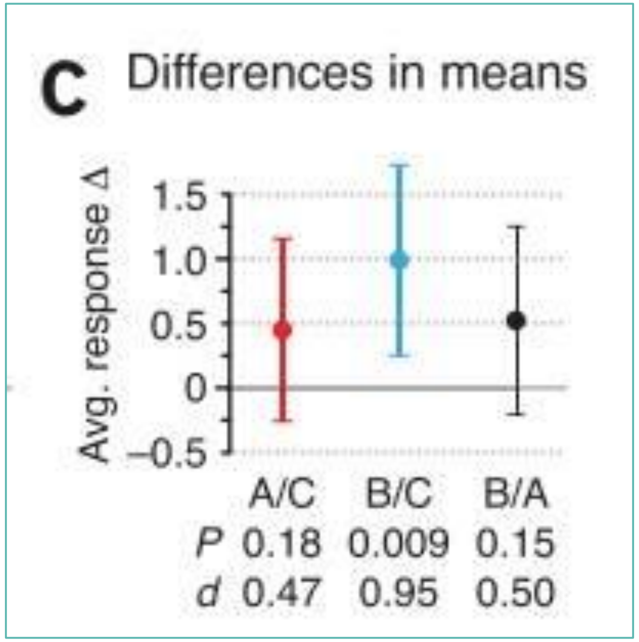
- What do you need to know before beginning an analysis?
  - Research question
  - How were the measurements recorded?
  - How were the treatments assigned?
  - What was the experimental unit?
  - What was the measurement unit?
  - Was the study carried out in the way it was originally designed?

A good analysis should…

- Involve data cleaning
- Involve data exploration
- Use an appropriate statistical method
- Verify that the method's assumptions are met
- Describe the results in appropriate language

Report example

- Reporting a single factor experiment with post-hoc comparisons
  - Krzywinski and Altman, Nature methods, 2014



P-values and effect sizes

- A p-value is easily misinterpreted.
  - A low p-value does not necessarily mean that a finding is of major biological interest. If the sample size is large enough, a tiny effect can result in a low p-value.

p-hacking and *p*-harking

- *p*-hacking is the misreporting of true effect sizes. It occurs when researchers try out several statistical analyses and then selectively report those that produce significant results.
  - Eg. recording many response variables and deciding which to report postanalysis.
  - Synonyms: data-dredging, snooping, fishing, significance-chasing and double-dipping
- *p*-HARKing is the acronym for '**h**ypothesizing **a**fter the **r**esults are **k**nown'. It indicates the phenomenon of constructing hypotheses after the data are analyzed, suggesting that only one hypothesis was tested while many were contemplated.

Balancedness

- Balance i.e. all factor-level combinations (cells) have the same amount of replication
- Imbalance destroys orthogonality ie the computation of the SS of one term does not depend on what other terms there are in the model.

Balancedness

- Balanced data are less susceptible to the effects of nonnormality and nonconstant variance
- Exact statistical methods are not available for imbalanced data, one must resort to approximate methods

- Special case: compare g treatments to a control treatment:

$$n_c = n_t\sqrt{g-1}$$

with n_c sample size control group

and n_t sample size treatment groups

Consequences of imbalance

- In more complex models, some null hypotheses have no exact F-test. Approximate tests are used and the denominator degrees of freedom must be adjusted with the Satterthwaite approach.
- Unbalanced data lead to complicated forms.
  - A preferred method over the ANOVA method is based on the restricted maximum likelihood (REML) approach
  - Wald-type F-tests must be used
  - Denominator df are estimated with the Kenward-Roger method
  - Adjusted means (least square means) must be used instead of arithmetic means. Lsmeans are identical to arithmetic means in the balanced case only.
  - Variance-covariance matrix can be not positive-definite

Missing factor level combinations

- The problems of unbalanced data are increased when there is no data for one or more of the factor-level combinations.
  - some of the least squares means are not estimable. Because statistical software such as SAS/STAT PROC MIXED will not compute least squares means (LSMs) that are inestimable, obtaining desired estimates and comparisons may become difficult.

Missing factor level combinations

- Possible solutions:
  - Truncating data to create a complete data structure
    - Loss of power
  - Fit unsaturated models
    - effects have been eliminated from the model even though they may account for a large portion of the response variability.
    - Consequently, estimates of fixed effects remaining in the model are unadjusted for eliminated effects and may be biased, particularly when data are unbalanced.
    - estimates of the covariance components are consistent only when the mean is correctly specified

References

- Fernandez G. , Horticultural Science, 2007
- Goodman S. et al. , Science Translational medicine, 2016
- Head M. et al., PLoS Biology, 2015
- Nettleton D. , The Plant cell, 2006
- Oehlert Gary W. (2000). A First Course in Design and Analysis of Experiments. Eds W. H. Freeman
- Spilke, J. et al., Journal of Agronomy and Crop Science, 2005
- Schaalje Bruce G., SASpaper 262-26, 2001
- Vanleeuwen D. M. , Amer. Soc. Hort. Sci, 2006
- Coursenotes from Prof. Lynn Eberly (http://www.biostat.umn.edu/~lynn/ph7406/notes/basicprinciples.pdf)