Modern Plant Breeding - "Regular" Level
Experimental designs for High Throughput phenotyping

Prof. Laurent Gentzbittel, Prof. Cécile Ben

Skolkovo Institute of Science and Technology

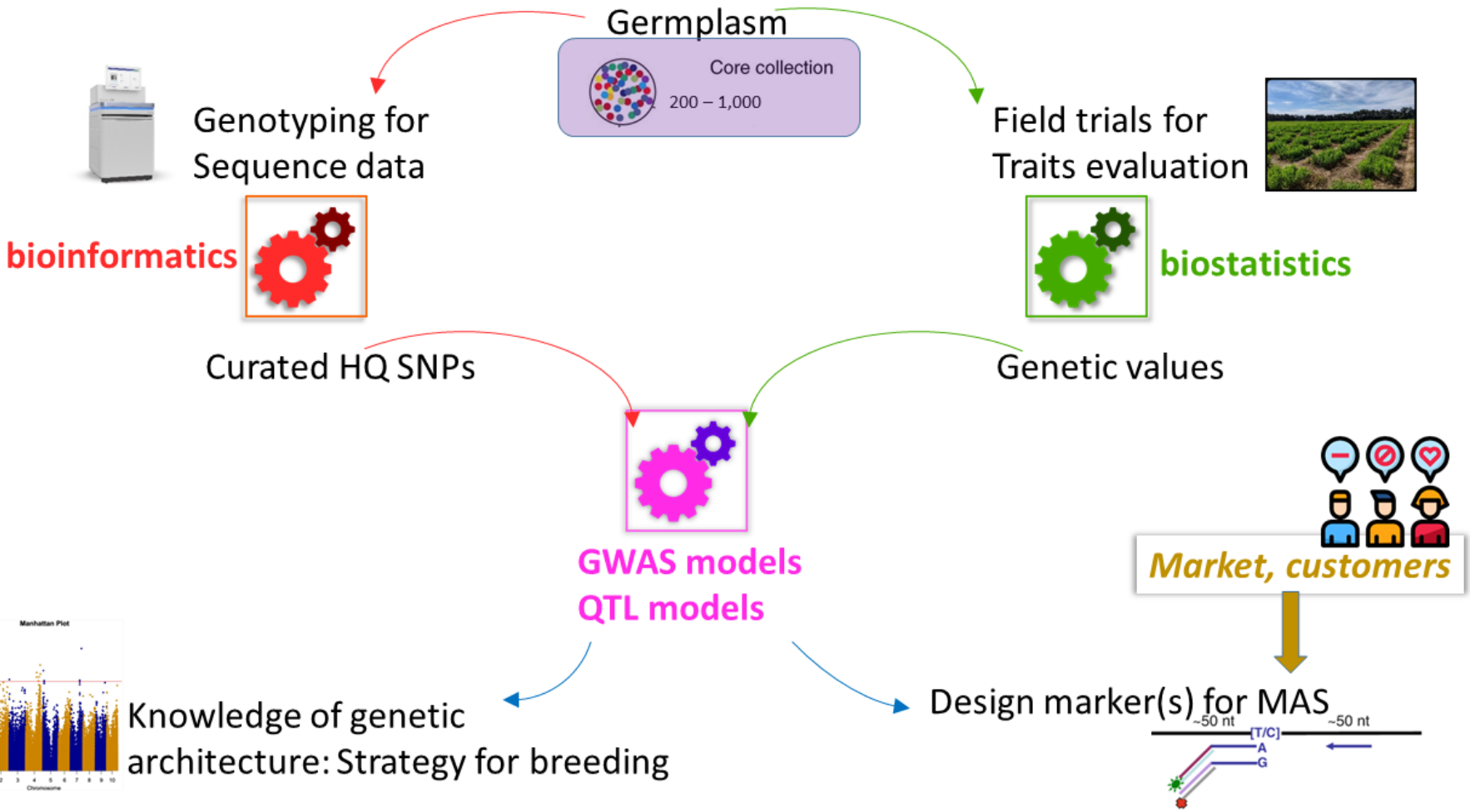
Moscow, 13 - 24 March, 2023

Syllabus

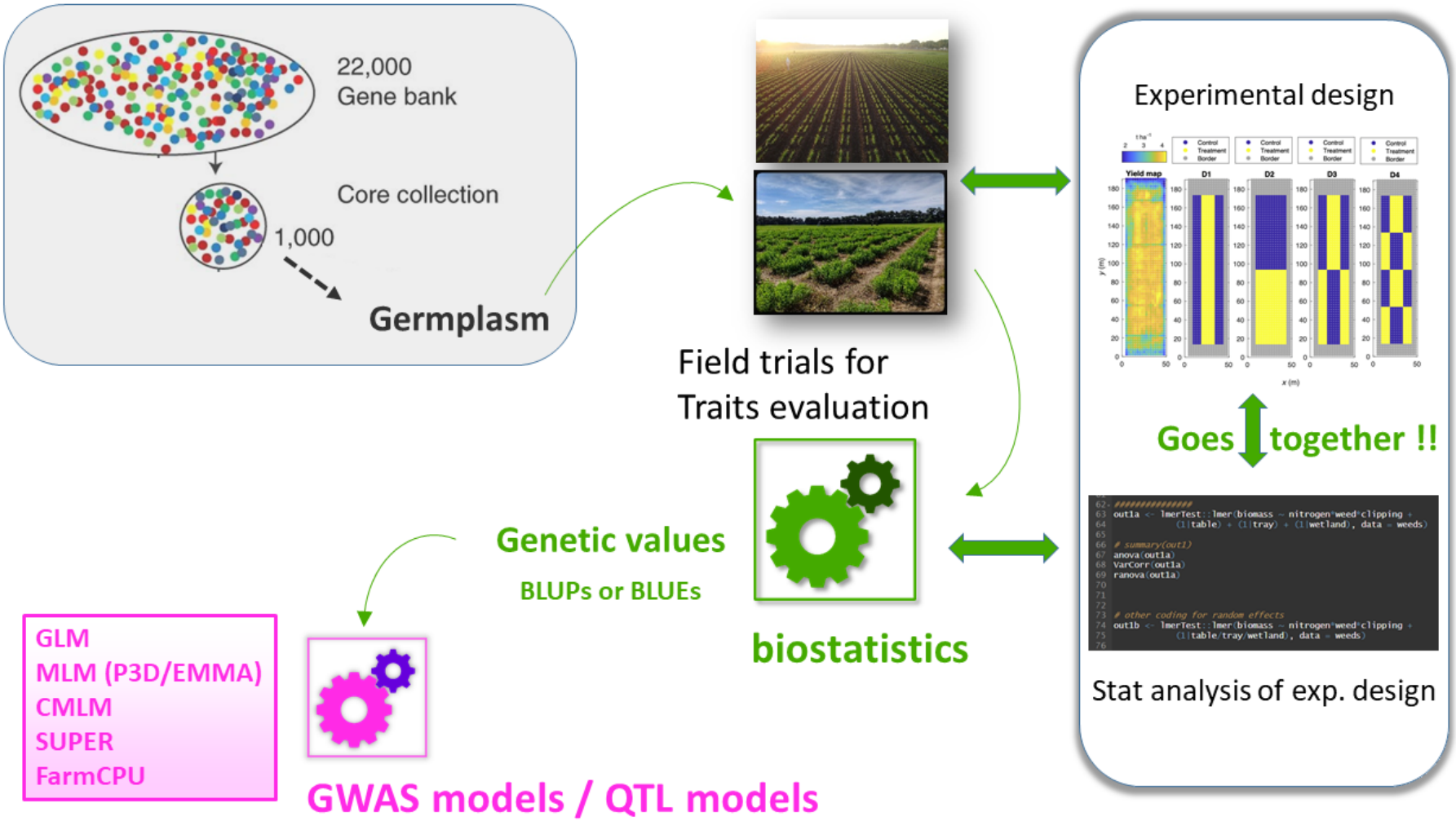
1. A succesful GWAS analysis requires a careful preparation of phenotypic data
2. Augmented block designs - analysis with fixed model
3. Analysis of Balance Incomplete Block Design - BIBD
4. Analysis of lattices

A succesful GWAS analysis requires a careful preparation of phenotypic data

The ‘value chain’ for a GWAS



How to make the best of a budget for experimental trials



The general paradigm of quantitative genetics and breeding

$$
\begin{aligned}
P &= G + E \\
&= \mu + a + d + i + E \\
V_P &= V_G + V_E \\
V_P &= V_a + V_d + V_E
\end{aligned}
$$

Additive value : $a = \sum a_i$ with $a_i > 0$ or $a_i < 0$

Dominance value : $d = \sum d_i$ with $d_i > 0$ or $d_i < 0$

Broad sense heritability : $H^2 = V_G / V_P$

Narrow sense heritability : $h^2 = V_a / V_P$

- for a GWAS, we are interested in the value of **G**, *NOT* in the value of **P**.

Estimating G as precisely as possible

$$P = G + E$$

P obs. = 8
shares of G and E
are “hidden”

After biostat. analysis

$P_{corrected} \approx 2$
G = 2
E = epsilon

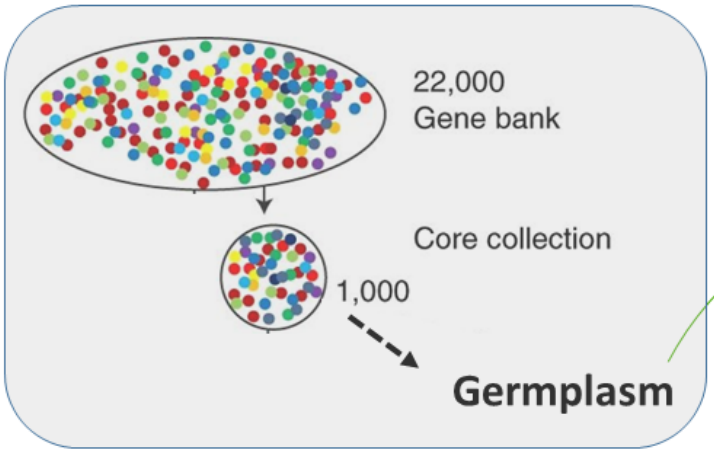
‘corrected for environmental effects”
G = 99% of $P_{corrected}$

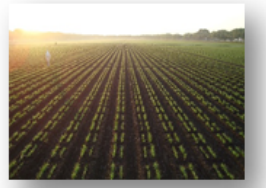
without biostat. analysis

$P_{uncorrected} = 8$
G = 2
E = 6

‘uncorrected for environmental effects”
G = 25% of $P_{uncorrected}$

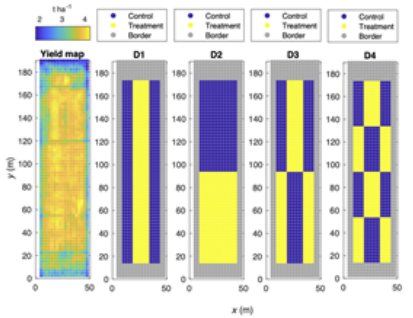
How to make the best of a budget for experimental trials





Field trials for
Traits evaluation

Experimental design



Goes together !!

```
outla <- lmerTest::lmer(biomass ~ nitrogen*weed*clipping +
            (1|table) + (1|tray) + (1|wetland), data = weeds)

# summary(out1)
anova(outla)
VarCorr(outla)
ranova(outla)


# other coding for random effects
outlb <- lmerTest::lmer(biomass ~ nitrogen*weed*clipping +
            (1|table/tray/wetland), data = weeds)
```

Stat analysis of exp. design

Genetic values

BLUPs or BLUEs

biostatistics

GLM
MLM (P3D/EMMA)
CMLM
SUPER
FarmCPU

GWAS models / QTL models

Some useful thoughts . . .

- *Essentially, all models are wrong, but some are useful*

  G. Box (1987) in: Empirical Model-Building and Response Surfaces p424

- *To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.*

  Sir Ronald A. Fisher in: Presidential Address to the First Indian Statistical Congress, 1938

- The three **R** :
  - **R**eplication,
  - **R**eduction in variability (mostly by *blocking*),
  - **R**andomisation

The "Block Design" : the reference for any field trials.

Aims of blocking: To reduce the residual variance by discriminationg the 'pure error' due to micro-environmental variation – σ_e^2 – AND some macro-environmental variation – the variation due to `blocks`.

- presence of blocks of ''equal size", **each of which contains all the treatments**;
- variability within each block should be minimized and variability among blocks should be maximized;
- levels of the main factor(s) need to be randomized within each block (independency of the data);
- There is *NO* interaction between the block and the studied factor(s).

in summary:

- blocks are **homogeneous within-**
- blocks are **heterogeneous between-**

Augmented block designs - analysis with fixed model

Augmented Block designs: when all treatments are not tested in each block

- Introduced by Federer (1956)
- Controls (check varieties) are replicated in a standard experimental design : plan en bloc
- New treatments (genotypes) are not replicated, or have fewer replicates than the checks – they **augment** the standard design
- Provide an estimate of standard error that can be used for comparisons
  - Among the new genotypes
  - Between new genotypes and check varieties

How to implement an augmented block design

- Area is divided into blocks
  - These are incomplete blocks because they will contain only a subset of the entries
  - Two or more check varieties are assigned at random to plots within the blocks
- the same check varieties appear in each block

little is lost if you want to place one check systematically - a block marker

- Most efficient when block size is constant
- Checks are replicated **in each block**, but new entries are *not*

Applications

- Early stages in a breeding program
- Evaluation of large biodiversity collections
  - May be insufficient seed for replication
  - Using a single replication permits more genotypes to be screened
- Participatory plant breeding or biodiversity evaluation (common gardens)
- Want to evaluate promising genotypes (or other technologies) in as many environments as possible

Advantages of Augmented Block designs

- Observations on new genotypes can be adjusted for field heterogeneity (blocking)
- **Unreplicated designs can make good use of scarce resources**
- Fewer check plots are required than for designs with systematic repetition of a single check
- Flexible – blocks can be of unequal size

Drawbacks of Augmented Block designs

- Considerable resources are spent on production and processing of control plots
- Relatively few degrees of freedom for experimental error, which reduces the power to detect differences among treatments

- **Unreplicated** experiments are inherently **imprecise**, no matter how sophisticated the design

How much blocks ?

- At least 10 df for residual variance
- the df for residual are: $(r-1)(c-1)$
  c : number of checks per block
  r : number of blocks = number of repeat for a check
- The minimum number of blocks ought to be $\geq \left(\frac{10}{c-1}\right) + 1$
- Example with 4 checks :
  $\left(\frac{10}{4-1}\right) + 1 = 4.33 \approx 5$ blocks
- Each block contains at least $c+1$ plots (ie at least one new accession is assessed per block)

Analysis

- The residual error is estimated via the analysis of checks as in a RCBD
- MSE, σ_e^2 is used to build the standard error to compare the means of the accessions.
- The trait value of the new accessions will be "corrected'' for the block effect

ANOVA table for the checks :

Variation	ddl	SS	MS	EMS
blocks	r-1			
checks	c-1			
Residual	(r-1)(c-1)		σ_e^2	σ_e^2

Standard errors to compare the means

- difference between two checks
  $S_c = \sqrt{2\sigma_e^2/r}$
- difference between the adjusted means of two new entries within a same block
  $S_d = \sqrt{2\sigma_e^2}$
- difference between the adjusted means of two new entries in two different blocks
  $S_v = \sqrt{2(c+1)\sigma_e^2/c}$
- difference between the adjusted means of a new entry and a check
  $S_{vc} = \sqrt{(r+1)(c+1)\sigma_e^2/rc}$

c=number of different checks per block ; r=number of blocks=number of replicates of a check
To compute confidence interval, we use Student t distribution with $(r-1)(c-1)$ df

Augmented Block Design – Example 1

8 new varieties are tested along with $c = 4$ reference varieties (checks)
The trial has 3 blocks. Unfortunatly this is sub-optimal: the df for residual are $(r-1)(c-1) = 6$ df ie < 10 df
$r = 3$; $c = 4$ thus the t distribution for mean comparisons will get $(r-1)(c-1) = 6$ df

- At your keyboards: `AugmentedBlockDesign_exemple1.R`

Augmented Block Design – Example 2

We aim to test 30 new entries, using $c = 3$ standard genotypes (checks).

- Suitable minimum number of blocks :

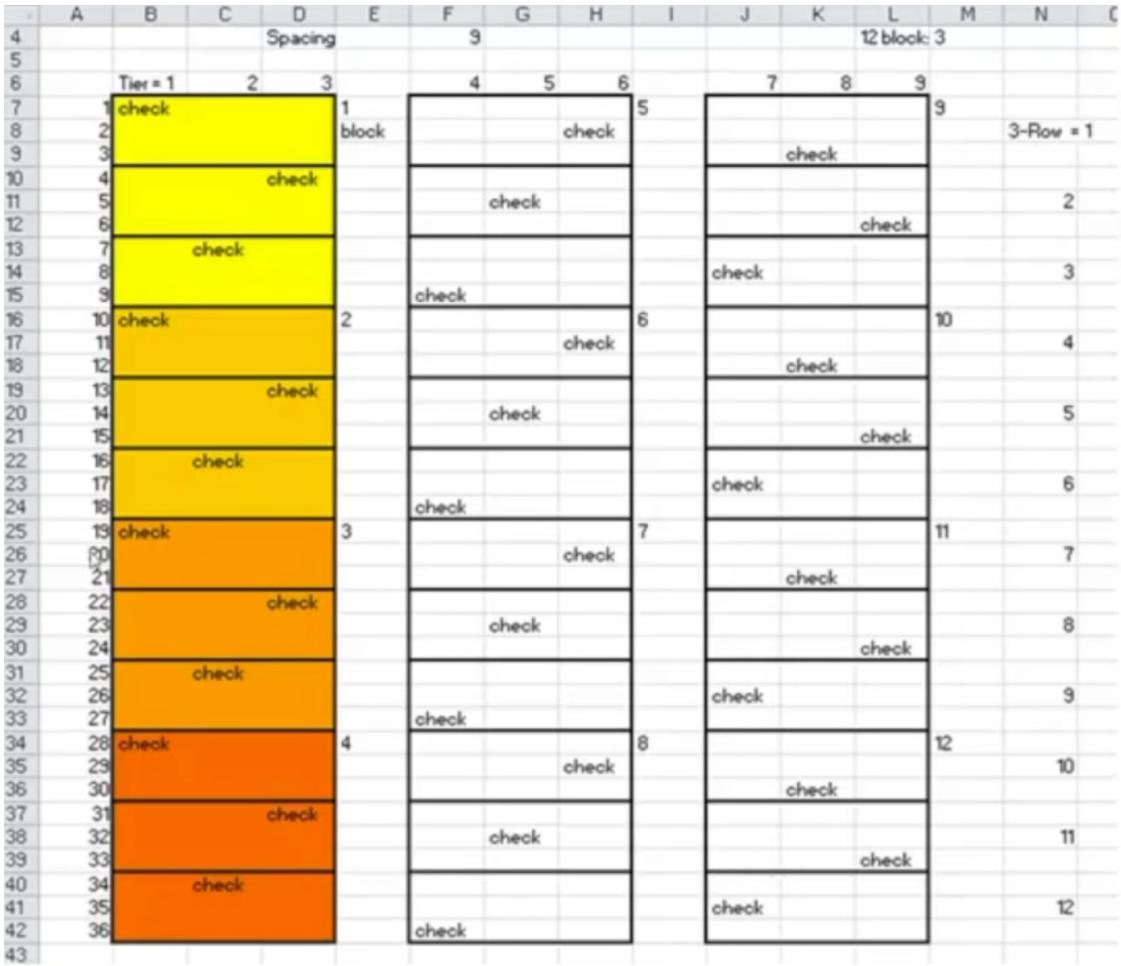
$\frac{10}{c-1} + 1 = \frac{10}{2} + 1 = 6$

- Number of new accessions per block : $30/6 = 5$
- Total number of plots, given the 6 blocs : $(5 + 3)x6 = 48$

▶ At your keyboards: `AugmentedBlockDesign_exemple2.R`

Augmented Block Design can be used to correct for spatial heterogeneity

- Based on a 'Latin Square' concept
- it is possible to correct for
  - 'column',
  - 'row'
  - and 'quadrant' effect



▶ see Advanced Level, using MLM and BLUPs

Analysis of Balance Incomplete Block Design - BIBD

Issues of available space for performing experiments . . .

1. Not enough available space

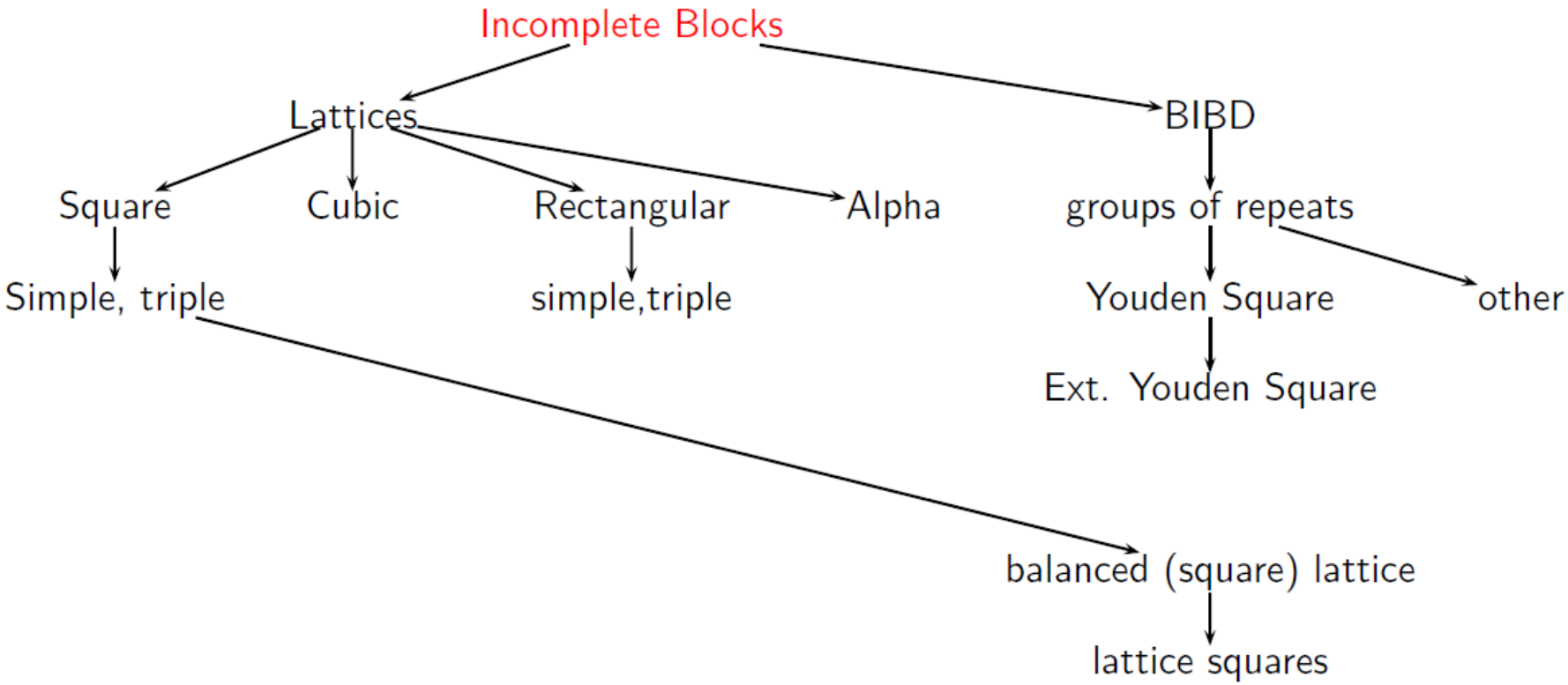
- 6 brands of tires for tractor to test
  A block is a couple tractor + driver – but there are only 4 wheels
- Six different processes for extracting avocado oil.
  There is a large fruit to fruit variability, so block on fruit – but each fruit is only big enough to extract amount of oil to test four processes.
- 250 varieties to test but each block can only handle 70 varieties.

2. Heterogeneity within blocks

- In breeding programs, the number of treatments may be large
- As blocks get larger, the conditions within blocks become more heterogeneous

The family of Incomplete Designs

Solutions to manage spatial issues in the blocks . . . . or variability within the blocks: Just make use of incomplete blocks !



What defines a Balanced Incomplete Block Design:

The basic prototype of all incomplete block designs is the BIBD. We have:

- g treatments
- b blocks
- k units per block
- each treatment used r times
- $b.k$ total units
- therefore, $bk = rg$

In addition, **each pair of treatments occurs together in the same number of blocks**.

$$\lambda = r(k-1)/(g-1)$$

BIBD require connections between blocks

- A simple case:

A B	A C	C B

versus

A B	A B	C C

Both have $g = 3$, $k = 2$, $r = 2$, $b = 3$
but left side is BIBD and right side is not.

- Why this design is not a BIBD ?

A B	A C	B C	D E	D F	E F

Example of a BIBD for 7 treatments in blocks of 3 experimental units

1	2	4
2	3	5

3	4	6
4	5	7

1	5	6
2	6	7

1	3	7

Compared to an Augmented Block Design, each pair of treatment is compared with the same accuracy.

- $g = 7$ treatments
- $b = 7$ blocks
- $k = 3$ units per block
- each treatment used $r = 3$ times
- $bk = 21$ total units

$$bk = rg \quad \lambda = r(k-1)/(g-1) = 1$$

Statistical analysis of BIBD 1/2

The ANOVA model is simple :

$$y_{ij} = \mu + \beta_j + \alpha_i + \epsilon_{ij}$$

with β_j the effect of the blocks and α_i the effects of treatments / varieties, ...

CAUTION !!

1. The design is unbalanced: the effects of blocks and treatments are NOT independant. ***PUT*** **`treament`** **at the end of ANOVA syntax.**
2. Therefore, a part of the variation among treatments is included in the variation among blocks
3. It will be required to compute treatment means that are 'adjusted for the block effect'.

Statistical analysis of BIBD 2/2

intrablock analysis : All estimates are based on differences from within blocks.
we assume blocks are fixed.

interblock recovery analysis :there is some information about the treatments in block SS
we assume blocks are random (see "Advanced level")

▶ A BIBD is alway less efficient than an RCBD with the same number of repeats
The *efficiency factor*, compared to a RCBD, is defined as :

$$E = \frac{1 - \frac{1}{k}}{1 - \frac{1}{g}} = \frac{\lambda g}{rk}$$

Limits of BIBD - 1

- If $\lambda = r(k-1)/(g-1)$ is not a whole number, then no BIBD exists for that set of parameters.
- It does not exist a BIBD design for some combinations of $kb = rg$.

example :
with $g = 5$ treatements , $b = 5$ blocks of size $k = 3$ plots

we have $r = 3$, but therefore $\lambda = \frac{3(3-1)}{5-1} = 3/2$ is not an integer.

Limits of BIBD - 2

- For some combinations of g and k, the number or repeats r and/or of blocks b may be too large to be manageable:

exemple:
Let's attempt to work with $g = 12$ treatments, with blocks of size $k = 7$ plots.

We need $\lambda = \frac{r(k-1)}{g-1} = 6r/11$ to be an integer. So r is a factor of 11 called $11.m$
also

$$bg = rk \quad \text{thus} \quad b = rg/k = (11.m)\frac{12}{7}$$

also ought to be an integer, so m is a factor of 7
$\triangleright$ let choose $m = 1 \times 7 = 7$ then

- $b = (11 \times m)\frac{12}{7} = (11 \times 7)\frac{12}{7} = 132$ blocks
- $r = 11 \times m = 77$
- for a total number of $bk = rg = 924$ plots !

Analysis and creation of BIBD

- At your keyboard : BIBD1.R

- At your keyboard : BIBD2.R

BIBD with 'groups of repeats'

Let's consider the following BIBD :

block 1 :	A	B
block 2 :	A	C
block 3 :	A	D
block 4 :	B	C
block 5 :	B	D
block 6 :	C	D

with :

- $g = 4$ entries
- $b = 6$ blocks
- $k = 2$ units per block
- each entry is assessed $r = 3$ times

and $\lambda = \frac{r(k-1)}{g-1} = \frac{3(2-1)}{3} = 1$.

We may arrange the blocks in the field alike:

Replicate 1 :

bloc 1 :	A	B
bloc 6 :	C	D

Replicate 2 :

bloc 2 :	A	C
bloc 5 :	B	D

Replicate 3 :

bloc 3 :	A	D
bloc 4 :	B	C

A BIBD is said resolvable when it can be divided into replicates such that each replicate (i.e. each group of blocks) contains each treatment exactly once.

Replicates of BIBD allow for controlling another source of variability 1/2

Let's consider the following BIBD : :

block 1 :	A	B	C
block 2 :	A	B	D
block 3 :	A	C	D
block 4 :	B	C	D

with :

- $g = 4$ entries
- $b = 4$ blocks
- $k = 3$ units per block
- each entry is assessed $r = 3$ times

and $\lambda = \frac{r(k-1)}{g-1} = \frac{3(3-1)}{3} = 2$.

Here, the number of incomplete blocks is a multiple of the number of entries. $b = m.g$

Replicates of BIBD allow for controlling another source of variability 2/2

It is possible to re-arrange the design in the field alike :

block 1 :	A	B	C
block 2 :	B	A	D
block 3 :	C	D	A
block 4 :	D	C	B

Here have $b = g$ thus $r = k$ (because $bk = rg$ is a propety of BIBD)

This design is called *Youden square* or *incomplete Latin square*.
A Youden square has the number of rows equal to the number of treatments and the number of columns less than the number of treatments. Each treatment occurs once and only once in each column and the treatments must be balanced across the row

Analysis and creation of Youden Square design

- At your keyboard : YoudenSquare.R

Analysis of lattices

Lattice designs

- Incomplete set of treatments in each block
- Ideal for experiments with large number of treatments (treatments = lines, genotypes, varieties

Advantages

- Increase homogeneity within block
- Increase precision

Disadvantages

- Fixed number of treatments and replications
- Complex analysis and complex generation of plot design for the field
- Unequal precisions in treatment comparison : these are most often NOT balanced designs

Uses of Lattice designs

- **in large experiments** - reduce the number of treatments in a block. In plant breeding exp. the number of genotypes is often large
- **in growth chambers and greenhouse** - there may be insufficient space within each block to put all lines

The lattices : IBD organised in replicates with contraints on g and k

Lattices are resolvable designs of *IBD* where it exists a relation between the number of entries g and the number of plots per block k.

1. A Square Lattice can be used when $g = k^2$
2. A Cubic Lattice Cubique can be used when $g = k^3$
3. A Rectangular Lattice can be used when $g = s.(s-1)$ and $k = s-1$
4. An α-Lattice design may be used when $g = s.k$
   so there are s blocks per replicate and $b = s.r$ blocks at total

Where:

- g = number of entries
- s = number of blocks within a replicate (block number)
- k = number of plots within a block (block size)

▶ Some lattice designs are *balanced* ie **B***IBD*

Example 1 of a Balanced Square Lattice

▷ Example with $g = 3^2 = 9$ entries (the number of varieties ought to be a perfect square). Here $k = 3$.

(1)	1	2	3	(4)	1	4	7	(7)	1	5	9	(10)	1	8	6
(2)	4	5	6	(5)	2	5	8	(8)	7	2	6	(11)	4	2	9
(3)	7	8	9	(6)	3	6	9	(9)	4	8	3	(12)	7	5	3

This is a balanced (square) lattice. This is a BIBD with:

- $g = k^2 = 3^2 = 9$ entries
- $b = k(k+1) = 3 \times (3+1) = 12$ blocks
- $r = k + 1 = 4$ assessment of each entry, in 4 replicates
- $\lambda = 1$
- $E = k/(k+1) = 3/4 = 0.75$

Example 1 of a Balanced Square Lattice

▷ The creation of Balanced Square Lattice requires to use $k-1$ *orthogonal Latin Squares* of size $k \times k$. Example: A balanced square

lattice for $g = 16 = 4^2$ genotypes (ie $k = 4$) will be with :

- $g = k^2 = 4^2 = 16$ entries
- $b = k(k+1) = 4(4+1) = 20$ blocks
- $r = k + 1 = 5$ assessment of each entry, ie 5 replicates
- $\lambda = \frac{r(k-1)}{g-1} = \frac{5(4-1)}{15} = 1$
- $E = k/(k+1) = 4/5 = 0.80$

All these constraints limit their use in practice.

Example 2 of a Simple Square Lattice

Here, we will give-up with the constraint of getting a balanced design. A *Simple Square Lattice* has two replicates

▷ Example with $g = 3^2 = 9$ entries. Here $k = 3$.

Rep 1

(1)	1	2	3
(2)	4	5	6
(3)	7	8	9

Rep 2

(4)	1	4	7
(5)	2	5	8
(6)	3	6	9

This design is called a simple (square) lattice.

This is NOT a BIBD

the `agricolae` package generates Simple Square Lattice.

Analysis and creation of Simple Square Lattice design

- At your keyboard : SimpleSquareLattice.R

Example 3 of a Triple Square Lattice

▷ Example with $g = 3^2 = 9$ entries. Here $k = 3$.

Rep 1

(1)	1	2	3
(2)	4	5	6
(3)	7	8	9

Rep 2

(4)	1	4	7
(5)	2	5	8
(6)	3	6	9

Rep 3

(7)	1	6	8
(8)	2	4	9
(9)	3	5	7

This design is a triple (square) lattice.

This is NOT a BIBD

the `agricolae` package generates Triple Square Lattice

Example of a α-Lattice

$\triangleright$ the α-Lattice has less constraint.
Let's define s as the number of blocks in each complete replication and consider k plots per block. We need only $g = sk$.

Example with $s = 4$ and $k = 4$ so $g = sk = 16$

Replication (Super block) 1

Block 1				Block 2				Block 3				Block 4			
13	5	2	8	3	12	14	4	7	16	10	11	15	6	1	9

Replication (Super block) 2

Block 5				Block 6				Block 7				Block 8			
2	3	11	1	7	12	5	9	13	6	16	4	15	10	8	14

Replication (Super block) 3

Block 9				Block 10				Block 11				Block 12			
9	14	11	2	15	3	8	16	13	7	1	12	4	6	5	10

Example of a α-Lattice

Replication (Super block) 1

Block 1				Block 2				Block 3				Block 4			
13	5	2	8	3	12	14	4	7	16	10	11	15	6	1	9

Replication (Super block) 2

Block 5				Block 6				Block 7				Block 8			
2	3	11	1	7	12	5	9	13	6	16	4	15	10	8	14

Replication (Super block) 3

Block 9				Block 10				Block 11				Block 12			
9	14	11	2	15	3	8	16	13	7	1	12	4	6	5	10

α-lattice designs are **NOT** BIBD.

- α-lattice designs are partially balanced designs
- some pairs of varieties are never together in the same incomplete block (0), other pair of varieties are together in the same incomplete block once (1), others are together twice (2), and others more (3, 4, ...).
- The *more balanced* the α- lattice design is the better: most of the α-lattices are (0, 1) or (0, 1, 2).

Example of a α-Lattice

The generation of α-Lattice design is complex and 'mysterious'. It involves several steps of permutation. The linear model for α-Lattice is :

$$Y_{ijk} = \mu + \tau_i + \gamma_j + \rho_{k(j)} + \epsilon_{ijk}$$

with:

- Y_{ijk} is the phenotype of ith treatment in jth replication and k block
- τ_i is the treatment effect $treatmenteffecti = 1, 2, \ldots k$
- γ_j replicate effect $j = 1, 2, \ldots r$
- $\rho_{k(j)}$ is the block within replicate effect $k = 1, 2, \ldots s$
- ϵ_{ijk} is the random error

For example, a design for $g = 200$ varieties can be produced in $s = 20$ blocks per replicate, each with $k = 10$ treatments per block.

the `agricolae` package generates α-Lattice

Analysis and creation of α-Lattice

A spring oats trial. There were 24 varieties in 3 replicates, each consisting of 6 incomplete blocks of 4 plots. Planted in a alpha design.

- At your keyboard : AlphaLattice1.R

Matrix notation of the two-way ANOVA model

$$\underset{(n,1)}{\boldsymbol{Y}} = \underset{(n,p)}{\boldsymbol{X}} \cdot \underset{(p,1)}{\boldsymbol{b}} + \underset{(n,1)}{\boldsymbol{e}}$$

	Sun	Shade	StrongShade
V1	112	86	80
V2	90	73	62
V3	123	89	81

- at the whiteboard

des codes

$$y_{crijk} = \mu + Row_r + Col_c + Strain_i + \eta_{k(cri)} + Genotype_j + Genotype_j \cdot Strain_i + \epsilon_{crijk}$$

$$y_{crijk} = \mu + Row_r + Col_c + Strain_i + \eta_{k(cri)} + Genotype_j + Genotype_j \cdot Strain_i + \epsilon_{crijk}$$

This slide has two columns

left

right

This slide also has columns

contents. . .

contents. . .