HOS 6236 Molecular Marker Assisted Plant Breeding

Fall 2017

Last Class:

Important breeding concepts

Survey plant breeding methods

Todays Class:

Recurrent selection methods and BLUP

Plant Breeding

Plant Breeding is a simple three step process:

1. Create or assemble a germplasm population with enough variation

2. Identify the superior individuals

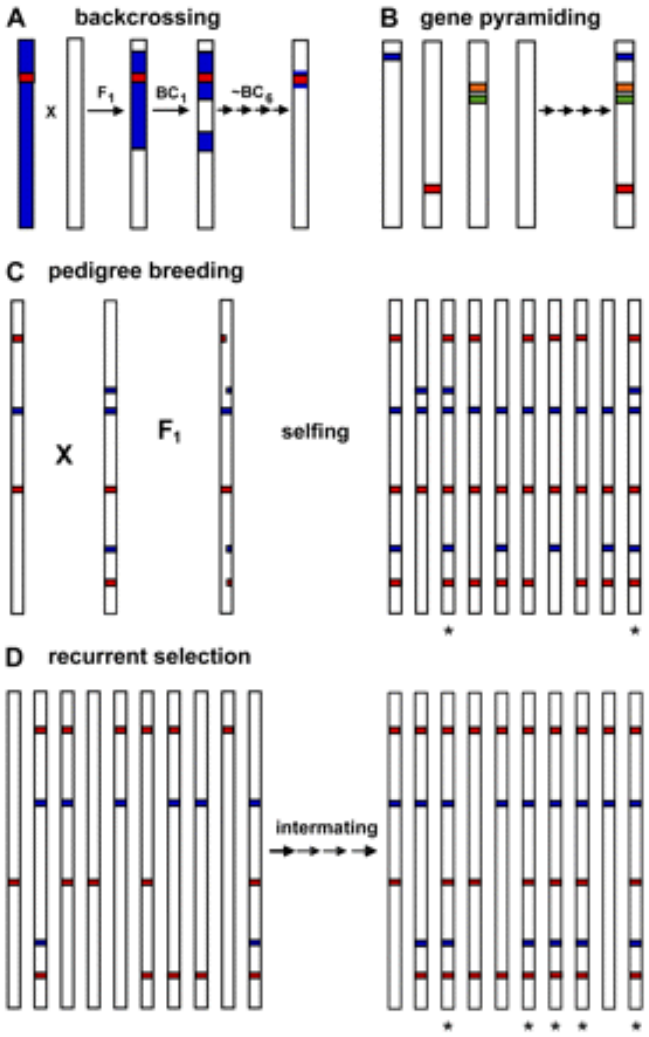
3. Develop cultivars with the superior individuals

Methods for Plant Breeding

Backcrossing/Introgression

Pedigree Method

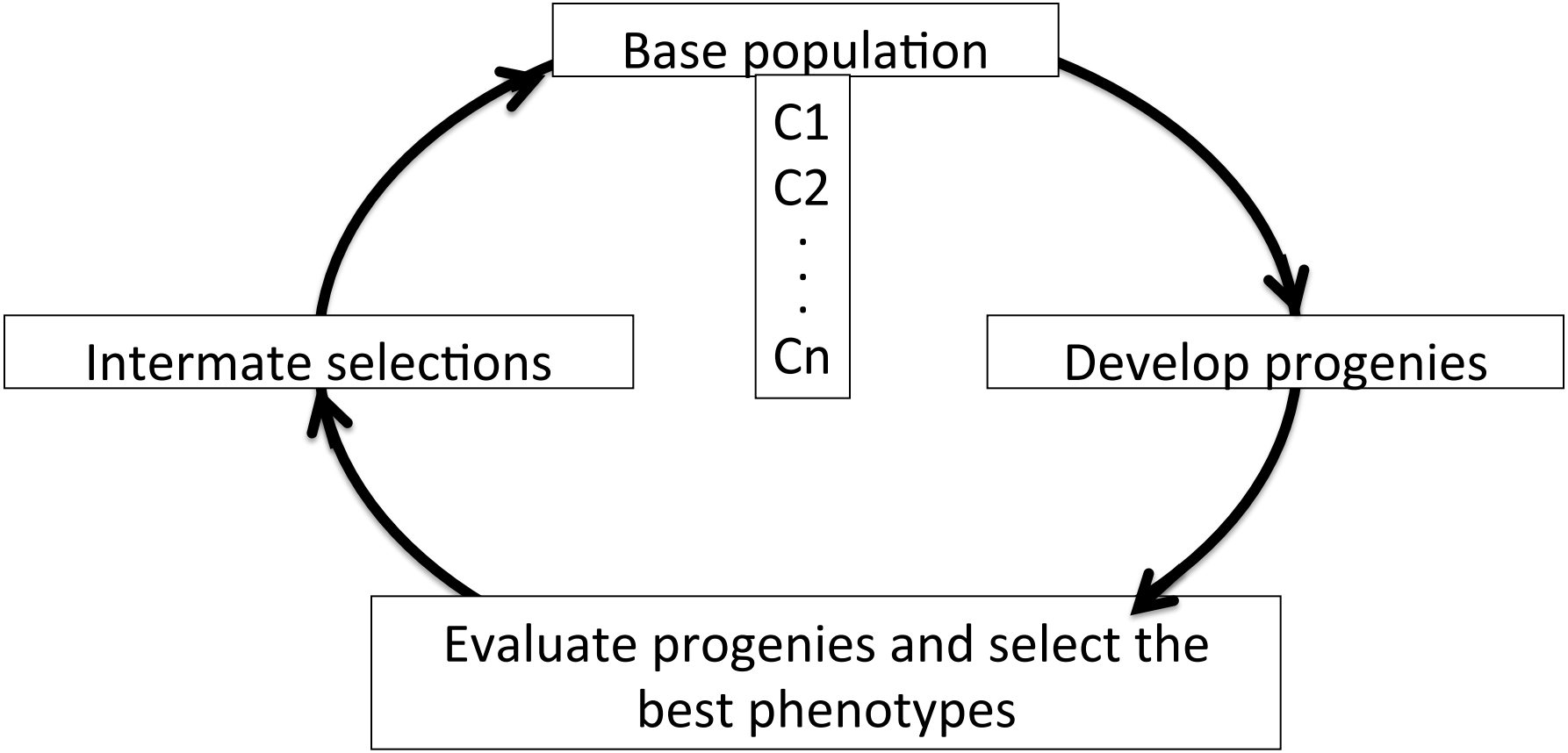
Recurrent Selection



Moose and Mumm 2008: Plant Physiol 147

Recurrent Selection

Cyclical and systematic technique in which desirable individuals are selected from a population and inter-mated to form a new population, maintaining the variability of the new population

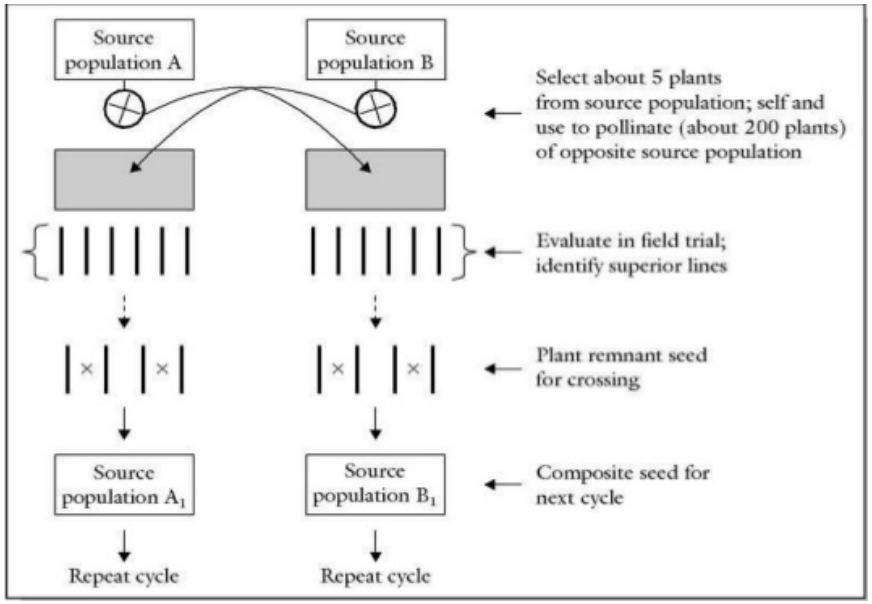


Reciprocal Recurrent Selection (RRS)

A simultaneous method to improve General Combining Ability (GCA) and Specific Combining Ability (SCA), proposed by Comstock et al (1949).

GCA= ability of a genotype individual to combine well with many other individuals

SCA= ability of a genotype individual to combine well with another specific individual



Acquaah 2006. Wiley-Blackwell

How to select the best individuals

Phenotypic Selection – Selection based on the phenotype. Based on visual observation by a trained breeder.

Advantages?

Disadvantages?

Genotypic Selection – Selection based on the breeding value predicted based on statistical methods (ex: mixed models, Bayesian methods).

Advantages?

Disadvantages?

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

g_j → average additive effect of genes an individual receive from both parents (breeding value)

Each parent contributes a sample half of its genes to its progeny. The average effect of of this sample is the general combining ability (GCA) of the parent → half of the breeding value

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

Breeding value (BV) of progeny then is the sum of the GCA of both parents.

Since the GCA is a function of the genes transmitted from parents to progeny, it is the only components that can be selected for.

Dominance and epistasis assumed to be insignificants.

Basic Linear Mixed Model

Yield = Environment Effect+ Genetic Effect + Residual

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j sum of additive (g_a), dominance (g_d), and epistasis (g_i)

e_{ij} random error of the ij^{th} observation

It is assumed that *y* follows a multivariate normal distribution (MVN), implying that traits are determined by infinitely many additive genes of infinitesimal effect, infinitesimal model (Fisher, 1918).

Thus BV for individual $i \rightarrow a_i = g_a = 1/2a_f + 1/2a_m + m_i$

m_i= *Mendelian sampling*

Pedigree and Relationship Matrices

- Why worry about the pedigree in genetic analyses?

  Statistically, random genetic effects (i.e. BLUPs) are not independent and their matrix of correlations or co-variances (**G** or **A**) needs to be specified.

  Genetically, it is important to consider information about relatives as they will share some alleles, and therefore their response is correlated.

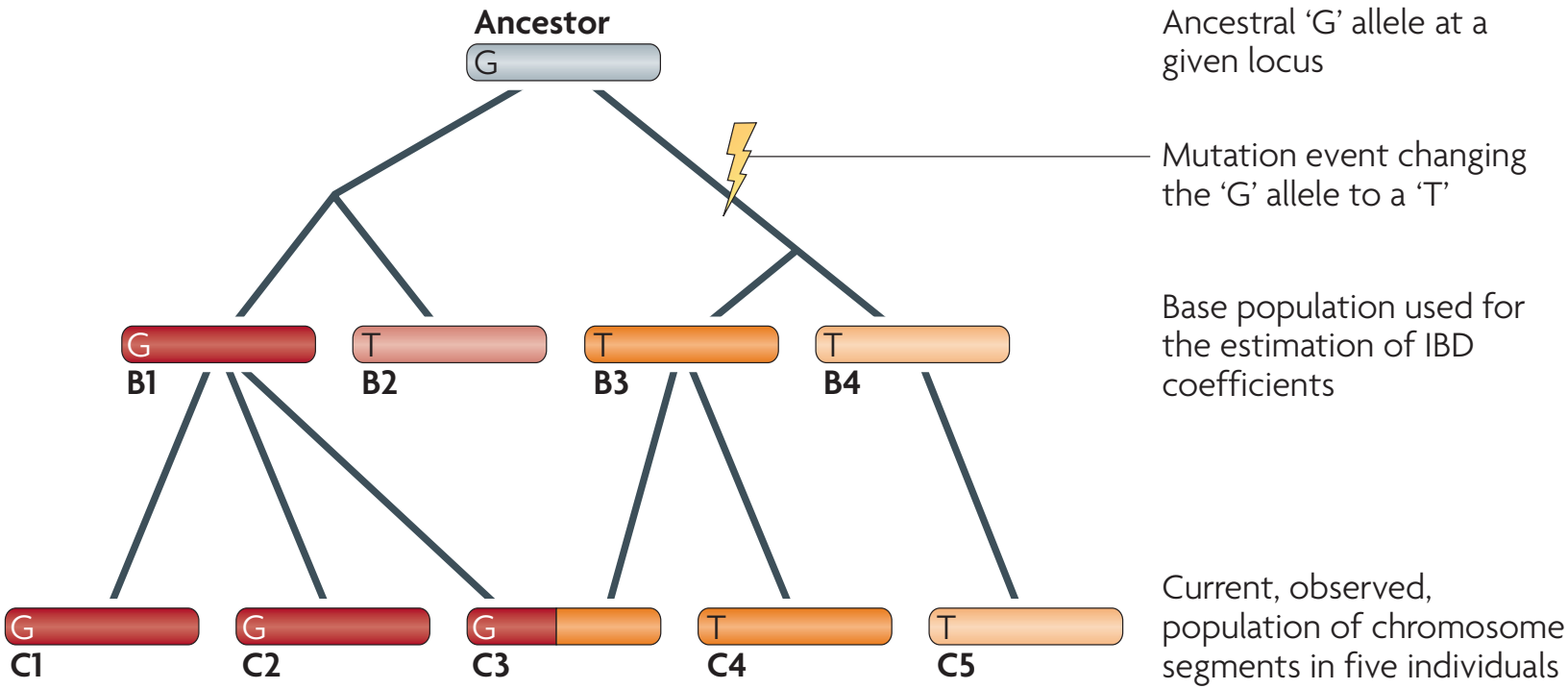
- How to incorporate this information?

  Genetic relationships (pedigree) can be calculated using genetic theory (expected values) or molecular information (e.g. SNPs), and included into the linear mixed model by specifying a Relationship matrix

- Are there other benefits?

  Many. It is a more efficient use of the information about individuals, but also genetic values of individual not tested, but with relatives tested, can be *predicted* and selected.

Pedigree and Relationship Matrices



Ancestor is the point of coalescent for current alleles C1-C5.
Identity by descent (IBD) of current alleles can be defined respect to B1-B4, thus
G allele in C1-C3 are IBD
T allele in C4 and C5 are identity by state (IBS)
The chromosome segment C2 and C3 are IBS as well.

Powell et al 2010

Pedigree and Relationship Matrices

- The additive genetic relationship between two individuals is twice their IBD (aka coancestry or kinship coefficient)

- The matrix that indicates the additive genetic relationship among individuals is the numerator relationship matrix (**A**).
- Properties of **A:**

  - Is symmetric

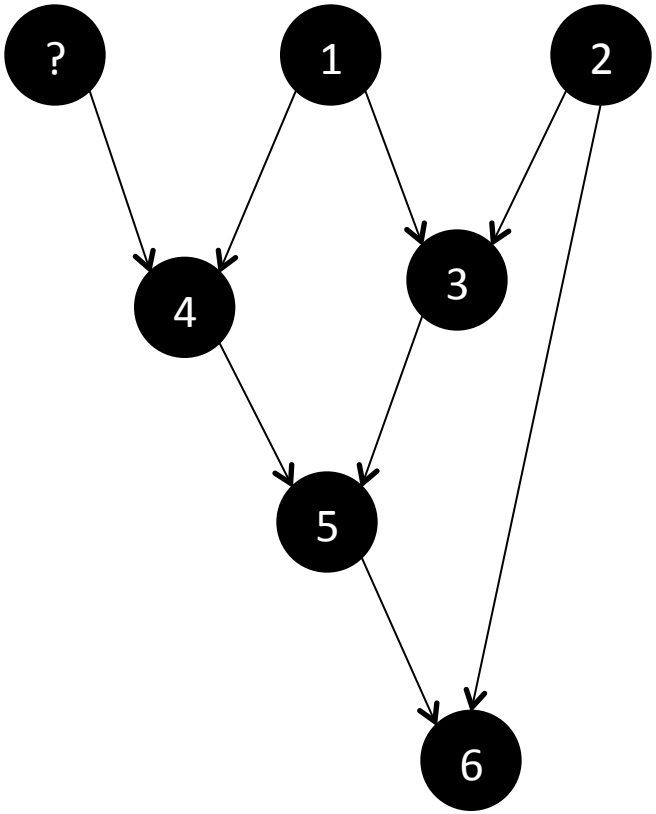
  - Diagonal elements is equal to $1+F_i$ (F_i is the inbreeding coefficient on individual i)

  - Can be computed by different methods, the recursive method (Henderson 1976) is simpler.

Pedigree and Relationship Matrices

Example

Pedigree of a group of individuals:



Individual	Male	Female
3	1	2
4	1	Unknown
5	4	3
6	5	2

Pedigree and Relationship Matrices

Numerator relationship matrix (A)

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1.00 & 0.00 & 0.50 & 0.50 & 0.50 & 0.25 \\ & 1.00 & 0.50 & 0.00 & 0.25 & 0.625 \\ & & 1.00 & 0.25 & 0.625 & 0.563 \\ & & & 1.00 & 0.625 & 0.313 \\ & & & & 1.125 & 0.688 \\ & & & & & 1.125 \end{bmatrix} \end{array}$$

- Linked to the concept of **identity by descent**.
- **Diagonal** $a_{ii} = 1 + F_i$ (inbreeding coefficient on individual *i*)

  Twice the probability that two gametes taken at random from animal *i* will carry identical alleles by descent.
- **Off-diagonal** a_{ij} numerator of the coefficient of relationship between animal *i* and *j*.
- Several algorithms are available in ASReml to obtain this matrix.

Pedigree and Relationship Matrices

CALCULATING THE A MATRIX

- Let $\mathbf{A} = \{a_{ij}\}$ be the relationship matrix.
- Let $a_{i,-j}$ the the *i-th* row of **A** except for the *j-th* element.
- Assume the relationship matrix for the base individual is known (e.g. unrelated, non inbred). This will for a base matrix (e.g. identity)
- The row of the relationship matrix for the progeny of two parents is generates as the average of the relationship matrix rows for the parents:

$$a_{i,-j} = (a_{s,-i} + a_{d,-i})/2$$

- The diagonal element, $a_{i,i}$ of this new individual is:

$$a_{i,i} = 1 + a_{s,d}/2 = 1 + F_i$$

where F_i is the inbreeding coefficient.

Pedigree and Relationship Matrices

Construction / Check

- Pedigree information is associated with proper management and validation/check of data.
- Individuals need to be ordered by generation (e.g. parents need to be defined before progeny).
- All parents need to be defined in pedigree file (the inclusion of founder parents is optional).
- All individuals present in dataset (i.e. levels associated with pedigree file) need to be defined in pedigree file.
- Individuals can be defined as male or female parents (but this should be checked if is not biologically possible).

The Mixed Model Equation

Consider a model with block as fixed and variety as random effects.

yield = μ + block + variety + error

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

g_j random effect of the j^{th} variety, $E(g_j) = 0$, $V(g_j) = \mathbf{A}\sigma_g^2 = \mathbf{G}$

e_{ij} random error of the ij^{th} observation, $E(e_{ij}) = 0$, $V(e_{ij}) = \mathbf{I}\sigma^2 = \mathbf{R}$

$Cov(g_j, e_{ij}) = 0$

Variances need to be estimated first with one method: OLS, method of the moments, ML, REML, or Bayesian

Random or Fixed effect

- **Mixed models** extend the linear model by allowing a more flexible specification of the errors (and other random factors). Hence, it allows for a different type of inference and also allows to incorporate *correlation* and *heterogeneous variances* between the observations.

- **Fixed effects:** are those factors whose levels are selected by a nonrandom process or whose levels consist of the entire population of possible levels. Inferences are made *only* to those levels included in the study. Hint: all levels of interest are in your data set.

- **Random effects:** a factor where its levels consist of a random sample of levels from a population of possible levels. The inference is about the population of levels, not just the subset of levels included in the study.

- Mixed linear models contain both *random* and *fixed* effects.

The mixed model equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad E\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

X (n x r) design matrix for fixed effects

β (r x 1) vector of fixed effects

Z (n x t) design matrix for random effects

g (t x 1) vector of random effects

e (n x 1) vector of random errors

G (t x t) matrix of variance-covariance of random effects

R (n x n) matrix of variance-covariance of random errors

The mixed model equation

$$\mathbf{G} = \begin{array}{c} \\ g_1 \\ g_2 \\ \cdots \\ g_t \end{array} \begin{array}{c} \begin{array}{cccc} g_1 & g_2 & \cdots & g_t \end{array} \\ \begin{bmatrix} \sigma^2_g & & & 0 \\ & \sigma^2_g & & \\ & & \cdots & \\ 0 & & & \sigma^2_g \end{bmatrix} \end{array} = \sigma^2_g \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \cdots & \\ 0 & & & 1 \end{bmatrix} = \sigma^2_g \mathbf{I}_t$$

$$\mathbf{R} = \begin{array}{c} \\ e_{11} \\ e_{12} \\ \\ e_{rt} \end{array} \begin{array}{c} \begin{array}{cccc} e_{12} & e_{12} & \cdots & e_{rt} \end{array} \\ \begin{bmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \cdots & \\ 0 & & & \sigma^2 \end{bmatrix} \end{array} = \sigma^2 \mathbf{I}_{rt}$$

The mixed model equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad E\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

Assumptions

- Random effects: $E(\mathbf{g}) = \mathbf{0}$, $\mathbf{V}(\mathbf{g}) = \mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$
- Deviations: $E(\mathbf{e}) = \mathbf{0}$, $\mathbf{V}(\mathbf{e}) = \mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$
- **g** and **e** independent.

hence, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$$

Note: normality assumptions can be made about **g** and **e**.

$$\mathbf{g} \sim \text{MVN}(0, \mathbf{G}) \quad \text{and} \quad \mathbf{e} \sim \text{MVN}(0, \mathbf{R})$$

Variance Component Estimation

- Variance components need to be estimated before obtaining estimates of fixed/random effects and performing any type of inference.

$$\hat{\mathbf{G}} = \mathbf{G}(\hat{\theta})$$
$$\hat{\mathbf{R}} = \mathbf{R}(\hat{\theta})$$
$$\Rightarrow \hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}'+\hat{\mathbf{R}}$$

- Restricted/residual maximum likelihood (REML) is a likelihood-based method used to estimate these variance components and is based assuming that both **g** and **e** follow a multivariate normal distribution.
- The REML variance component estimates are later used to estimate the solutions of fixed and random effects.
- Henderson (1950) derived the Mixed Model Equations (MME) to obtain the solutions of all effects:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \qquad \text{BLUE} \rightarrow \text{EBLUE}$$
$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad \text{BLUP} \rightarrow \text{EBLUP}$$

Breeding Value Prediction and BLUP

Definition

- The **average effect** of the parental *alleles* passed to the offspring determine the mean genotypic value of its offspring, or
- The **genetic value** of an individual (or cross) judged by mean value of its progeny.

  - Sum of average effects across loci (theoretical, now molecular).
  - Mean value of offspring (practical).

- Not equivalent concepts if interaction between loci is present or if mating is not at random.

Estimation

- By **BLUP** (Best Linear Unbiased Predictor), i.e. the *prediction* of the random effects from linear mixed models.

Breeding Value Prediction and BLUP

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$\hat{\mathbf{g}}$	vector of random effect predictions.
$\hat{\mathbf{G}}\mathbf{Z}' = \mathbf{C}'$	covariance matrix between observations and random (genetic) effects to be predicted.
$\hat{\mathbf{V}}$	variance-covariance matrix for the observations.
$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	individual observations 'corrected' by fixed effects.

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\hat{g}_i = [\sigma_a^2 / \sigma_p^2] \times (y_i - \bar{y})$$

$$\hat{g}_i = h^2 \times (y_i - \bar{y}) \quad \rightarrow \Delta\text{Gain}$$

Note: the expression changes depending of what model is being evaluated (**y**).

Random or Fixed effect

Hypothesis of interest

Fixed effects: $H_0: \mu_1 = \mu_2 = \ldots = \mu_t$

$H_1: \mu_i \neq \mu_j$ for some i, j in the set 1 … t

(i.e. is there a significant treatment effect)

Fixed effects could be estimated by:

Ordinary Least Square Method (OLS) in ANOVA, or

Best Linear Unbiased Estimator (BLUE) in BLUP/REML

Test statistic:

Usually are F-test or t-test (P-value from ANOVA)

Normally a multiple comparison on the means is performed to find what levels of the fixed factor (treatment) are different and to rank them!

Random or Fixed effect

Hypothesis of interest

Random effects:

$$H_0: \sigma_g^2 = 0$$
$$H_1: \sigma_g^2 > 0$$

(i.e. is there a significant variation due to the random effects)

Variances for random effects can be estimated by:

- OLS in ANOVA
- Maximum Likelihood (ML)
- Restricted/Residual Maximum Likelihood estimator (REML, normally in BLUP)
- Bayesian methods.

Test statistic:

Chi-square through the likelihood ratio test (LRT)

Levels (treatments ~ cultivars) of random factors can be predicted through BLUP in our case. Thus the cultivars can be rank and select.

Testing Variance Component – Random Effects

LRT: likelihood ratio test

•Used to compare nested models and is **valid if the fixed effects are the same** (under REML).

•Examples:

$H_0: \rho = 0$ against $H_0: \rho \neq 0$

$H_0: \sigma^2_g = 0$ against $H_0: \sigma^2_g > 0$

•Test Statistic: $d = 2\,[\,\log L_2 - \log L_1] \sim \chi^2_{r2\text{-}r1}$

$\log L_1$= log(Likelihood model 1)

$\log L_2$= log(Likelihood model 2)

r2-r1 = difference on random terms between model 1 and 2

Hypothesis	**P-value**
Two-sided	$\text{Prob}(\chi^2_{r2\text{-}r1} > d)$
One-sided	$0.5(1 - \text{Prob}(\chi^2_1 \leq d))$

Testing Variance Component – Random Effects

Critical values

$r_2 - r_1$	$\alpha = 0.05$		$\alpha = 0.01$	
Δdf	**Two-sided**	**One-sided**	**Two-sided**	**One-sided**
1	3.84	2.71	6.63	5.41
2	5.99	4.61	9.21	7.82
3	7.81	6.25	11.34	9.84
4	9.49	7.78	13.28	11.67
5	11.07	9.24	15.09	13.39

Goodness-of-fit statistics

- When models are not nested AIC and BIC can be used to select/rank models

$$\text{AIC} = -2\times\text{logL} + 2\times t$$
$$\text{BIC} = -2\times\text{logL} + 2\times t\times\log(v)$$

t number of variance parameters in the model

v residual degrees of freedom, $v = n - p$

Testing Variance Component – Random Effects

Testing Genetic variation

$$H_0: H^2 = 0 \quad \text{against} \quad H_0: H^2 > 0$$

Model with Variety

```
7 LogL= 51.7370     S2= 0.47653E-01     66 df   0.5809      1.000

Source                Model  terms     Gamma     Component    Comp/SE   % C
Variety                  12     12  0.580868      0.276798E-01   1.81   0 P
Variance                 72     66   1.00000      0.476526E-01   5.24   0 P
```

Model without Variety

```
2 LogL= 44.8781     S2= 0.75332E-01     66 df    1.000

Source                Model  terms     Gamma     Component    Comp/SE   % C
Variance                 72     66   1.00000      0.753324E-01   5.74   0 P
```

Testing Variety

$$H_0: \sigma^2_g = 0 \quad \text{against} \quad H_0: \sigma^2_g > 0$$

$$d = 2\,[51.737 - 44.878] = 13.72\,,\ \Delta df = 1$$

$$\chi^2_{0.05} = 2.71,\ \text{p-value} < 0.001$$