Modeling and Data Simulation

By Thomas Lübberstedt, William Beavis, Walter Suza (ISU)

Introduction

The main objective of plant breeding is to develop new cultivars that are genetically superior to those presently available across a range of environmental conditions. However, to a large extent, conventional breeding relies heavily on phenotypic selection and the skill of the breeder. Increasing production of genomic data and better methods to phenotype plants provide an opportunity to evaluate important traits in plants. Computer simulation can help to utilize the large and diverse pool of genetic data to build appropriate models to predict the performance of testcrosses based on pre-existing information, and to compare different and establish optimal selection methods in plant breeding. In this module, computer simulation tools available to plant breeders and geneticists will be introduced. This module will include examples of computer simulation for crop genetic improvement. In the last part of this module, you will learn how to conduct a simple simulation study.

Objectives

- Familiarize with genetic simulation tools
- Familiarize with simulation modeling
- Learn to design simulation experiments for plant breeding

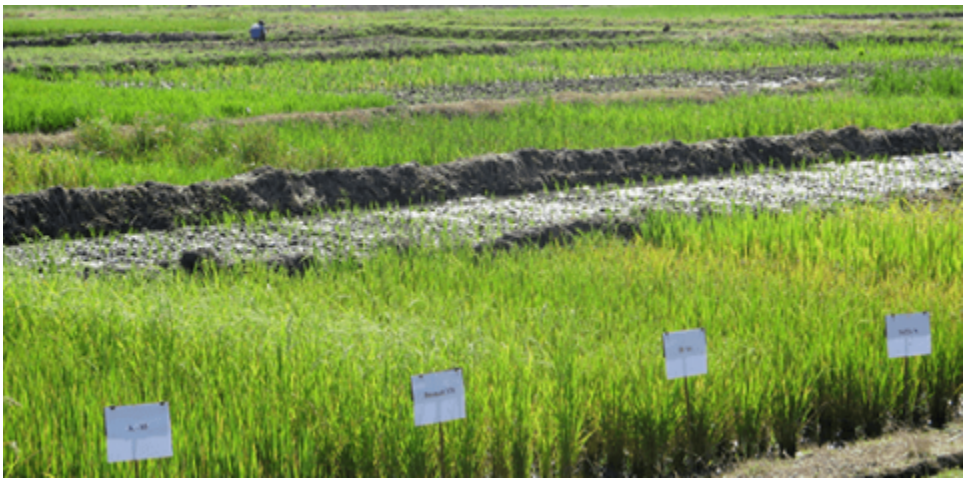

Fig. 1 A field test plot in Uganda. Photo by Iowa State University.

Genetic Simulation Tools

Methods and Processes

Natural or artificial methods and processes are modeled for purposes of predicting unknown outcomes. In plant breeding, simulation models are used to choose among proposed breeding methods because experimental evaluation of breeding methods is time and resource limited. Modern computers are designed to possess greater computational power and data storage space at a reduced price. With the recent explosion in production of genomic data, custom designed programs will provide opportunity for data analysis and simulation to improve plant breeding methods. Examples of publicly available simulation software for plant breeding, software functionality, and assumptions made in modeling are summarized in the next pages.

A. Plabsoft

Plabsoft is a computer program used to analyze data and build simulations based on various mating systems and selection strategies. Plabsoft uses the following model:

$$G = \sum_{S \subseteq N} X_S$$

where:
G = genotypic value
$\mathbf{X_S}$ = genetic haplotype effect at a subset of loci S
N = loci set

An article describing how population simulation and data analysis can be conducted using Plabsoft is found here: http://link.springer.com/content/pdf/10.1007%2Fs10681-007-9493-4

B. QU-GENE

QU-GENE is used to estimate epistatic and G x E effects using the E(N:K) genetic model.

Where:
E = the number of types of environments
N = the number of genes
K = level of epistasis.

Parentheses in the model indicate that different N:K genetic models can be “nested” within types of environments.

More information on the QU-GENE platform is found here: **http://bioinformatics.oxfordjournals.org/content/14/7/632.long**

The following link contains information on use of computer clusters for large QU-GENE simulations: **http://bioinformatics.oxfordjournals.org/content/17/2/194.long**

C. MBP

MBP is used for optimizing resource allocation to maximize genetic gain in breeding of hybrid maize using doubled haploid techniques. MBP uses the following model:

$$\sigma_t^2 = \sigma_{GCA}^2 + \sigma_{SCA}^2/T$$

Equation 1

$$\sigma_t^2 = \sigma_{GCA}^2 + \sigma_{SCA}^2/T$$

where:
σ^2 = estimated genetic variance between test cross progenies
σ_{GCA}^2 and σ_{SCA}^2 = derivatives of additive and dominance variance estimates
T = the number of testers

Read more about MBP software here: http://jhered.oxfordjournals.org/content/99/2/227.full.pdf+html

D. GREGOR

GREGOR is used to predict the mean result of mating and selection in plant breeding. GREGOR is implemented in the MS-DOS environment and does not require use of empirical data. All inputs including individual, trait, and marker data are simulated by the program. GREGOR can create files that are compatible for Mapmaker/Mapmaker QTL programs. **http://jhered.oxfordjournals.org/content/84/3/237.extract**

E. PLABSIM

PLABSIM is used for simulation of marker-assisted backcross methods. **http://jhered.oxfordjournals.org/content/91/1/86.long**

F. GENEFLOW

GENEFLOW provides a platform for determining the nature and structure of genetic diversity by integrating pedigree, genotype, and phenotype data. Simple statistical analyses, such as ANOVA, regression, t tests and correlations are supported in GENEFLOW.

Go to this link to access GENEFLOW: **http://www.geneflowinc.com**

G. COGENFITO

The composite genotype finder tool (COGENFITO) is a web-based program used as a search tool for identification of specific genotypes (Fig. 2).

Fig. 2 COGENFITO is available through MaizeGBD at http://archive.maizegdb.org/Cogenfito.php.

Summary of Functionality and Assumptions of Computer Software Programs

Software: Plabsoft

- **Assumptions**: Absence of selection in the base population; random mating; infinite population size; no crossover interference
- **Models**: Quantitative genetic model; count location model
- **Functionality**: Integrates population genetic analyses and quantitative genetic models for estimating genetic diversity; tests HWE and calculates LD; haplotype-block-finding algorithms to predict hybrid performance

Software: QU-GENE/QuLine

- **Assumptions**: No mutation; no crossover interference; all random terms normally distributed
- **Models**: E(NK) model; Infinitesimal model
- **Functionality**: Employs simple to complex genetic models to mimic inbred breeding programs, including conventional selection and MAS

Software: MBP

- **Assumptions**: Timely staggered breeding cycles; no epistatic and maternal effects; no correlated response in test cross performance; infinite population size to calculate selection intensity
- **Models**: Quantitative genetic model for optimization; Infinitesimal model
- **Functionality**: Optimizes hybrid maize breeding schemes based on DH lines and maximizes the expected genetic gain per year by means of quantitative genetic model calculations under the restriction of a given annual budget

Software: GREGOR

- **Assumptions**: No crossover interference; no epistatic effect
- **Models**: Quantitative genetic model
- **Functionality**: Predicts the average outcome of mating or selection under specific assumptions about gene action, linkage, or allele frequency

Software: PLABISM

- **Assumptions**: No crossover interference
- **Models**: Random-walk algorithm to simulate crossovers during meiosis

- **Functionality**: Simulates marker-assisted introgression of one or two target genes using backcrossing

Software: GENEFLOW

- **Assumptions**: Diploid inheritance
- **Models**: Genotype; Pedigree; Population and Report modules; optional Multiplex and Germplasm
- **Functionality**: Studies nature and structure of genetic diversity

Software: COGENFITO

- **Assumptions**: Maize only
- **Models**: Security modules, Genome model limited to marker maps in MaizeGDB
- **Functionality**: Screens marker data from a given genetic mapping population to identify line with user-defined informative haplotypes

Applying Computer Simulation

Computer Simulations

Computer simulations were used as early as 1957 to solve theoretical problems in population genetics that are intractable using conventional algebraic and statistical approaches (Fraser and Burnell, 1970). Substantial time and field resources are needed to conduct field experiments to compare breeding efficiency from different selection strategies to predict cross performance using available gene information. The power of computer simulation is the ability to sample as many conditions as possible beyond the breeder's capability of solving them by hand. Taking advantage of the speed and efficiency of sampling by computers, breeders have found a tool that can be used to test models and provide more confidence in the performance of the model in the field environment. The major applications of computer simulation in crop genetic improvement are indicated in Fig. 3.

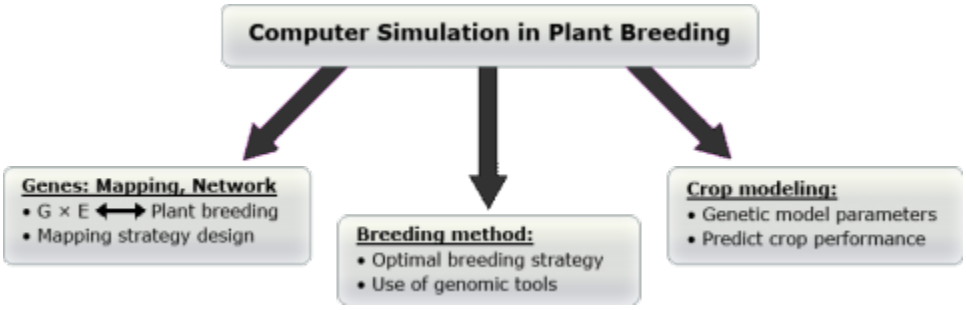

Fig. 3 Applications of computer simulation in crop genetic improvement. Adapted from Li et al., 2012.

Examples of Application of Computer Simulation in Plant Breeding

Example 1: Evaluating plant breeding strategies

Chapman et al. (2003) simulated the S1 recurrent selection method for sorghum in three drought environment types in Australia. The assumption was that 15 genes influence yield in sorghum by controlling several traits including, transpiration efficiency coefficient, flowering time, osmotic adjustment, and stay green traits (Chapman et al. 2003). In this work, QU-GENE was linked with Agricultural Production Systems sIMulator (APSIM) program (Fig. 4) to simulate breeding population and the corresponding trait values for each genotype. As mentioned earlier, QU-GENE helps determine gene effects, G x E interactions, and epistasis (Podlich and Cooper, 1998). Therefore, combining QU-GENE with APSIM helps determine the importance of the interactions detected by QU-GENE on yield in target environments.

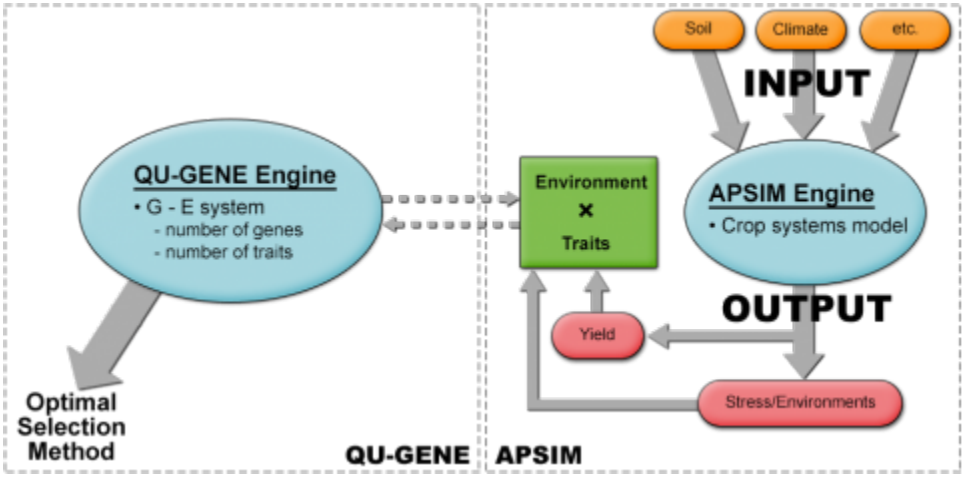

Fig. 4 Linkages between QU-GENE and APSIM for simulation of S1 recurrent selection of sorghum for adaption to drought conditions. Gene information and expression states in target population environments (TPE) are entered in QU-GENE to simulate breeding population and trait values. Trait values are entered into APSIM to predict yield value in TPE. ETs = drought environment types encountered in the target population environments (TPE). MET = multienvironment trial. Adapted from Chapman et al., 2003.

Findings

The data in Fig. 5 suggest that for different combinations of traits being tested in particular environments, the fixation of certain traits may not occur until one or more other traits have been improved. While in Fig. 5a the rate of gene fixation is similar, in Fig. 5b, the genes are fixed at different rates. To the breeder, it is important to fix all desirable alleles at the same rate so that desirable level of homozygosity is attained in earlier generations (See the Crop Genetics module on **Population Genetics**).

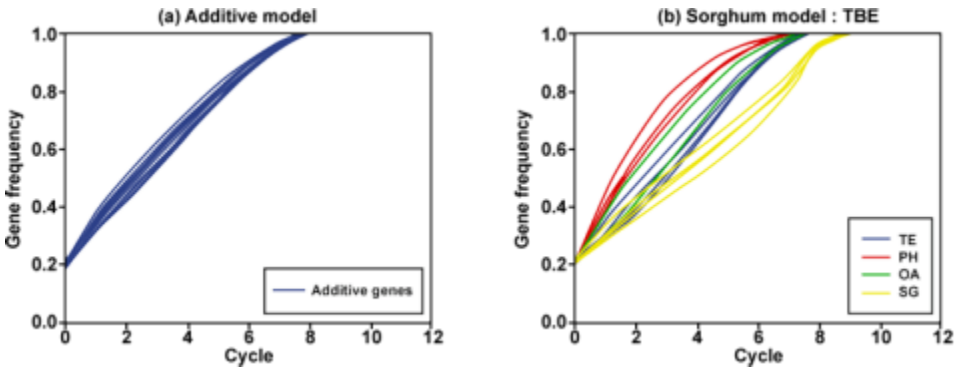

Fig. 5 The rate of fixation of additive alleles for (a) a 15-gene additive model generated by QU-GENE and (b) the 15 additive gene and APSIM model for transpiration efficiency coefficient (TE), flowering time (PH), osmotic adjustment (OA) and stay green (SG) in target population environments (TPE). Adapted from Chapman et al., 2003.

The work by Chapman et al. (2003) can be found here: https://www.crops.org/publications/aj/abstracts/95/1/99

Example 2: Efficiency of Marker-Assisted Selection

Example 2: To study the efficiency of marker-assisted selection

Hospital et al. (1997) investigated the relative efficiency (RE) of marker-assisted selection (MAS) based on an index consisting phenotypic value and molecular score of individuals (**Cluster Analysis, Association & QTL Mapping**). In this example, the phenotypic value of (P_i) of individual **i** was computed as the sum of its genotypic (G_i) and environmental (E_i) values:

$$P_i = G_i + E_i$$

One of the assumptions is that the environmental value is a random normal variable with mean 0 and variance σ^2_E. The genetic value was computed as:

$$G_i = \sum_{q=1}^{nq} x_q \Theta_{iq'}$$

Where:

X_q = effect of QTL *q*

Θ_{iq} = the number of favorable alleles carried by individual *i* at locus *q*

nq = total number of QTL (for this study 25 QTLs were considered)

Finding 1

1. The genetic variances at the QTL in the original F_2 follow a geometric series
2. There is no genetic interference in recombination

Findings

1. The relative efficiency of MAS depends on population size (Fig. 6). At low heritabilities, the larger the population size, the higher the RE of MAS.

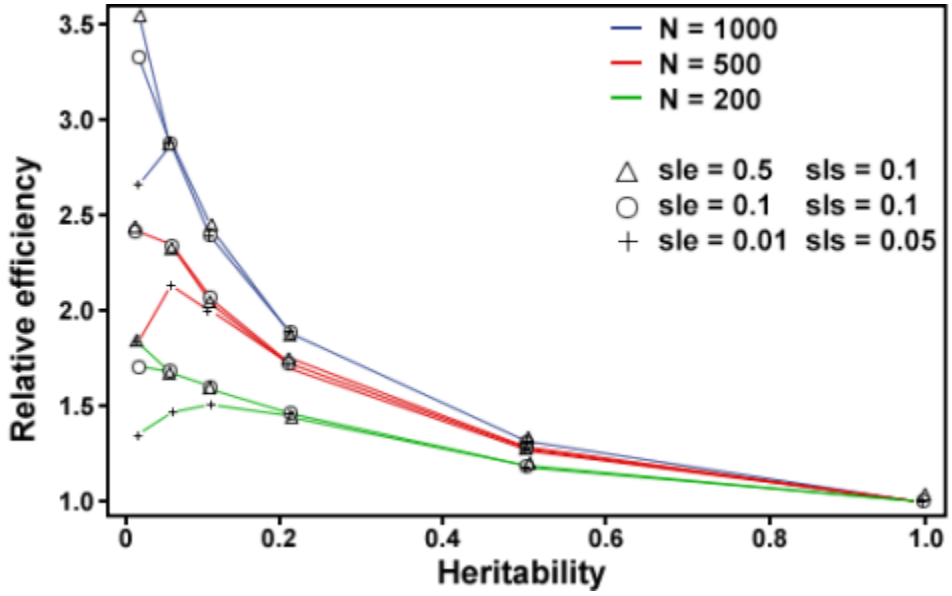

Fig. 6 Relative efficiency of MAS in the first generation. RE is indicated in the y-axis at a different heritability values in the x-axis. Simulations were performed for three population sizes (N), and three significance levels (sle and sls) for each heritability value. Each data point is on average over 300 replicates for N = 1000 and N = 500, and over 1000 replicates for N = 200 Adapted from hospital et al., 1997.

Finding 2

2. MAS is less efficient than phenotypic selection in the long term (Fig. 7).

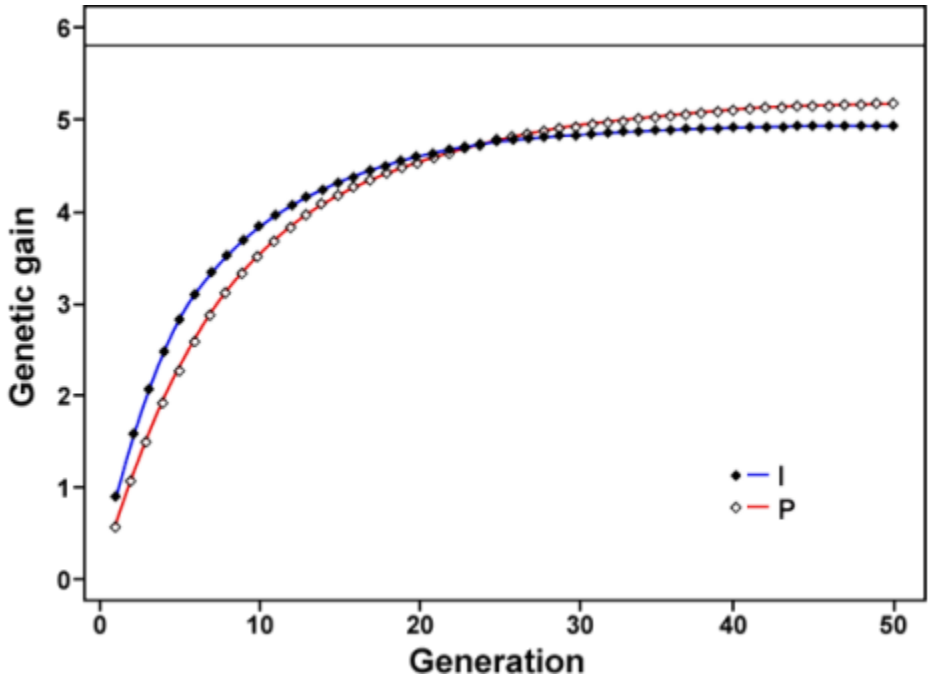

Fig. 7 Responses to phenotypic and MAS over several successive generations. I = marker-phenotype index, and P = phenotypic selection. Horizontal line at y-value 5.82 shows the maximum possible genetic gain for given QTL effects. Adapted from Hospital et al., 1997.

The work by Hospital et al. (1997) can be found here: https://link.springer.com/article/10.1007%2Fs001220050679?LI=true#

How to Design a Simulation Experiment

New Breeding Methods

The future success of plant breeders will depend upon their abilities to propose and evaluate new breeding methods. The motivation to succeed will rely on the breeder's ability to predict cross performance by developing and validating new statistical methods, and evaluating new breeding processes. This will require application of models to simulate the methods or processes, and to evaluate the methods based on appropriate criteria, for example, accuracy, power, precision, efficacy, and efficiency (e.g., genetic gain).

Models are used to represent, describe and quantify natural phenomena, and can be arbitrarily simple depending upon their purpose. For example, consider two cultivars (1 and 2) of a crop species. Our task is to (a) describe how the two cultivars might be the same and/or different, and (b) how to test whether the two cultivars are the same. The following statistical model can be used to compare quantitative differences (e.g., yield) of the two cultivars (Table 1).

$$Y_{ij} = \mu + C_i + \varepsilon_{(i)j}$$

where:

Y_{ij} = observation for the i^{th} cultivar entry at the j^{th} location

μ = an overall mean

C_i =an effect due to the i^{th} cultivar entry

$\varepsilon_{(i)j}$ = a random error associated with the response of i^{th} cultivar entry at the j^{th} location

i = 1

j = 1

Table 1 Observed yield data for cultivars 1 and 2.

Cultivar	1	2	3	4	5	Total	Mean
1	19	14	15	17	20	85	17
2	23	19	19	21	18	100	20
						185	18.5

Assumptions of the Model

Assumptions of the model

1. Effects are additive
2. Errors are normally distributed, homogeneous, and independent

In a field experiment it would be possible, as a result of randomization for all the plots with one of the cultivars to be grouped together in one corner of the experimental plot (Fig. 8). With spatial variability in soil fertility and moisture content possible this might lead to misleading results.

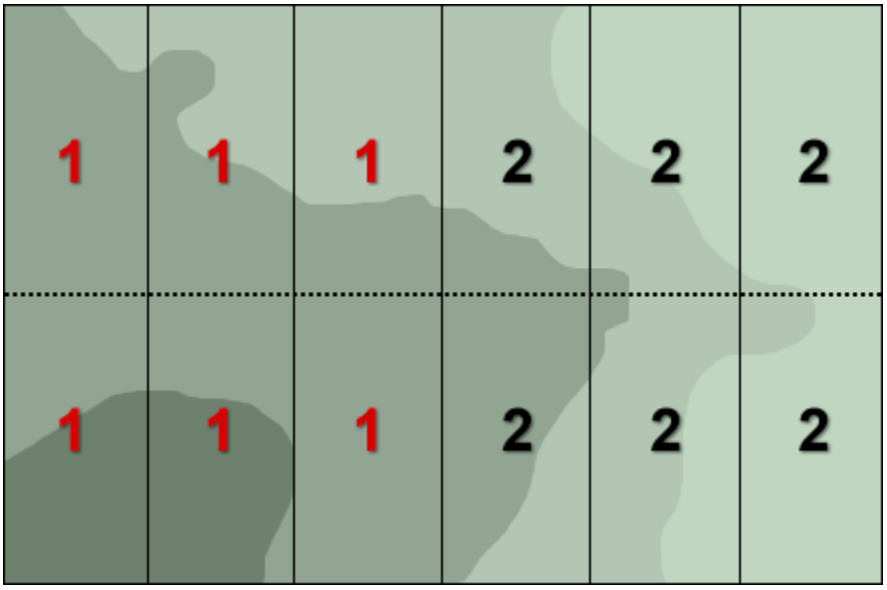

Fig. 8 A soil map describing variability in soil fertility across the test field. Cultivars 1 and 2 are grown in six replications overlaying opposite sides of the field.

One of the remedies to address such spatial field variability (Fig. 8) is to group the units (blocks) such that units in the same group are as similar as possible, and then allocate at random each cultivar to one unit each of the groups (Fig. 9).

New Field Experiment Design

The new design (Fig. 9) allows the application of the following model:

$$Y_{ijk} = \mu + B_j + C_i + \varepsilon_{(ij)k}$$

where:

Y_{ijk} = observation for the k^{th} replicate of the j^{th} block of the i^{th} cultivar

μ = an overall mean

C_i = an effect due to the i^{th} cultivar entry

$\varepsilon_{(ij)k}$ = a random error associated with the response of the k^{th} replicate of the i^{th} cultivar in the j^{th} block

B_j = an effect due to the j^{th} block

k = 1

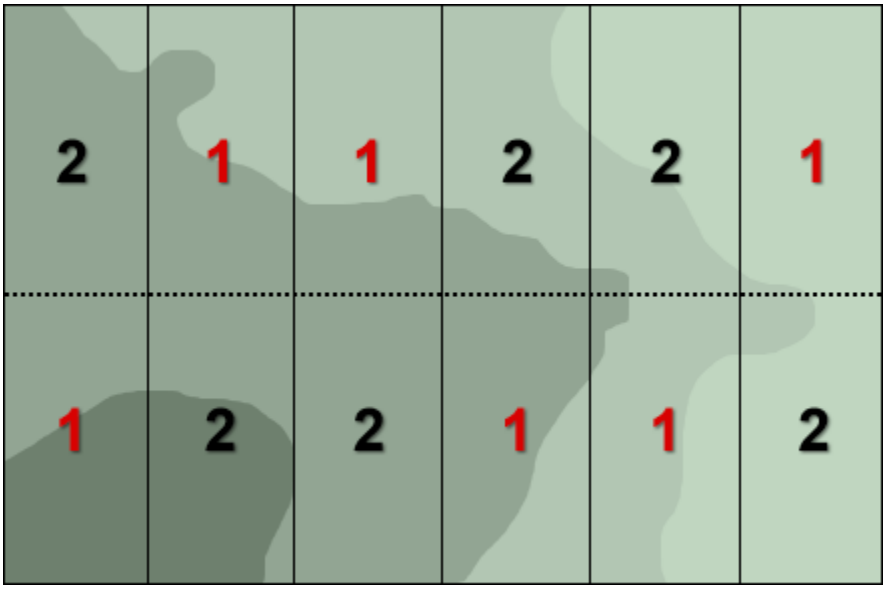

Fig. 9 Possible forms of grouping of plots for cultivar field trials.

An Example for Simulating a Double Haploid (DH) Population in Excel

Goal: create phenotypic values for 30 DH genotypes in Excel; a model for phenotypic performance of these lines includes the population mean, a single gene with additive effect of +1 or -1 (G), equal environmental effect (E = +1) for all 30 DH genotypes, no genotype x environment interactions (GxE), and a normally distributed error.

Thus the model is Phenotype = Mean + Genotype + Environment + GxE + Error.

Excel Exercise:

To create a simulated population of 30 DH genotypes in Excel, these are the steps:

1. In column A (Lines), provide line numbers 1-30. Type a “1” in field A4 and a “2” in A5. Mark both fields with the mouse, and drag down the bottom right corner of the box around fields A4 and A5 to field A33. This will create numbers 1-30 in sequence within this column in fields A4-A33.
2. In column B (Environment, Env): type a “1” in field B4. Mark this field and drag down to B33. All fields will in this case show a value of 1.
3. In column C (Mean): type a “150” (bushels per acre). Proceed like in column B, so that all 30 DH genotypes get the same mean value of 150.
4. In column D (Genotype, G), add the following command in field D4: “= IF(RAND()<0.5,-1,1)”. “RAND()” will generate random numbers in the interval of 0 to 1. Thus, the expression “IF(RAND()<0.5, -1,1)” will generate a value of -1, when a random number below 0.5 is generated (in 50% of the cases). This expression will generate a value of +1 in the other 50% of the cases. By entering this command in field D4, and then dragging down to D33, random numbers -1 or 1 will be added in the fields D4 to D33.
5. In column E (GxE): type a “0” in field E4. Mark this field and drag down to E33. All fields will in this case show a value of 0.
6. In column F (Error), add the following command in field F4: “= NORM.INV(RAND(),0,1)”. his command will create normally distributed random numbers. The Excel NORMINV function calculates the inverse of the Cumulative Normal Distribution Function for a supplied value of x, and a supplied distribution mean (0 in this case) & standard deviation (1 in this case). This information and further useful information on functions in Excel can be found under the Excel “Help function”. When opening this Help function by clicking on the “?” symbol, information on functions can be accessed in various ways, e.g., by searching an alphabetical list of functions.
7. The Phenotype can be determined in column I, by adding the following command in field I4: “=SUM(C4:F4)”. This will add for DH genotype 1 the values in fields C4 to F4, which according to the

model adds up to the Phenotype of this genotype. By dragging down to I33, this summation will be conducted for all 30 DH genotypes.

8. Additional, new simulations of Phenotypes for 30 DH genotypes are obtained by marking fields I4-I33, and copying those into a new column (e.g., K4-K33). By repeating this copy and paste step, multiple sets of 30 DH genotypes can be simulated in short time.

Possible Uses

How could this be used?

Assume, a genetic marker for the gene with additive effect of -1 or +1 is available and co-segregating with that gene. It could be evaluated, how often a t-test would indicate a significant difference between the two genotype classes, in other words, it would enable to determine the power of detecting a gene with this effect, in a DH population of this size. Generally, respective simulation studies can be used to determine the power of detecting a known effect, and thus help to design proper experiments in terms of population size, number of environments, etc. The limitation is, that simulation studies have to make assumptions about unknown effects.

Reflection

The **Module Reflection** appears as the last "task" in each module. The purpose of the Reflection is to enhance your learning and information retention. The questions are designed to help you reflect on the module and obtain instructor feedback on your learning. Submit your answers to the following questions to your instructor.

1. In your own words, write a short summary (< 150 words) for this module.
2. What is the most valuable concept that you learned from the module? Why is this concept valuable to you?
3. What concepts in the module are still unclear/the least clear to you?

References (1)

Chapman, S., M. Cooper, D. Podlich, and G. Hammer. 2001. Evaluating plant breeding strategies by simulating gene action and dryland environment effects. Agron J. 95: 99-113. https://www.crops.org/publications/aj/abstracts/95/1/99

Frisch, M., M. Nohn, and A. E. Melchinger. 2009. PLABSIM: Software for simulation of marker-assisted backcrossing. J. Heredity 91: 86-87. http://jhered.oxfordjournals.org/content/91/1/86.long

Gordillo, G. A., and H. H. Geiger. 2008. MBP (Version 1.0): A software package to optimize maize breeding procedures based on doubled haploid lines. J Heredity 99: 227-231. http://jhered.oxfordjournals.org/content/99/2/227.full.pdf+html

Hospital, F., L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais. 1997. More on efficiency of marker-assisted selection. Theor. App. Genet. 95: 1181-1189. http://link.springer.com/article/10.1007%2Fs001220050679?LI=true#

Jiankang Wang. 2012. Modelling and Simulation of Plant Breeding Strategies, Plant Breeding, Dr. Ibrokhim Abdurakhmonov (Ed.), ISBN: 978-953-307-932-5, InTech, DOI: 10.5772/27863. http://www.intechopen.com/books/plant-breeding/modelling-and-simulation-...

Li, X., C. Zhu., J. Wang, and J. Yu. 2012. Computer simulation in plant breeding. Advances in Agronomy 116: 219-264. http://ac.els-cdn.com/B9780123942777000063/1-s2.0-B9780123942777000063-main.pdf?_tid=2fd828d2-882c-11e2-aabf-00000aab0f27&acdnat=1362773593_725a623d50eb80d897303414217ab007

Maurer, H. P., A. E. Melchinger, and M. Frisch. 2008. Population genetic simulation and data analysis with Plabsoft. Euphytica 161: 133-139. http://link.springer.com/content/pdf/10.1007%2Fs10681-007-9493-4

Micallef, K. P. M. Cooper, and D. W. Podlich. 2001. Using clusters of computers for large QU-GENE simulation experiments. Bioinformatics 17: 194-195. http://bioinformatics.oxfordjournals.org/content/17/2/194.long

Podlich, D. W., and M. Cooper. 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14: 632-653. http://bioinformatics.oxfordjournals.org/content/14/7/632.long

Sun, X., T. Peng., and R. H. Mumm. 2011. The role and basics of computer simulation in support of critical decisions in plant breeding. Mol Breeding 28: 421-436. http://link.springer.com/content/pdf/10.1007%2Fs11032-011-9630-6

Tinker, N. A., and D. E. Mather. 1993. GREGOR: Software for genetic simulation. J. Heredity 84: 237.

http://jhered.oxfordjournals.org/content/84/3/237.extract

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Molecular Plant Breeding Modeling and Data Simulation Author: Thomas Lübberstedt, William Beavis, and Walter Suza (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Lübberstedt, T., W. Beavis, and W. Suza. 2016. Modeling and Data Simulation. *In* Molecular Plant Breeding, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.

Source URL: https://pbea.agron.iastate.edu/course-materials/molecular-plant-breeding/molecular-plant-breeding-0?cover=1