Cheat sheet on how to calculate *predicted* genetic gains

Excellence in Breeding Platform

Donors and stakeholders of breeding organizations expect to receive values for different terms of the breeders' equation every year in order to assess the progress of pipelines towards the goals of greater genetic gains and variety turnover in the target population of environments

This cheat sheet provides a suggested template and guidelines to calculate and report these metrics.

Optimizing breeding schemes

Giovanny Covarrubias-Pazaran
g.covarrubias@cgiar.org

25/08/2021

It is suggested to use a basic spreadsheet that records the following data for each trait and reporting date.

Glossary

STG 1 Stage 1: The first year a cohort of germplasm is tested for yield (in addition to other traits) in a multi-location field trial.

TPE Target population of environments.

SE Selection environments.

KPI Key performance indicator.

Basic details

Reporting Date	Center	Crop	Region	Continent	Pipeline name	Trait name	Trait unit	Year
dd/mm/yyyy	ACRONYM					e.g. Drought tolerance	e.g. score 1-10	2020

> Average breeding cycle time

Cycle time	Idealized cycle time
Age of parents	Time from crossing to recycling

> Estimated genetic breeding value

μ base	μ selected		
STG 1	**STG 1**	**STG 2**	**STG 3**
μ_0	μ_1	μ_2	μ_3

> Selection heritability and accuracy

Heritability (H^2)			SE-TPE correlation		
STG 1	STG 2	STG 3	STG 1	STG 2	STG 3

> Selection intensity

Proportion advanced		
STG 1	STG 2	STG 3

> Genetic variance

Genetic variance (σ^2_g)		
STG 1	STG 2	STG 3

Average breeding cycle time

Cycle time	Idealized cycle time
Mean age of parents at time of crossing	Time from crossing block to recycling

The 1st KPI refers to the average age of the parents (mother and father) of individuals entering STG1.

Example

Consider a line in stage 1 in 2021 with pedigree A/B//C (using Purdy pedigree notation), where the double cross (A/B//C) was made in 2015.

The Mother is an F1 hybrid between two pure lines, here A/B:

A A direct introduction selected as a parent, perhaps due to novel genetic variance for an important trait. For example, this may be a landrace or a line from the germplasm bank. The age is 50 years and A contributes 25% of the lineage of the line in question.

B An introduced line that was selected for the 2015 crossing block after performing very well in the stage 4 trial in 2014. It was in stage 1 in 2011. The age of B is 10 years (2021-2011) and this parent contributes 25% of the lineage of the line in question.

The Father is a pure line, C:

C C is a recycled line. The cross that originated this line was made in 2008. The age of this parent is 2015-2008 = 7 years. This parent contributes 50% of the lineage of the line in question.

The age of this line is calculated as:

$$(0.25 \times 50) + (0.25 \times 10) + (0.5 \times 7)$$
$$= 12.5 + 2.5 + 3.5$$
$$= 18.5 \text{ years}$$

When the age of each line in stage 1 is calculated and averaged across the population of individuals, the mean breeding cycle length can be calculated for the cohort of stage 1 individuals.

Troubleshooting

Is it necessary to obtain the age of each individual in the pedigree at stage 1?

No, only the age of the father and mother is required. In the case that one of the parents is an F1 hybrid between two lines then we need to know the age of the lines. In case of hybrids in clonal crops we just need the age of the mother and father since all individuals are considered hybrids.

What is the age of a parent that is introduced (i.e., not a recycled parent) and does not have any detailed performance data?

This can be regarded as 50 years. For example, it might be a donor line carrying a trait of interest but its estimated breeding value for all traits in the product profile is not known.

What is the age of a parent that is being introduced (i.e., not a recycled parent) after being extensively tested?

The age can be regarded as the current year (the year when that line entered Stage 1).

What is the age of the parent that has been recycled?

This should be calculated as the span of time that it took from making the cross that originated that individual to the time it was recycled.

Performance of parents (Estimated genetic/breeding value)

μ base	μ selected		
STG 1	STG 1	STG 2	STG 3

In order to understand the performance of the individuals selected as parents, we need a reference point. In this case, the reference proposed is the average performance of the entire STG1 population (μ.base column). Then we need to record the performance of the selected parents at STGx (where x is a number greater or equal than 1) for a given trait (e.g., yield) (μ.selected column).

Example

Let us assume that the parents of the crossing block come from individuals selected at STG1, STG2 & STG3 of testing. If we evaluate 1000 individuals in STG1, we must record the mean yield of the entire STG1 population of 1000 individuals (μ.base column), then select a small portion to be moved to STG2. Finally, the mean yield of the even smaller portion selected to be parents in the next crossing block is recorded (μ.selected) column (Figure 1).

Troubleshooting

What is the surrogate of merit that should be reported?

The breeding value is preferred for inbred crops (based on a pedigree or marker-based model), while genetic value is preferred for non-inbred crops. If it is not possible to calculate breeding value, genetic value should suffice. The intercept (overall mean) should be added to the estimate.

How to ensure that values of individuals in STGx+1 are comparable to STG1??

Use of checks to connect years or use the pedigree to connect the data.

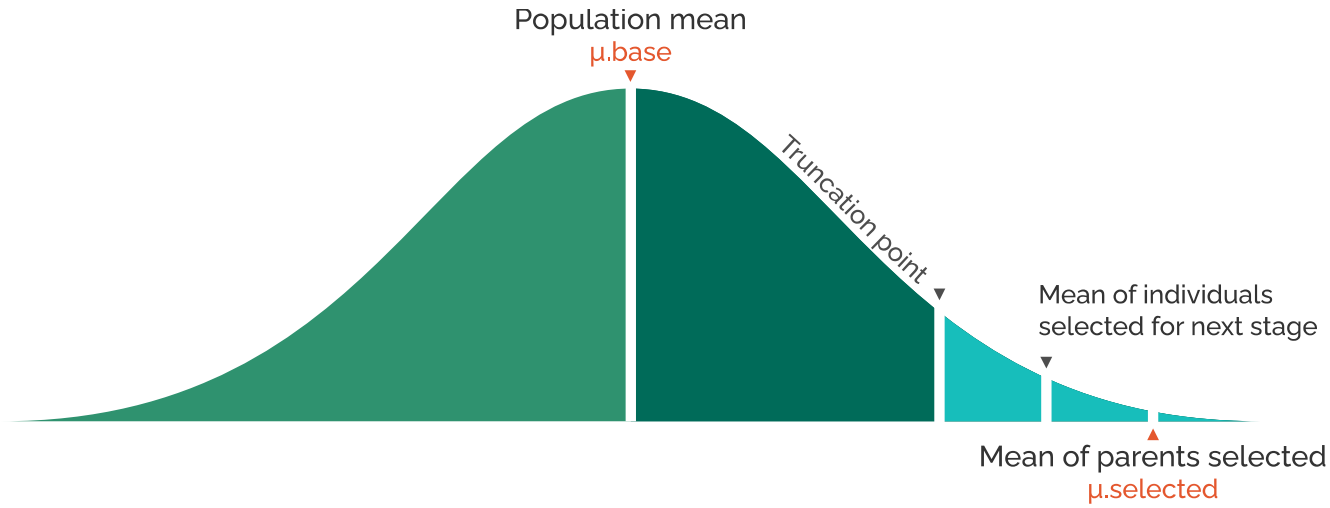

Figure 1. Example of means to be captured for the STGn of evaluation to assess performance.

Broad sense heritability and correlation

Heritability (H^2)			SE-TPE correlation		
STG 1	STG 2	STG 3	STG 1	STG 2	STG 3

Here, we calculate the heritability of trials and the correlation between the sets of trials that collectively make up the "selection environments" and the TPE, characterized by the farms where varieties aim to be grown. Initially, this analysis may only be possible for advanced stages of testing (e.g., stage 3 or 4 tested in the TPE). Ideally, it is expected that both early and late stages take place in the TPE. Report the average of entry-mean heritabilities (or reliability) of the different environments explored at each stage, and record in parentheses the number of environments used in the calculation.

Example

Assume that the STG1 trials were evaluated in two environments with an augmented design and 10% checks. A diagonal mixed model is fitted for the trait of interest.

In ASReml-R v4 language this is written as:

```
Fixed = Trait ~ Env

Random = ~diag(Env):Geno   OR
diag(Env):vm(Geno, A) # A is the
relationship matrix

Residual = ~dsum(~units | Env)
```

We then use the Cullis *et al*. (2006) formula:

$$H^2 = 1 - \frac{PEV_{\mu}}{2\sigma_g^2}$$

Where PEV is the predicted error variance and σ^2_g is the genetic variance in a pedigree-based, marker-based or simple genetic model. An R script for calculating H^2 is available at: gitlab.com/excellenceinbreeding/module2.

The genetic correlation of SE-TPE can be extracted from factor analytic models. If the STGx was performed in the TPE then this correlation is assumed to be 1.

Troubleshooting

Should I report phenotypic or genetic correlation between the TPE and SE?

It is preferable to report genetic correlation. In case this cannot be estimated, phenotypic correlation should be provided.

What if I don't have material connectivity between STG1 and the TPE?

You should provide the same correlation you provide for later stage trials with the TPE but adding a note that this is the case.

How is an average H^2 obtained for each stage?

Request the support from your Biometrician to calculate H^2, using a diagonal model to obtain a H^2 value for each environment where the STGn material was evaluated. The weighted average of these values is reported. It is also possible to use across-environment predictions to obtain a cross-environment H^2.

Which H^2 formula should I use?

The Cullis *et al*. (2006) formula is used to calculate broad or narrow sense heritability with variance components coming from a diagonal model and then average. The variance component changes due to the availability of pedigree, marker or relationship data, and the value will be interpreted as narrow or broad sense heritability. Otherwise, use the regular entry-mean broad-sense heritability based on ratios of variance components. Note which method was used and whether kinship information was used.

Selection intensity

Proportion recycled/tested		
STG 1	STG 2	STG 3
20/1000	15/200	5/20

Proportion advanced/tested		
STG 1	STG 2	STG 3
200/1000	20/200	5/20

Provide, as a ratio, the number of parents recycled relative to the number of individuals tested in each stage (proportion recycled/tested column), and as a ratio the number of individuals advanced from one stage to the other (proportion advanced column).

Example

Assume that STG1 contains 1000 individuals. First, we advance 20% of the best individuals from STG1 to STG2 (best 200), the best 10% from STG2 to STG3 (best 20) and so on: this is captured in the "proportion advanced" columns as a ratio (200/1000 and 20/200). Although the best 20% individuals from STG1 are moved to STG2, it is not certain that all will become parents. E.g., from the best 200 moved from STG1 to STG2, 20 are used as parents in the crossing block.

Genetic variance

Genetic variance (σ^2_g)		
STG 1	STG 2	STG 3
20	15	10

This refers to the average genetic variance across the n environments where the material from STGx was evaluated (where x is a value equal or greater than one).

Example

Assume we evaluated our STG1 trials in two environments with an augmented design and 10% checks. We will fit a diagonal model for the trait of interest.

In ASReml-R v4 language that is:

```
Fixed = Trait ~ Env

Random = ~diag(Env):Geno   OR
diag(Env):vm(Geno, A) # A is the
relationship matrix

Residual = ~dsum(~units, Env)
```

We will then take the variance components for genotypes for each environment and average them to put it in the table.

References

Gustavo de Los Campos, Daniel Sorensen, and Daniel Gianola. "Genomic heritability: what is it?" PLoS Genet 11.5 (2015): e1005048.

Cullis, Brian R., Alison B. Smith, and Neil E. Coombes. "On the design of early generation variety trials with correlated data." Journal of Agricultural, Biological, and Environmental Statistics 11.4 (2006): 381-393.

excellenceinbreeding.org/toolbox/tools/eib-breeding-scheme-optimization-manuals

CGIAR | Excellence in Breeding Platform