Cheat sheet on how to calculate *realized* genetic gains

CGIAR | Excellence in Breeding Platform

Optimizing breeding schemes

In this document we provide recommendations on how to set experiments, organize the data and analyze it in order to provide accurate estimates of realized genetic gain.

Contact: **Giovanny Covarrubias-Pazaran**
g.covarrubias@cgiar.org

Version: 19/03/2021

There are three important aspects to be considered when calculating the realized genetic gain:

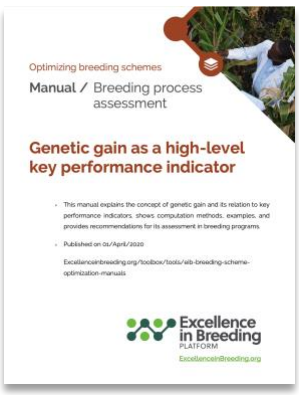
1. The experimental design of the trials (e.g., number of locations, check strategy, etc.).
2. The extraction and organization of the data to be analyzed (stage, pipelines, etc.).
3. The statistical analysis of the data (traits, model, etc.).

In this document, it is assumed that the desired future state is a program that has an experimental design that maximizes connectivity (i.e., with a check strategy), where pipelines are organized by concrete market segments so that data can be easily retrieved by pipeline, where data is analyzed in a way so that connectivity is maximized to account for the genotype by year effect, and where genotype means are accurately adjusted for all other nuisance factors.

It is also assumed that the calculation of realized genetic gain uses data generated by the regular stage-gate process and no additional trials are run (i.e., era trials) but that analysis of historical data is the standard.

The future and current state may differ. For example, a program may not be following the recommendations in this document yet (e.g., not using a check strategy, the proper design, etc.) but can use this as reference to know where to go in the future and how to analyze the current data available.

For a more thorough review of the topics discussed in this document, see the EiB manual on *Genetic gains as a key performance indicator*.



Annex 1. Recommendations for a proper experimental design

A regular breeding program follows a stage-gate approach when it comes to testing for population improvement (early stages) and advancement of products (late stages) (**Figure 1**). Below, we present a protocol on how to maximize connectivity in a trial run for a particular stage to maximize connectivity.

Figure 1. Graphical representation of an example stage-gate approach used in a plant breeding program.

1. Generate the list of materials (entry list) to be tested in each stage and extracted from the database (inventory). The entry list should include:
    a. The list of *new materials* (current cohorts) to be tested.
    b. At least two *genetic gain checks* that are as most "dynamically stable" as possible (genotypes that should be present in all the trials run by the program despite of the region or time).
    c. At least three to four *current checks* (commercial varieties or competing products currently being grown in the wide target market and to be beaten according to the replacement strategy in the product profile, and that show a GxE pattern similar to the cohorts being tested) that are renewed at the rate of one per year.
    d. Add the number of *local checks* needed [materials locally grown due to their specific preference (e.g., resistance, performance, etc.)] that can be renewed as desired.
2. Generate the field books for each testing stage making sure you do the following:
    a. Move the entry list including the new materials and all types of checks mentioned in point 1 to the field book.
    b. Generate the experimental design required for the testing stage ensuring that all checks have at least 2 replicates (We recommend augmented and p-rep designs for early and intermediate generation variety trials and alpha designs for late-stage trials). 5% to 10% of plots allocated to checks in each trial will suffice.
    c. Try to cover as many locations of the target population of environments (TPE) as possible spreading your seed packets using sparse testing methodologies if the seed supply is limited. Checks should be present in all locations (not follow sparsity).
3. Run the trials and store the phenotypic data collected in the database for posterior analysis.

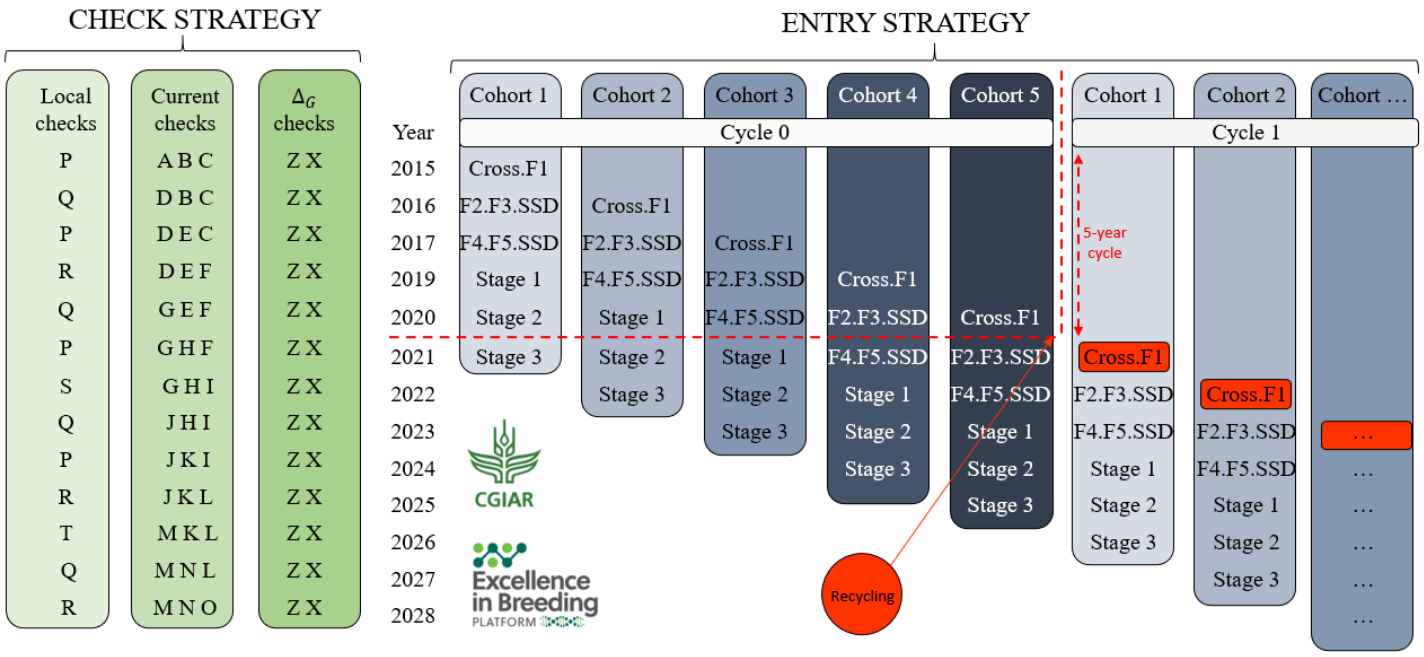


Figure 2. Graphical representation of an optimal check strategy to maximize connectivity among breeding trials across years. On the rights a five-year cycle time program (recycling at stage 2) is shown yielding five cohorts (columns) across multiple years (rows) and stages within a year. On the left, a check strategy where steady genetic gain checks (present in all trials and years), variable current checks (renewed at the rate of one per year), and local checks is displayed.

Annex 2. Recommendations for a proper organization and extraction of the data

The data produced for genetic gain calculation can come from the regular stage-gate testing approach or trials explicitly run for this purpose. Below, we provide some recommendations for users to know how to structure and deal with the data before moving to the analysis.

1. Data Generation

Analysis of historical data

1. Approach your database manager
2. Gather data records from the pipeline of interest
3. Split the data by market segment targeted (TPE + product features) (e.g., Drought conditions environments late maturity).
4. Keep the split data for a time period of interest (e.g., last 10 years).
5. Keep the split data for a given germplasm stage(s) of interest.[1]
6. Keep the split data only for the trait(s) of interest (e.g., yield, or an index of traits if economic weights are available).
7. Hand the batches of data to your Biometrician.

Analysis of era trial data

1. Go to your germplasm bank manager.
2. Gather the list of materials for the pipeline you are interested in.
3. Split the list of materials by market segment targeted (TPE + product features) (e.g., Drought conditions environments late maturity).
4. Keep the list of materials for the time period of interest (e.g., last 10 years).
5. Keep the list of materials for a given germplasm stage(s) of interest.
6. Decide the traits to phenotypes in the experiment and if will be considered by separate or in an index (if economic weights are available).
7. Hand the list of materials to your Biometrician to run an experimental design.
8. Plant the experiment and harvest the phenotypes.

[1] From product development perspective use late-stage trials (e.g., Stage3 & On-Farm). From a population improvement perspective use early-stage trials (e.g., Stage 1 & Stage 2)

2. Analysis

1. Ask your Biometrician to fit a multi environment trial (MET) analysis with a mixed model accounting for nuisance (spatial, environment, year) and genotype (genotype, genotype by year, genotype by location, genotype by year by location) factors. Extract adjusted means for all genotypes.
2. Ask your Biometrician to merge the year of origin of the genotypes to their adjusted means produced from the MET analysis to fit a linear model of the form adjusted.mean~year.origin.
3. Take the slope of the regression and its standard error (to measure uncertainty). Present these two metrics as the realized genetic gain to stakeholders (e.g., donors and other organizations) and plot the regression to present it as the genetic trend with confidence intervals.

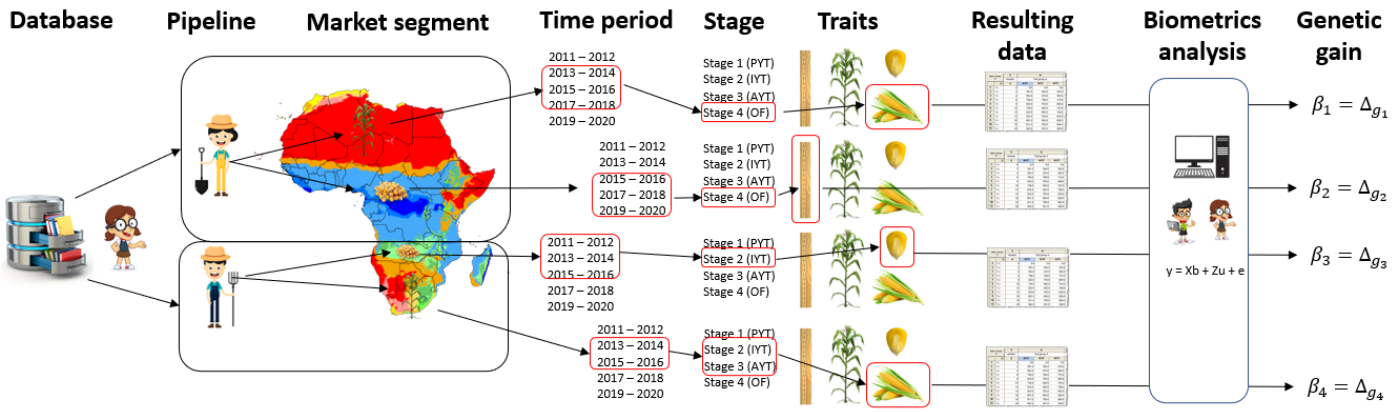


Figure 3. Graphical representation of major steps to calculate realized genetic gain using historical data. It is critical to split the data in a meaningful way to interpret properly the resulting value of genetic gain.

Annex 3. Recommendations for phenotypic analysis to calculate genetic gain

Even when trial data is produced and stored properly, proper statistical analysis is required to ensure an accurate estimate of realized genetic gain and trend. Below, we propose the following steps to perform such statistical analysis.

1. Retrieve the trial data as described in **Annex 2**. That is, extract the data for a specific pipeline targeting one or more market segments, and subset the data for specific stages of testing (preferably Stage 3 and/or Stage 4) and time period (minimum 5 years). Decide the trait to focus on for the analysis.
2. Perform a single year single location analysis to remove outliers and typos in the data.
3. Perform a single year single location analysis to identify trials with low h^2 and remove experiments with h^2 lower than 0.2.
4. With the data cleaned perform a one-stage analysis to fit all genetic and nuisance terms in order to obtain adjusted means for all genotypes. In ASReml-R language the model has the form:

```
model <- asreml(data= yourData,
fixed =    Trait ~ Location + Year
           + Genotype +
           isCheckAsFactor,
random =   ~ Genotype:Year +
           Genotype:Location +
           Genotype:Year:Location
           + YearAsFactor,
residuals = ~ dsum(~units |
           Year:Location),
)
```

   Genotypes should be fit as fixed to properly calculate the genetic value. Year (numeric format) will account for the non-genetic trend. The random terms will ensure that year effects are accounted for and that across-years-locations predictions are adjusted.
5. Obtain adjusted means for all genotypes in the historical datasets. In ASReml language that is:

```
predictions <- predict(model,
classify= "Genotype")$pvals
```

6. Merge the adjusted means (predictions) with the year of origin of the material and fit a model to calculate the genetic gain. In ASReml language that is:

```
finalModel <- asreml(data=
predictions,
fixed = predicted.value ~
        yearOfOrigin    )
```

7. The slope from the final model can be interpreted as the rate of response to selection or genetic gain.

The connectivity of the analysis can be improved by using data from multiple testing stages (e.g., Stage 1 and Stage 2 together) as opposed to data from a single testing stage. Using the pedigree or relationship to fit the model can increase the connectivity of the data. Sample scripts can be found at: gitlab.com/excellenceinbreeding/module2.