**Strategies to Improve Adaptation in Plant Breeding with Complex Genotype by Environment Interactions**

The objective of this project is to devise and evaluate various experimental design strategies used in plant breeding. In particular, we are interested in identifying how many replicates of each genotype should be grown in each environment to maximise the efficiency of evaluation and thence gauge the extent of genotype by environment (GxE) interaction. If we can do this well, then we will i) increase the accuracy of our selections and increase genetic gain, but also ii) improve the adaptation of plant genotypes amid climate change.

Take a look at the breeders equation and explain why improving the accuracy will result in higher genetic gain. Take a look at this paper <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1129591/full>  
It talks about the accuracy of selection being equal to the accuracy of the predictions multiplied by the alignment between the dataset and the target population of environments (TPE). The TPE represents the set of environments that plant breeders target for production, i.e. the farming conditions where their genotypes are well-adapted and perform well.

The designs you will be looking at are usually referred to as “sparse testing”, which means that not all genotypes are grown in all environments. Take a look at these papers:  
<https://academic.oup.com/g3journal/article/10/8/2725/6048674>

<https://www.mdpi.com/2073-4425/14/4/927> Figure 1 in particular

<https://link.springer.com/article/10.1007/s00122-022-04085-0>

The general steps of the project are (for you to shape/refine):

**Step 1: Devise a set of experimental design (sparse testing) strategies to evaluate**

* For example, assume there are genotypes evaluated in environment, with replicate in this environment and plots in total. This strategy would look like the following (shaded means grown, unshaded means not grown):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Environment 1 | Environment 2 | Environment 3 | Environment 4 | Environment 5 |
| Genotype 1 |  |  |  |  |  |
| Genotype 2 |  |  |  |  |  |
| Genotype 3 |  |  |  |  |  |
| Genotype 4 |  |  |  |  |  |
| Genotype 5 |  |  |  |  |  |
| … |  |  |  |  |  |
| Genotype 999 |  |  |  |  |  |
| Genotype 1000 |  |  |  |  |  |
| Total plots | 1000 |  |  |  |  |

All genotypes are allocated to environment 1.

* Lets say we now go to 2 environments, but still only have one replicate of each genotype. The next strategy may look like the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Environment 1 | Environment 2 | Environment 3 | Environment 4 | Environment 5 |
| Genotype 1 |  |  |  |  |  |
| Genotype 2 |  |  |  |  |  |
| Genotype 3 |  |  |  |  |  |
| Genotype 4 |  |  |  |  |  |
| Genotype 5 |  |  |  |  |  |
| … |  |  |  |  |  |
| Genotype 999 |  |  |  |  |  |
| Genotype 1000 |  |  |  |  |  |
| Total plots | 500 | 500 |  |  |  |

Half of the genotypes are allocated to each environment.

* If we now go to 4 environments, the next strategy may look like the following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Environment 1 | Environment 2 | Environment 3 | Environment 4 | Environment 5 |
| Genotype 1 |  |  |  |  |  |
| Genotype 2 |  |  |  |  |  |
| Genotype 3 |  |  |  |  |  |
| Genotype 4 |  |  |  |  |  |
| Genotype 5 |  |  |  |  |  |
| … |  |  |  |  |  |
| Genotype 999 |  |  |  |  |  |
| Genotype 1000 |  |  |  |  |  |
| Total plots | 250 | 250 | 250 | 250 |  |

A quarter of the genotypes are allocated to each environment.

* By going to more environments you are sampling more of the TPE, and hence capturing more of the variability in farming conditions. Generally, the more environments you sample, the better your selection accuracy will become, up to a point where there is little improvement with more environments.
* You can do the same for 5, 10 and 20 environments, with 1/5, 1/10 and 1/20 of the genotypes in each environment. This is a good start.
* An extension of this would be if we have plots in total, so two replicates of each genotype. The question becomes; should we have a balance between replication within environments and going to more environments? This is a question you will try answer.

**Step 2: Design a simulation to produce relevant data**

* Simulate genetic values that capture GxE interaction
* Simulate plot errors that capture spatial trend
* Generate phenotypes by combining genetic values and plot errors according to the various sparse testing strategies devised in Step 1.

**Step 3: Calculate expected accuracies for each strategy**

* Line-mean (average genotype effect across environments) and line-specific (genotype effect within environments) accuracies

**Step 4: Fit various machine learning and statistical models to obtain predictions of genotype effects**

Some examples include:

* Best linear unbiased prediction (BLUP)
* Pedigree best linear unbiased prediction (PBLUP)
* Genomic best linear unbiased prediction (GBLUP)
* Factor analytic genomic best linear unbiased prediction (FA-GBLUP)
* Machine learning

Note that BLUP is fitted within a linear mixed model

**Step 5: Make recommendations of best sparse testing strategy and prediction model**