Checklist for Genomic Prediction

# Review and QC field trial data

Purpose:

* Download the data
* Become familiar with data
* Check it to ensure all variables are within expected ranges.
* Make preliminary choices about the data to use for GS.
* Generate hypotheses about the sources of variation in the data.

Inputs:

* "Raw" field trial data

Expected outputs:

* "Cleaned" field trial data
* Hypotheses about sources of variation in the data

## Checklist: Cleaned data

* Download phenotypes/metadata from DB in reproducible / recorded way
* Population / Group
* Convert plant-basis to plot-basis
* Make sure the experimental design variables are all used in a homogenous way that is compatible with intended downstream analysis
	+ Check that "numberBlocks" reported on DB matches observed "Nblock"
	+ Check that "numberReps" reported on DB matches observed "Nrep"
* Select trials, if necessary, using reproducible methods
* Check Meta data for: row/col designs, plot spacing/size/numb. Planted info
* For the trials where >0 plants were harvested, did any have plots with more plants harvested than were planted / expected to be harvested?
	+ If many were above the expected NOHAV, it might indicate the design of the trial is different than meta-data indicated, warranting additional checking.
	+ Set missing when MaxNOHAV>ExpectedMaxNOHAV
* Traits and TraitAbbreviations
* QC Trait values
	+ Set missing Disease Severity values <1 or >5
	+ Set missing disease incidence <0 or >1
	+ DM
		- Set missing DM==0
		- Set missing DM >65 and/or <4 (alternative)
		- Check which methods were used to measure this. For plots with multiple methods used, decide on a way to choose between them or average them. Create a consensus column for “DM”, which should have a minimum of missing values. Remove all other DM traits.
	+ Yield traits (RTWT, RTNO, SHTWT, FYLD, TOPYLD)
		- Set missing when: Value==0 | NOHAV==0 | is.na(NOHAV)
		- NOHAV==0
	+ Remove non-integer values of RTNO
* Assign genos to phenos
	+ Genomic prediction relies on the genotyped samples in the phenotype data. By whatever means, must maximize the number of plots with SNP marker records *and* ensure only one marker record points to each plot/germplasmName.
* Check the genotyping rate for each trial
	+ Calculate the prop. Of the unique “germplasmName” in the DB also have at least one genotyping record (e.g. GBS sample)
	+ Make sure genotyping rates are as you expect
	+ Consider excluding trials with very low genotyping rates
* Select one DNA sample per “germplasmName"
* Often there are more than one.
* It may actually be ok for one DNA sample to point to multiple germplasmName (not all clone names are successfully merged / synonymized)
* But ideally we don’t want many DNA pointing to same data point, for downstream mixed modeling.
* Calculate Additional Traits
	+ Harvest Index
	+ PerArea Calculations (FYLD and TOPYLD)
		- FYLD=RTWT/(ExpectedMaxNOHAV\_netPlot\*m2\_perPlant)\*10
		- TOPYLD=SHTWT/(ExpectedMaxNOHAV\_netPlot\*m2\_perPlant)\*10
	+ Season-wide mean pest/disease incidence/severity
* Remove outliers
	+ Extreme values for FYLD (e.g. FYLD>100, TOPYLD>300)
* Save a “cleaned trial” dataset

# Preliminary Analysis of Field data

A.k.a Exploratory Data Analysis phase

Purpose:

* Test hypotheses generated in the previous step.
* Detection of statistical outliers
* Analyze each trial to check for genetic vs. non-genetic variability (H2)
* Determine the "best" model for the data

Inputs:

* "Cleaned" field trial data

Expected outputs:

* "Curated" field trial data (e.g. trials w/o variability removed)
* A plan for modelling the data during genomic prediction
* Based on computation requirements and the size / structure of the data, decide whether to use one- or two-stage prediction downstream?
* Possibly: BLUPs for validation prediction accuracy
* If planning two-step genomic prediction: BLUPs, de-regressed BLUPs, PEV, Weights for second-stage

## Checklist: Curated data

* Log-transform yield traits (e.g. FYLD is almost always heteroskedastic unless you do this). Change trait-name to e.g. logFYLD so this can’t be forgotten.
* Analyze each trial.
	+ Check for and remove outliers
		- Example: Studentized residual>4
	+ Re-run “outlier-free” models.
	+ Estimate genetic and error variance for each trial.
	+ Consider removing trials or investigating them further.
		- Example: Remove trials where p-value for chi-square on genetic variance was <= 0.2.
* Save “curated” dataset
	+ “Curated” == only trials and data points we have decided to keep for analysis

# Review and QC of SNP marker data

Purpose:

* All phenotypic records which *should* be genotyped, have one DNA record ID assigned.
* Remove “poor” or “marginal” SNPs post-imputation
* Assess the population genetic structure, esp. any divergence between "training" samples and selection candidates ("test")
* Construct a genomic relationship (kinship) matrix

Inputs:

* Imputed allele-dosage matrix
* "Cleaned" field trial data (list of sample names and other relevant meta-data)

Expected outputs:

* Filtered SNP dosage matrix
* Kinship matrix for prediction

## Checklist: Kinship matrix

* Remove extraneous samples
	+ E.g. If no phenotypes *and* not selection candidates
* Remove SNP with low imputation accuracy (AR2<0.3)
* Remove SNP with low MAF (<1% MAF)
* PCA on SNP matrix
	+ Check variance explained by first several PCs.
		- Extremely high variance explained by the first PC could indicate a batch effect from genotyping or imputation and should be checked.
	+ Make a scatter plot of PC scores, usually on PC1 vs. PC2 and PC3 vs. PC4. Color coding where relevant. For example, to highlight the training population vs. the selection candidates.
		- If TP and candidates appear highly diverged, consideration about whether/how to proceed should be made. Accuracy is likely to be poor.
	+ Save key results
* Construct genomic relationship matrix and save.

# Evaluate genomic prediction accuracy

Purpose:

* Estimate and maximize the expected accuracy of prediction
* Decide whether to include a trait using GS
* Tests to do depend on the prediction scenario. See checklist.

Inputs:

* Kinship matrix
* Curated trial data
* Possibly BLUPs

Expected outputs:

* Prediction model with best accuracy selected
* Selection accuracy for each SI trait evaluated
* Information on h2/accuracy prompt alteration in selection plan? Whether to include a trait we can't predict well?

## Checklist: Prediction accuracy

GEBV vs GETGV?

Cross-validation

Cross-generation prediction

* + GEBV or GETGV?
	+ Predicting progeny?
		- Data to test cross-generation prediction accuracy?
	+ Cross-validation with each dataset/population to assess the maximum accuracy of prediction for un-evaluated candidates in that dataset/population.

# Genomic prediction

Purpose:

* Generate GEBV for selection candidates and make selections for crossing block.

Inputs:

* Curated trial data
* Kinship matrix

Expected outputs:

* GEBV
* h2 and related model outputs
* Selection index
* Selections

## Checklist: Get GEBVs

# Genomic selection

Purpose:

* Generate GEBV for selection candidates and make selections for crossing block.

Inputs:

* Curated trial data
* Kinship matrix

Expected outputs:

* GEBV
* h2 and related model outputs
* Selection index
* Selections

## Checklist: Selection Indices

# Final Checklist: Cassavabase