Genomic mate selection for outbred clonal crops: predicting offspring variance in additive and total merit

Marnin D. Wolfe[1], NextGen Cassava Breeding, Jean-Luc Jannink[1,2]

1. Section on Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA; 2. USDA-ARS, R.W. Holley Center for Agriculture and Health, Ithaca, NY, USA;

MOTIVATION & OBJECTIVES

Diverse crops ranging from staples (e.g., cassava) to cash crops (e.g., cacao) are both outbred and clonally propagated. In these crops, exceptional genotypes can be immortalized and commercialized as clonal varieties. Genomic truncation selection (TS) evaluates parents based on breeding value (i.e. the mean value of their *unselected* offspring). Predictions can include non-additive effects in clonal crops to select candidates with high total genetic merit for variety development **(Wolfe et al. 2016)**. Improvements over truncation selection are possible by selecting crosses instead of parents. By predicting the variance in a cross using a **genetic map**, **phased haplotypes**, and **genome-wide marker effects (Lehermeier et al. 2017; Allier et al. 2019; Bijma et al. 2020)** mate-selection criteria like the mean of *selected* offspring (aka the "usefulness criterion", UC) can be derived**.**

Overall objective: improve on TS by deriving optimizing schemes for population improvement (mating) and clone testing efforts (variety development).

As a first contribution, in this poster, we:

1. Retrospectively analyze empirical data comparing predicted and realized variances from a cassava genomic selection program
2. Prospectively evaluate the interest of possible future crosses in terms of additive and total merit

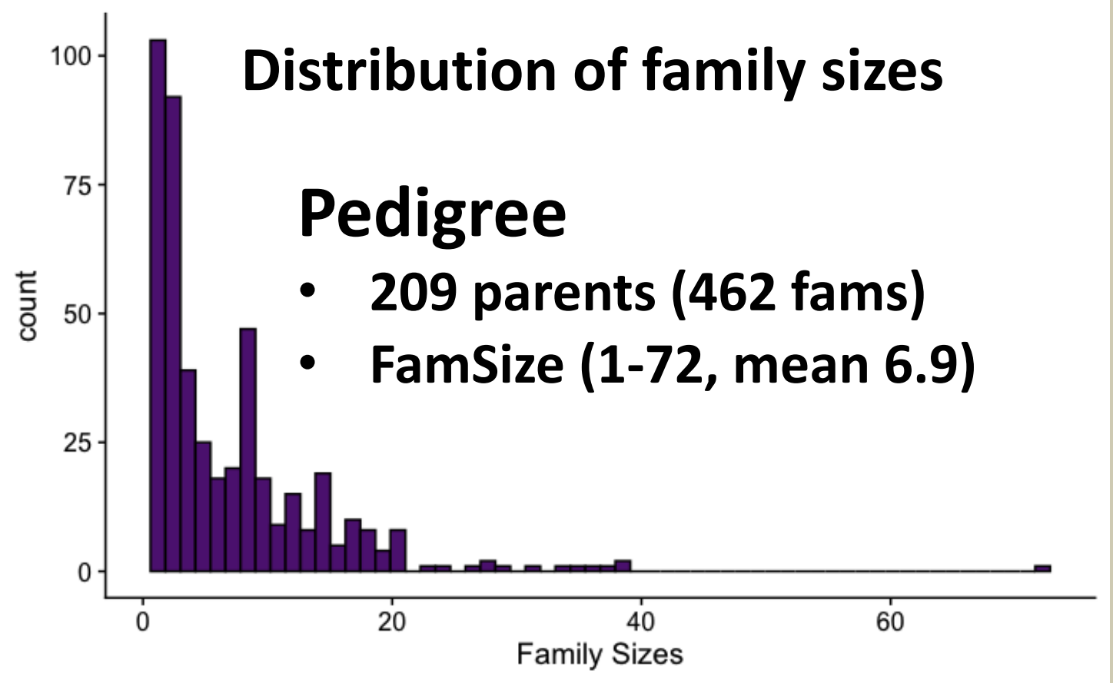
Pedigree, Haplotypes, Recombination and Training Data

Data from Chan et al. 2020. *In Review*.

Curation, Imputation and Phasing Details

- Pedigree verified by AlphaAssign (Whalen et al. 2018)
- Technical replicate GBS samples checked with BIGRED (Chan et al. 2018) and reads merged
- Keep sites <70% missing data and mean read depth<120
- Keep individuals <80% missing
- Impute/phase with SHAPEIT2 → duoHMM (O'Connell et al. 2014).

Cycle	Nentries	Nparents	Nfamilies
C0	9	14	9
C1	1524	75	120
C2	1196	86	198
C3	470	77	137



Distribution of family sizes

Pedigree

- **209 parents (462 fams)**
- **FamSize (1-72, mean 6.9)**

Genotyping data

Derived from genotyping-by-sequencing (GBS)

- **3199 clones**
- **23657 SNPs**

Phenotype data

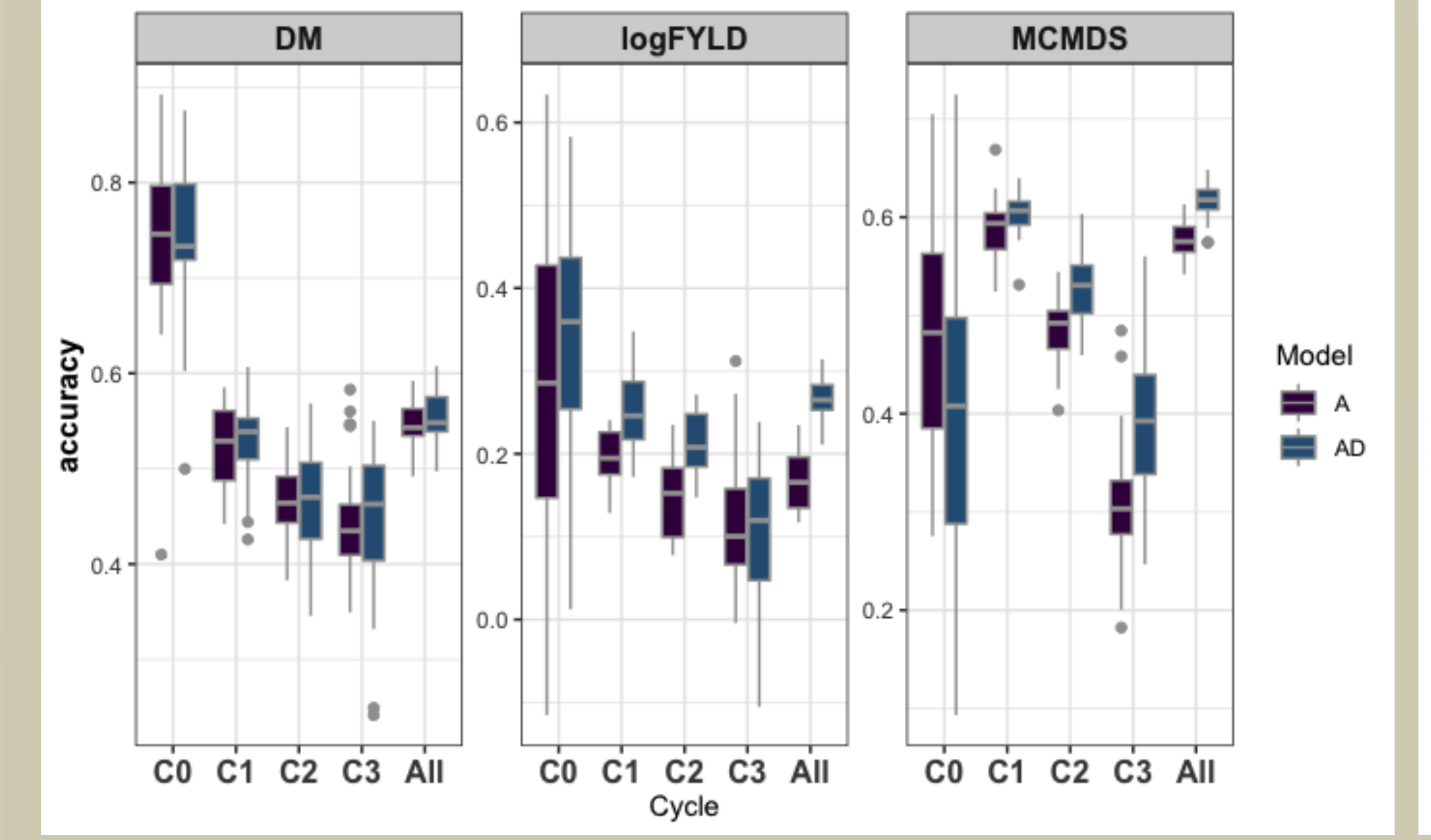
BLUPs from IITA cassava breeding, 2012 to present

Details, code and data: https://wolfemd.github.io/IITA_2019GS/

TraitAbbrev.	Trait	H^2
DM	% Dry Matter	0.44
logFYLD	log(Fresh Root Yield) tons per hectare	0.47
MCMDS	Mosaic Disease Season-wide mean severity score (scale: 1 to 5)	0.76

Including dominance consistently improves prediction accuracy

5-fold cross-validation replicated 5 times per group

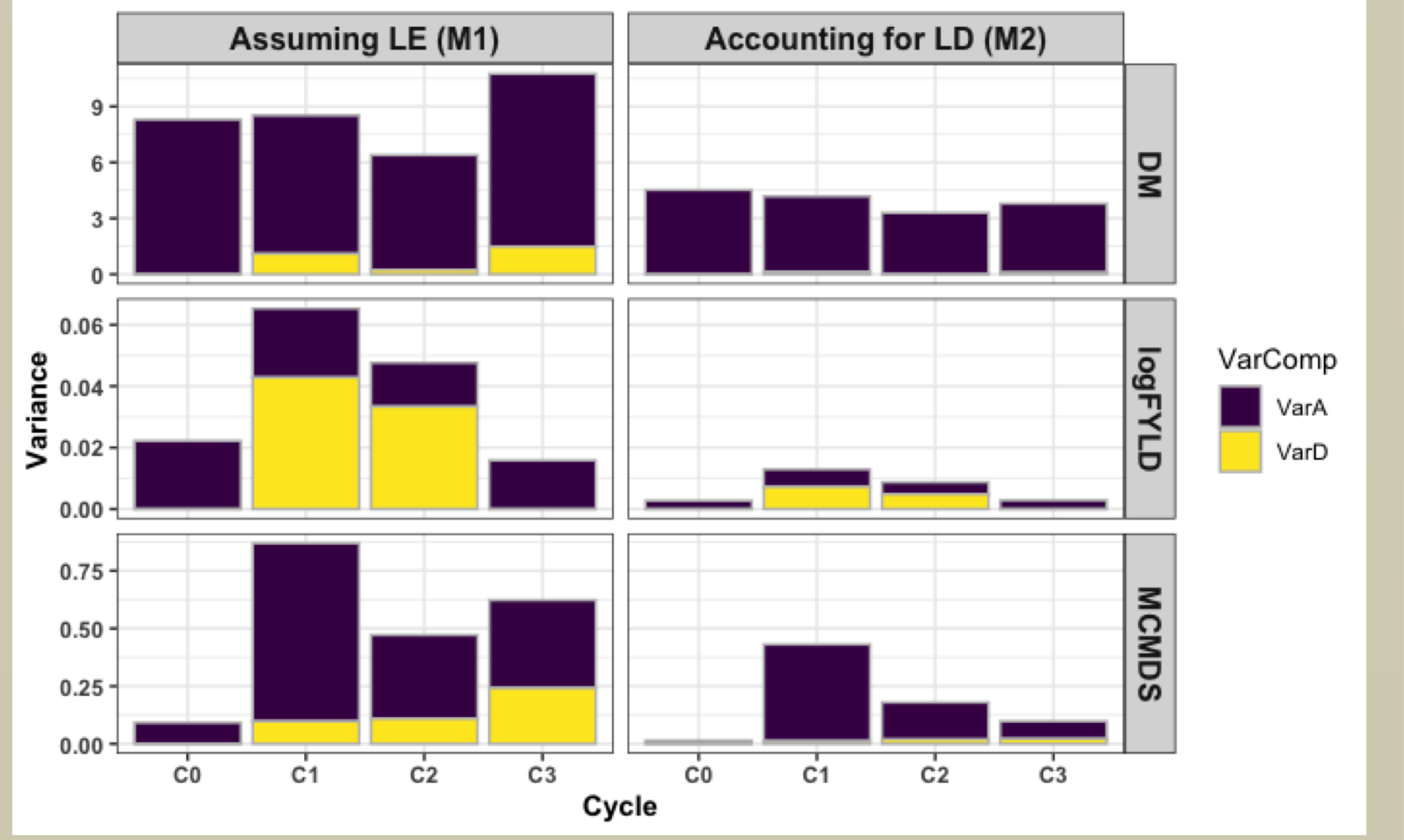


- GBLUP using *sommer* mixed-model solver in R.
- A = Model with additive component only
- AD = Model with additive + dominance component

Large dominance variance in GS progeny

Even after accounting for LD using M2.

M2 matches method of cross variance prediction.

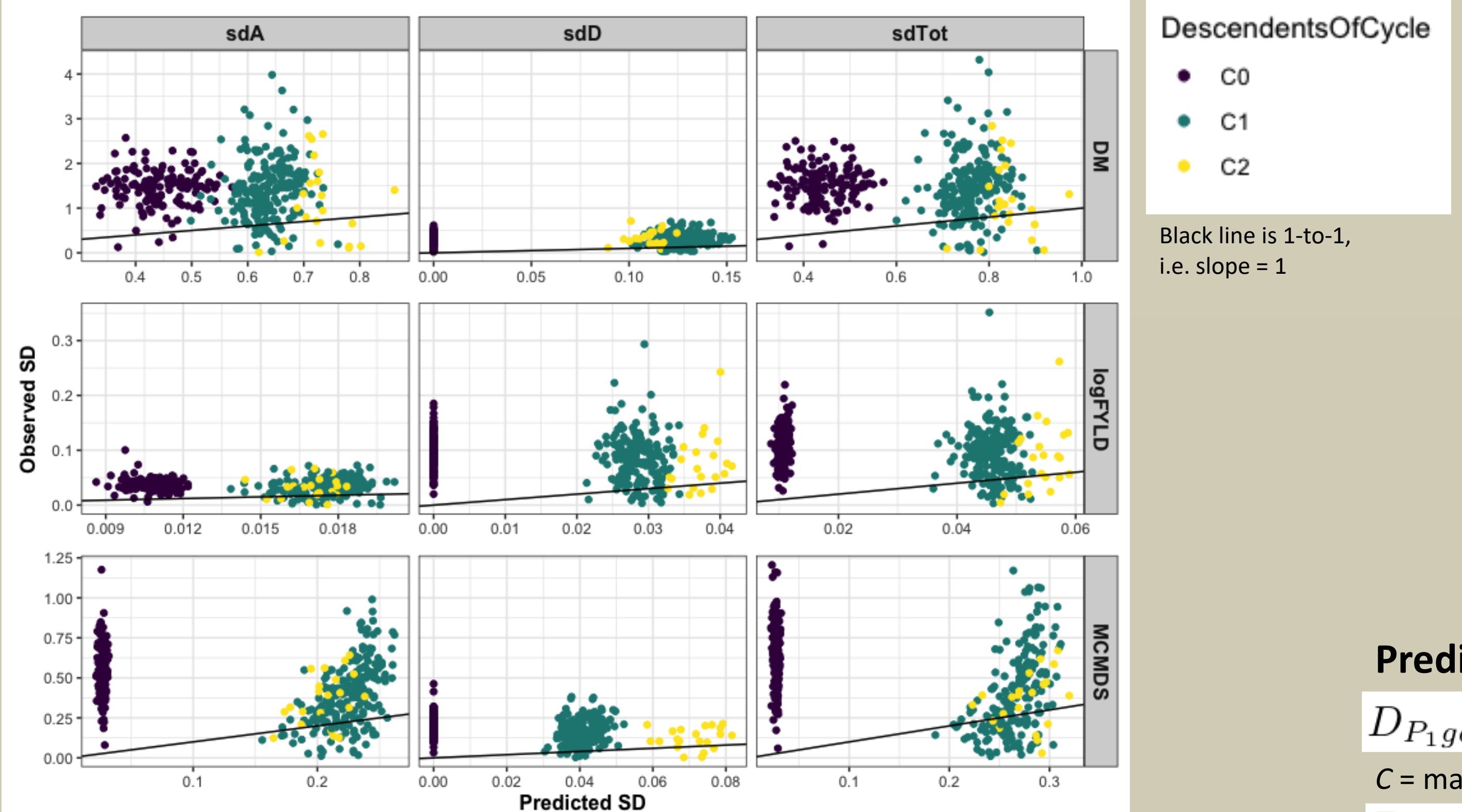


	Assuming LE (M1)	Accounting for LD (M2)
Additive	$\sigma^2_a(M1) = 2\sum p(1-p)\alpha^2$	$\sigma^2_a(M2) = \alpha^T D\alpha$
Dominance	$\sigma^2_d(M1) = \sum(2p(1-p))^2d^2$	$\sigma^2_d(M2) = d^T D^2 d$

- Full GBLUP model (no hidden phenotypes) using additive + dominance model.
- M1 refers to genetic variance components from GBLUP.
- M2 genetic variance accounting for LD as in Lehermeier et al. 2017 (see formulae at left).
- *p* are allele frequencies
- *α* are additive marker effects back-solved from GBLUP, equiv. to SNP-BLUP effects
- *d* are dominance marker effects
- D is variance-covariance matrix among markers (i.e. linkage disequilibrium)

REFERENCES:
Wolfe et al. (2016). *G3*. https://doi.org/10.1534/G3.116.033332
Lehermeier et al. (2017). *J. An. Br. Gen*. https://doi.org/10.1111/jbg.12268
Allier et al. (2019). *F. Gene.* https://doi.org/10.3389/fgene.2019.01006
Bijma et al. (2020). *Genetics*. https://doi.org/10.1534/genetics.119.302643
Chan et al. (2019). BioRXiv. https://www.biorxiv.org/content/10.1101/794339v1.full

Correspondence between predicted and realized variances



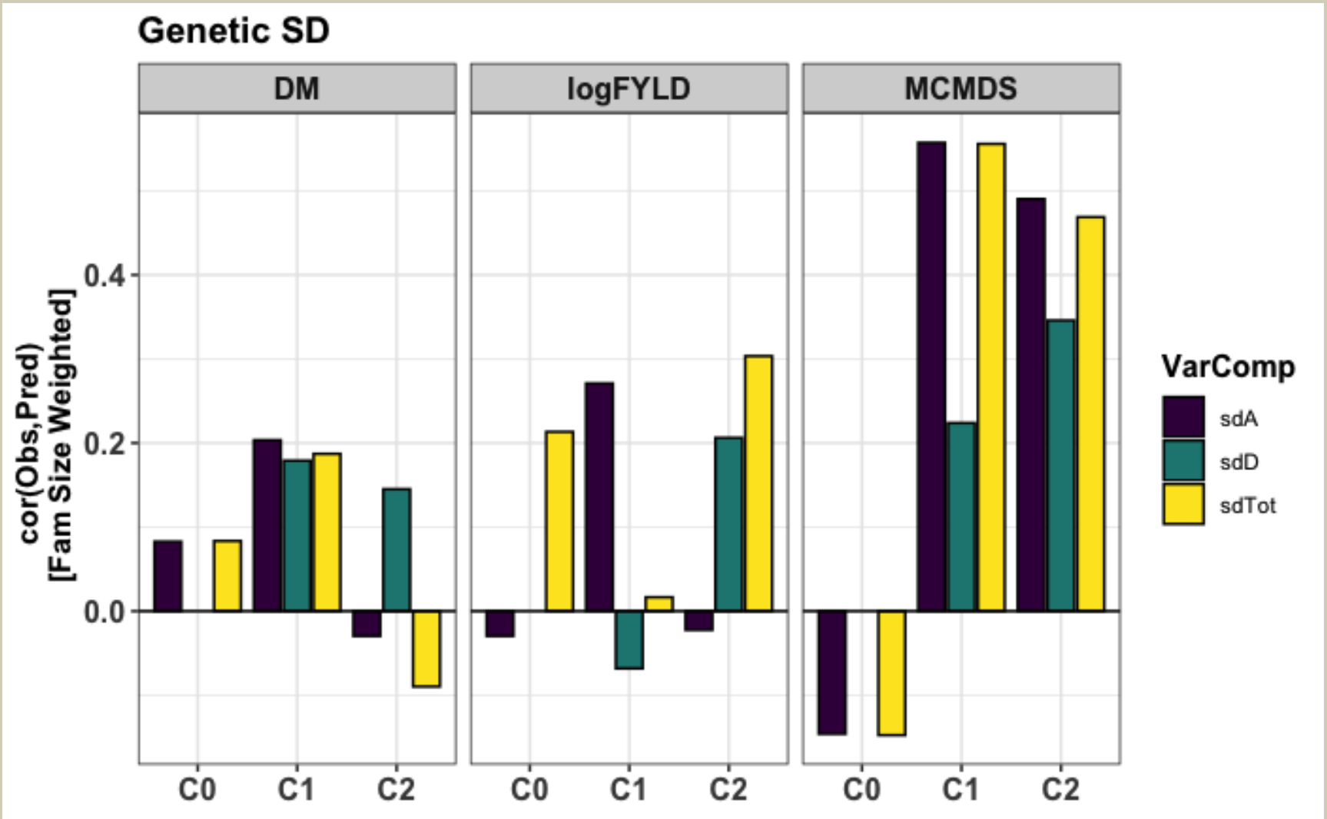
Black line is 1-to-1, i.e. slope = 1

Validation Data (y-axis):
- from the model with training data from all cycles.
- sd(GEBV), sd(GEDD), sd(GETGV)
- GEBV = genomic estimated breeding value
- GEDD = g. e. dominance deviation
- GETGV = GEBV + GEDD

Training data used for predictions (x-axis):
- C1: DescendentsOfCycle==C0 → TP = C0
- C2: DescendentsOfCycle==C1 → TP = C0+C1
- C3: DescendentsOfCycle==C2 → TP = C0+C1+C2

Correlation of Pred-Obs SD

Weighted by family size



Predicting outbred cross variance

$D_{P_1 gametes} = (1-2c)$ **LD matrix for P1 gametes**

C = matrix of pairwise recombination frequencies derived from genetic map

$$D_{P_1 gametes} + D_{P_2 gametes} = D_{OffspringGenotypes}$$

Predicted additive variance

$$\sigma^2_a = \alpha^T(D_{ProgenyGenotypes})\alpha$$

Predicted dominance variance

$$\sigma^2_d = d^T(D_{ProgenyGenotypes})^2 d$$

Prediction of all possible crosses reveals new opportunities

Evaluating all possible crosses of parents in pedigree

- 209 parents → 43219 crosses
- Only 462 actual families

Prediction of family means

- Only 462 actual families

$$\mu_a = \frac{GEBV_{P1}+GEBV_{P2}}{2}$$

Cross Usefulness Criterion (UC)

- Equivalent to the mean of the *selected fraction* of the progeny of a cross
- AKA the "Superior Progeny Mean"
- *i* = standardized selection intensity (set to 2 in this analysis)
- *h* = selection accuracy (assumed h=1 in this analysis)

UC_{parent} $UC_a = \mu_a + ih\sigma_a$

UC_{tot} $UC_{tot} = \mu_a + ih\sigma_{tot}$

$$\sigma_{tot} = \sigma_a + \sigma_d$$

RESULTS

- Red boxes on the top row highlight regions of interest where novel crosses are suggested
- Family means and variances *were not* strongly associated
- Different crosses may be indicated for logFYLD to exploit σ_{tot} vs. σ_a
- But strong correlation between UC_{parent} and UC_{tot} indicate family mean GEBV is main driver of variation in UC

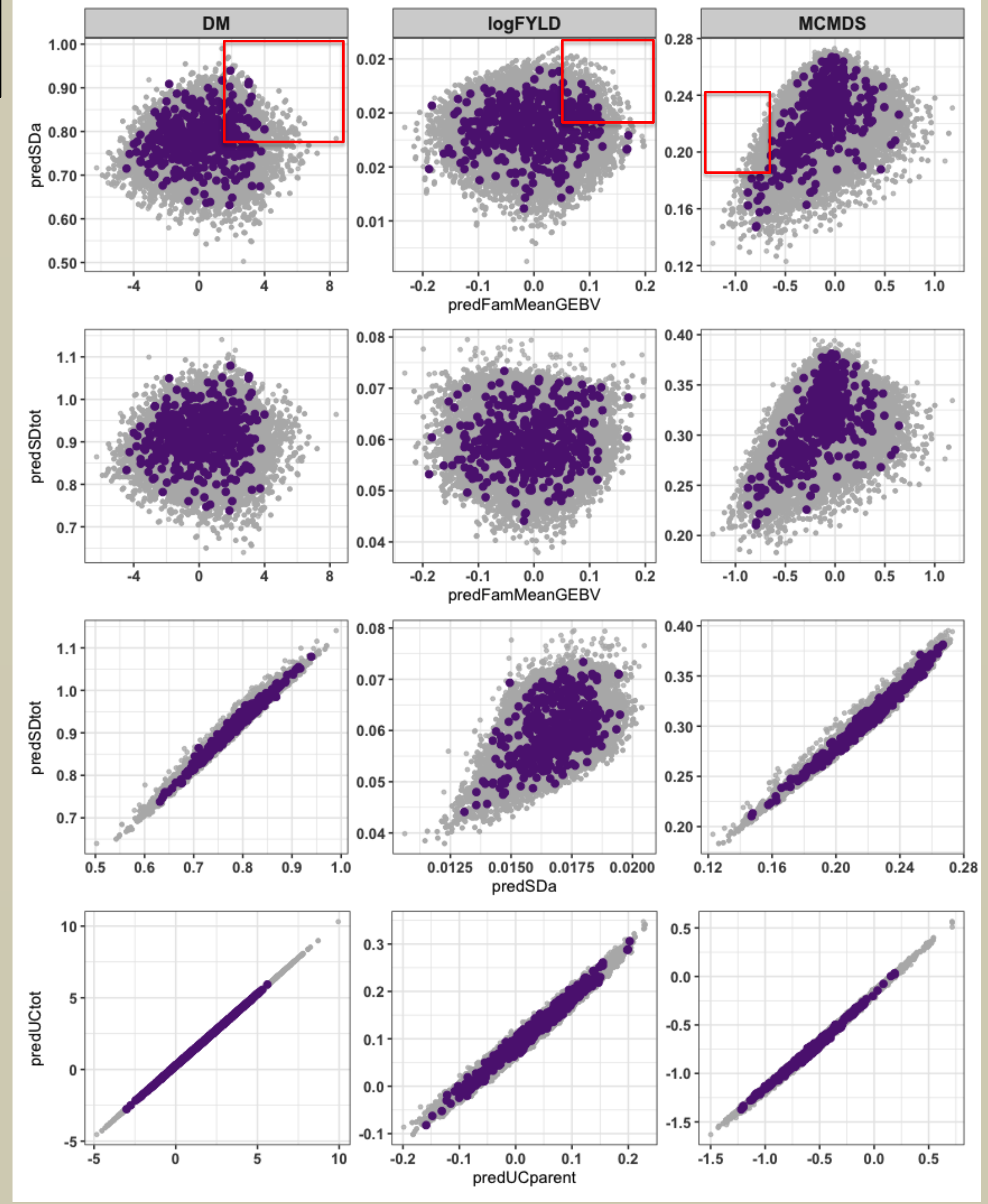
- **Original crosses made**
- **New potential crosses**

Predicted Family μ_a vs. σ_a

Predicted Family μ_a vs. σ_{tot}

Predicted Family σ_a vs. σ_{tot}

Predicted "Usefulness Criteria" UC_{parent} vs. UC_{tot}



KEY CONCLUSIONS

- Preliminary analysis of cassava breeding data highlights the potential utility of cross variance prediction for optimizing mating schemes in outbred, clonal crops.
- Dominance and total variance can be predicted in addition to the additive component

Future directions

- Assess variance prediction accuracy with *in silico* recombination
- Predicting covariances and multi-trait *index* selection
- Breeding scheme simulation

ACKNOWLEDGEMENTS: Support from BMGF and UKaid (Grant 1048542) and CGIAR-RTB. Thanks to Ariel W. Chan for curated genotypes, pedigree and genetic map. Thanks to the NextGen Cassava Breeding team, esp. the IITA Cassava Breeding team for data and motivation supporting for this work.