Modeling white mold with more than genomics

Alencar Xavier
Quantitative Geneticist at Corteva Biostatistics
Adjunct professor at Purdue University

Outline

1. **Introduction**
   - White mold
   - Model and data sources
2. **Environmental data**
   - Features and models
   - Models with GxE
   - Random regression
3. **Secondary traits**
   - High-throughput phenotyping
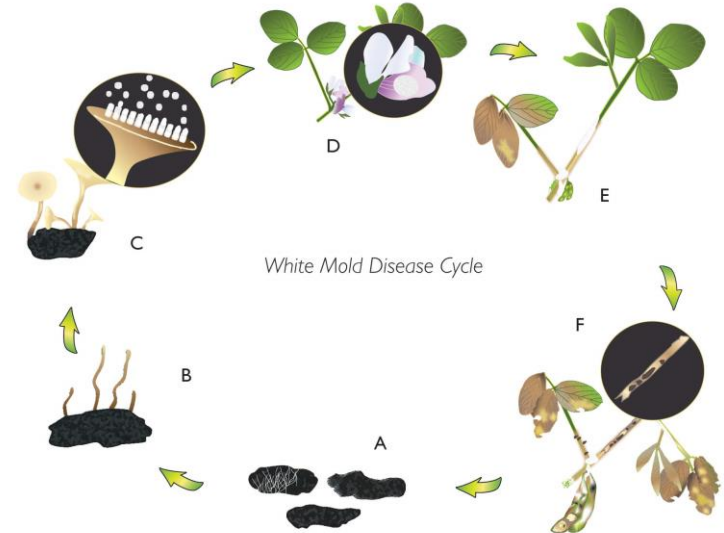   - Identification of key traits
4. **Conclusion**

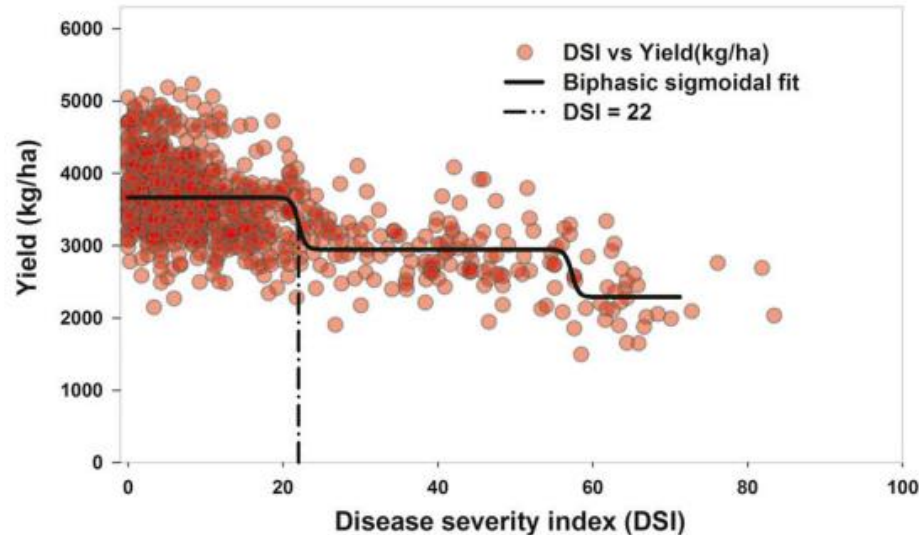http://www.sclerotia.org/lifecycle/apothecia

Sclerotinia sclerotiorum

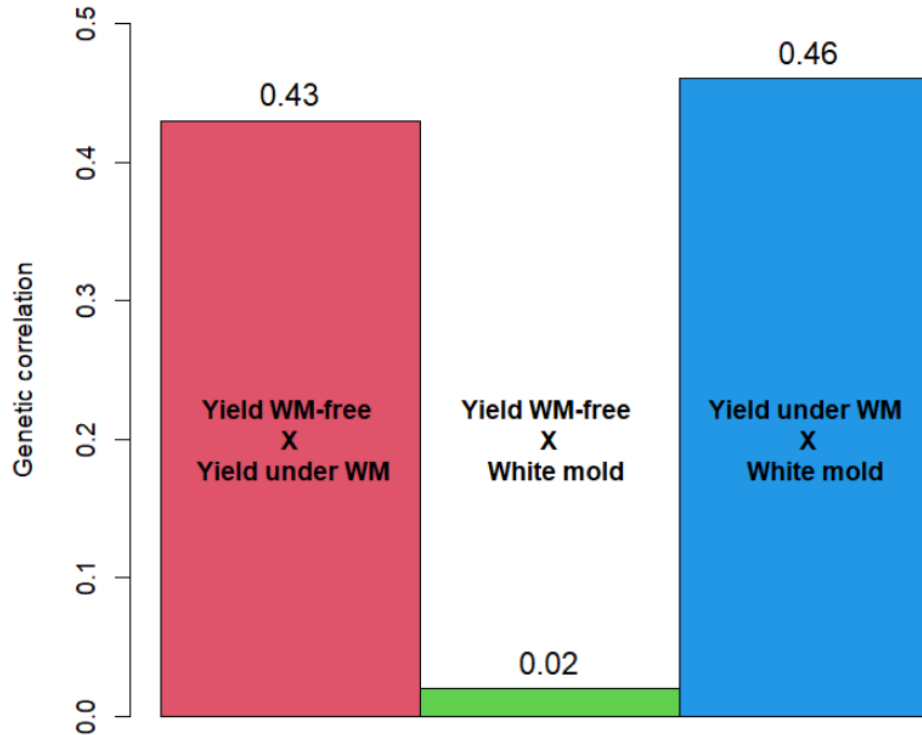https://www.pioneer.com/CMRoot/Pioneer/US/agronomy/cropfocus/soybeans/White-Mold-Soybeans.pdf

White Mold Disease Cycle

Source: https://soybeanresearchinfo.com/soybean-disease/white-mold/

White mold can decrease yield by up to 60%

(Cunha et al. 2010, 10.1111/j.1365-3059.2010.02279.x)

Fall et al. (2018) Phytopathology 108(4)
https://doi.org/10.1094/PHYTO-12-16-0446-R

WM and its genetic correlation to yield

Alencar.Xavier@Corteva.com
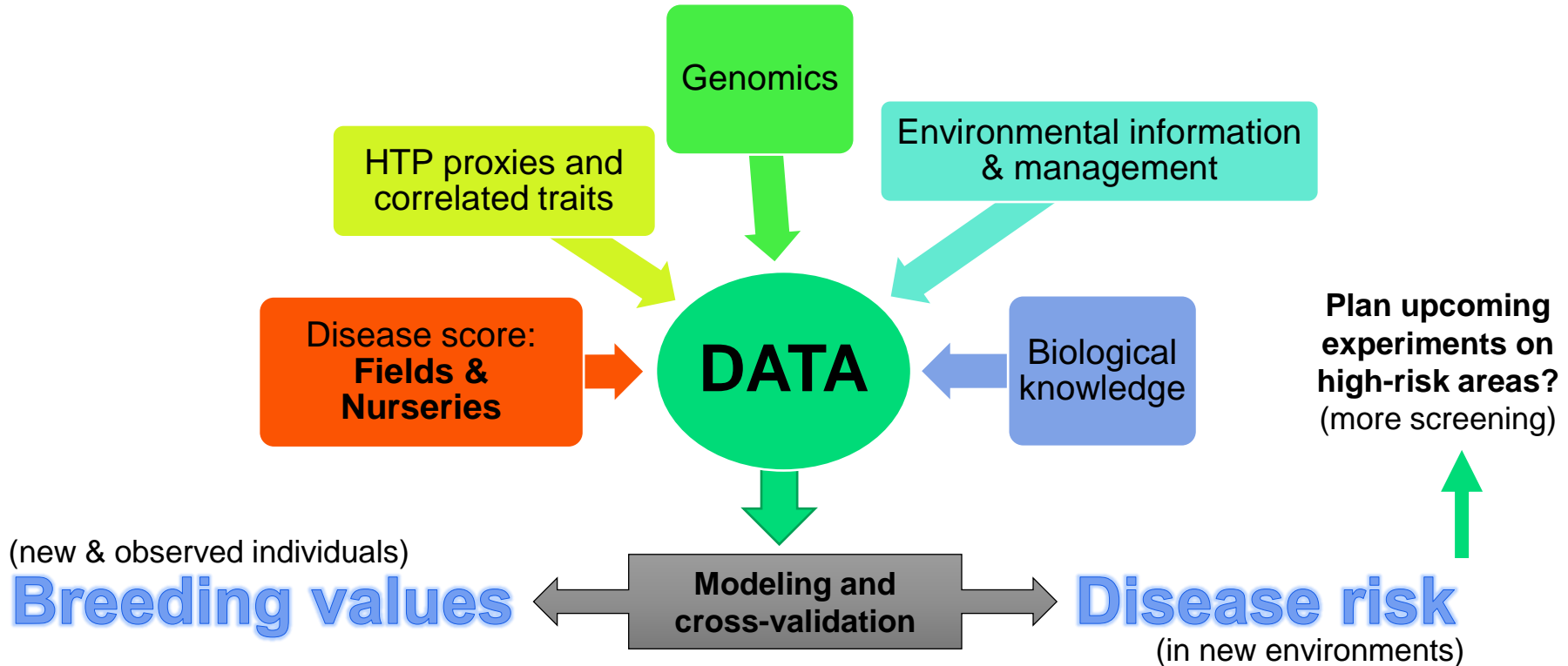Corteva Biostatistics, Quantitative Genetics group

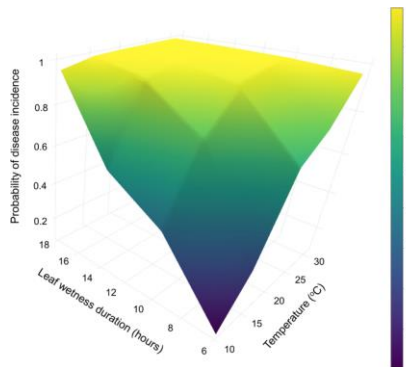# Topic 1: Models with environmental data

Disease principles
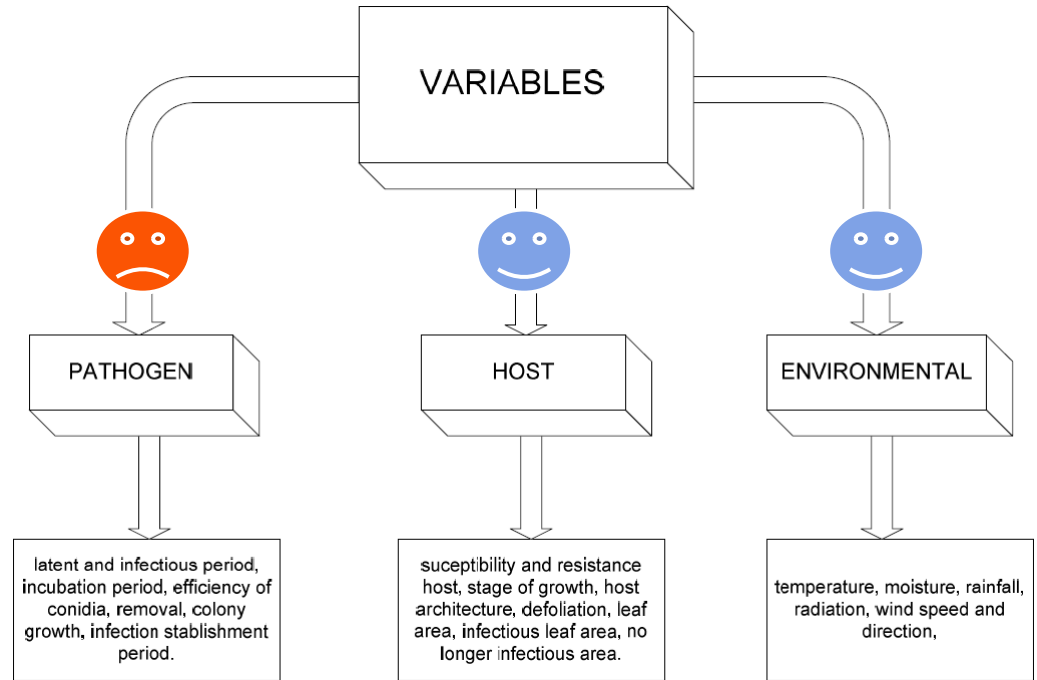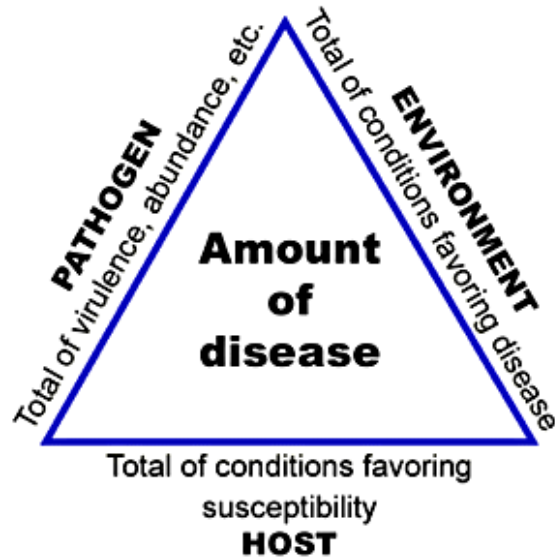
Contreras-Medina, L. M., et al. "**Mathematical modeling tendencies in plant pathology**." *African Journal of Biotechnology* 8.25 (2009).
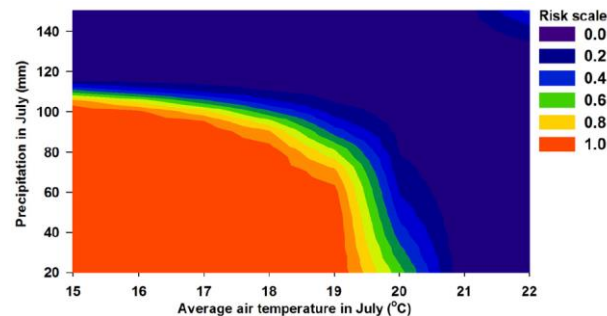
Model construction

Some things we know about white mold

- Conducive conditions: **Cold-wet**, dense canopy, high fertility

- Timing: Sclerotinia spores **infects flowers** (30+ days, R1-R4)

- Existing inoculum: Is the previous crop a host? beans, canola, alfalfa, sunflower

- QTLs reported on chromosomes: 1, **3**, 4, 7, **8**, 10, 11, 17, **19**, **20**
  - Iquira and François ([10.1186/s12870-014-0408-y](10.1186/s12870-014-0408-y)), Jing et al. 2021 (10.1016/j.ygeno.2020.10.042), Zhang et al. 2021 ([10.1007/s13353-021-00654-z](10.1007/s13353-021-00654-z)), Kandell et al 2018 (10.3389/fpls.2018.00505), Antwi-Boasiako et al. 2021 ([10.1007/s10681-021-02909-6](10.1007/s10681-021-02909-6))
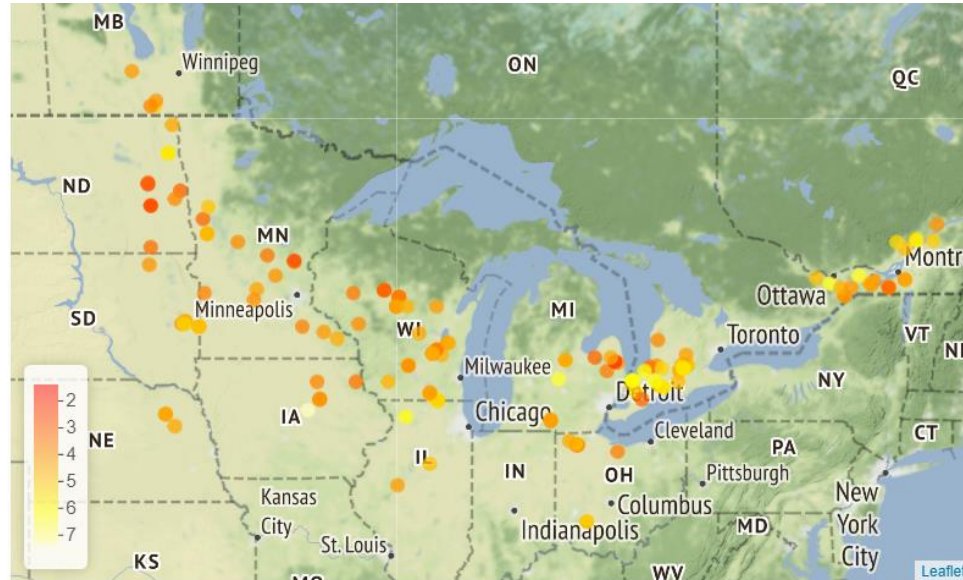

Shahoveisi et al. (2022) https://doi.org/10.1038/s41598-021-04743-1

Fall et al. (2018) https://doi.org/10.1094/PHYTO-12-16-0446-R
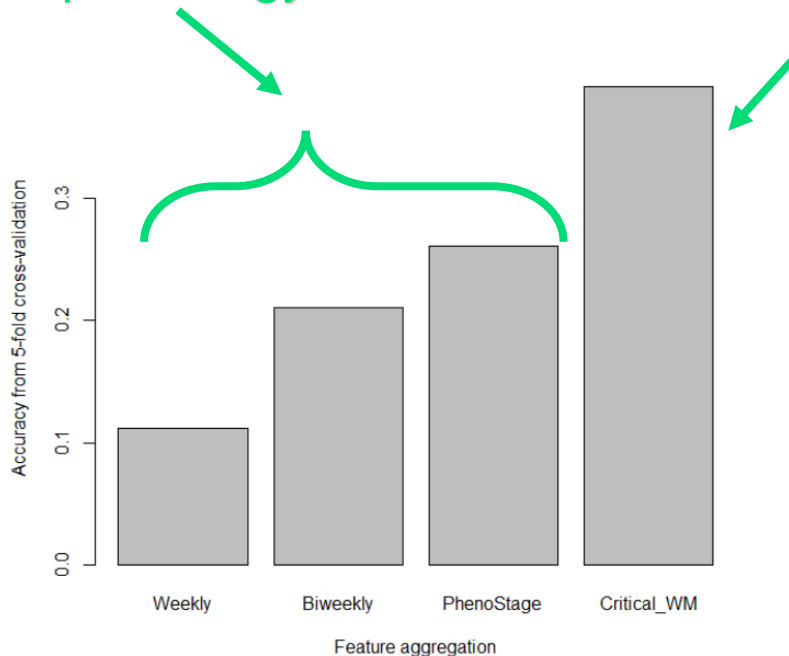
Dataset

23K observations
16K genotypes
166 environments
(over 7 years)

All-in approach vs tailored features
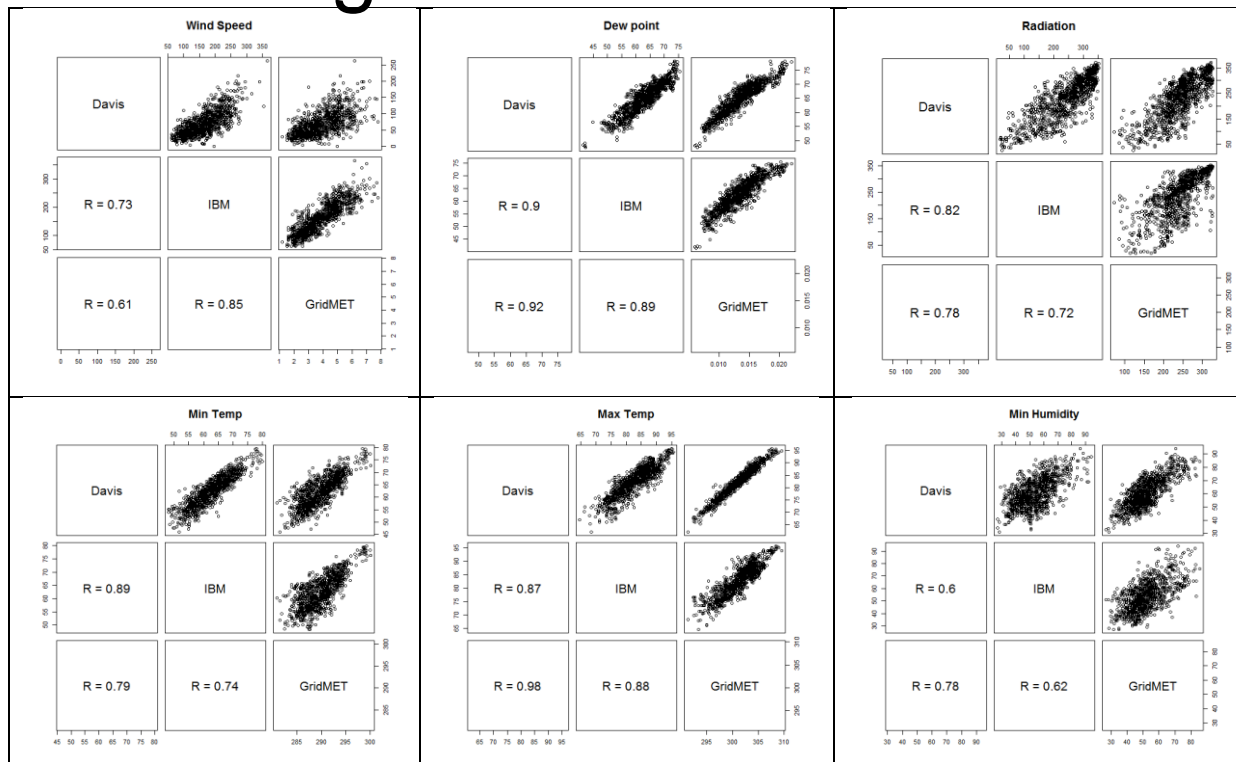
All env. features in the model, aggregated by time or phenology

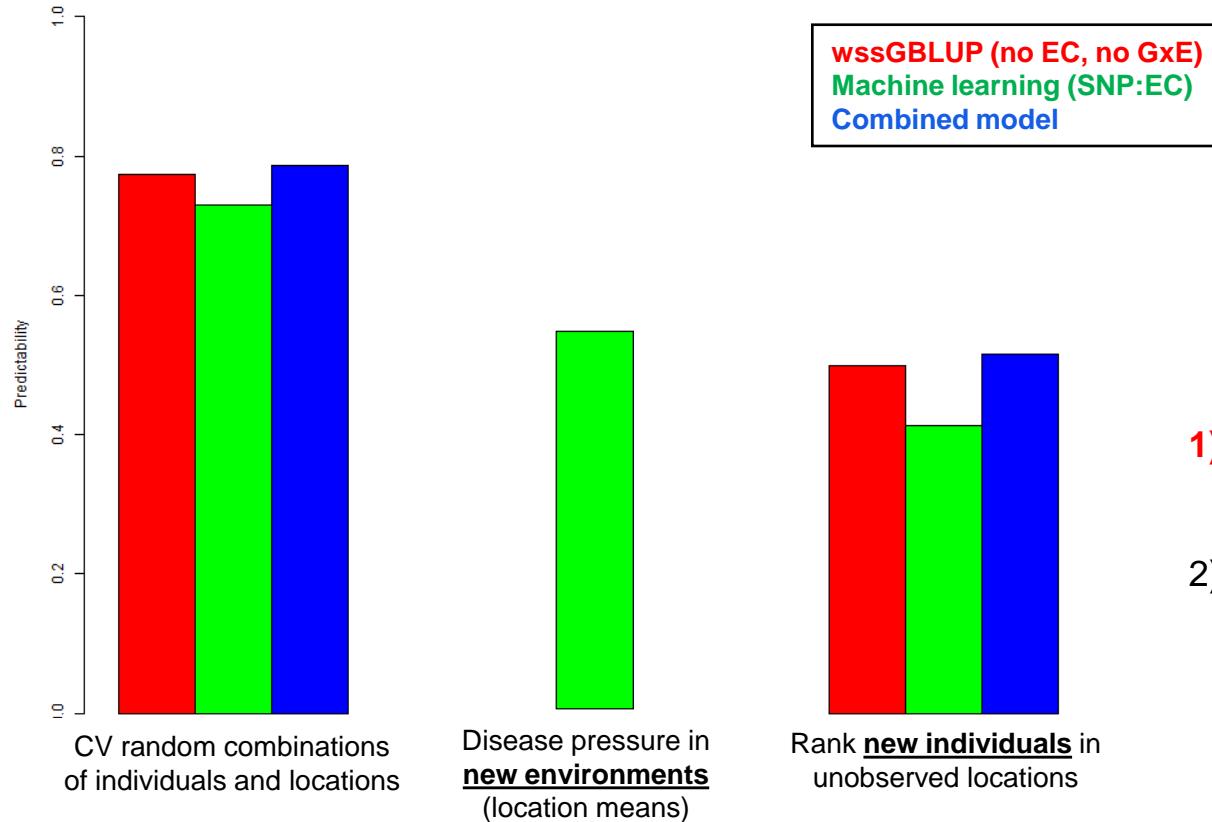Only features pertinent to WM, aggregated R1-R3

* Predictions of environmental means based on Gaussian kernel fit with REML

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group

(SIDE NOTE)
Variation among source of environmental data

Model testing and validation

wssGBLUP (no EC, no GxE)
Machine learning (SNP:EC)
Combined model
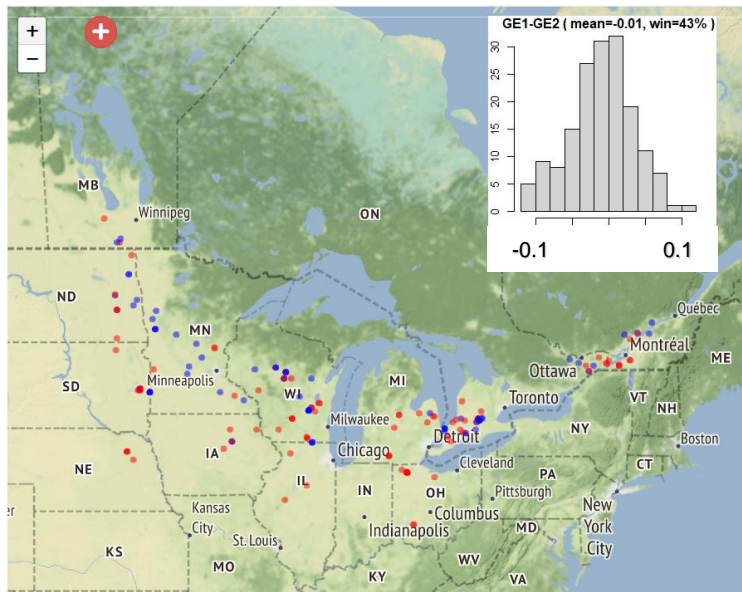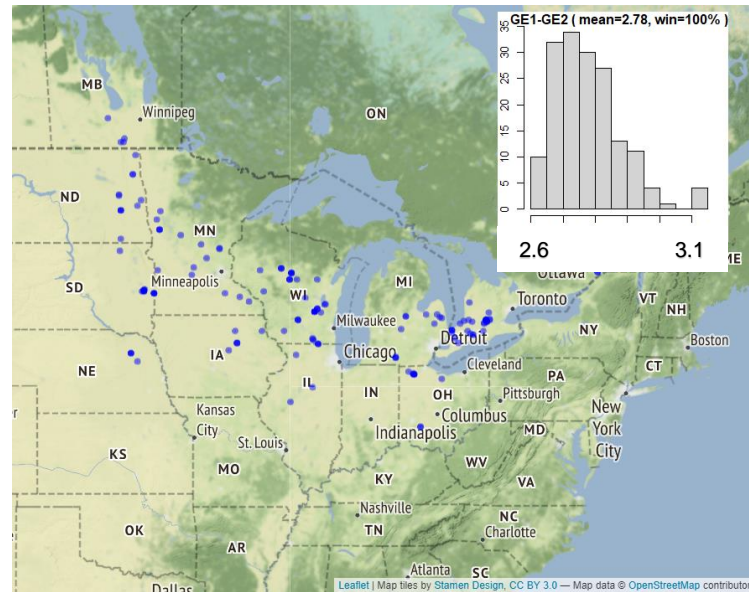
1) **Genetic architecture is largely additive**, minor GxE

2) ML models can predict OK the disease pressure ($r \cong 0.55$)

Product placement with GxE?

GxE depicts the reranking of genotypes with similar levels of disease resistance

Resistant cultivars will always outperform susceptible ones, despite GxE variation

Random Regression

For the random regression model

$$y = Xb + \sum_i^I Z_i u_i + e$$

$$V(Z_i u_i) = Z_i G Z_i' \sigma_i^2$$
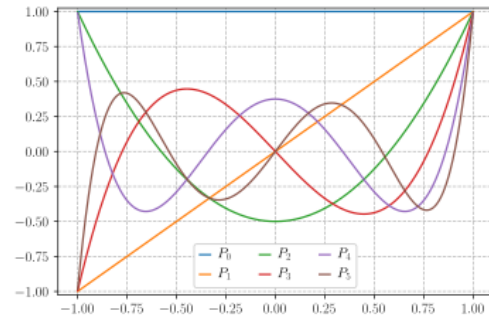$$V(u_{i,i'}) = 0$$
$$V(e) = I \sigma_e^2$$

What are the random effects?

X = Permanent environment → AOI means

$Z_{i=0}$ = Common genetic factor

$Z_{i \neq 0}$ = Polynomial of some order from some EC ⟶

Good when there are
few, but important ECs

Legendre Polynomials

Random regression uses specific environmental factors to modulate GEBVs and genetic variance

Avg. plot-level $h^2 = 0.28$

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group
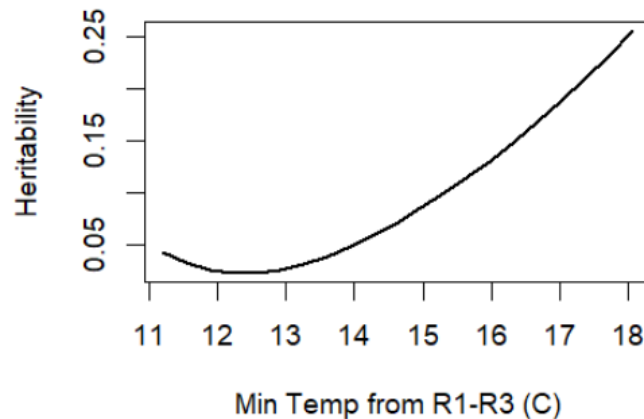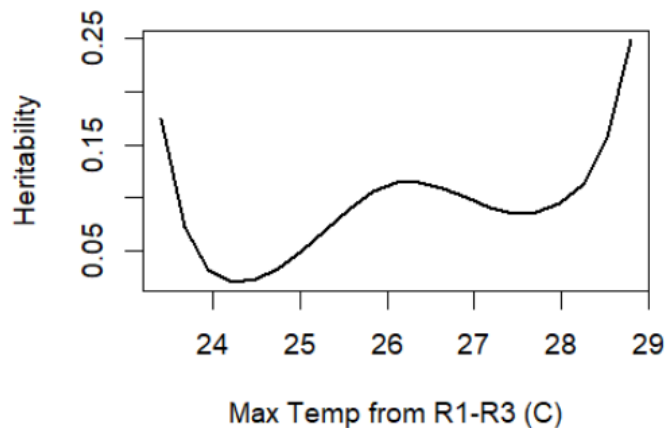
Topic 2: Models with secondary traits

High-throughput phenotyping

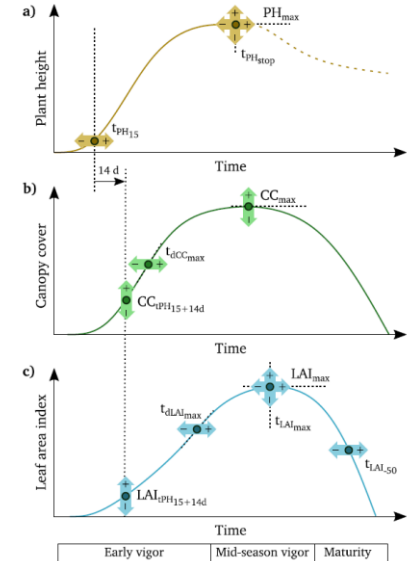Singh A.K. et al. (2021) High-Throughput Phenotyping in Soybean.
https://doi-org.ezproxy.lib.purdue.edu/10.1007/978-3-030-73734-4_7

But no direct HTP measurements of WM (yet)

Correlation(HTP, main trait)

Roth et al. (2022) High-throughput field phenotyping of soybean:
Spotting an ideotype. https://doi.org/10.1016/j.rse.2021.112797

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group

Identification of key traits

- ***Candidate traits***: Bandwidth transformations, ratios or models

  (e.g., Mfuka 2019 used NDVI; Vigier et al. 2004 used R_{636}-R_{685})

- ***Multivariate models***: Fit multiple correlated traits in the same model to capture their genetic covariance:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$
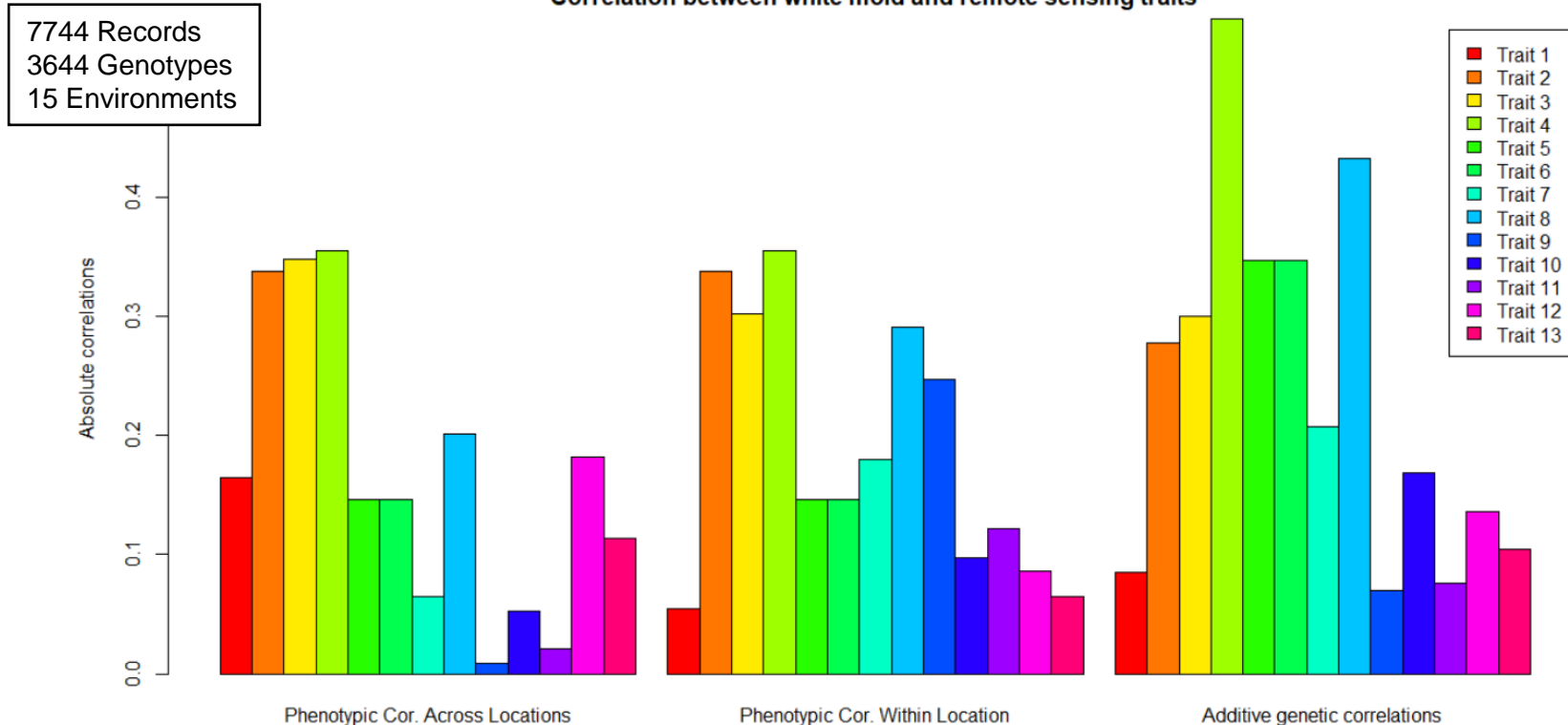
with

$$Var\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{bmatrix} Var(u_1) & \textbf{\textit{Cov}}(\textbf{\textit{u}}_1, \textbf{\textit{u}}_2) \\ \textbf{\textit{Cov}}(\textbf{\textit{u}}_1, \textbf{\textit{u}}_2) & Var(u_2) \end{bmatrix}$$

Key advantage

Leverage information from secondary and/or proxy traits through the genetic covariance

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group

CORTEVA agriscience™

Identification of key traits

Correlation between white mold and remote sensing traits

Candidate traits

7744 Records
3644 Genotypes
15 Environments

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group

Use of secondary traits

Empirical calculation
Formulas available in Wientjes et al. (2016) Genetics 202(2)
https://doi.org/10.1534/genetics.115.183269

$h_1^2 = h_2^2 = 0.5$
Pop: 20 FS + 10 HS

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Quantitative Genetics group

Use of secondary traits
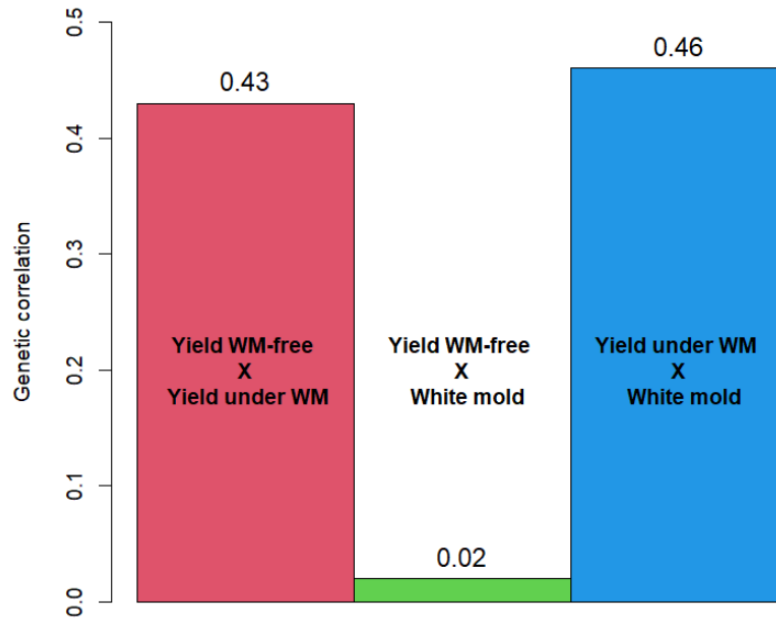
Alternatively, improve
yield under white mold
as a separate trait from yield

Thank you for your attention!

Remarks:

1) Biologically meaningful features are generally preferred over 'all-in' approach

2) GxE re-ranks genotypes with similar resistance levels (but not from "R" to "S")

3) Secondary traits from HTP can be used to increase GEBV accuracy

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com