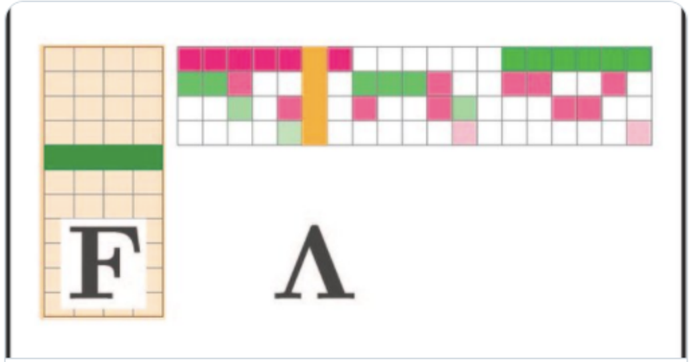
MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits

Daniel E Runcie[1], Jiayi Qu[2], Hao Cheng[3] and Lorin Crawford[4]

[1]Department of Plant Sciences, University of California Davis, Davis, CA, USA; deruncie@ucdavis.edu, [2]Department of Animal Sciences, University of California Davis, Davis, CA, USA; jyqqu@ucdavis.edu, [3]Department of Animal Sciences, University of California Davis, Davis, CA, USA; qtlcheng@ucdavis.edu, [4]Department of Biostatistics, Brown University, Providence, RI, USA; lorin_crawford@brown.edu

https://www.biorxiv.org/content/10.1101/2020.05.26.116814v2



deruncie/MegaLMM
R package for fitting high-dimensional multivariate linear mixed effect models - deruncie/MegaLMM
github.com

Daniel Runcie

UCDAVIS
DEPARTMENT OF PLANT SCIENCES

Background

How can we use data from high-throughput phenotyping to improve genetic value prediction?

Problem

Incorporating phenotype data from many traits at once is challenging

Solution

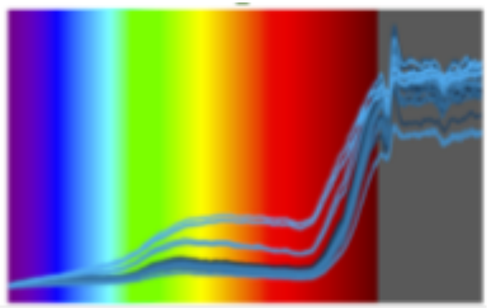
MegaLMM: Fast and Powerful multi-trait linear mixed effects models for an unlimited number of traits

Limitations and future directions

New technologies available to breeders

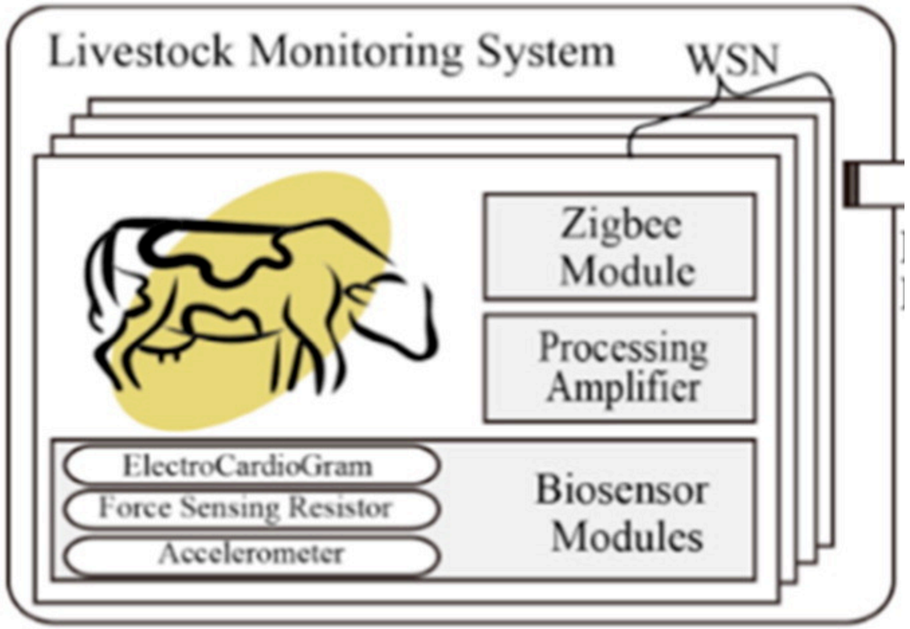
Drones

Hyperspectral cameras



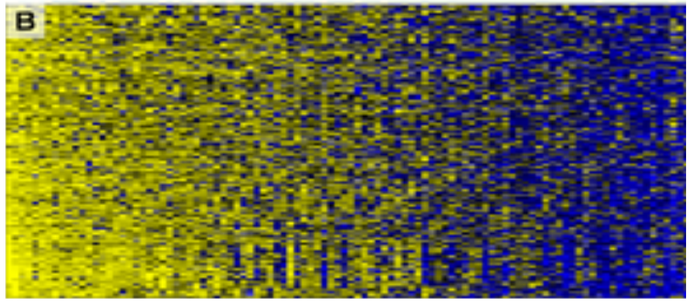
Lopez Cruz et al 2020

Wearable sensors



Neethirajan 2017

Gene expression / Metabolomics



High dimensional data

p >> n

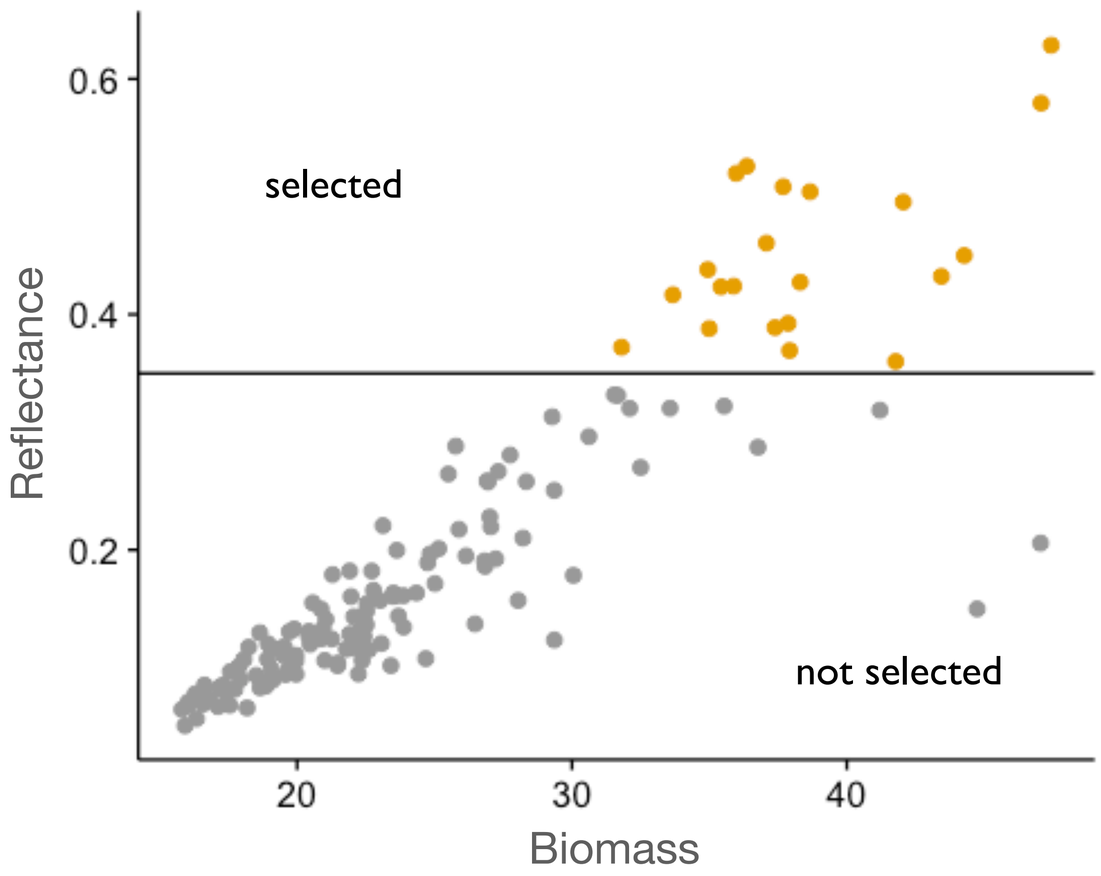
Highly correlated data

temporal,
spatial, etc

“Secondary traits”

Not of direct interest

Secondary traits improve prediction accuracy

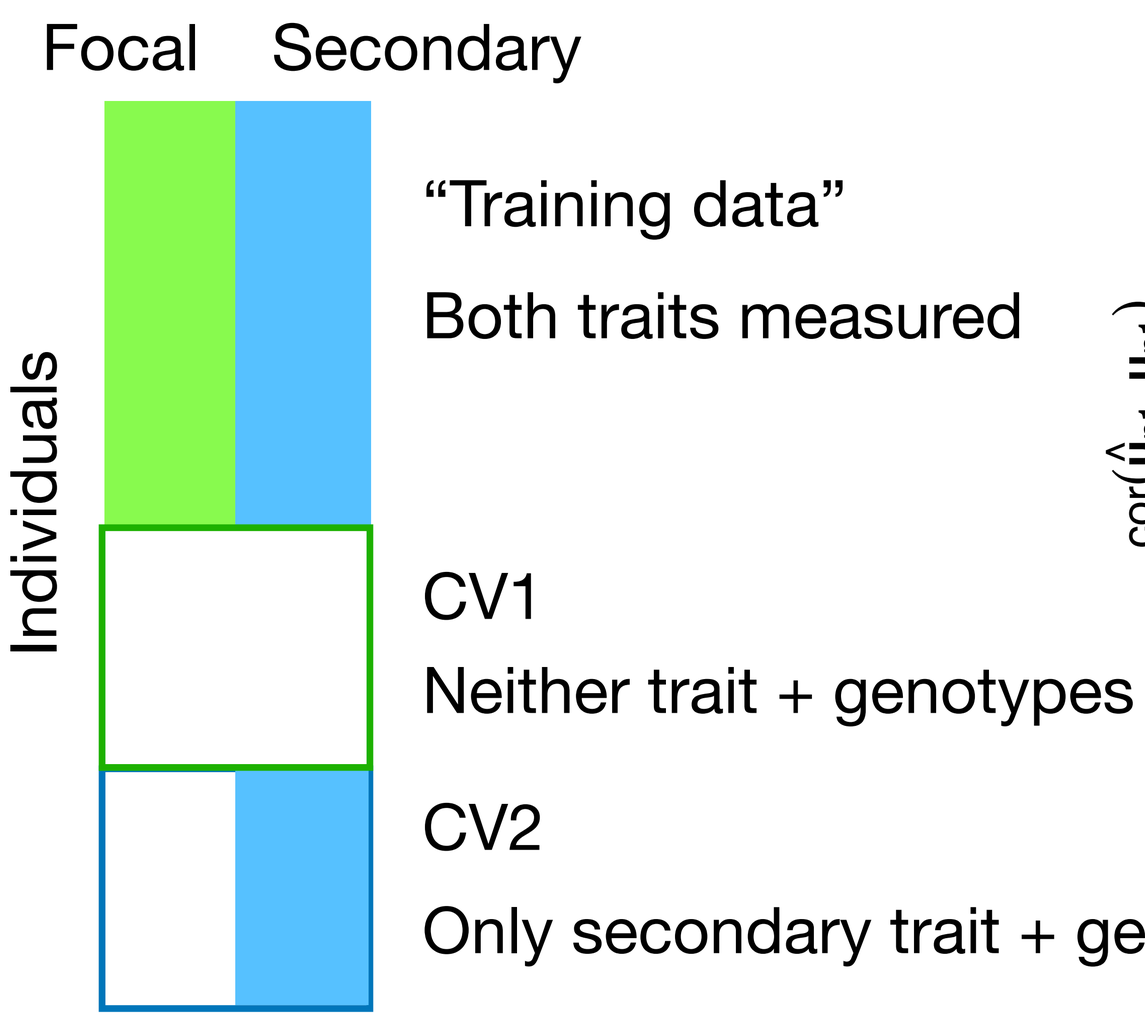


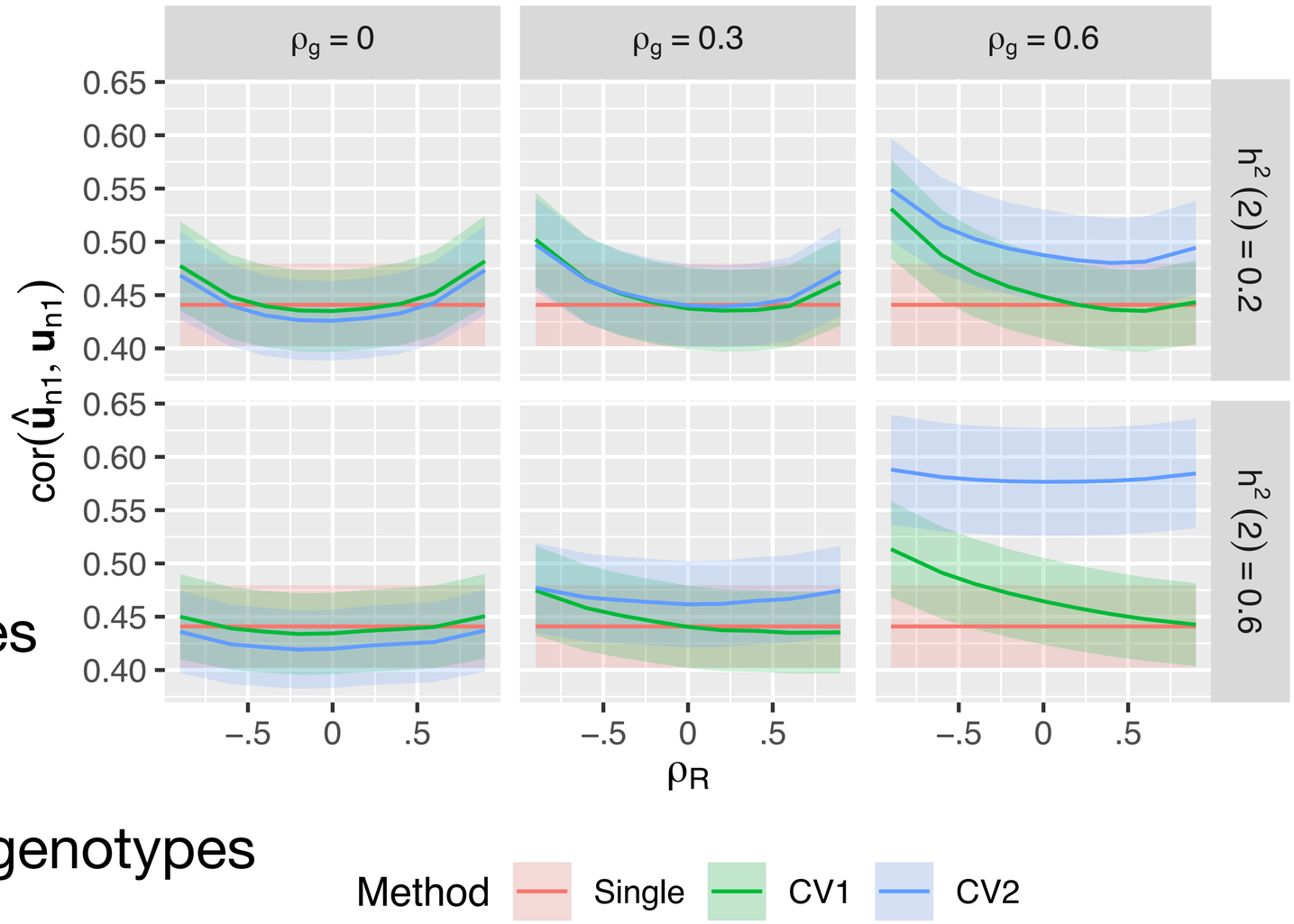
Rather than measure biomass directly, predict it based on other traits

Can be more efficient if other traits are:

- cheaper to measure
- faster to measure
- can be measured earlier in development

How to use secondary trait data

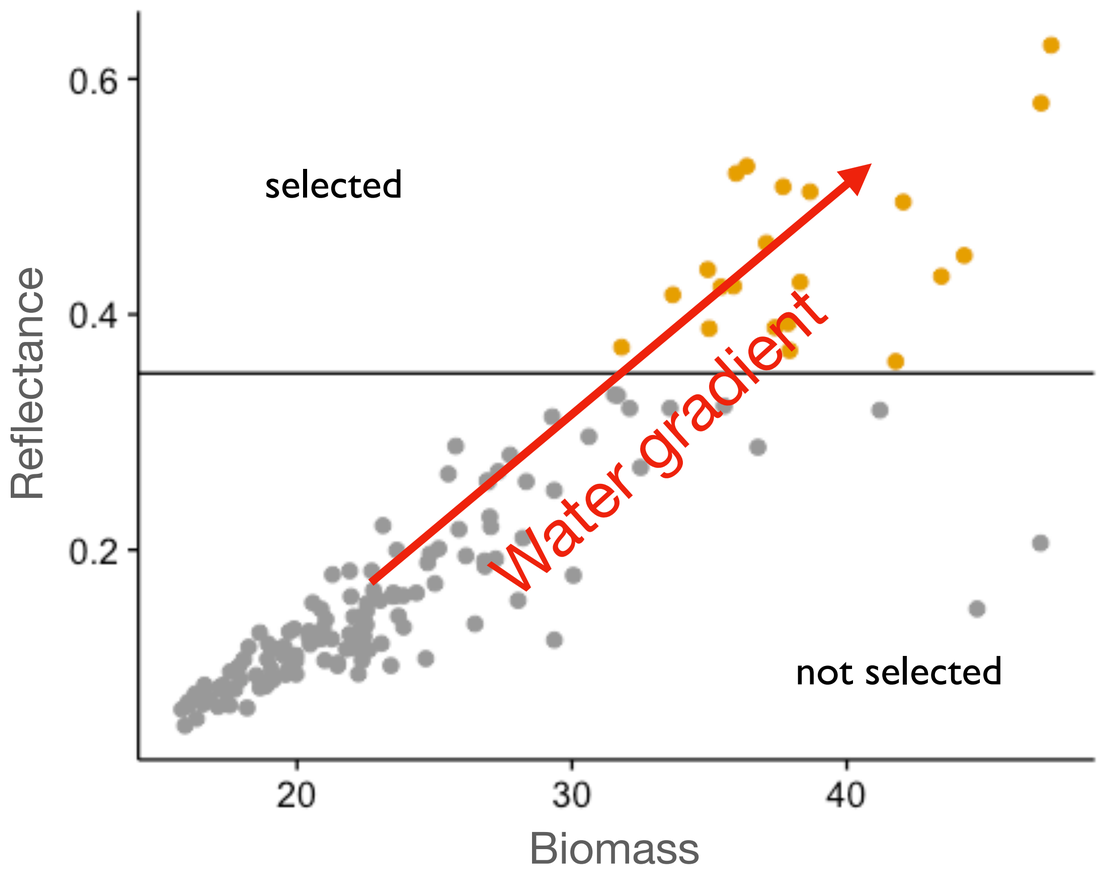




Runcie and Cheng 2019 *G3*

Problem: non-genetic correlations bias results

Need to separate **genetic** from **non-genetic** correlations



Estimate **G** and **R** in a multi-trait linear mixed model (MvLMM)

G

$$\begin{pmatrix} \sigma^2_{g1} & \sigma_{g12} \\ \sigma_{g21} & \sigma^2_{g2} \end{pmatrix}$$

R

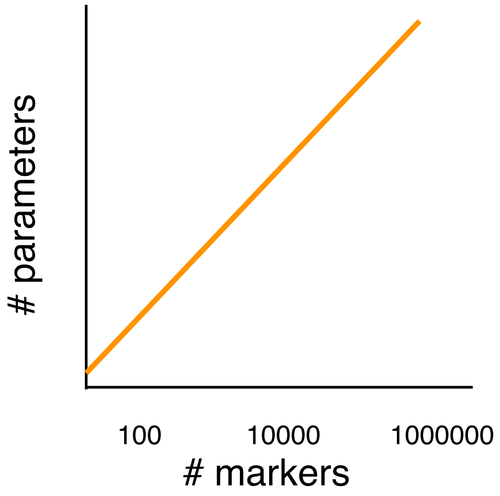
$$\begin{pmatrix} \sigma^2_{r1} & \sigma_{r12} \\ \sigma_{r21} & \sigma^2_{r2} \end{pmatrix}$$

Problem: **G** and **R** get very big!

with many traits, many more parameters than data points

Problem: too many parameters (and slow computation)

Single-trait Genomic Prediction



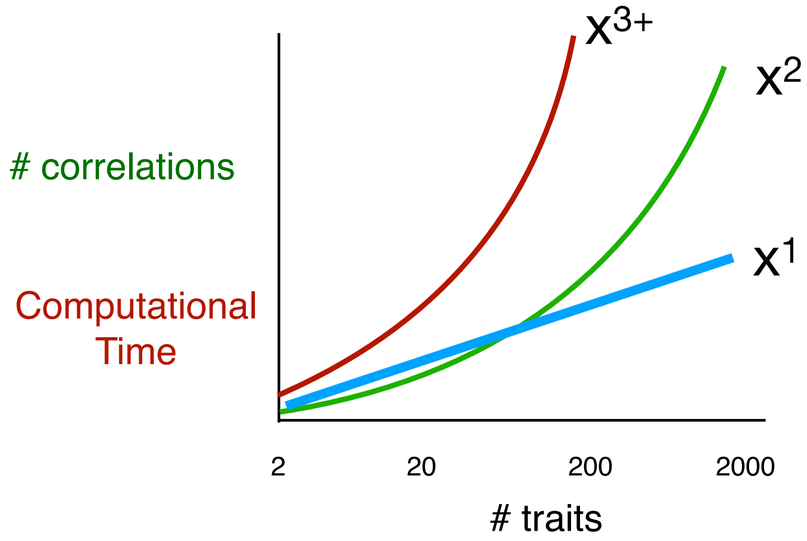
Regularized regression

GBLUP, rrBLUP

Bayesian Alphabet

RKHS

Multi-trait Genomic Prediction

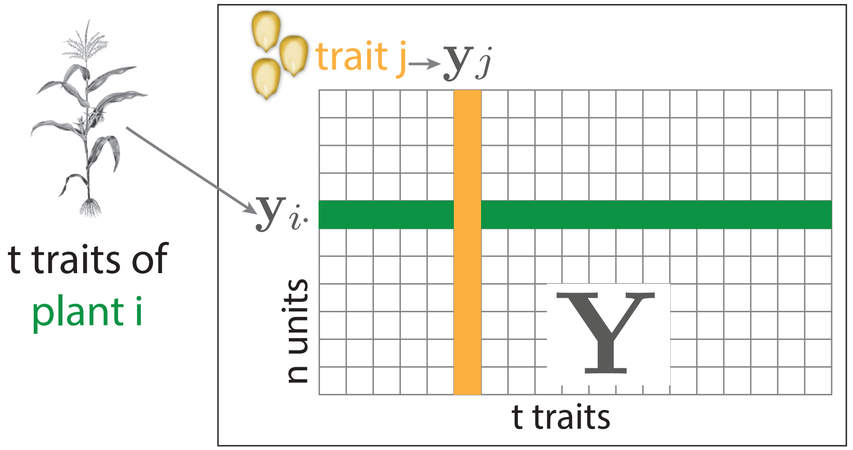


~~ASReml, sommer, MCMCglmm, GEMMA~~

Solution:

MegaLMM

MegaLMM: MvLMMs for an unlimited number of traits



$$= \mathbf{XB} + \mathbf{ZU} + \mathbf{E}$$

Blocks, markers

Genetic relatedness

Residuals

$$\mathbf{U} \sim \mathrm{MN}_{\mathbf{r}\times\mathbf{t}}(\mathbf{0}, \mathbf{K}, \mathbf{G})$$

Genetic values

Correlated across traits (columns) and individuals (rows)

$$\mathbf{E} \sim \mathrm{MN}_{\mathbf{n}\times\mathbf{t}}(\mathbf{0}, \mathbf{I}, \mathbf{R})$$

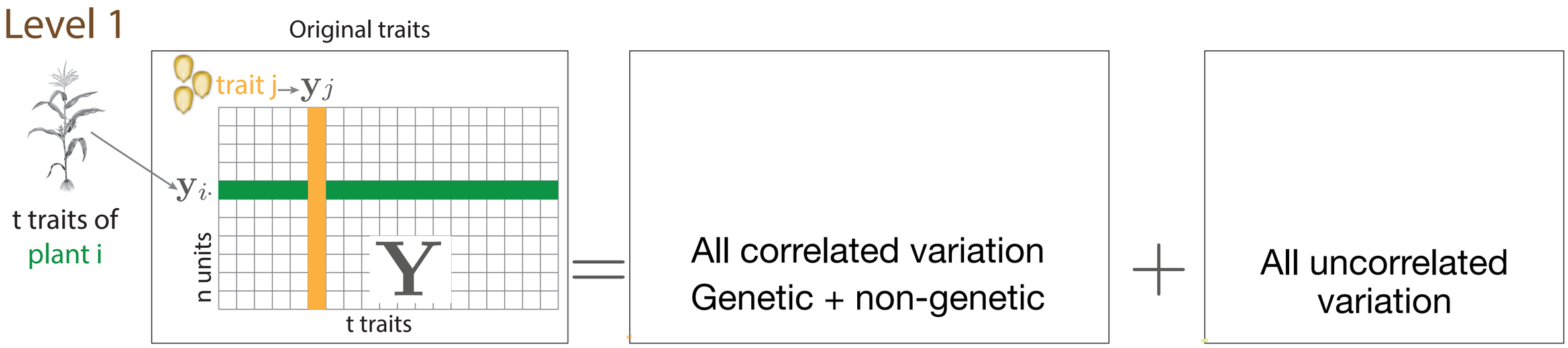
Residual values

Correlated across traits (columns)

Correlations mean we need to estimate whole matrices at once

G and **R** can be very large, so have too many parameters to estimate directly

Modeling strategy: two level model

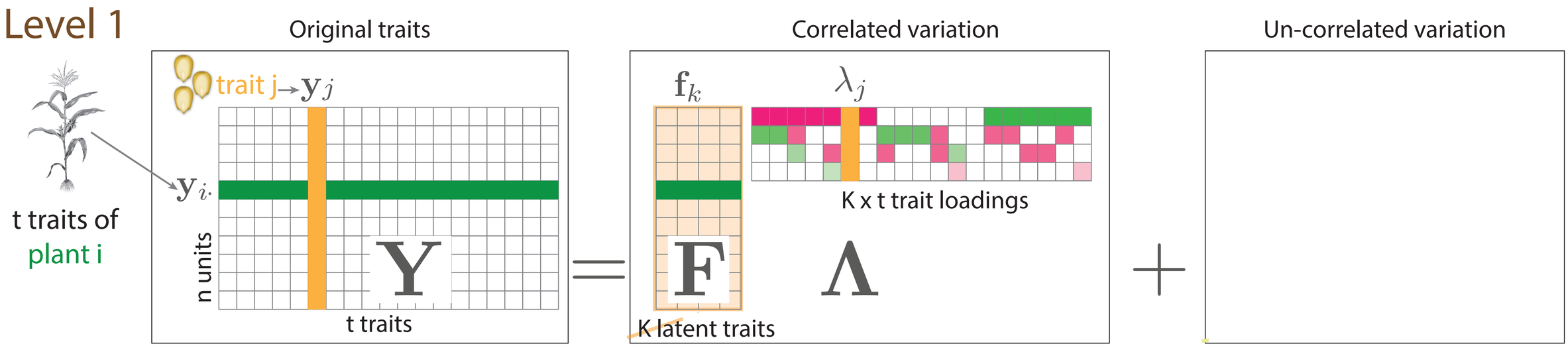


Level 1: Break up the trait matrix into two components

One contains all sources of variation shared among traits

The other contains all residual variation unique to each trait

Modeling strategy: two level model



Use a factor model to fit the correlated variation (regularization)

Introduce K latent traits called factors

like PCA: each row of **Λ** is a loadings vector of correlated traits

leave the residuals as is

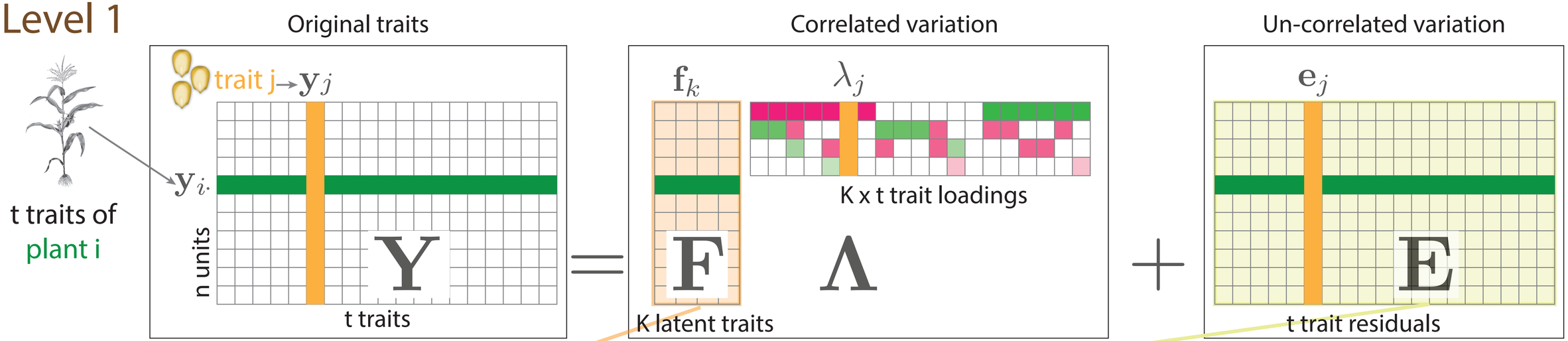
Result: K + t traits

all are uncorrelated!

correlations are modeled with Kt parameters instead of t^2

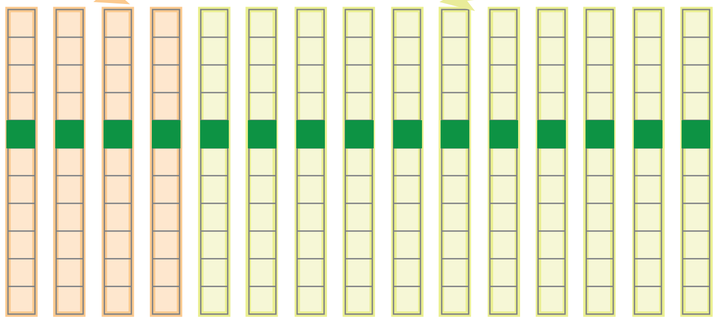
Modeling strategy: two level model

Level 1



Level 2

Independent
LMMs



M = K + t de-correlated traits

$$= \mathbf{X}\boldsymbol{\beta}_m + \mathbf{Z}\mathbf{u}_m + \boldsymbol{\epsilon}_m$$

Blocks, markers — Genetic relatedness — Residuals

Each of the K + t de-correlated traits is modeled with an independent linear mixed model

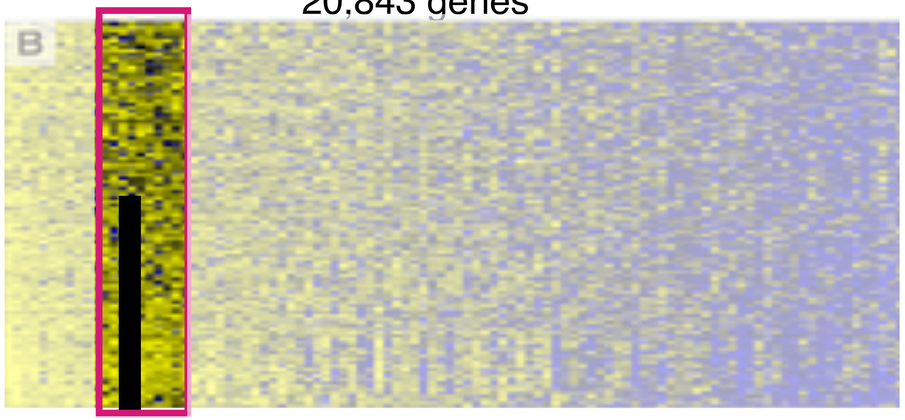
Novelty: factors are traits so represent all sources of variation

MegaLMM works and is fast

Gene expression from Arabidopsis (1001 genomes project)

20,843 genes

665 lines



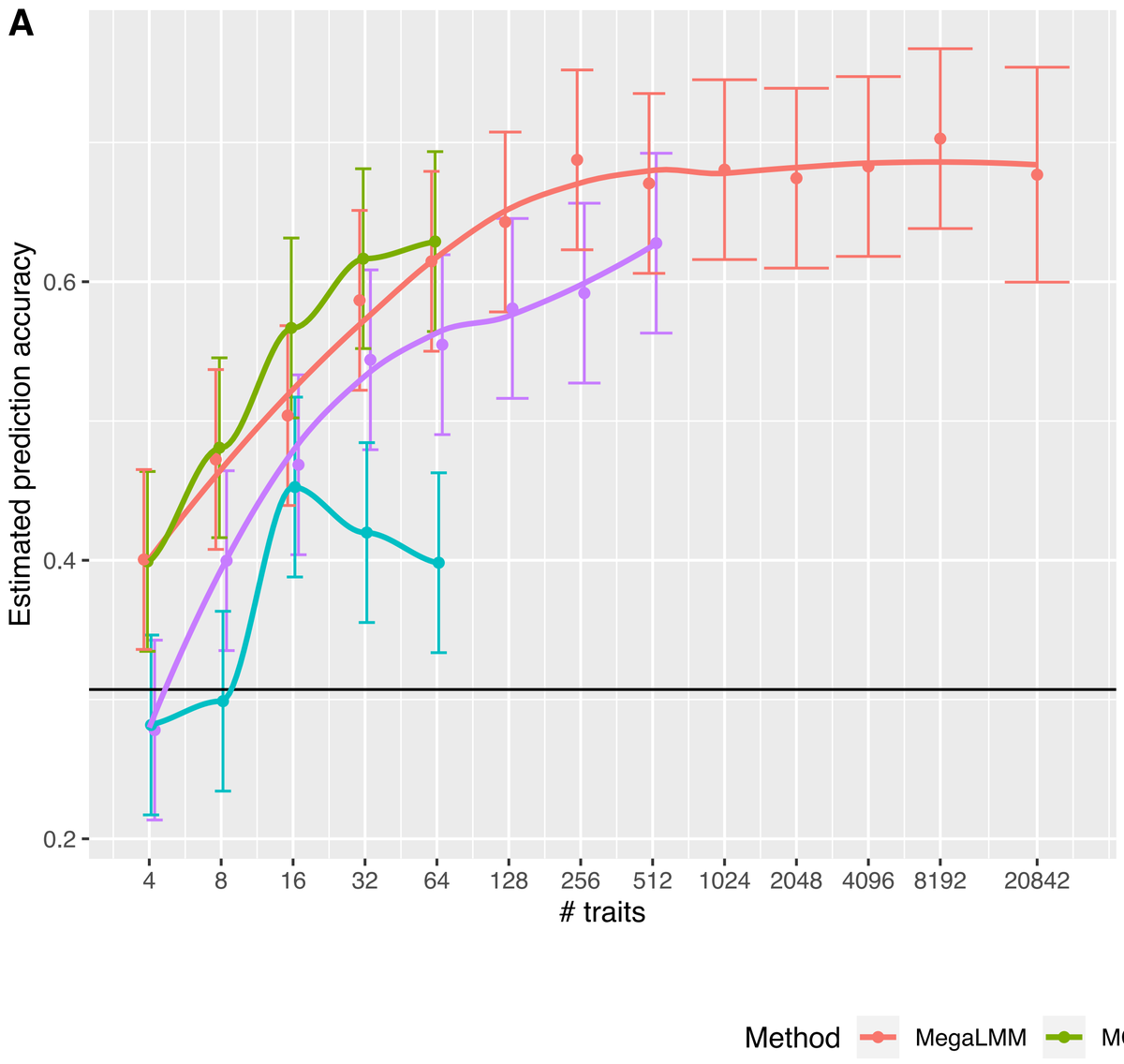
1. masked 50% of one gene
2. selected a set of other random genes
3. Used MvLMMs to predict genetic values of masked gene for masked individuals
4. Repeated multiple times with different genes
5. Measure accuracy of genetic value predictions for first gene

Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods

Daniel Runcie and Hao Cheng

G3: GENES, GENOMES, GENETICS *November 1, 2019 vol. 9 no. 11 3727-3741; https://doi.org/10.1534/g3.119.400598*

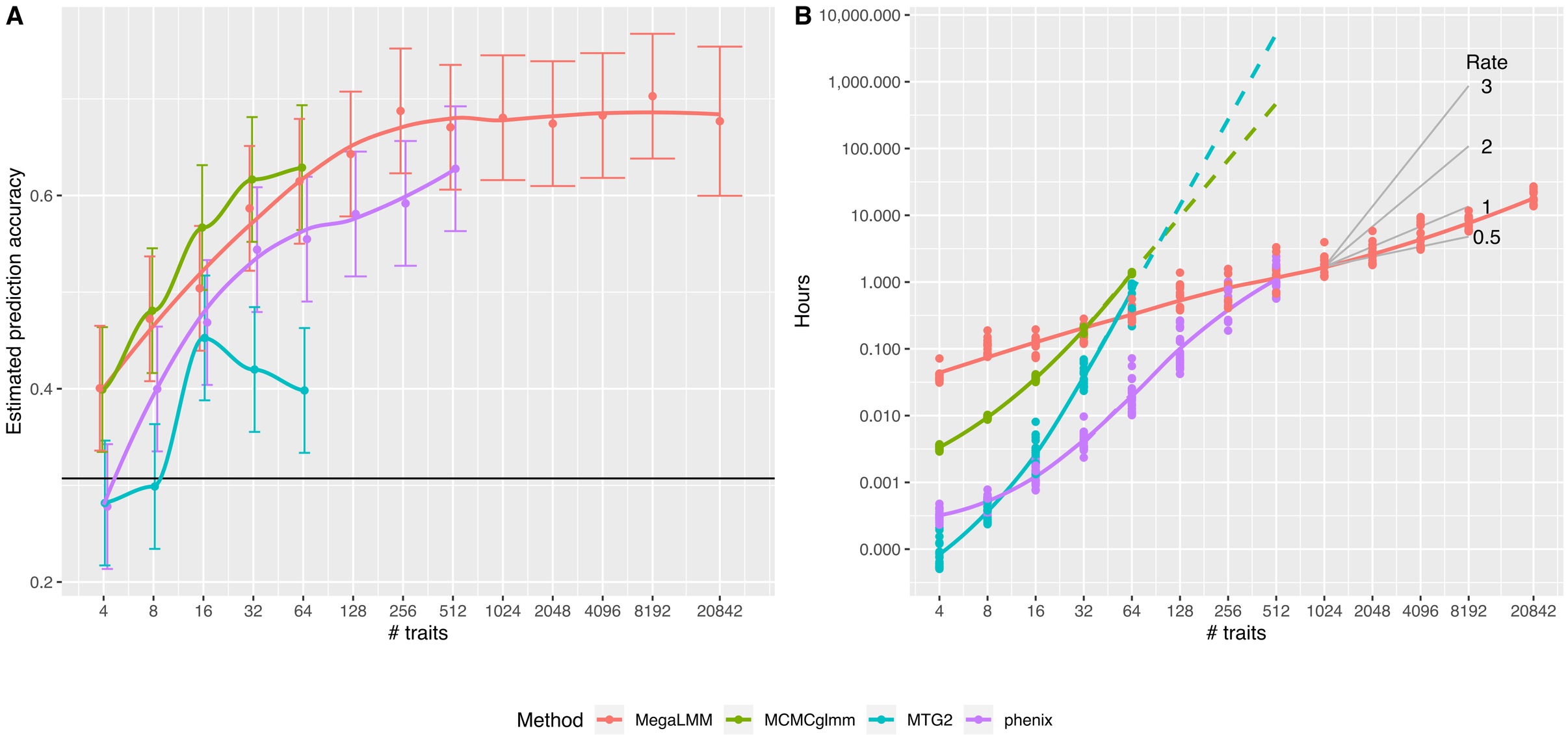
MegaLMM works



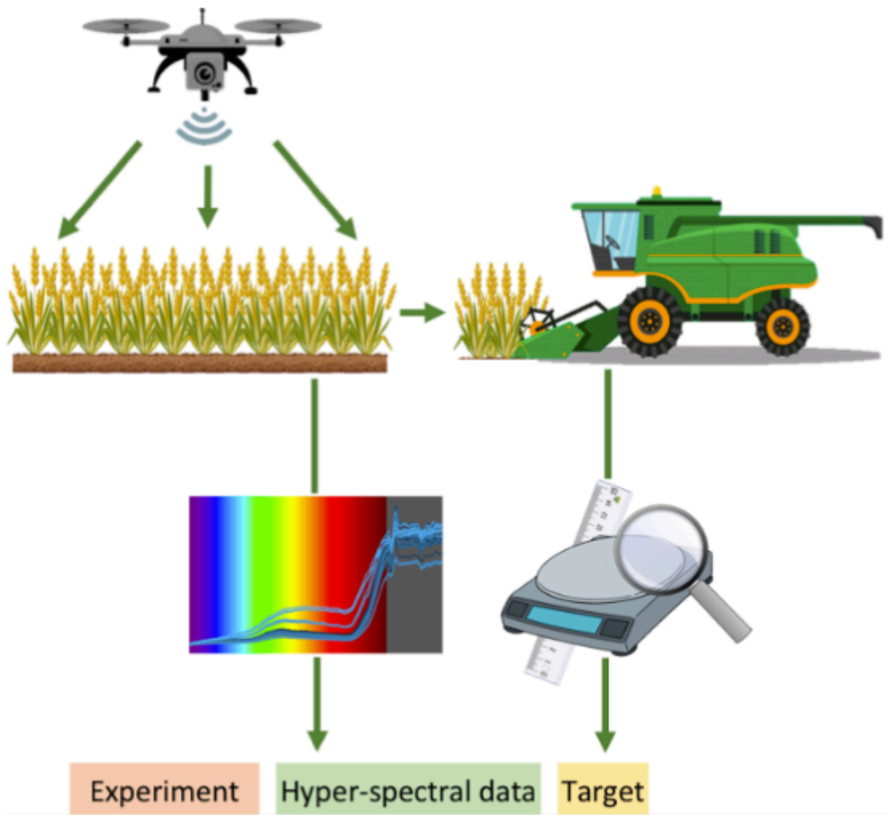
No decline in accuracy with more traits for Bayesian methods

Similar to results for #markers in genomic prediction

MegaLMM works and is fast



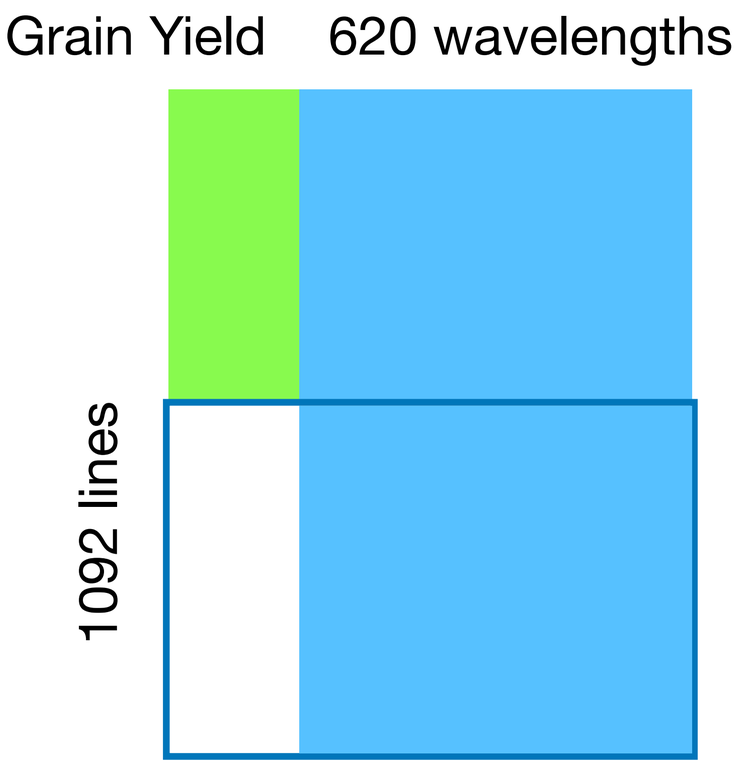
Case study 1: Wheat



Lopez Cruz et al 2020

CIMMYT Bread Wheat breeding program

Data from Krause et al 2019



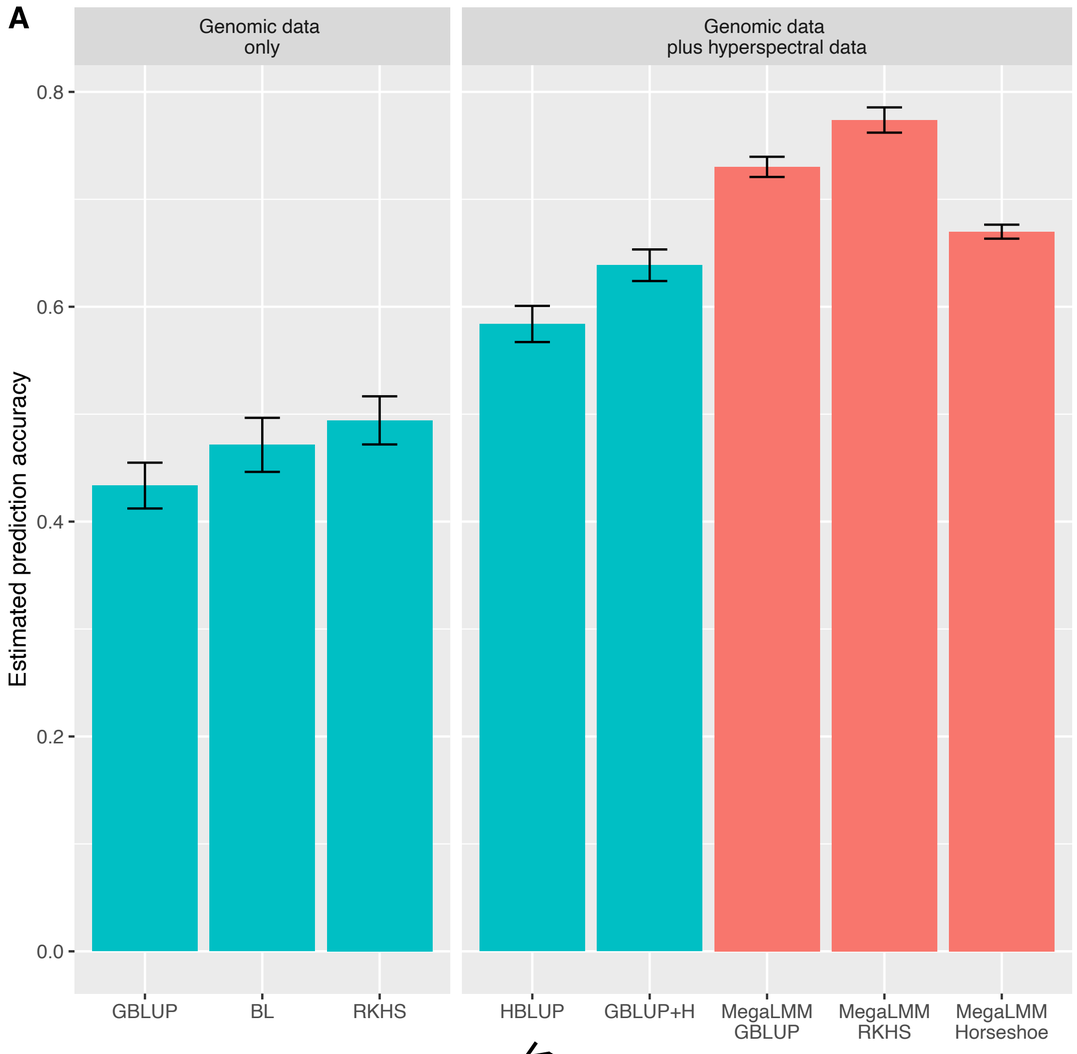
CV2 method

50:50 split

Compare:

- Single-trait genomic prediction
- Trait-assisted genomic prediction
  - H-matrix GBLUP
  - full MvLMM with MegaLMM

MegaLMM greatly outperforms other methods



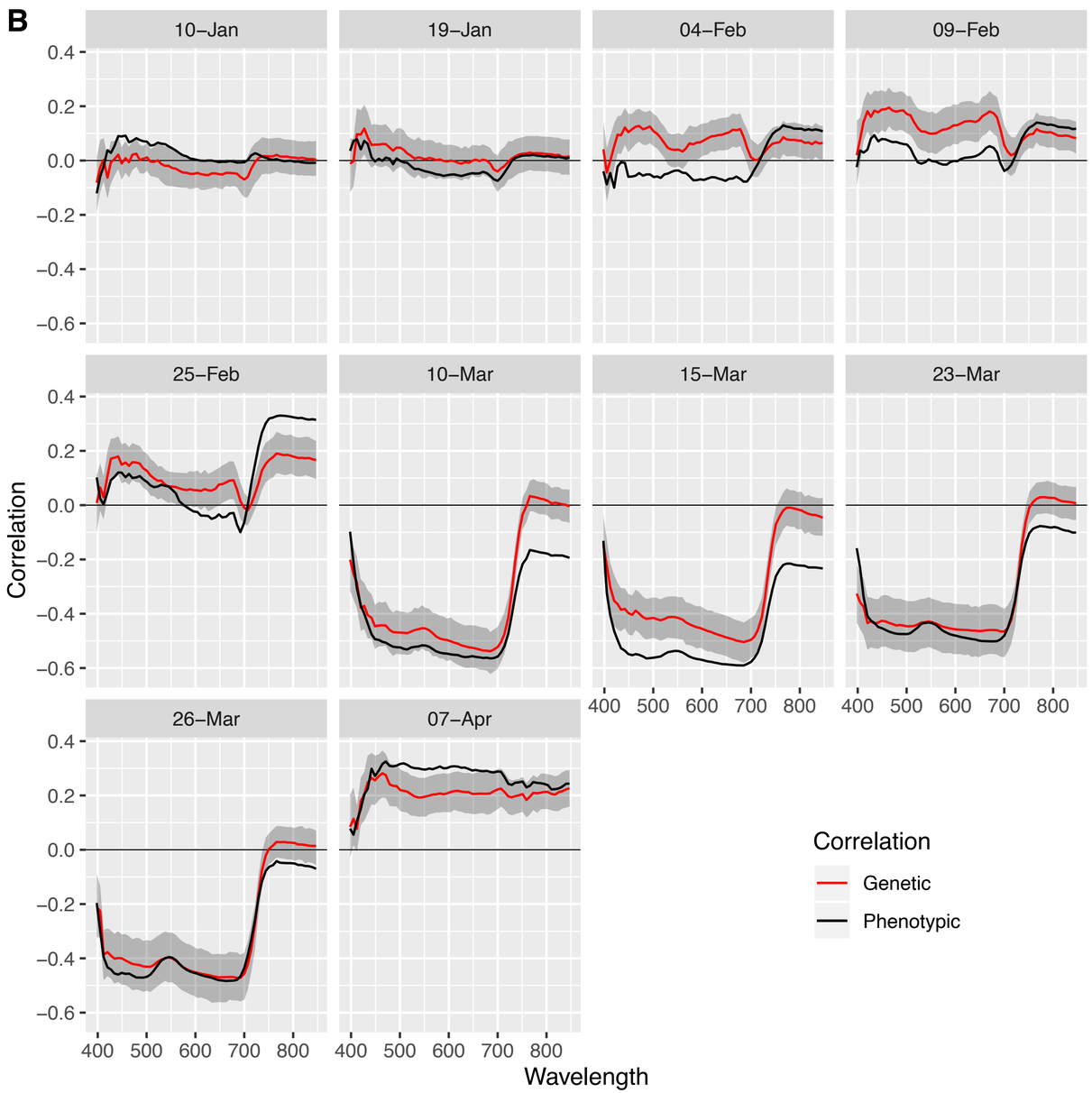
Krause et al 2019

Trait-assisted prediction works

Separating genetic from non-genetic correlations among traits is important

because yield and reflectance measured on the same individuals

Even subtle differences between **G** and **P** matter



Genetic and Phenotypic correlations with yield are mostly similar

But the difference is big enough to significantly reduce prediction accuracy if not accounted for in a MvLMM

Case study 2: Corn multi-environment trial

the Genomes to Fields initiative

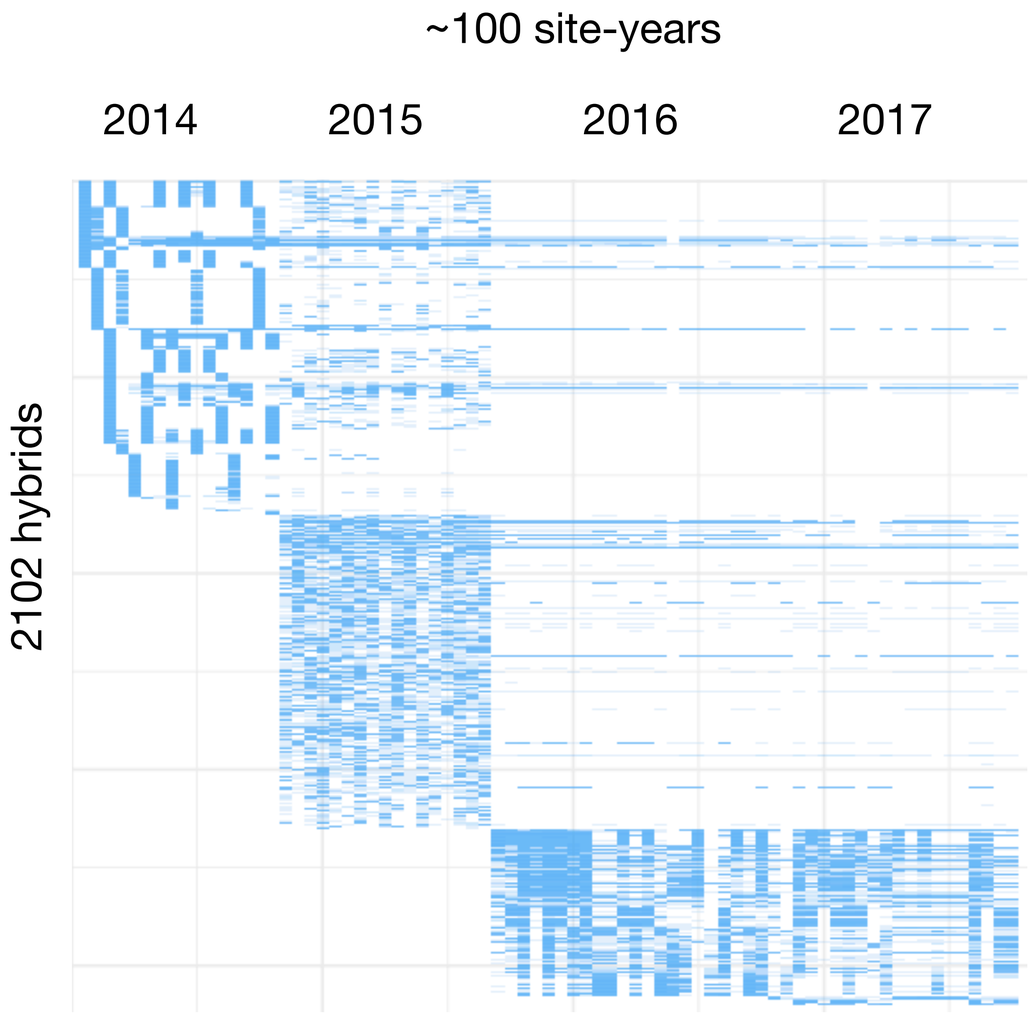
Goal: Genetic value predictions for each site-year

Days to Silking

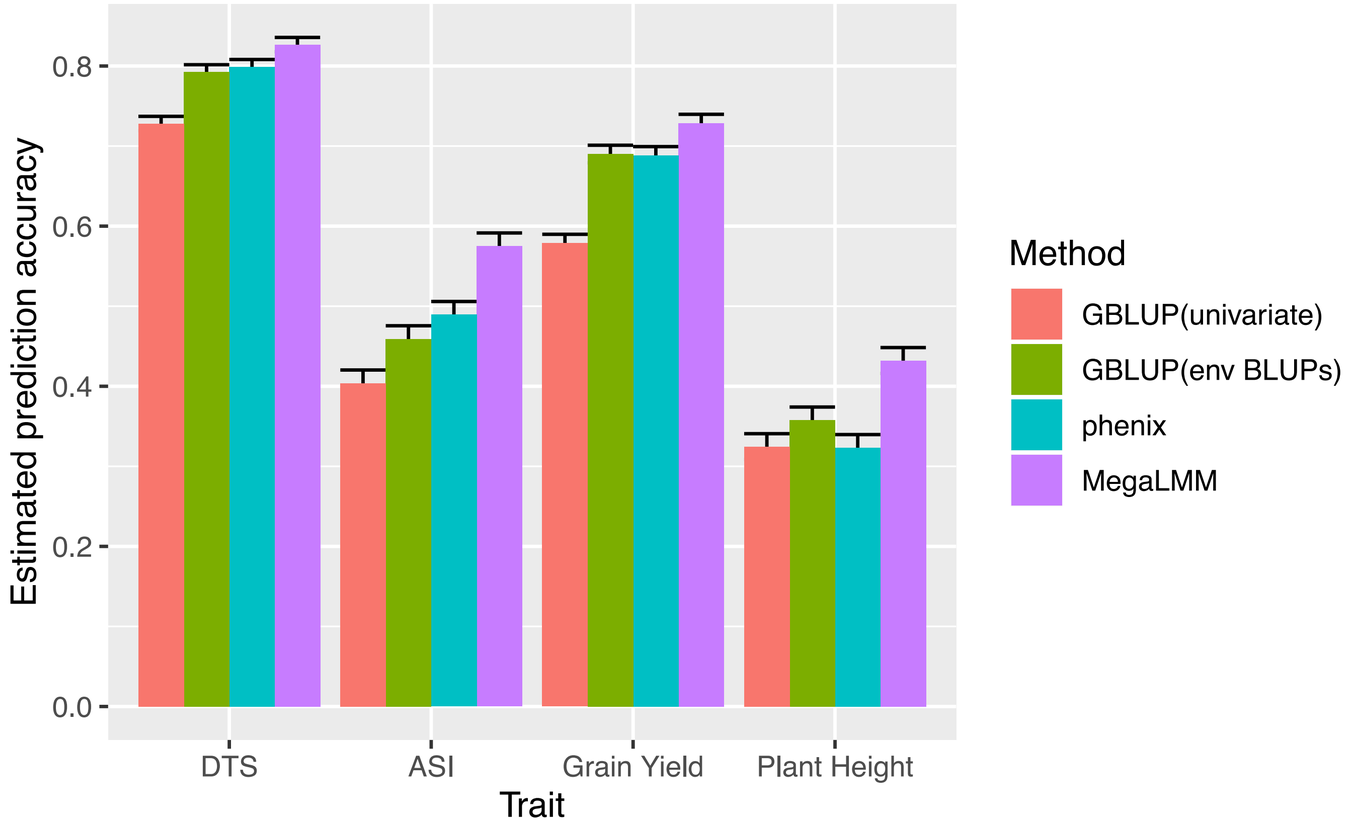
Anthesis-Silking interval

Grain Yield

Plant Height



MegaLMM greatly outperforms other methods

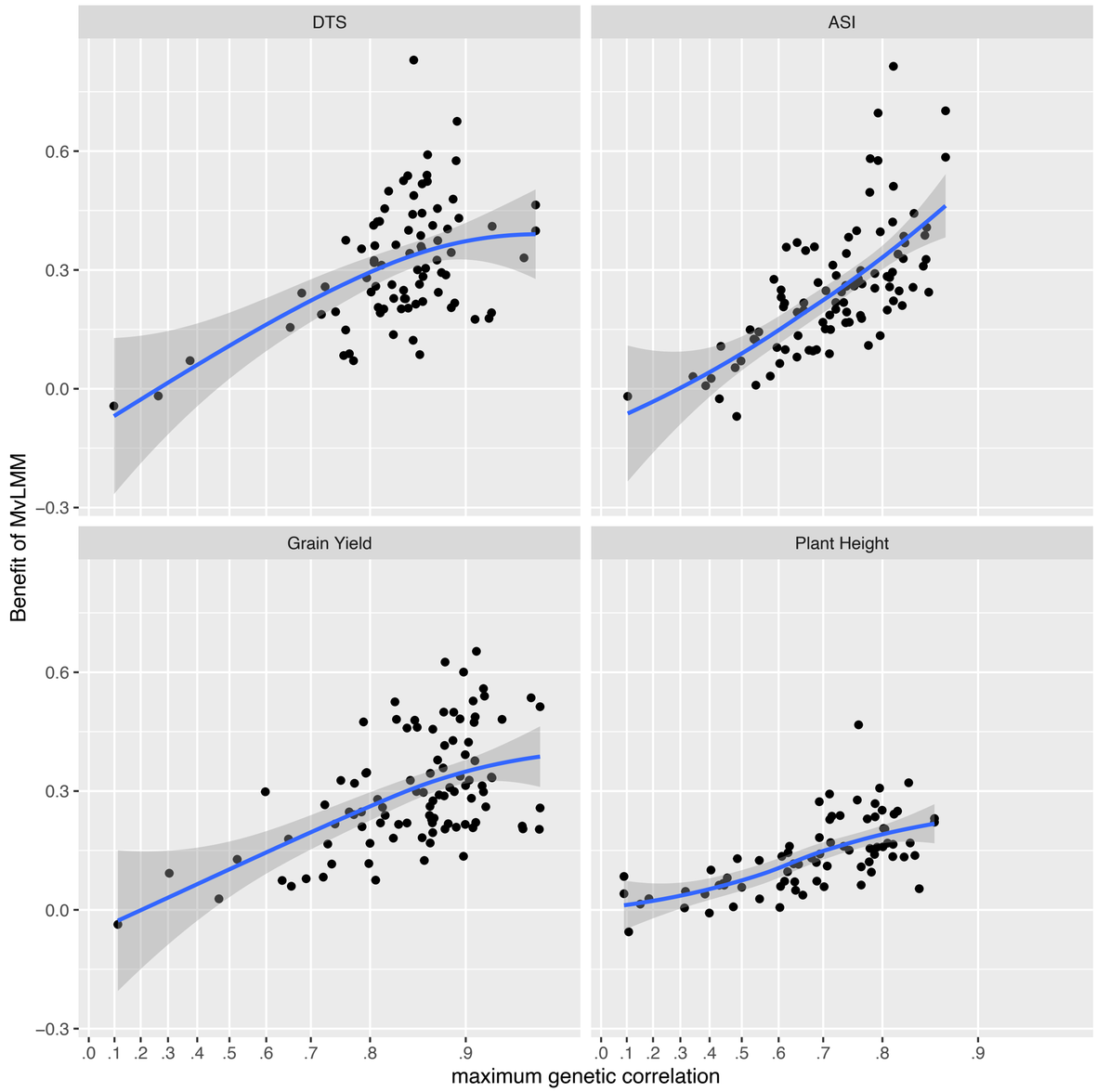


Results are average accuracies across ~100 site-years

MegaLMM almost always the best in every site-year for every trait

Improvement in some site:years even larger

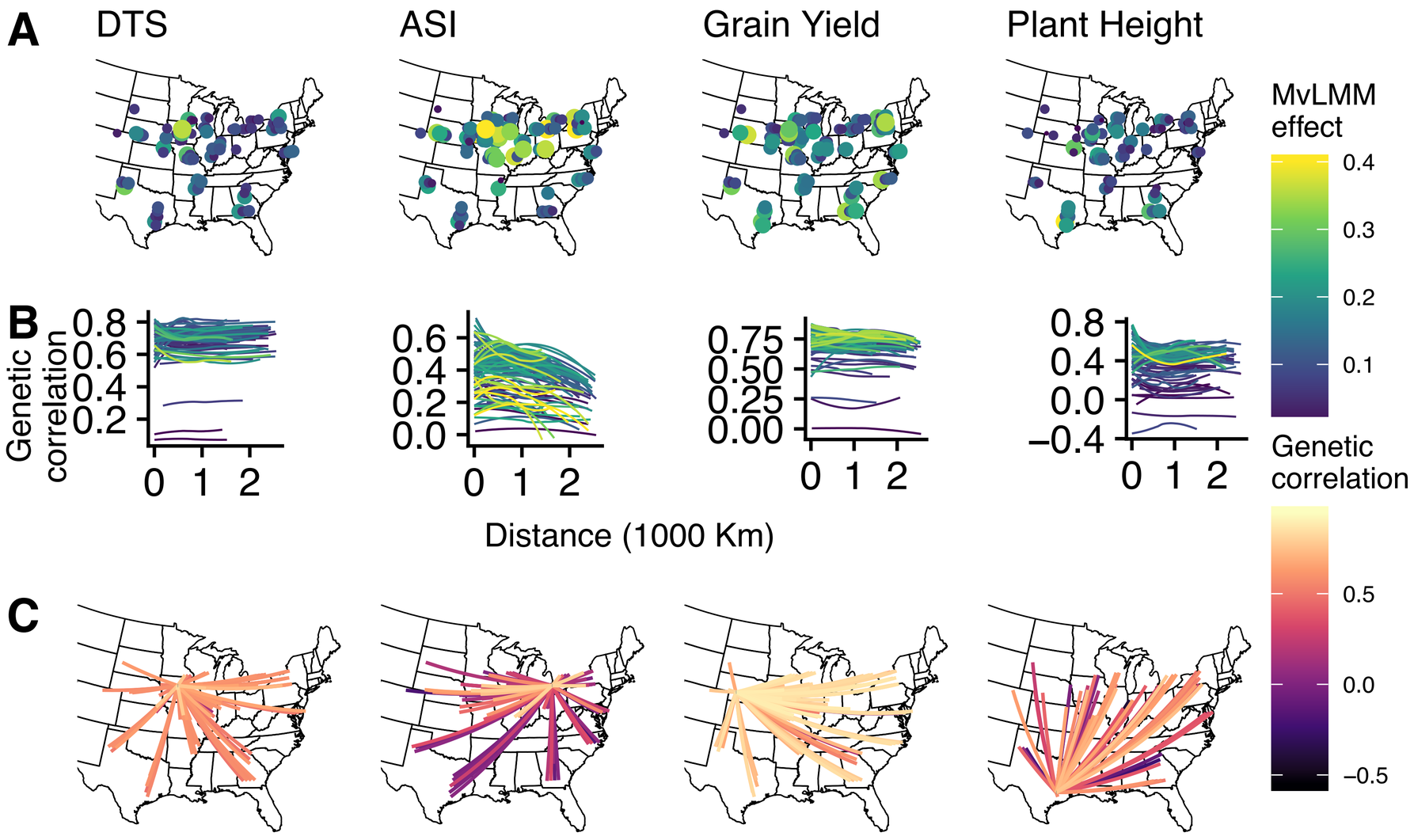
Why does multi-environment prediction work?



Site-years that have a similar partner show the biggest improvement

But the identity of the nearest partner field differs among sites, traits and years

Why does multi-environment prediction work?



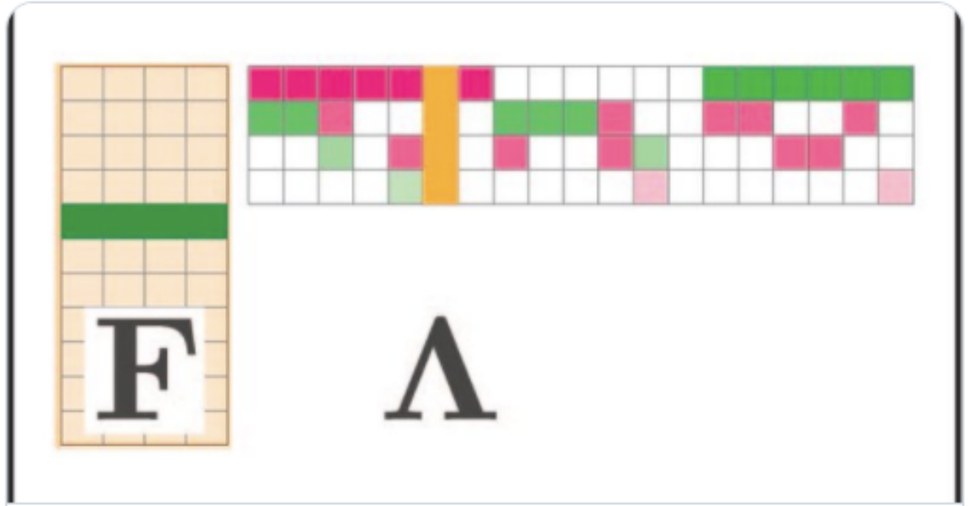
MegaLMM R package

lme4-style model specification

Multiple random effects

Flexible fixed-effect specification

(some) model diagnostics



deruncie/MegaLMM

R package for fitting high-dimensional multivariate linear mixed effect models - deruncie/MegaLMM

github.com

Future Directions

Can we allow more individuals (limit ~5k)?

Can we allow more random effects (limit ~3)?

Can we allow non-Gaussian traits?

Can we do selection on multivariate traits (shape, taste, quality)?

Acknowledgements

Runcie Lab

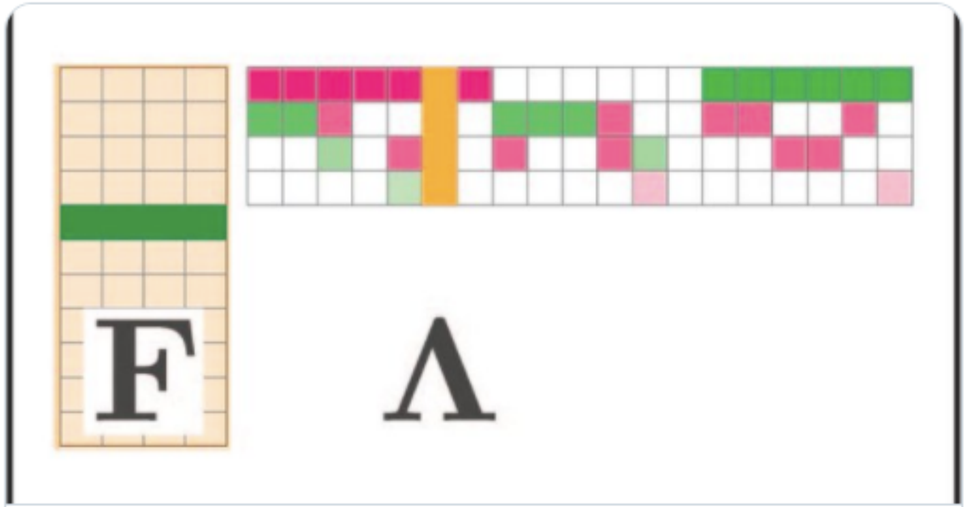
James Ta

Sarah Odell

Xin Li

Jerry Lin

Sayan Mukherjee



deruncie/MegaLMM

R package for fitting high-dimensional multivariate linear mixed effect models - deruncie/MegaLMM

github.com

Funding

NSF

USDA

UC DAVIS

DEPARTMENT OF PLANT SCIENCES