When are models too good to be true?

Accurately evaluating Phenomic Prediction as a tool for plant breeding

Daniel Runcie, Mitchell Feldmann, Fangyi Wang

UCDAVIS
DEPARTMENT OF PLANT SCIENCES

USDA

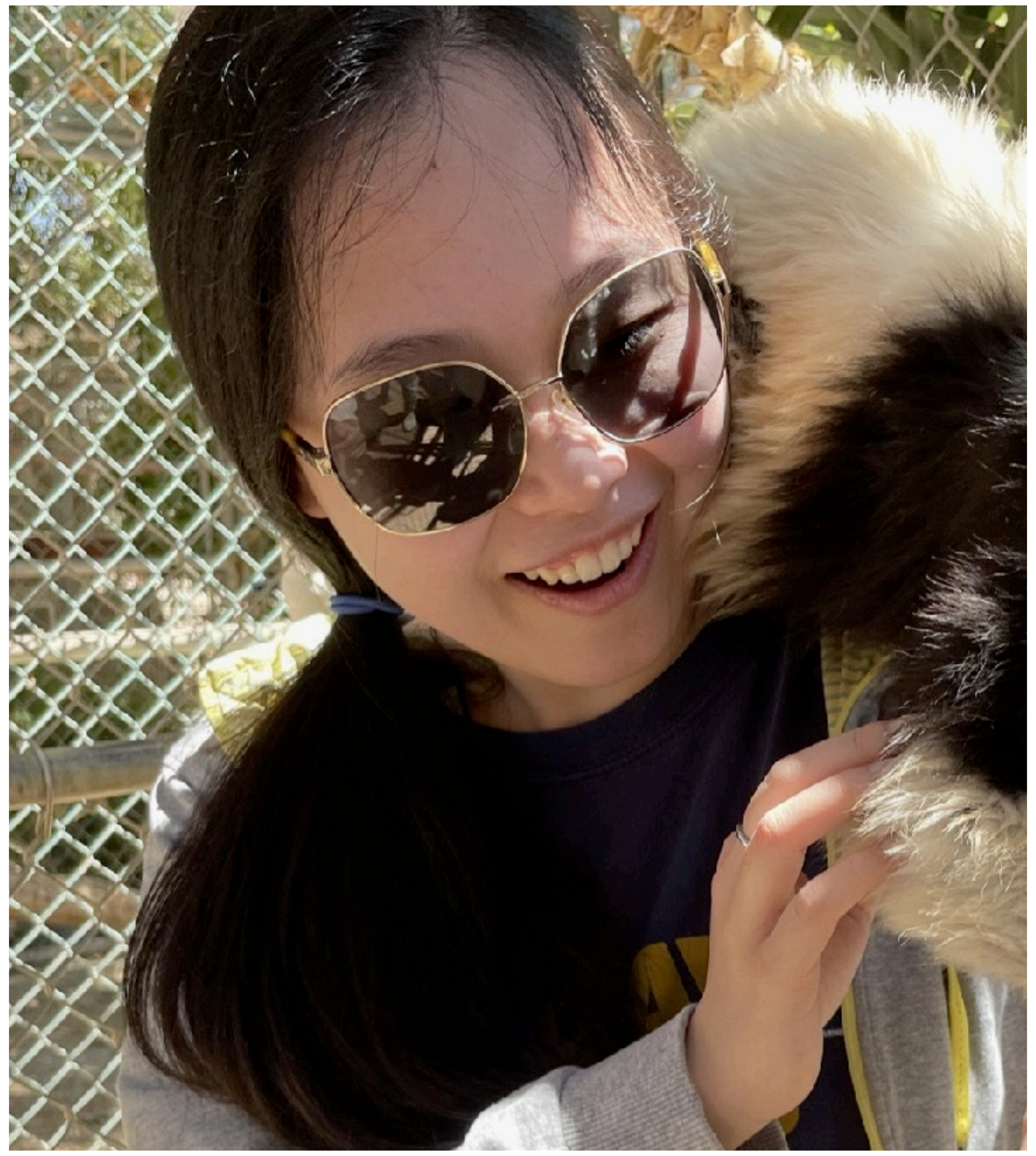
AG2PI
Agricultural Genome to Phenome Initiative

Delivering Resource Allocation Guidelines for Optimizing High-Throughput Phenotyping and Genotyping in Modern Breeding Programs

Daniel Runcie

Associate Professor

Quantitative Genetics and Statistics



Fangyi Wang

Graduate Student

Integrated Genetics and Genomics

Mitchell Feldmann

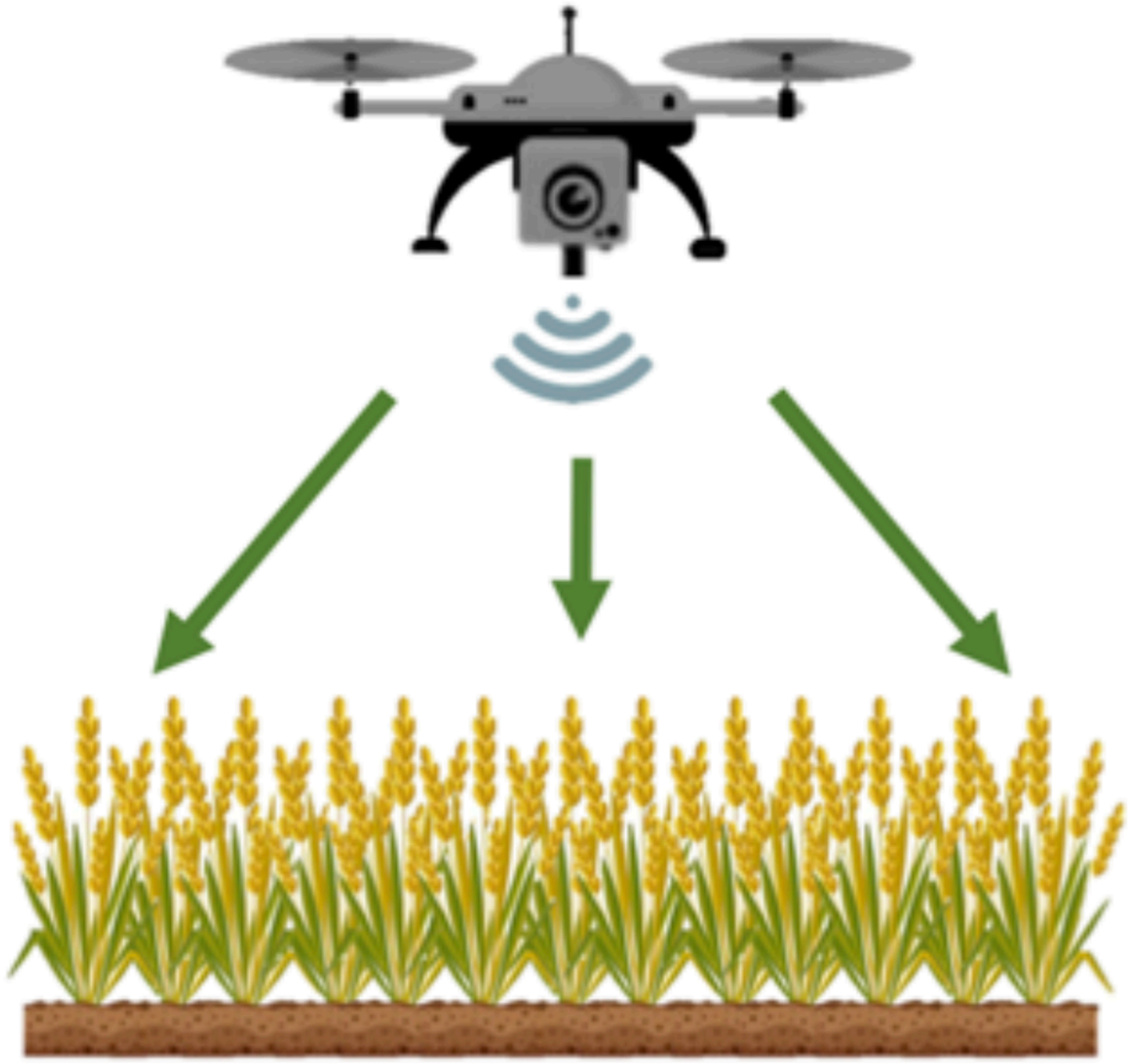
Assistant Professor

Director Elect of Strawberry Breeding program

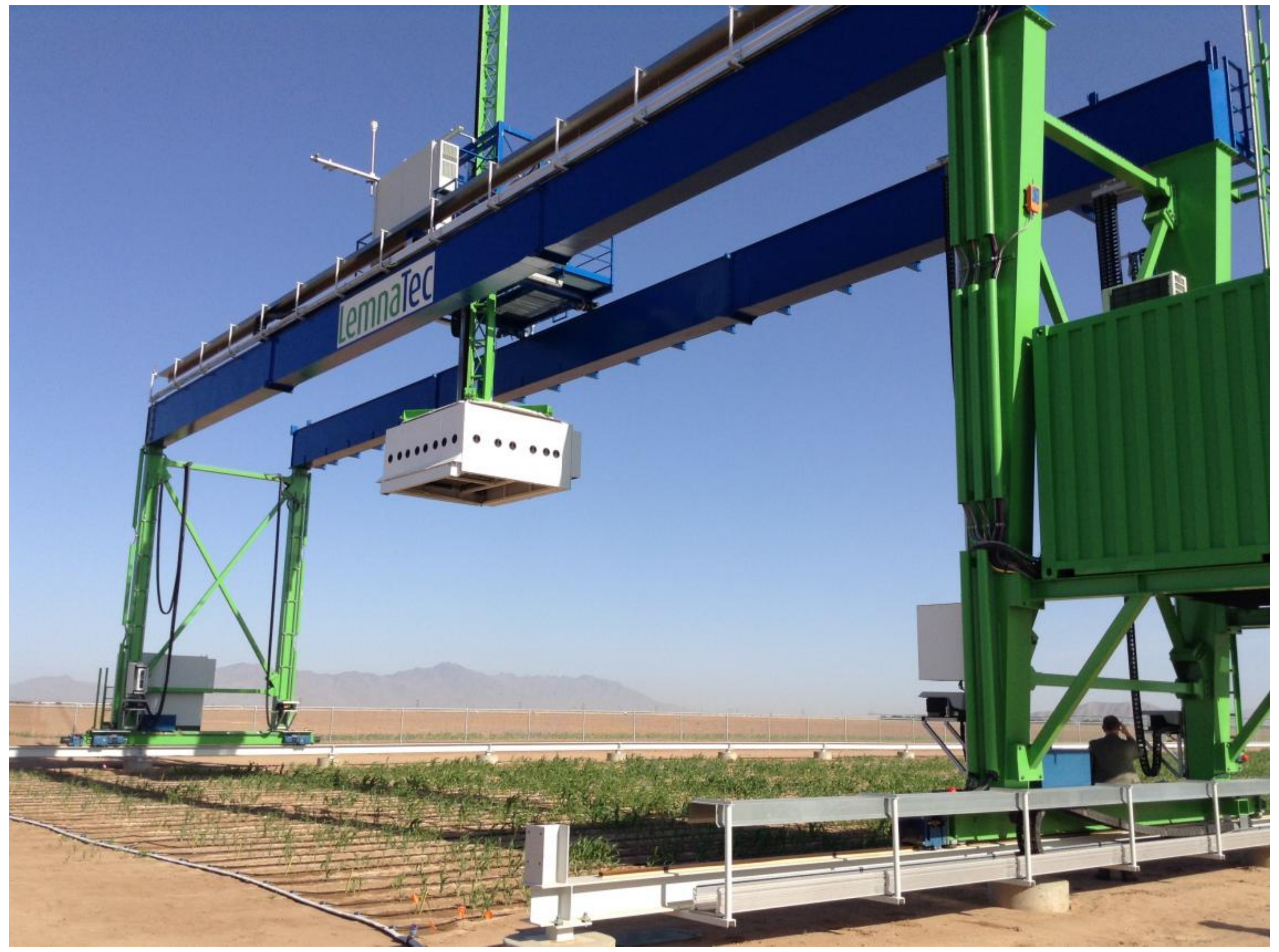
UCDAVIS
DEPARTMENT OF PLANT SCIENCES

AG2PI
Agricultural Genome to Phenome Initiative

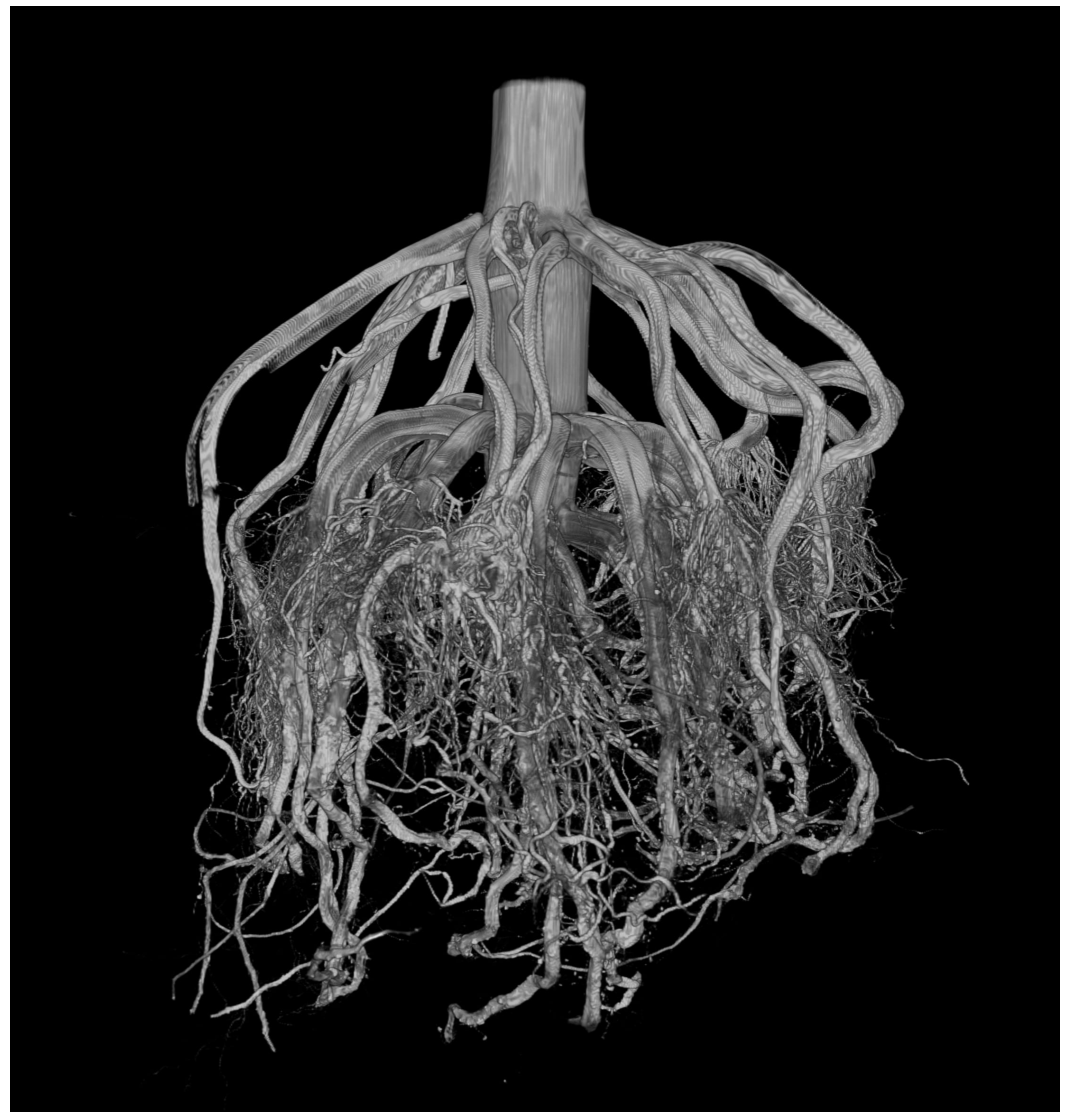
Many new technologies hold promise for improving breeding



Lopez Cruz et al 2020



University of Arizona

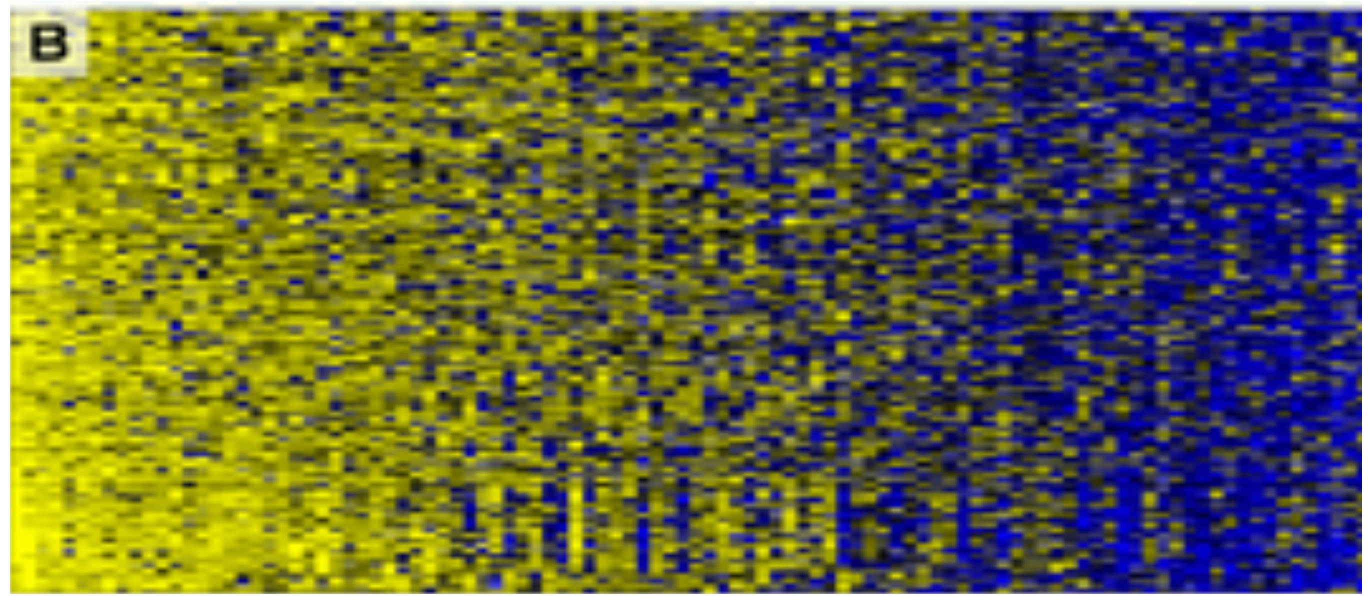


Chris Topp

Muncan et al 2022

Danforth Center

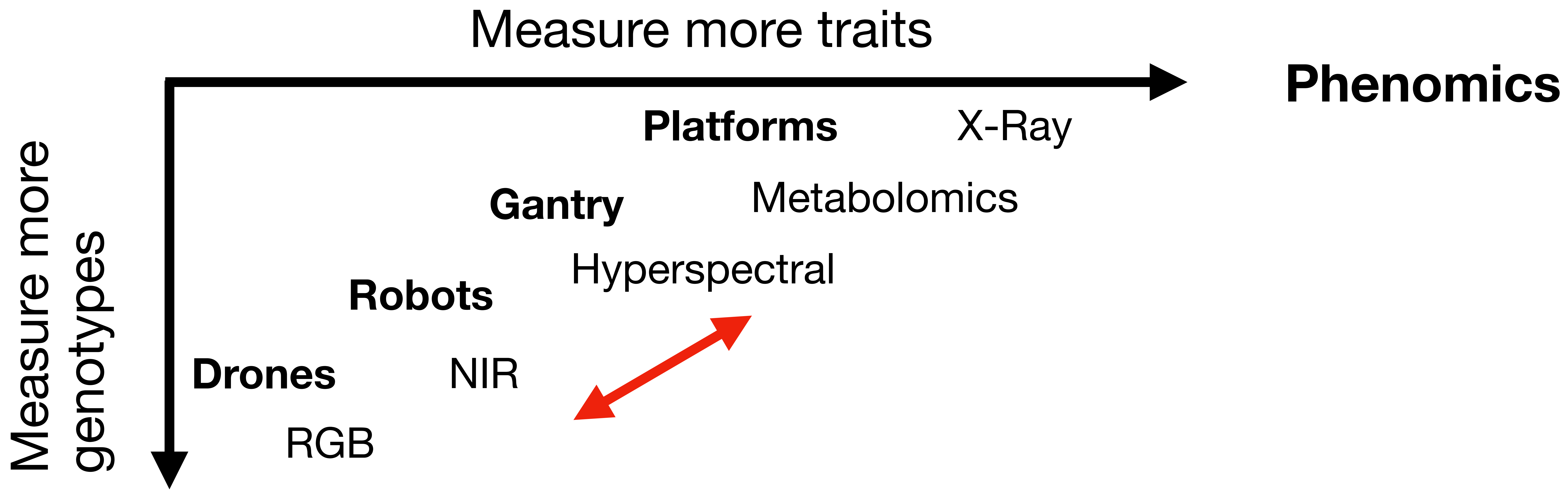
Gene expression / Metabolomics



Which technologies should we invest in and how should they be used?

Technologies are expensive

Require reduced investment in other aspects of a breeding program

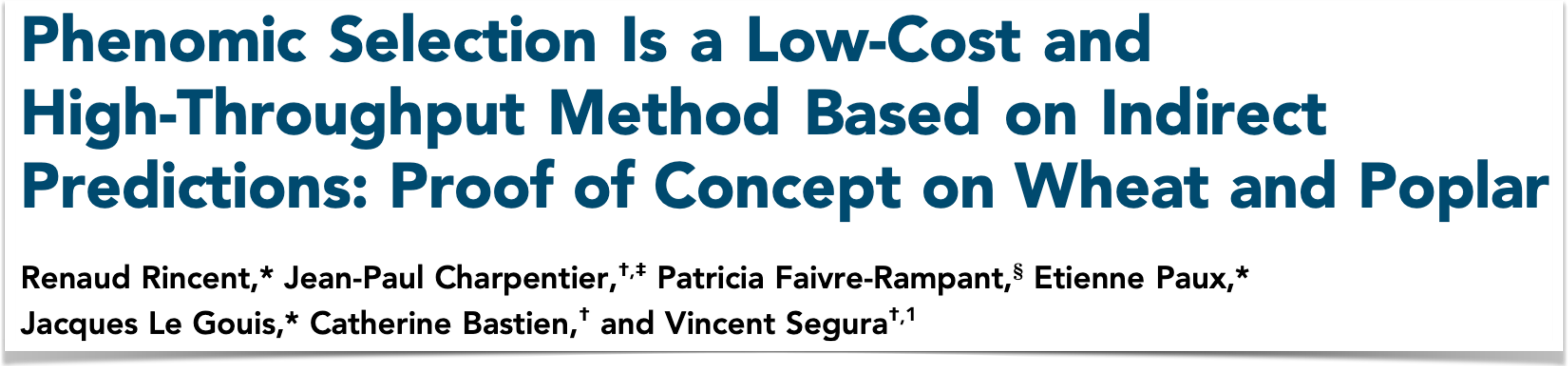


High Throughput Phenotyping (HTP)

Which dimension is more important?

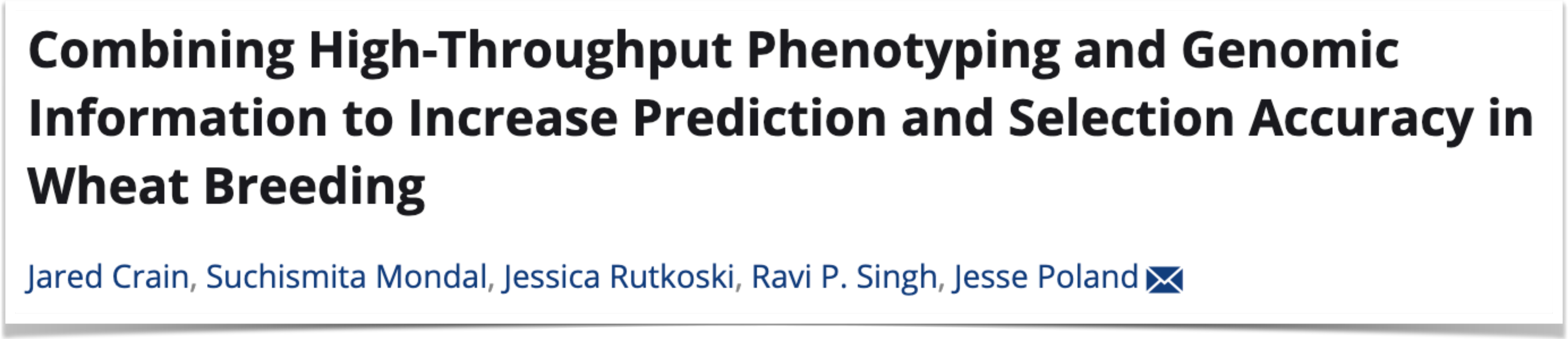
How do you use them in breeding decisions?

Idea: Use Phenomic data for Phenomic Selection



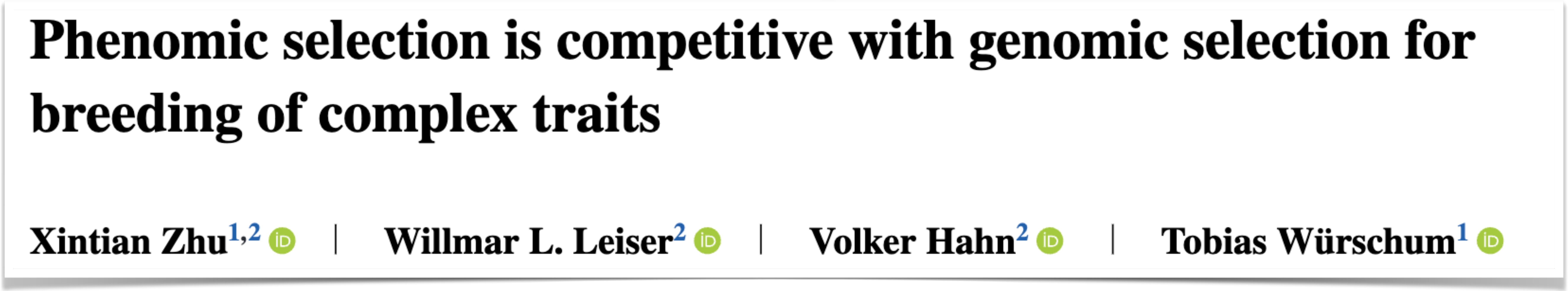
Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar

Renaud Rincent,[*] Jean-Paul Charpentier,[†,‡] Patricia Faivre-Rampant,[§] Etienne Paux,[*] Jacques Le Gouis,[*] Catherine Bastien,[†] and Vincent Segura[†,1]



Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding

Jared Crain, Suchismita Mondal, Jessica Rutkoski, Ravi P. Singh, Jesse Poland ✉



Phenomic selection is competitive with genomic selection for breeding of complex traits

Xintian Zhu[1,2] | Willmar L. Leiser[2] | Volker Hahn[2] | Tobias Würschum[1]

Idea: Use Phenomic data for Phenomic Selection

1) Genomic Selection is widely used and very successful

2) Genomic data is expensive

3) Phenomic data is cheaper than genomic data

4) We can predict difficult-to-measure traits as well using phenomic data as with genomic data

Can we use Phenomic Selection as a cheap and effective replacement for Genomic Selection?

Can Phenomic Selection be a cheap replacement for Genomic Selection?

Our conclusion: No … at least not in this way

1) The comparison between Phenomic Prediction and Genomic Prediction isn’t fair

2) This isn’t the right question to ask

Phenomics can complement, but can’t replace Genomic Selection

Phenomic Prediction’s use in is measuring traits, not genetic values

Outline

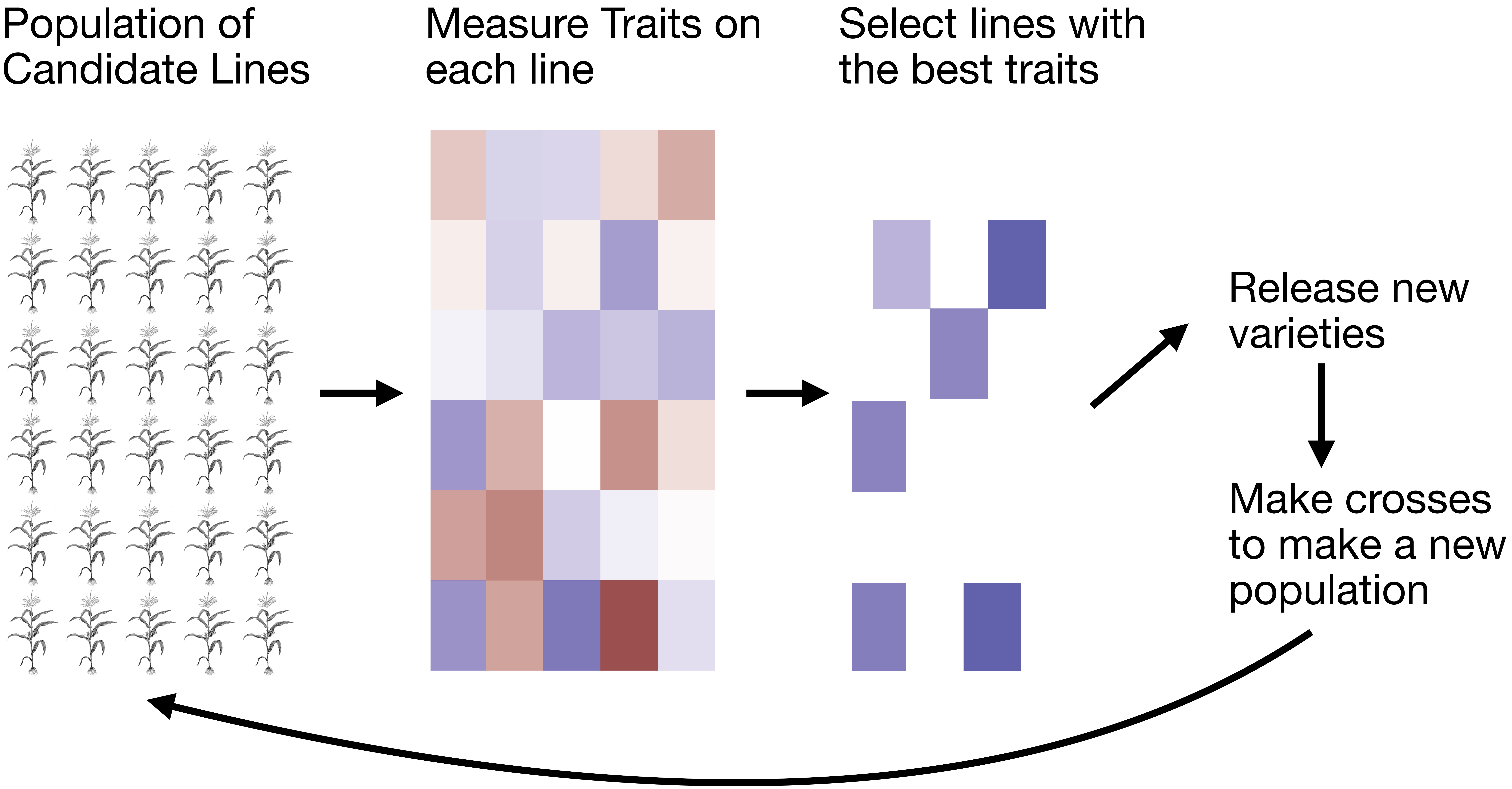
How does breeding work?

How can Genomic and Phenomic Selection fit into breeding programs?

Why are comparisons between Phenomic and Genomic Prediction misleading?

When is Phenomic Prediction most useful?

How does breeding work?



Population of Candidate Lines

Measure Traits on each line

Select lines with the best traits

Release new varieties

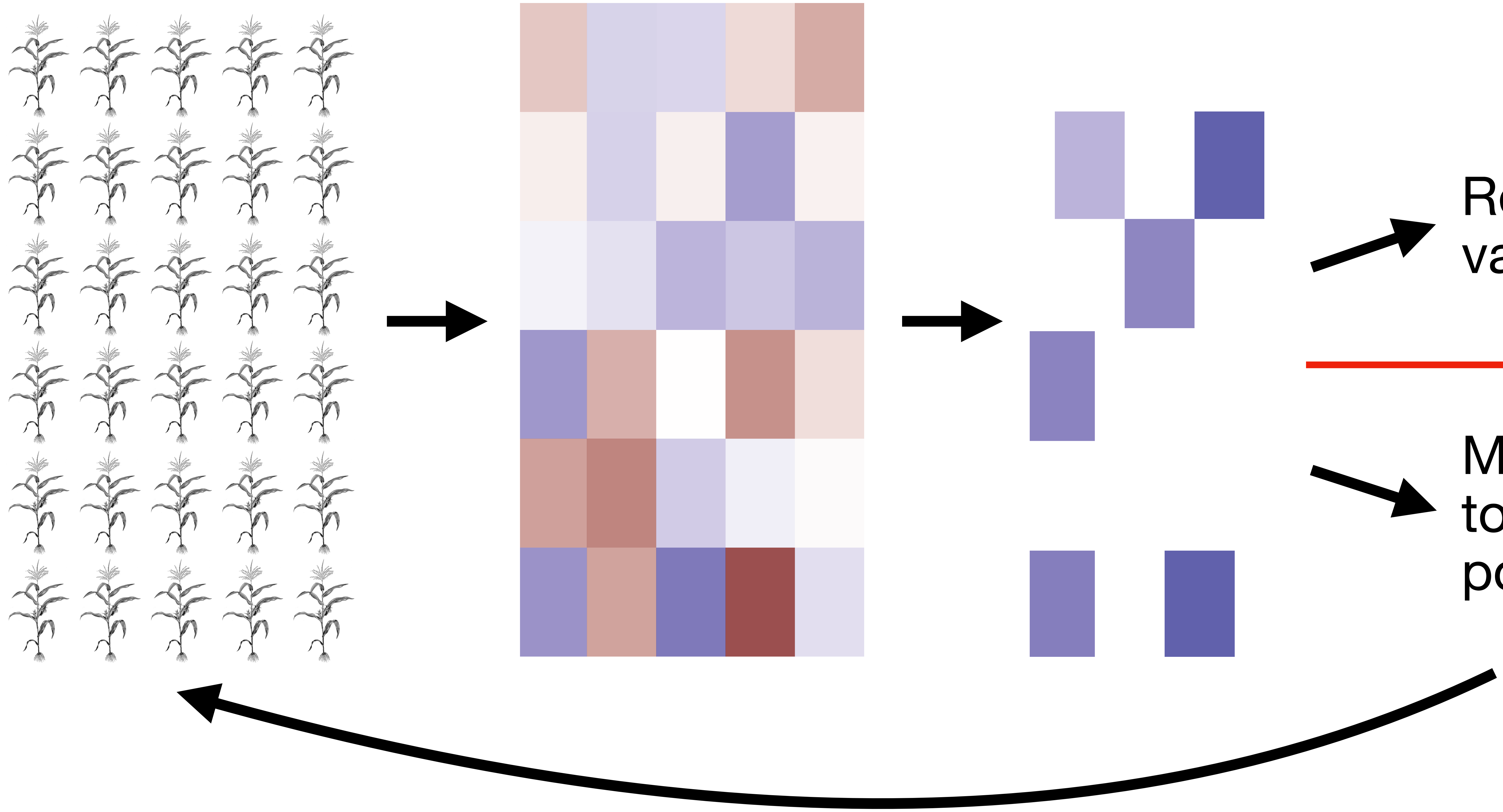
Make crosses to make a new population

How can we make breeding better?

Population of Candidate Lines

Measure Traits on each line

Select lines with the best **Genetic Values**



Averages over: environmental variation, GxE, measurement error

Requires lots of **phenotyping**

Release new varieties

Total Genetic Values

Make crosses to make a new population

Breeding Values

Excludes: Dominance, Epistasis

Requires **quantitative genetic models** of inheritance

Long-term success depends on improving the population

This is captured by the Breeder’s Equation

Rate of Gain

Intensity

Accuracy

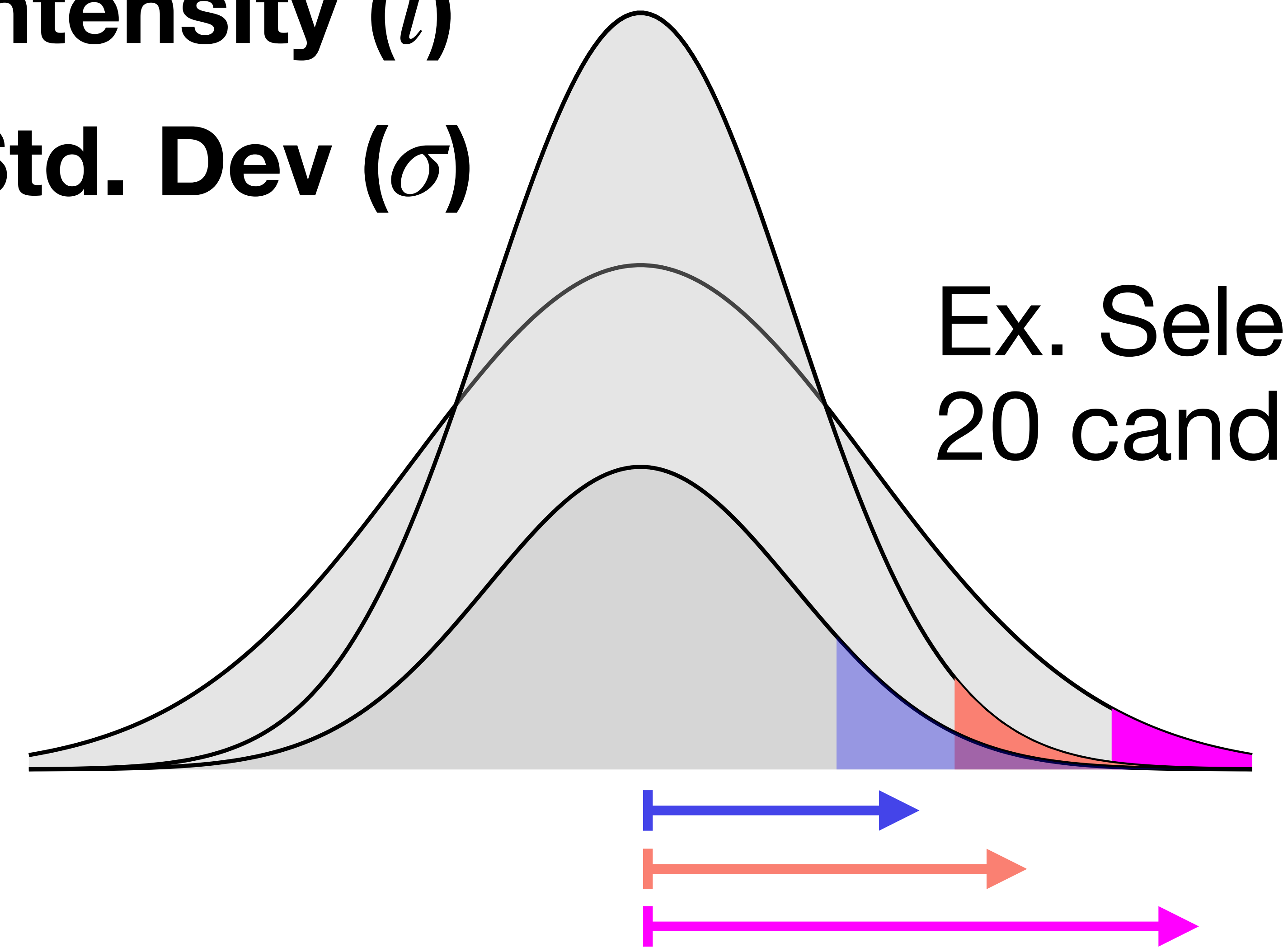
Std. Dev

Speed

$$\Delta g = i \cdot r \cdot \sigma \cdot 1/L$$

Intensity (i)

Std. Dev (σ)

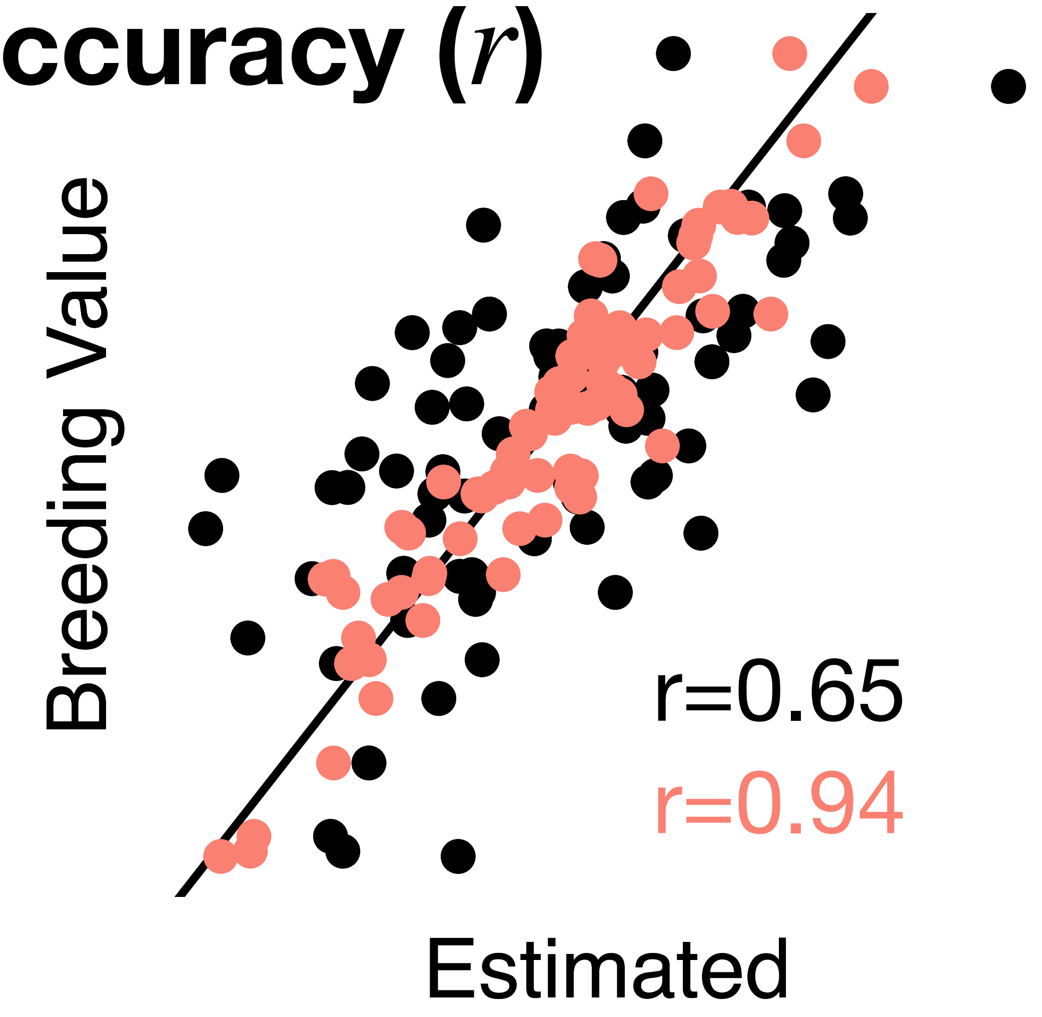


Ex. Select top 20 candidates

Measuring **more candidate lines** means faster gains

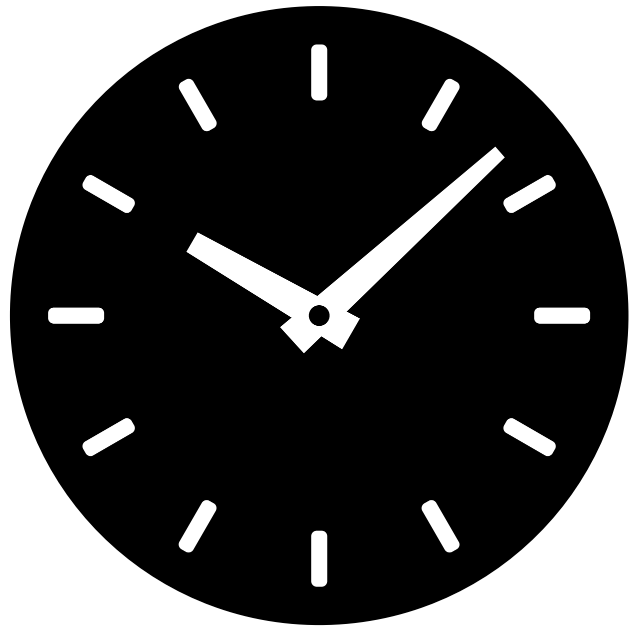
Higher variance **breeding values** means faster gains

Accuracy (r)



Higher **correlation** between estimated and actual **breeding values** means faster gains

Cycle Length (L)



Faster crossing decisions means faster gains

This is captured by the Breeder's Equation

Rate of Gain

Intensity

Accuracy

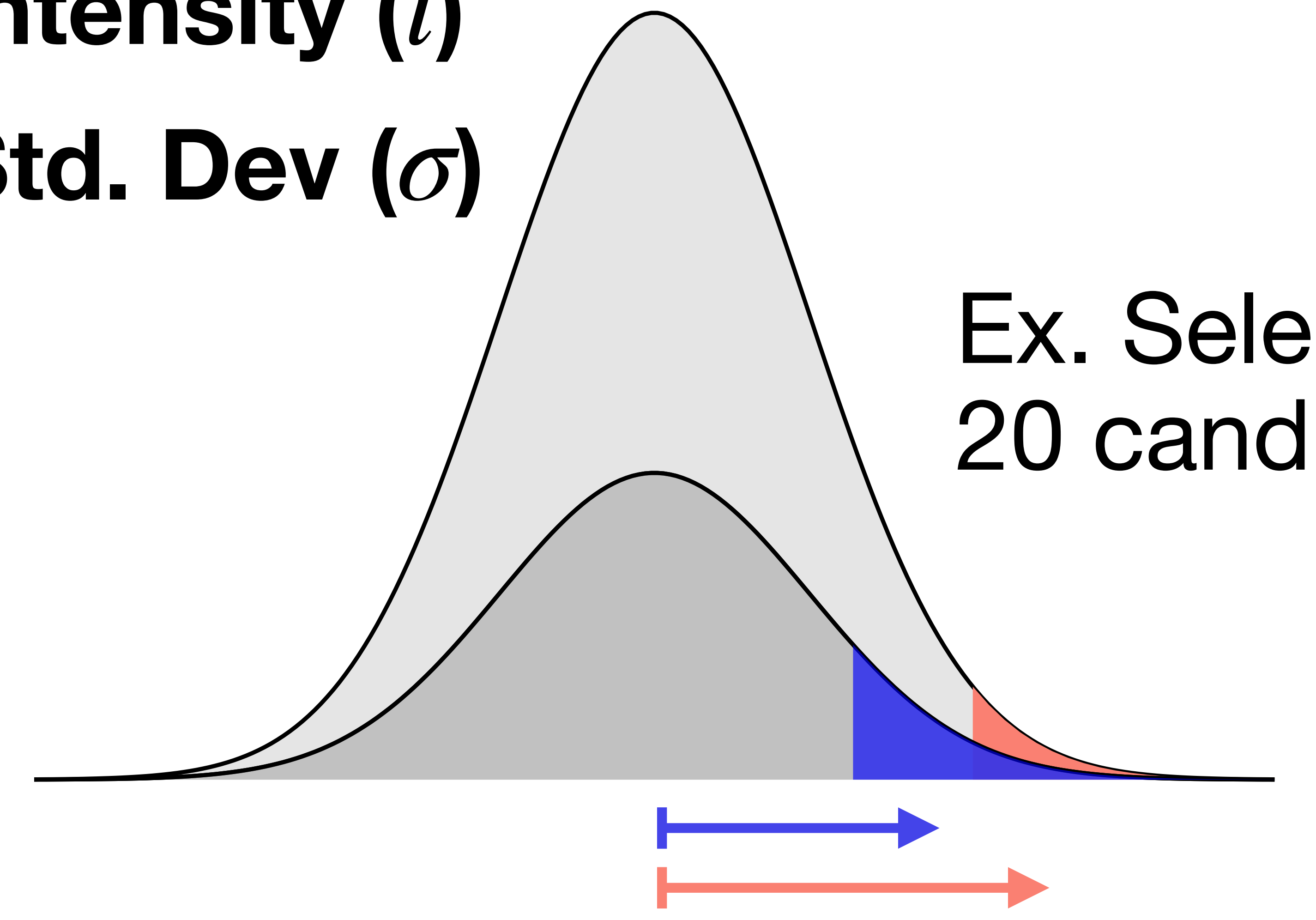
Std. Dev

Speed

$$\Delta g = i \cdot r \cdot \sigma \cdot 1/L$$

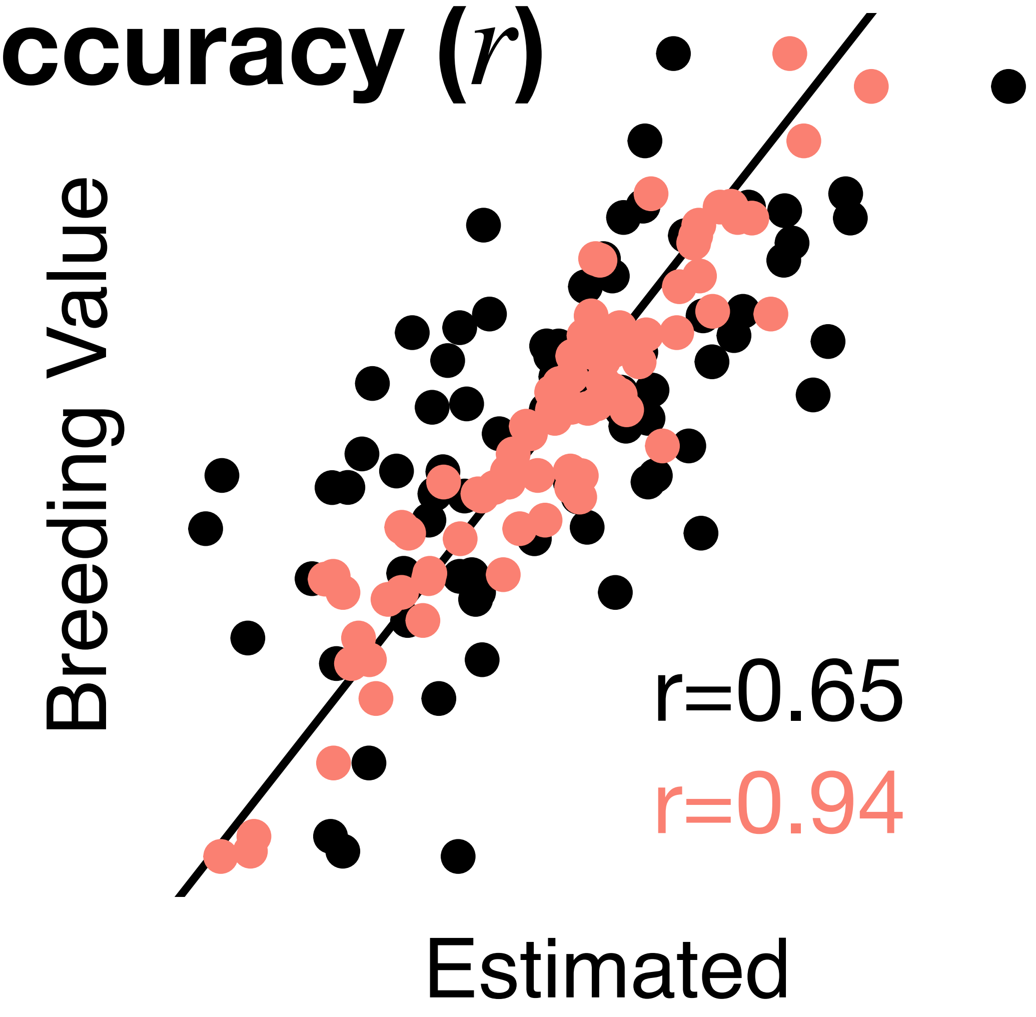
Intensity (i)

Std. Dev (σ)

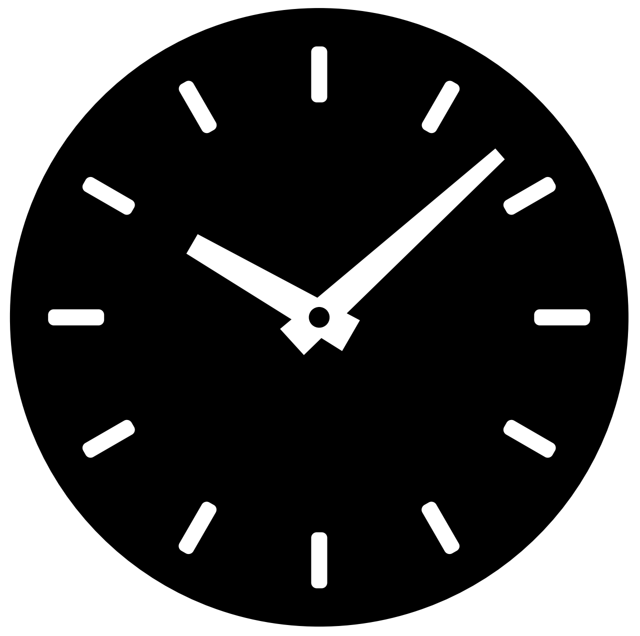


Ex. Select top 20 candidates

Accuracy (r)



Cycle Length (L)



How do **Genomic Selection** and **Phenomic Selection** affect each parameter?

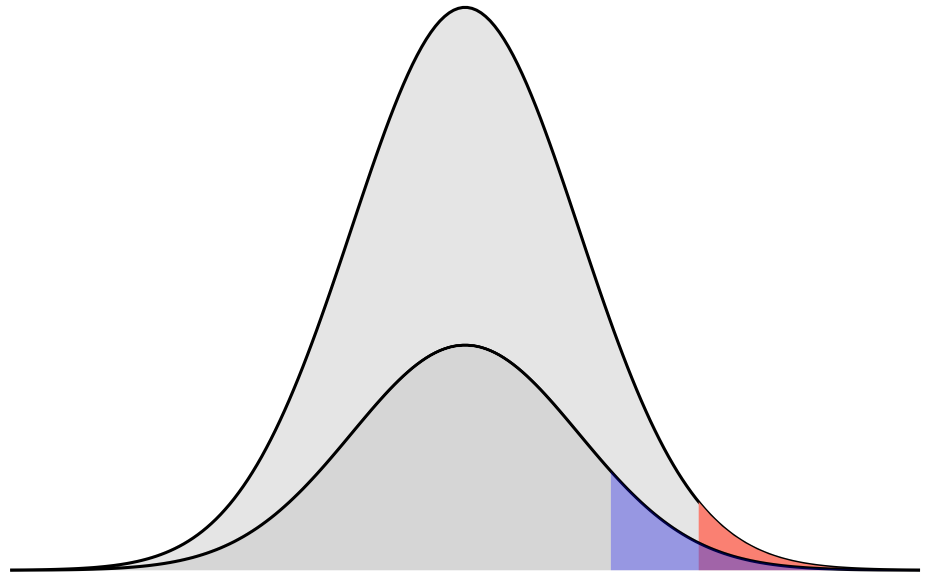
How do changes to one parameter affect the others?

Genomic Selection: Making crossing decisions based on genetic markers

Genomic Selection improves intensity, accuracy and speed

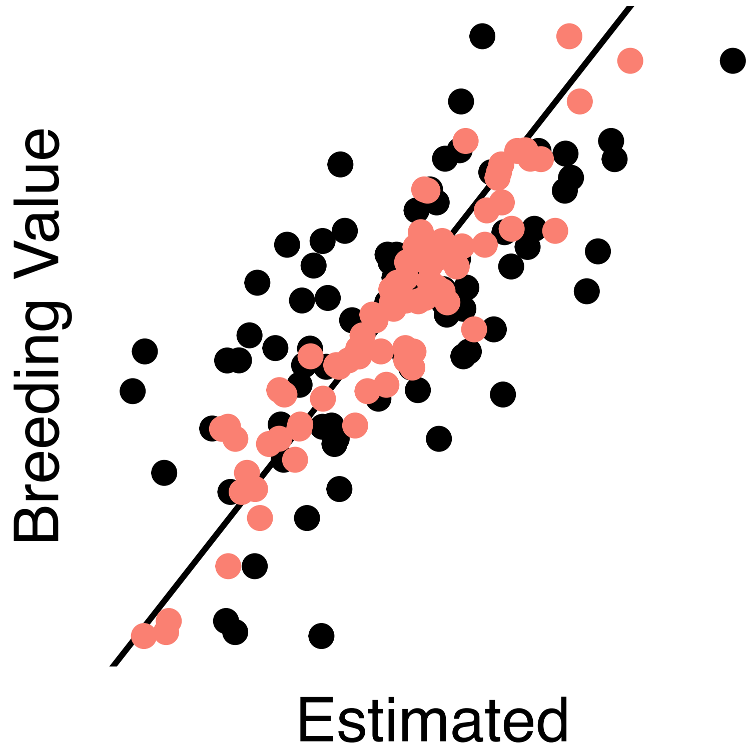
$$\Delta g = i \cdot r \cdot \sigma \cdot 1/L$$

Intensity (*i*)



If genotyping is cheap, you can evaluate more lines because they don't take field space

Accuracy (*r*)



If H^2 is low you can estimate genetic values of **alleles** instead of lines

Cycle Length (L)

You can make crossing decisions immediately without waiting for field trials

Main benefit of Genomic Prediction: Cycle Length

Gaynor et al 2017

Genomic Selection uses Genomic Prediction models

1) Measure target traits on some lines and train a **Genomic Prediction** model

Trait

Genetic markers

Parameters

Error

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

2) Use the model to predict **breeding values** of other lines

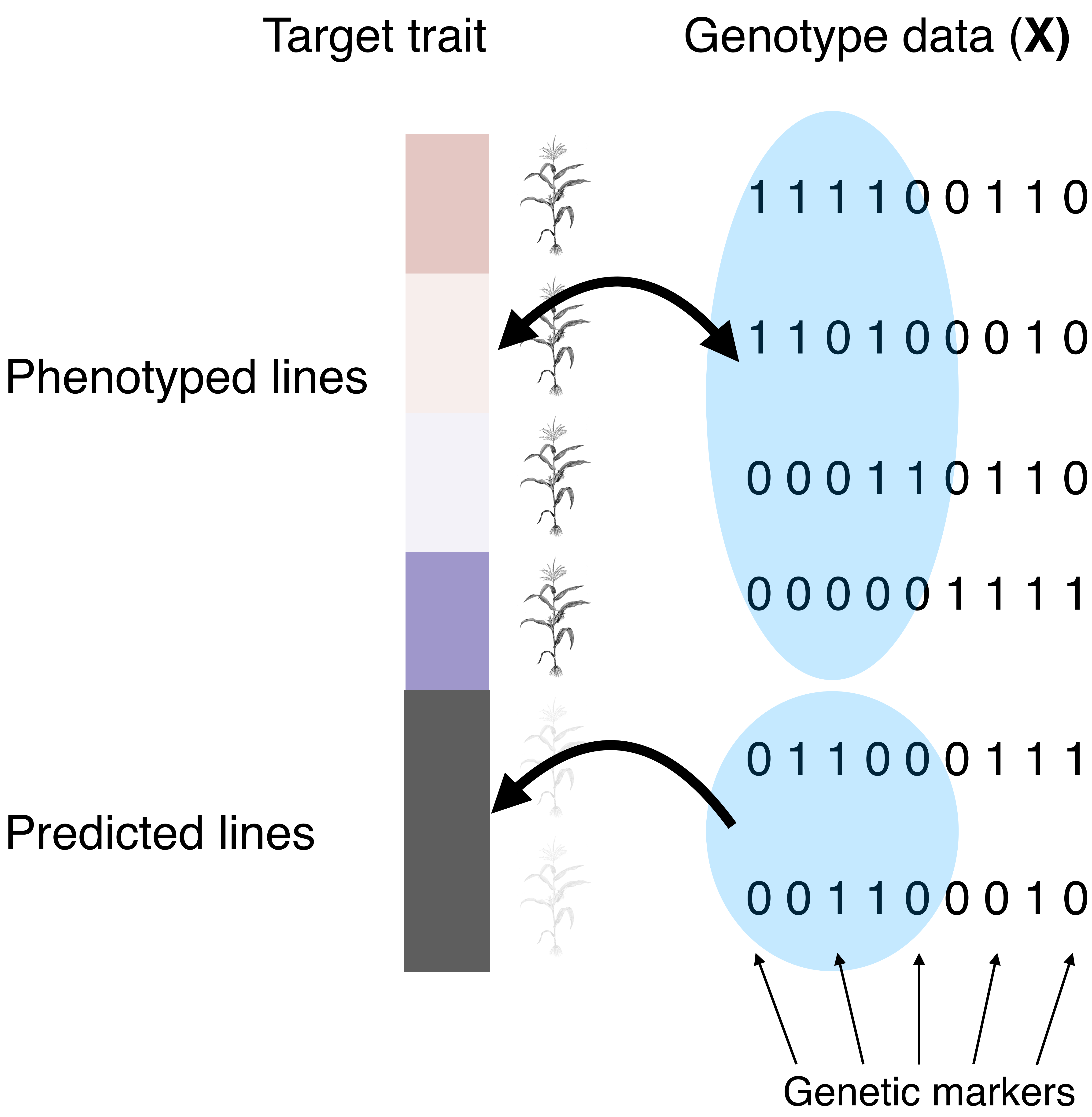
$$\hat{\mathbf{u}}_{\mathbf{g}} = \tilde{\mathbf{X}}\beta$$

Parameters learned from training

Predicted breeding values

Genotypes of new lines

3) Select lines based on $\hat{\mathbf{u}}_{\mathbf{g}}$ without phenotyping



Target trait

Genotype data (**X**)

Phenotyped lines

1 1 1 1 0 0 1 1 0

1 1 0 1 0 0 0 1 0

0 0 0 1 1 0 1 1 0

0 0 0 0 0 1 1 1 1

Predicted lines

0 1 1 0 0 0 1 1 1

0 0 1 1 0 0 0 1 0

Genetic markers

Genomic Prediction models are black boxes

Key ideas:

Trait

Genetic markers

Parameters

Error

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

Genotype data is **High Dimensional**

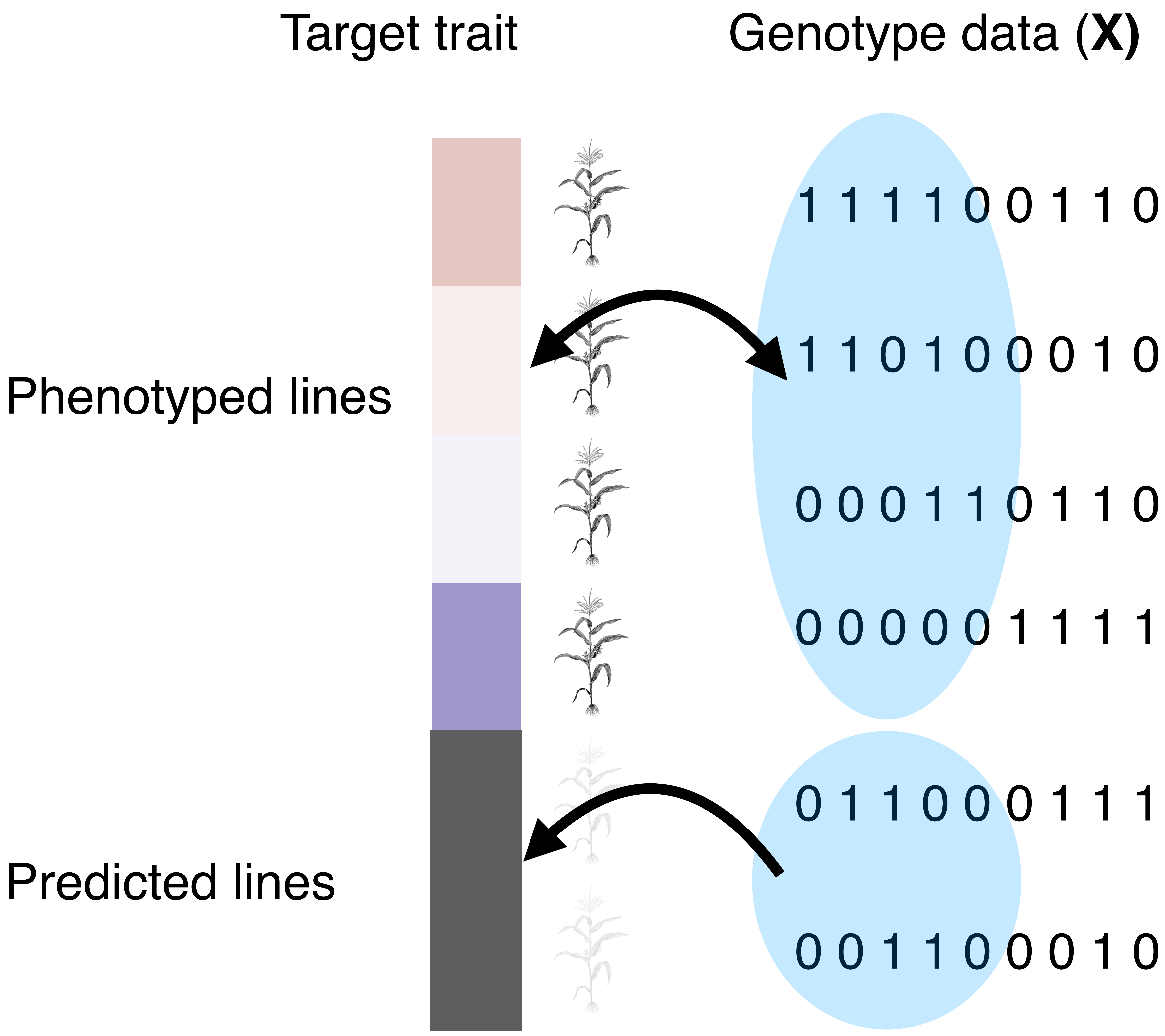
p >> n

We don’t have to know which features are useful beforehand

We expect some to be useful because of LD

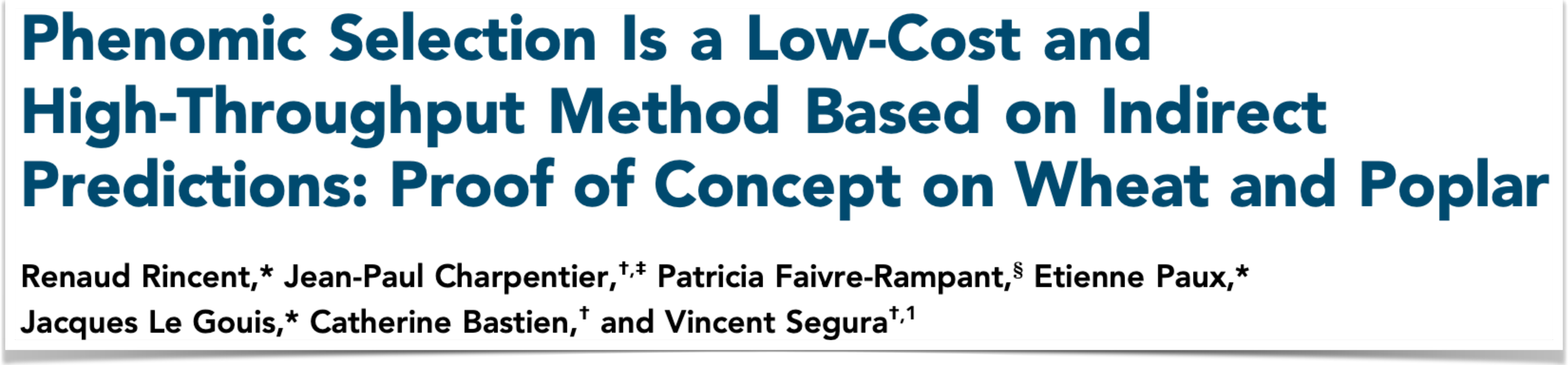
Models like rrBLUP, BayesB, RKHS work well

But… Genetic marker data is expensive



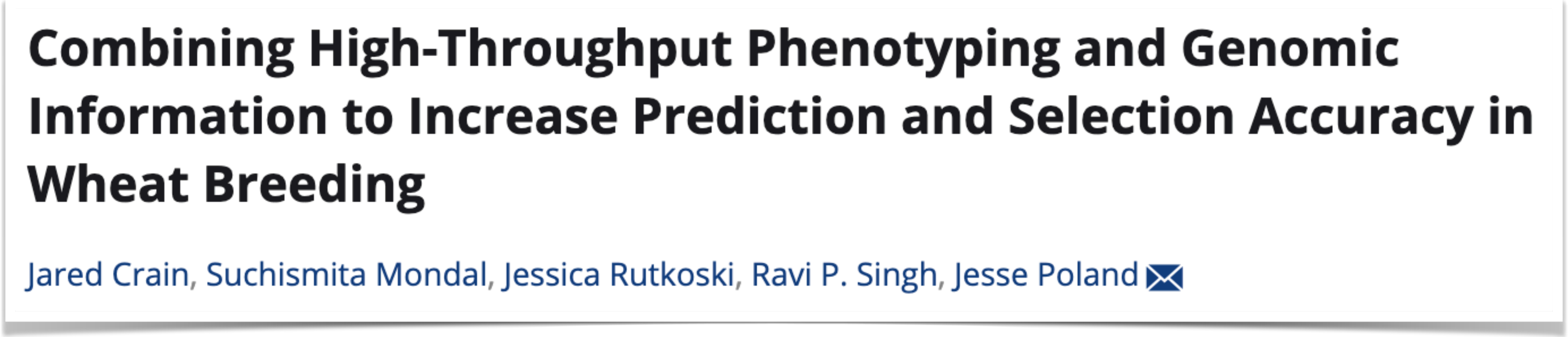
1 1 1 1 0 0 1 1 0
1 1 0 1 0 0 0 1 0
0 0 0 1 1 0 1 1 0
0 0 0 0 0 1 1 1 1
0 1 1 0 0 0 1 1 1
0 0 1 1 0 0 0 1 0

Can Phenomic Selection be a cheap replacement for Genomic Selection?



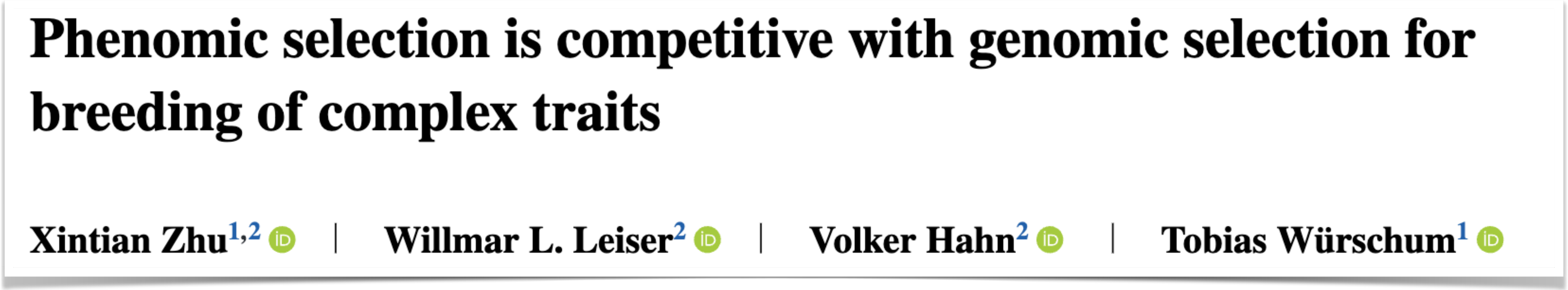
Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar

Renaud Rincent,* Jean-Paul Charpentier,[†,‡] Patricia Faivre-Rampant,[§] Etienne Paux,*
Jacques Le Gouis,* Catherine Bastien,[†] and Vincent Segura[†,1]



Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding

Jared Crain, Suchismita Mondal, Jessica Rutkoski, Ravi P. Singh, Jesse Poland



Phenomic selection is competitive with genomic selection for breeding of complex traits

Xintian Zhu[1,2] | Willmar L. Leiser[2] | Volker Hahn[2] | Tobias Würschum[1]

Phenomic Selection is deceptively similar

1) Measure target traits on some lines and train a **Phenomic Prediction** model

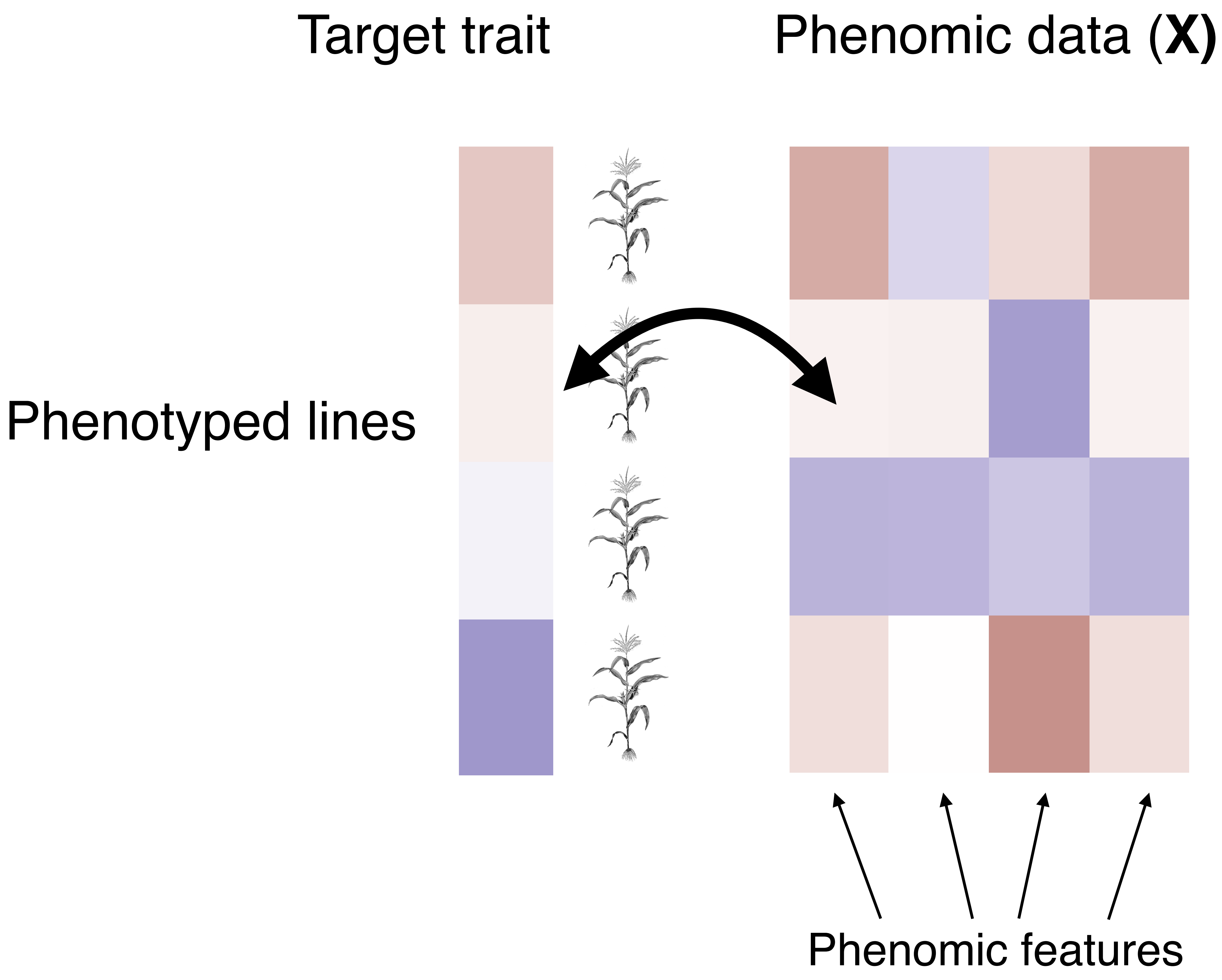
Trait

Phenomic features

Parameters

Error

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$



NIR wavelengths

Multiple timepoints

Multiple environments

Phenomic Selection is deceptively similar

1) Measure target traits on some lines and train a **Phenomic Prediction** model

Trait

Phenomic features

Parameters

Error

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

2) Use the model to predict **target trait** of other lines

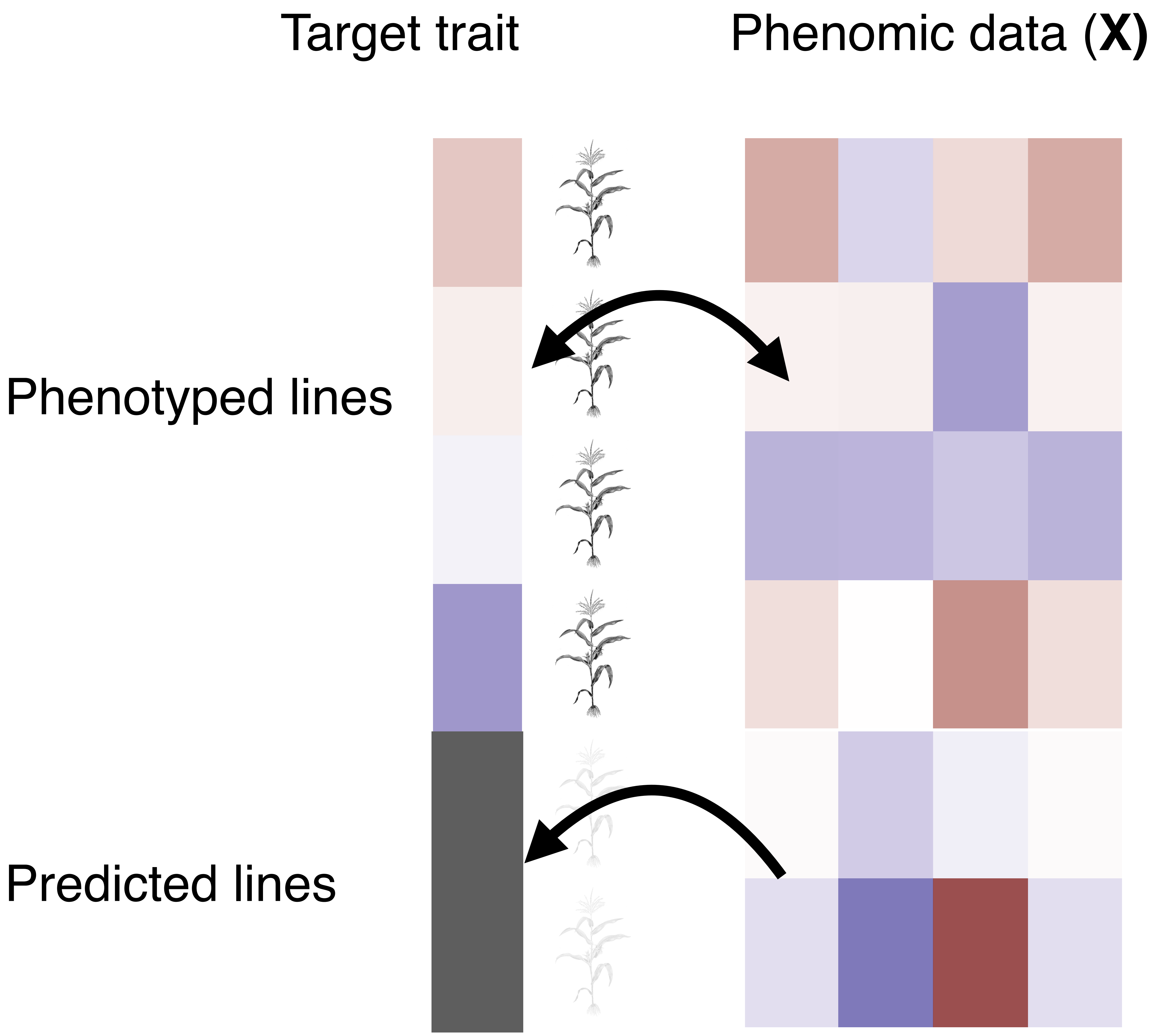
$$\hat{\mathbf{y}}_\mathbf{p} = \tilde{\mathbf{X}}\beta$$

Parameters learned from training

Predicted phenotypes

Phenomic features of new lines

3) Select lines based on $\hat{\mathbf{y}}_\mathbf{p}$ without measuring **target trait**



Phenomic Prediction is deceptively similar

The motivation is similar to Genomic Prediction

Trait

Phenomic features

Parameters

Error

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

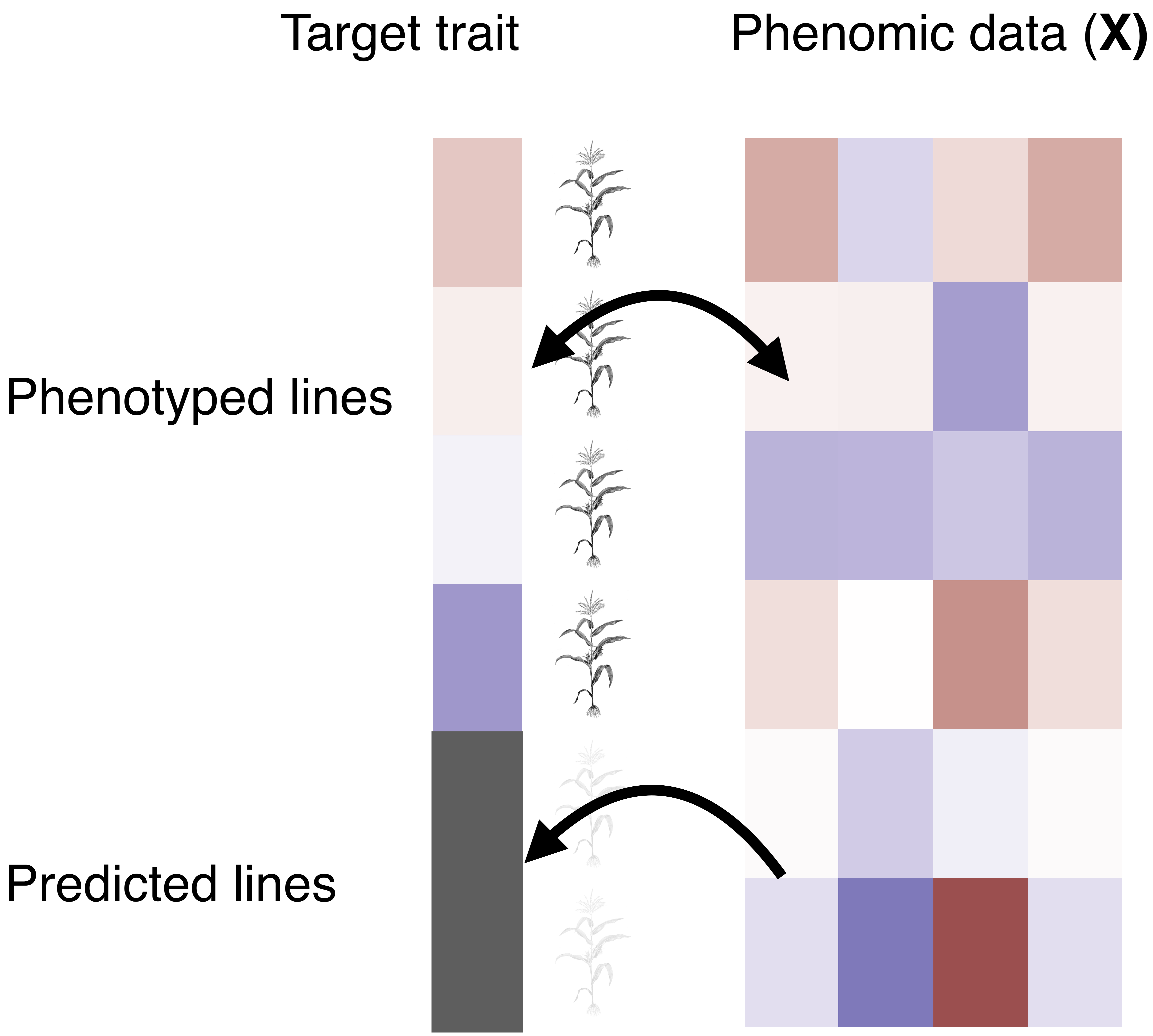
Phenomic data is **High Dimensional**

p >> n

We don’t have to know which features are useful beforehand

We expect some to be useful because of pleiotropy

Models like rrBLUP, Random Forrest work well



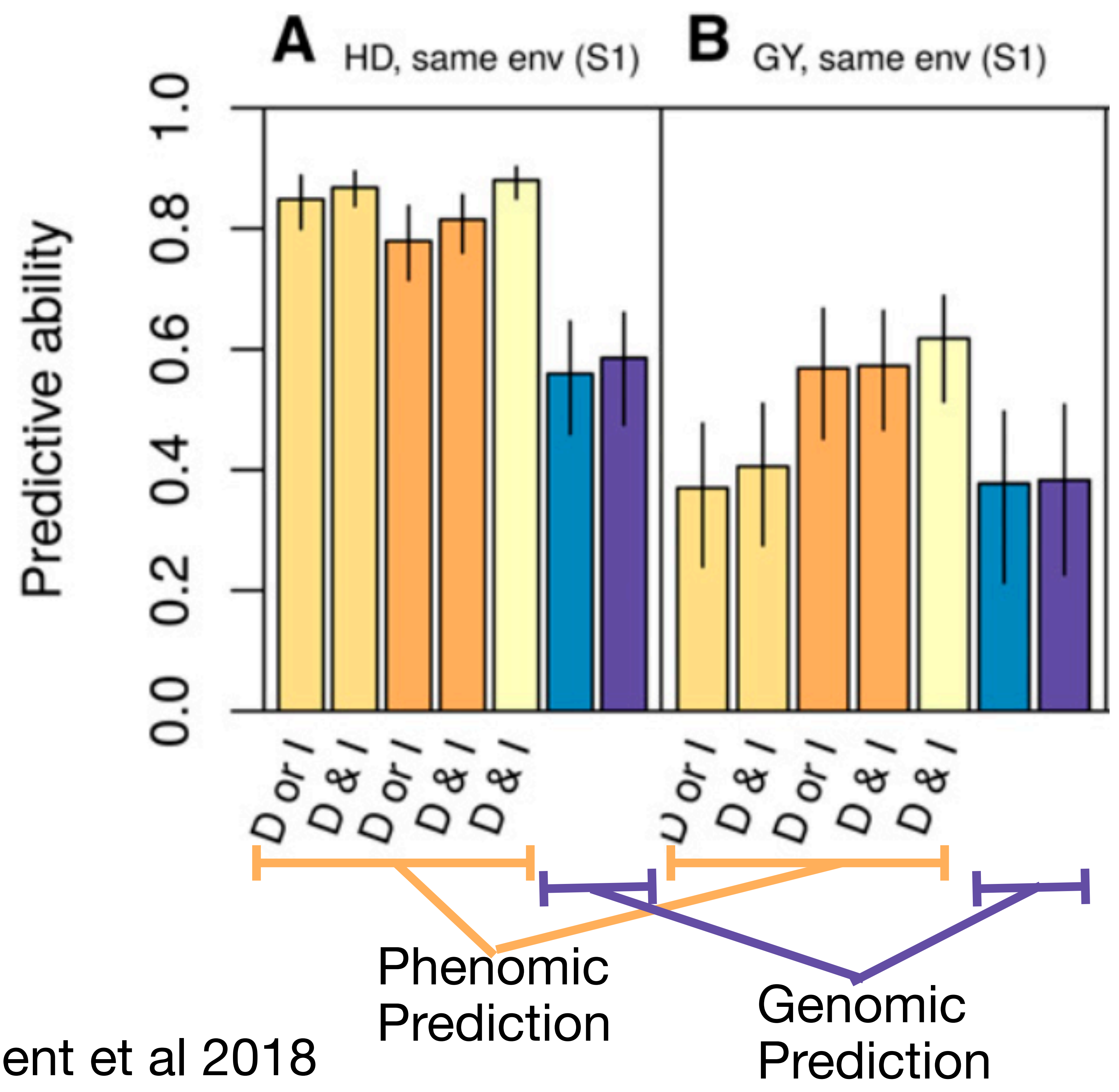
Target trait

Phenomic data (**X**)

Phenotyped lines

Predicted lines

Claim: Phenomic Prediction is competitive with Genomic Prediction



Phenomic Prediction

Genomic Prediction

Rincent et al 2018

Rate of Gain

$$\Delta g = \overset{\text{Intensity}}{i} \cdot \overset{\textbf{Accuracy}}{r} \cdot \overset{\text{Std. Dev}}{\sigma} \cdot \overset{\text{Speed}}{1/L}$$

Maybe, but…

1) This is not the **right** accuracy

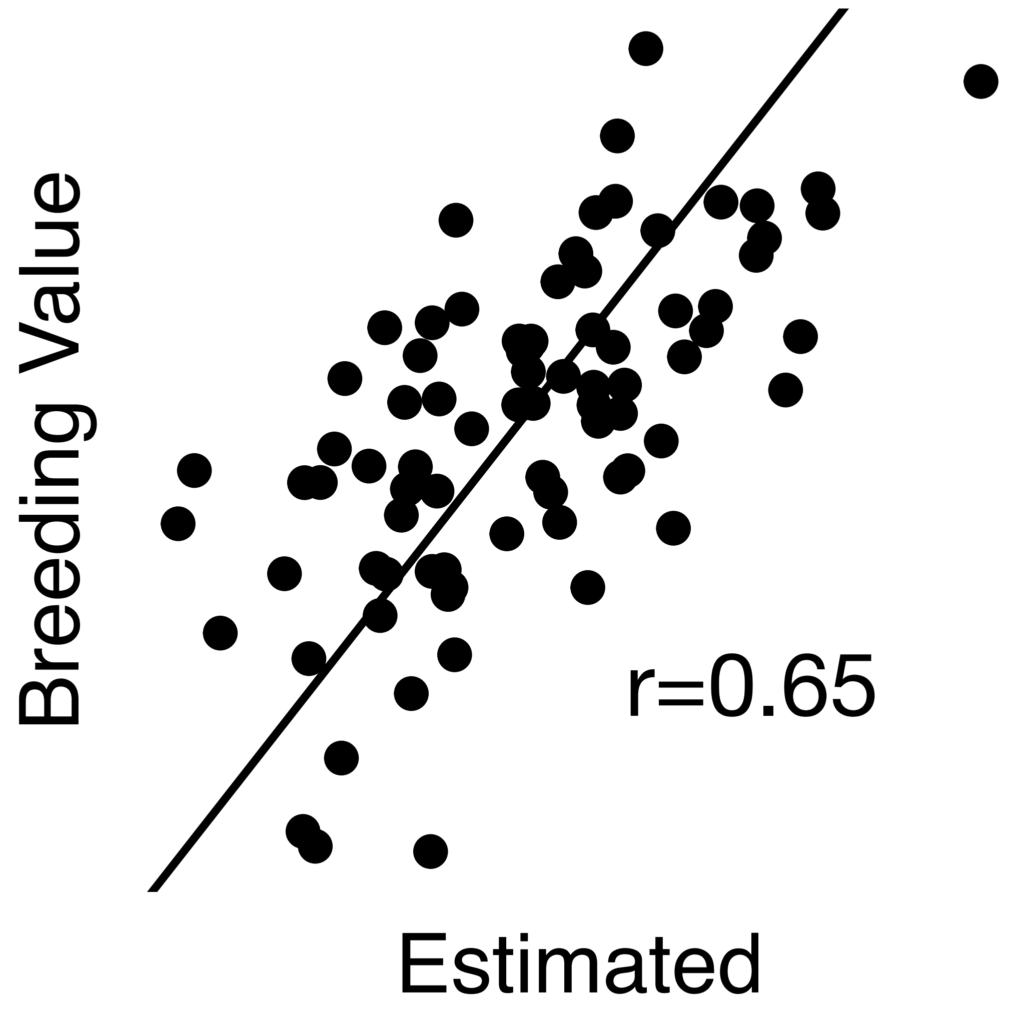
2) The parameters are **interrelated** and can’t be evaluated independently

Genomic Selection optimizes other parameters more effectively

Predictive Ability is not Accuracy

Accuracy: correlation of predictions with breeding values

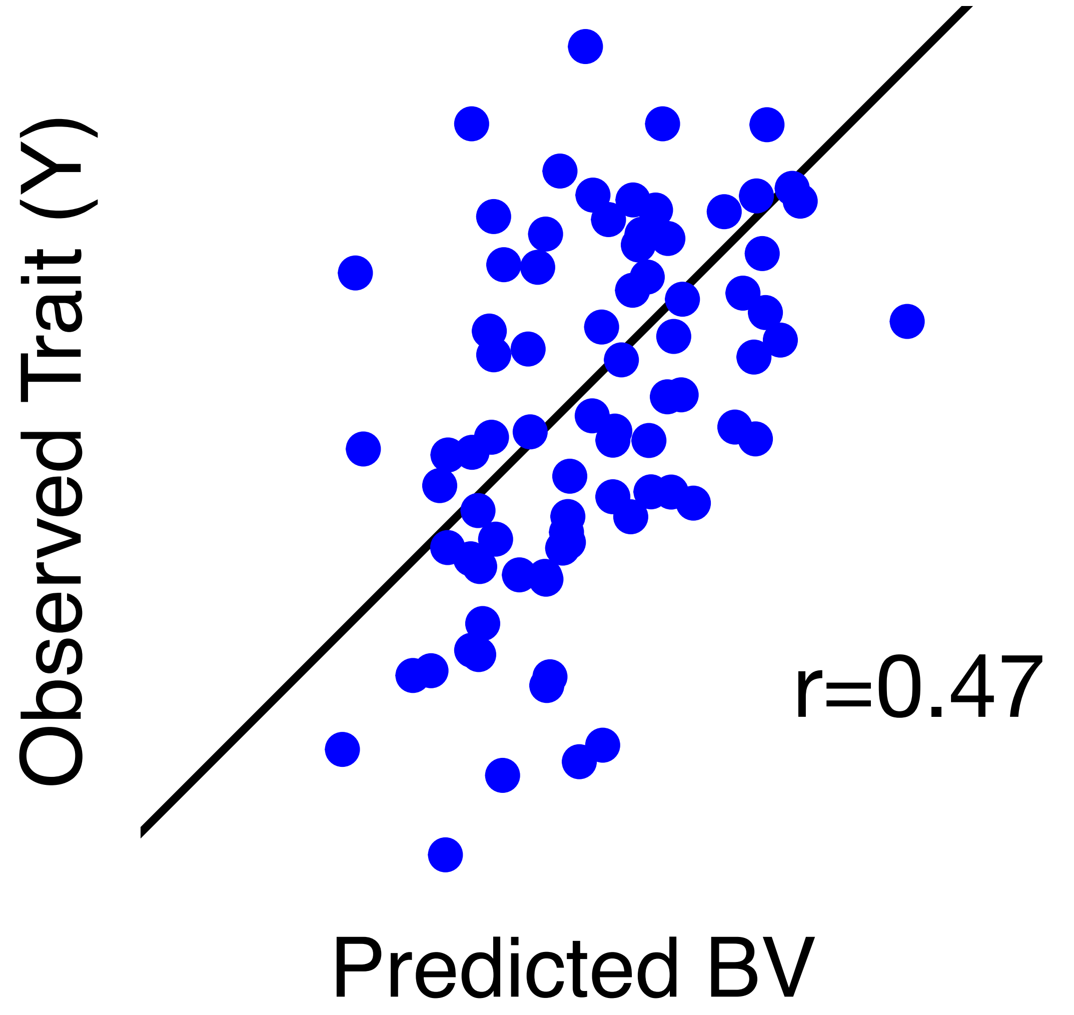
$$r = cor(\hat{u}, BV)$$



Predictive Ability: correlation of predictions with **observed traits**

Ould Estaghvirou *et al.* (2013)

$$r = cor(\hat{u}, Y)$$



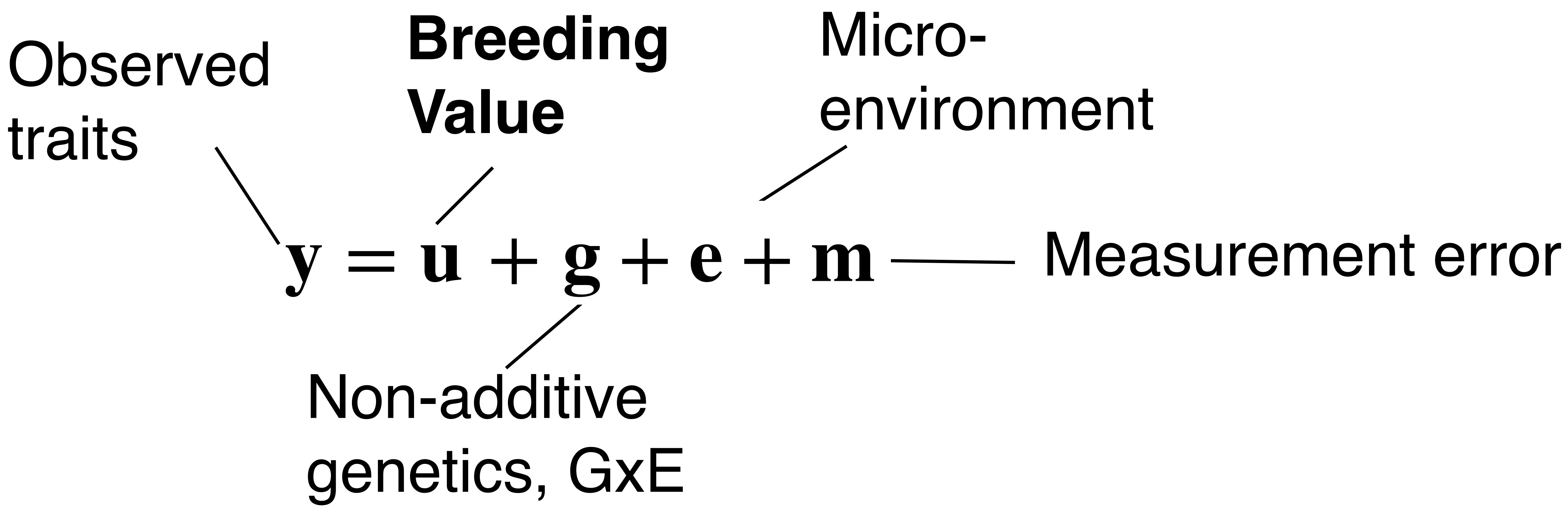
Problem: We can't **observe** breeding values

We can't measure accuracy directly

For Genomic Prediction: Ability < Accuracy

For Phenomic Prediction: Ability <?> Accuracy

Observed Traits are noisy estimates of Breeding Values



Observed traits

Breeding Value

Micro-environment

$$\mathbf{y} = \mathbf{u} + \mathbf{g} + \mathbf{e} + \mathbf{m}$$

Measurement error

Non-additive genetics, GxE

Genomic Prediction

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad \tilde{\mathbf{X}}\beta \rightarrow \hat{\mathbf{u}}$$

Only contains info from $\mathbf{u}$ for new lines

Phenomic Prediction

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad \tilde{\mathbf{X}}\beta \rightarrow \hat{\mathbf{y}}$$

Contains info from $\mathbf{u} + \mathbf{g} + \mathbf{e}$ for new lines

$\mathbf{g} + \mathbf{e}$ effects **contaminate** Predictive ability relative to Accuracy

Contamination is not removed by cross-validation

Runcie and Cheng 2019

Observed Traits are noisy estimates of Breeding Values

Observed traits

Breeding Value

Micro-environment

$$\mathbf{y} = \mathbf{u} + \mathbf{g} + \mathbf{e} + \mathbf{m}$$

Measurement error

Non-additive genetics, GxE

Genomic Prediction

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad \tilde{\mathbf{X}}\beta \rightarrow \hat{\mathbf{u}}$$

Only contains info from $\mathbf{u}$ for new lines

rrBLUP, BayesB,…

(RKHS also contains info from $\mathbf{g}$)

Phenomic Prediction

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \qquad \tilde{\mathbf{X}}\beta \rightarrow \hat{\mathbf{y}}$$

Contains info from $\mathbf{u} + \mathbf{g} + \mathbf{e}$ for new lines

If $\mathbf{y}$ is from the same plants

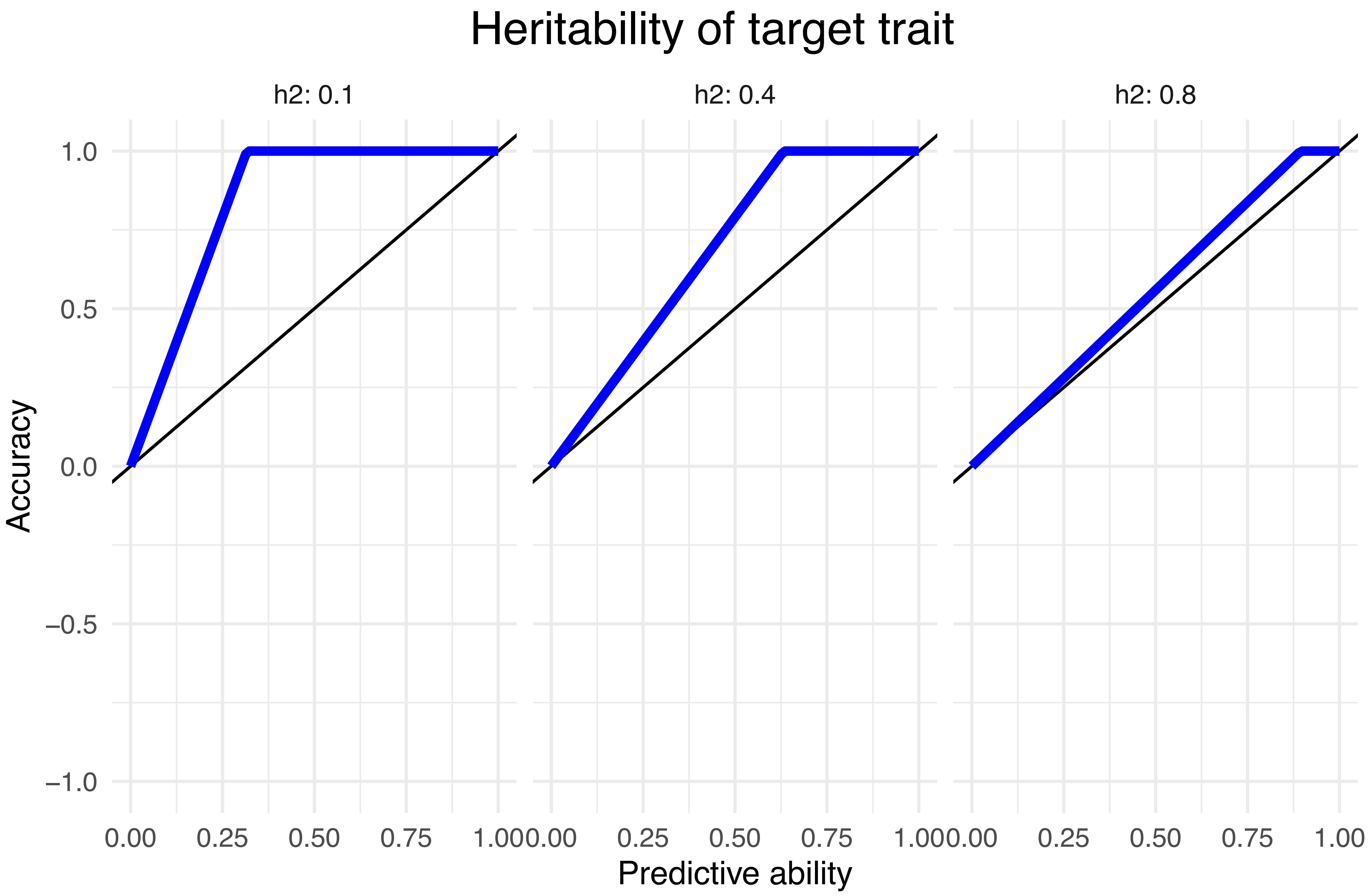
$\mathbf{u} + \mathbf{g}$ if $\mathbf{y}$ is from different plants

$\mathbf{g} + \mathbf{e}$ **contaminate** Predictive ability relative to Accuracy

Contamination is not removed by cross-validation

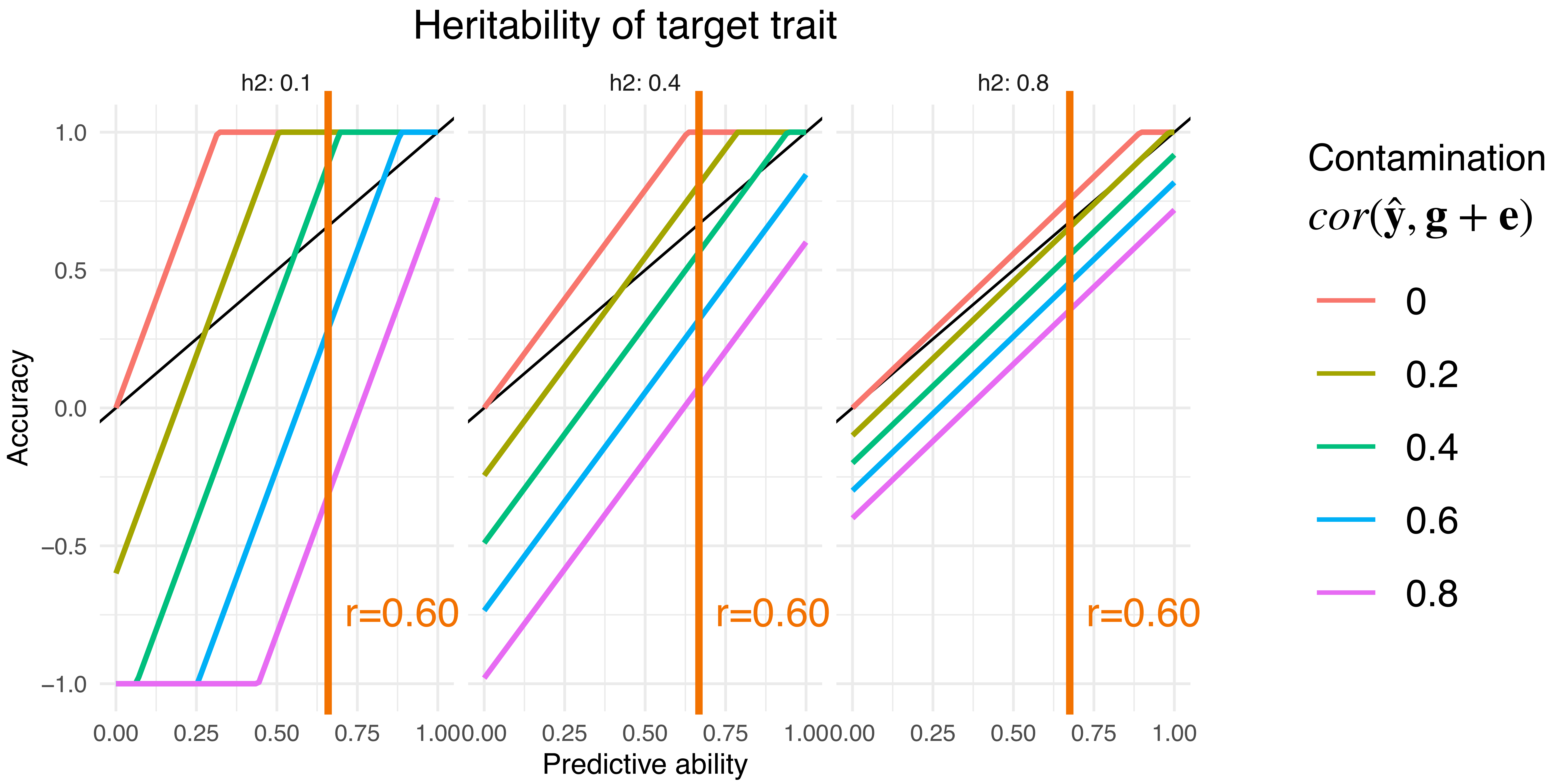
Runcie and Cheng 2019

For Genomic Prediction, Accuracy > Ability



$$\text{Accuracy} = \text{Ability} / \sqrt{h^2}$$

For Phenomic Prediction the relation between Ability and Accuracy is weak

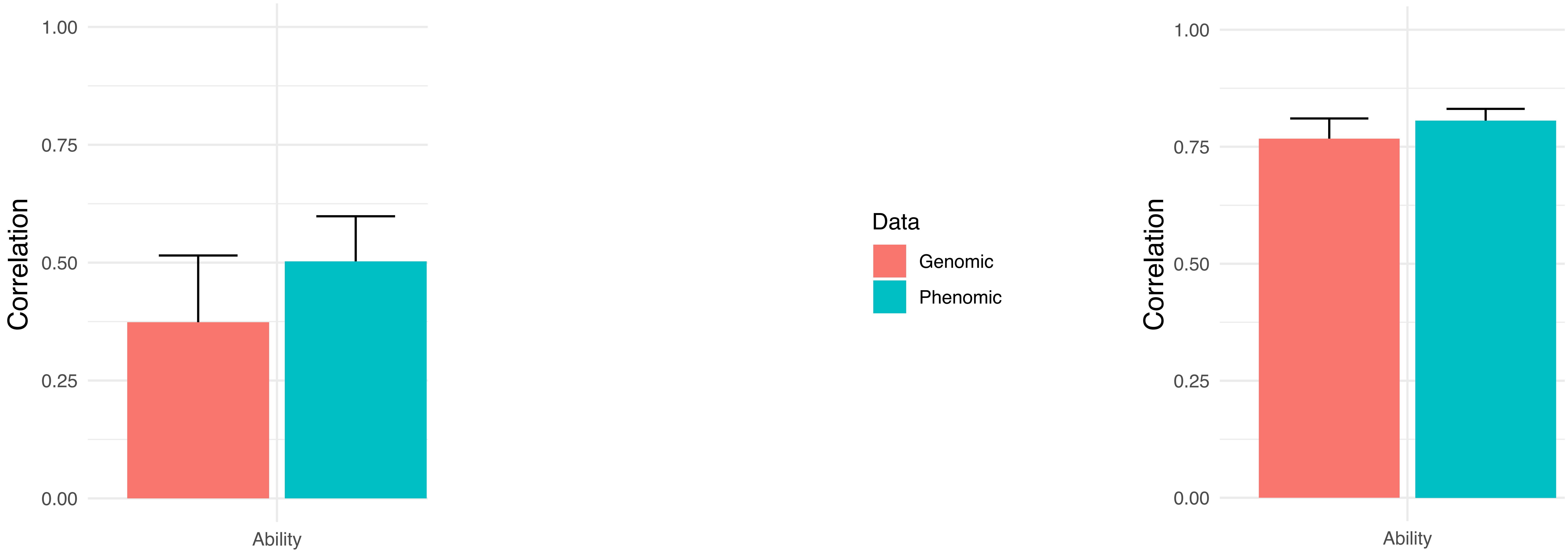


Predictive ability tells you little about Accuracy

When the contamination due to $\mathbf{g} + \mathbf{e}$ is strong

When h^2 of the target trait is low

Case Study - Rincent 2018



Wheat: GrainYield in Dry condition

Poplar: Circumference in SAV

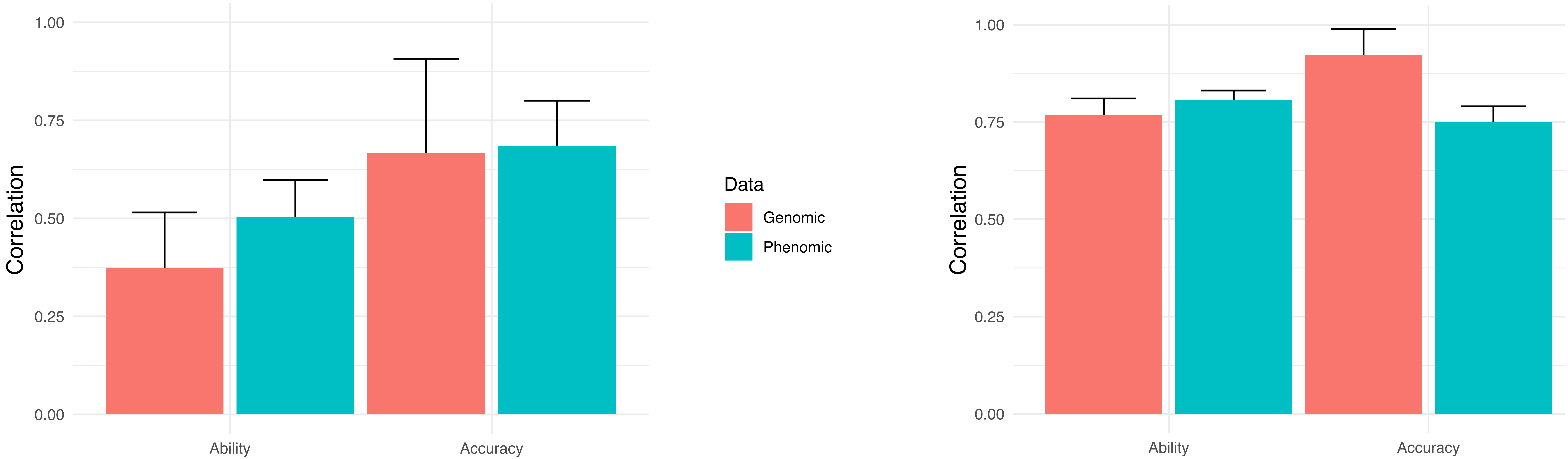
Observation: In both datasets, Phenomic Prediction “beats” Genomic Prediction

If scored with **Predictive Ability**

Case Study - Rincent 2018

Wheat: GrainYield in Dry condition

Poplar: Circumfrence in SAV



We developed an R function (soon to be R package) that can **estimate** accuracy

Using this function, Phenomic Prediction doesn’t beat Genomic Prediction in most datasets for **predicting breeding values**

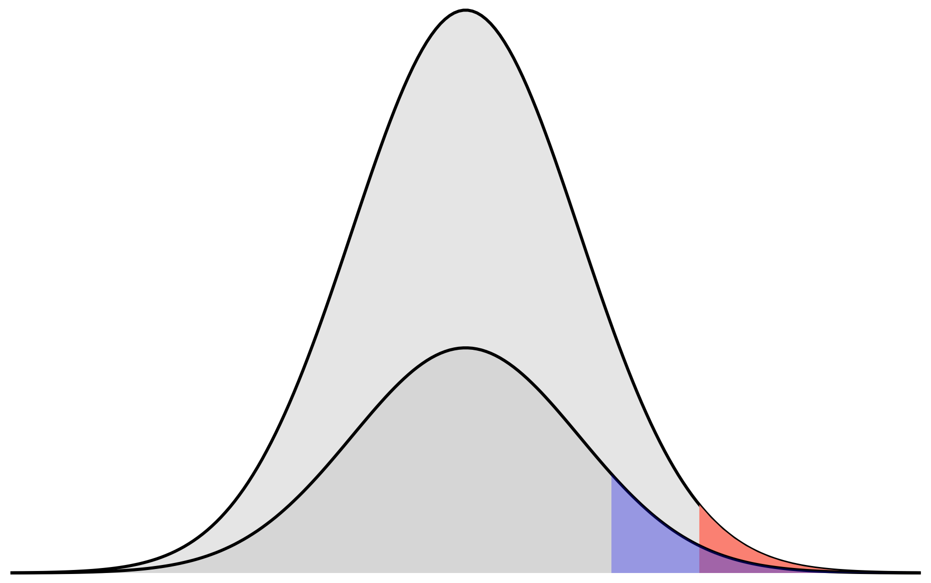
Does Phenomic Prediction need to be **more accurate?**

Rate of Gain

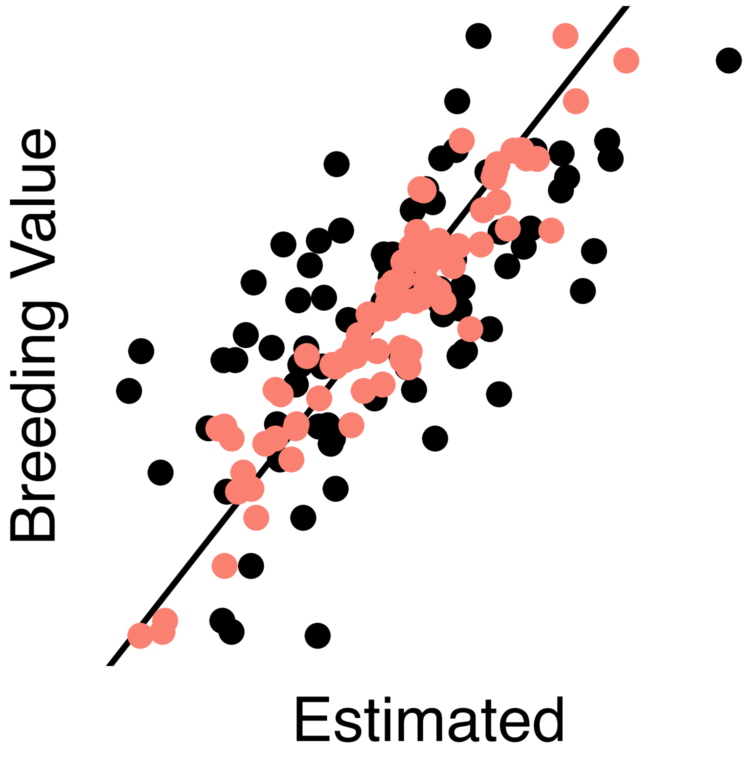
Intensity **Accuracy** Std. Dev Speed

$$\Delta g = i \cdot r \cdot \sigma \cdot 1/L$$

Intensity (*i*)



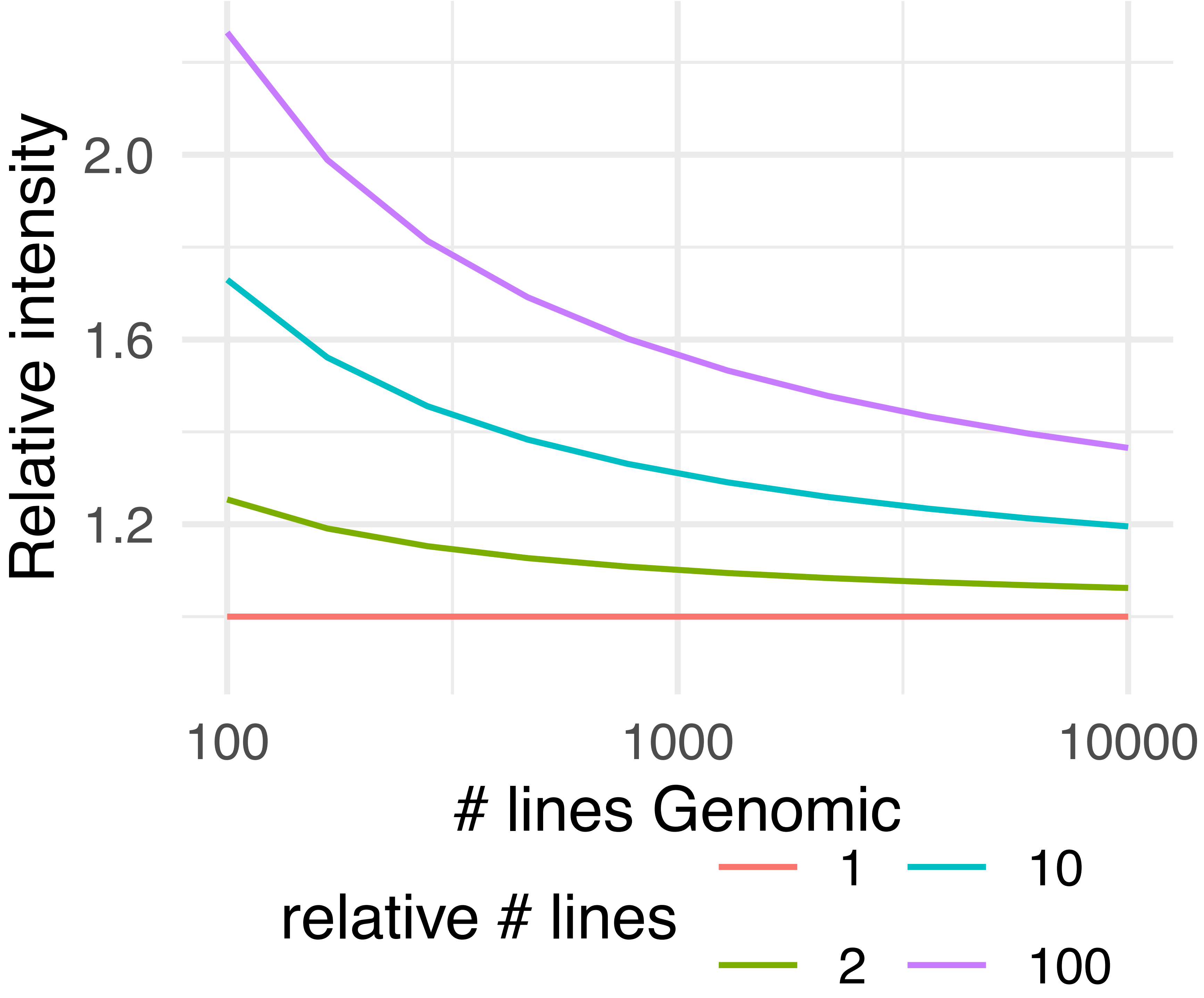
Accuracy (*r*)



Gain is a function of $i \cdot r$

Phenomic Prediction is cheaper, so more $$ to increase intensity

How much large populations are needed?



Need to increase by 10X-100X to get 50% higher gains

Genomic Selection can often increase 2x(+) in speed

Gaynor et al 2017

Where is Phenomic data most useful in breeding?

Rate of Gain

Intensity **Accuracy** Std. Dev **Speed**

$$\Delta g = i \cdot r \cdot \sigma \cdot 1/L$$

Genomic Selection is best because of speed

Genomic Selection limits intensity by cost

Accuracy of Genomic Prediction is limited by the training data quality:

Can we use Phenomic Prediction to get better data to train Genomic Prediction models?

Training data size (n)

Limited by cost of genotyping

heritability (h^2)

Measure each line more **accurately**

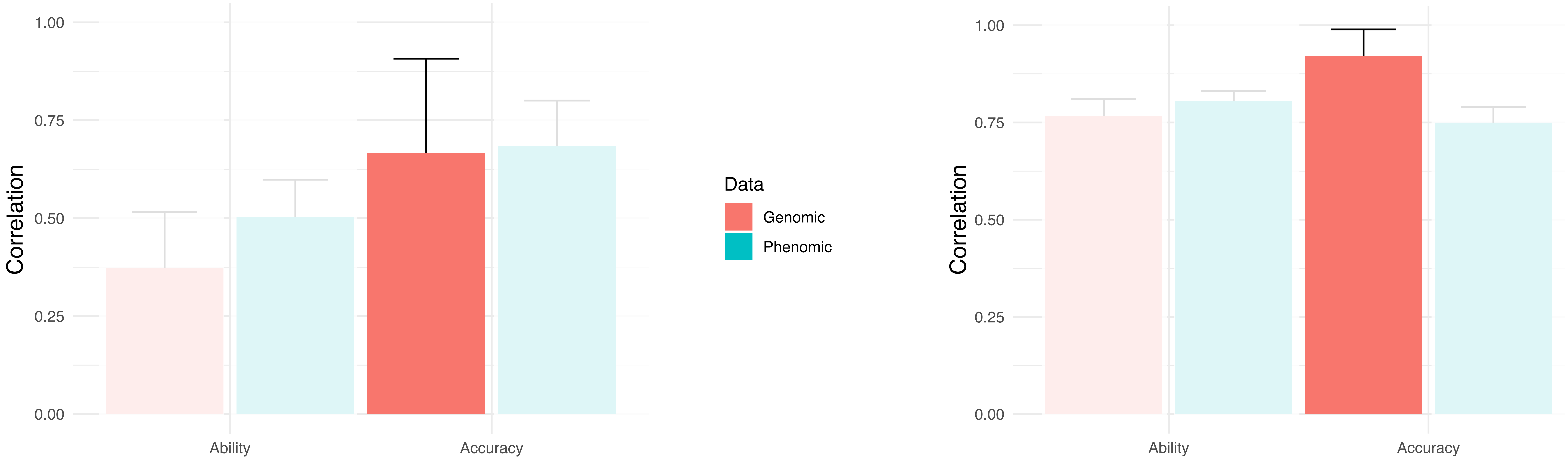
Measure each line **more times**

Measure **related lines** more times

Lane and Murray 2021

Prediction ability: $cor(\hat{\mathbf{y}}, \mathbf{y})$ is the right way to measure this!

Different measures of accuracy are useful



Wheat: GrainYield in Dry condition

Poplar: Circumfrence in SAV

Genomic Prediction: measure **Accuracy**: $cor(\hat{\mathbf{u}}, \mathbf{u})$

Use to predict genetic gain

Phenomic Prediction: measure **Ability**: $cor(\hat{\mathbf{y}}, \mathbf{y})$

Use to improve h^2

Don’t compare these two numbers

They’re measuring different things

Summary

High Throughput Phenotyping and Phenomics technologies are exciting

But are expensive

Phenomic Prediction works like Genomic Prediction

But it should be used differently in breeding

Use it to measure traits, not breeding values

Report accuracy as **Predictive Ability**, not **Predictive Accuracy**

Acknowledgements

UC Davis

Hao Cheng

Haixiao Hu

Steve Knapp

Funding

AG2PI
Agricultural Genome to
Phenome Initiative

USDA NIFA
United States Department of Agriculture
National Institute of Food and Agriculture

UCDAVIS
DEPARTMENT OF PLANT SCIENCES

AIFS AI INSTITUTE FOR FOOD SYSTEMS

NSF

California STRAWBERRY COMMISSION

hi-phen
INTEGRATED PLANT PHENOTYPING SYSTEMS