Day 4
A framework for simulating genotype by environment interaction

Simulation of multi-environment trials

Jon Bancic, Chris Gaynor, Gregor Gorjanc, Daniel Tolhurst

Lecture overview

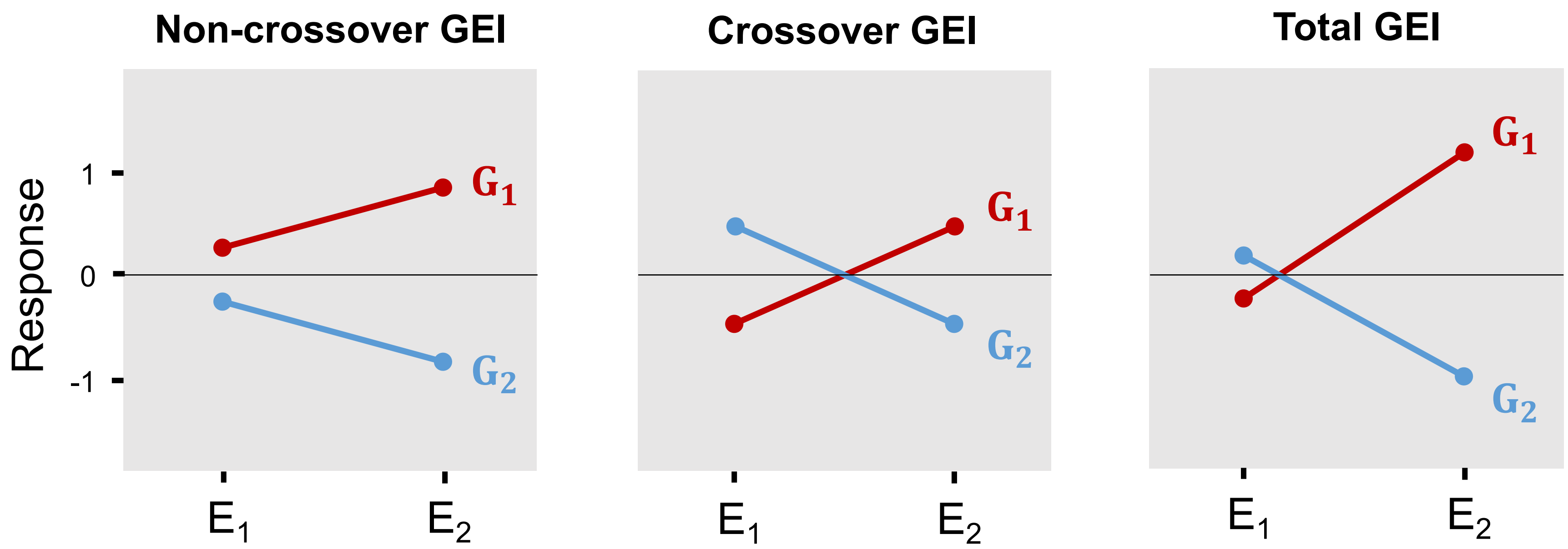
- Introduction
- Framework for simulating GxE interaction
- Framework application
  - Statistical model comparison
  - Breeding simulation

Why implement GxE into simulation?

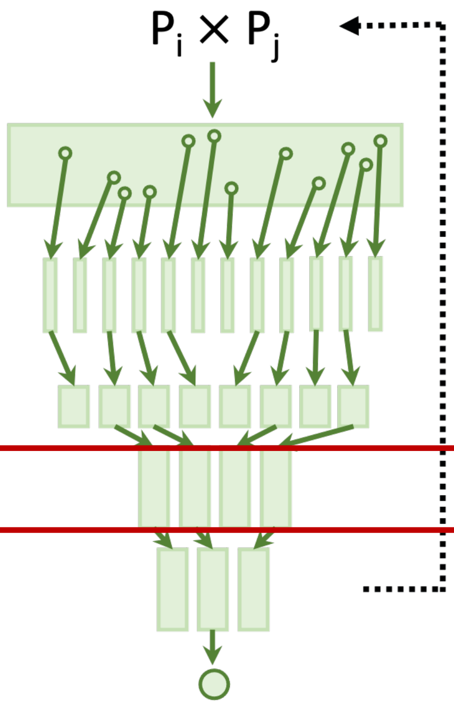
- Introduces more realistic structure and complexity to simulated field trial data
- Answer more targeted questions
  - What level of (partial) replication is required?
  - How many locations are required?
  - Where should material be deployed?
- Fine tuning a breeding pipeline
  - Comparison of breeding strategies, experimental designs and statistical analysis approaches in long-term

Genotype by environment (GxE) interaction

Genotype by environment (GxE) interaction complicates breeding

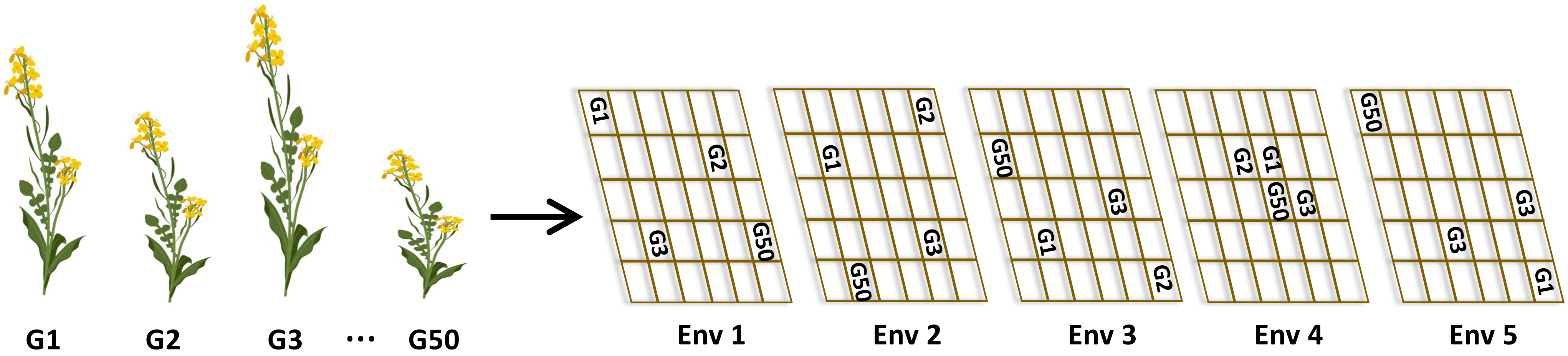


Plant breeding program



Year	Stage	Population	Env	h^2	Action
1	Crossing	100 crosses			Crosses with 100 parents
2	F_1/DH	100 families			Produce 100 DHs per cross
3	HDRW	10,000	1	0.10	Grow DHs
4	PYT	500	2	0.20	Field trial
5	AYT	50	5	0.50	Field trial
6	EYT	10	20	0.70	Field trial
7	Variety		1		Release inbred variety

MET: Multi-environment trials



Plant genotypes

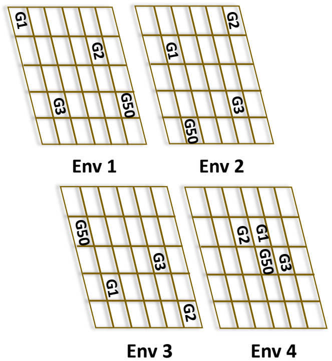
Field trials with pre-defined experimental design grown in selected environments

Multi-environment trial (MET) dataset

Id	Env	Block	Column	Row	Phenotype
1	1	1	1	1	1.20
2	1	1	5	2	1.07
3	1	1	2	3	0.75
⋮	⋮		⋮	⋮	⋮
50	1	1	6	3	1.19
⋮	⋮		⋮	⋮	⋮
50	5	1	1	1	0.77

Overview of plant breeding field trials

Experimental design & trials



Data collection

Id	Env	Block	Column	Row	Phenotype
1	1	1	1	1	1.20
2	1	1	5	2	1.07
3	1	1	2	3	0.75
⋮	⋮		⋮	⋮	⋮
50	1	1	6	3	1.19
⋮	⋮		⋮	⋮	⋮
50	5	1	1	1	0.77

Statistical analysis

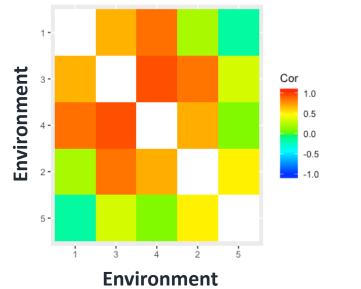
```
asreml(y ~ 1 + Env,
       random = ~ fa(Id, 3) + diag(Env)block,
       residual = ~ dsum(ar1(Col):ar1(Row)|Env),
       data = MET_df)
```

Interpretation

Candidate selection list

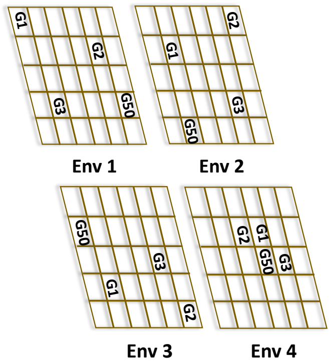
ID	Main effect	Stability	Rank
1	1.14	0.12	1
2	1.068	0.26	2
3	1.062	0.19	3
50	0.954	0.25	4

Between-environment genetic correlation matrix, C_e



Overview of plant breeding field trials

Experimental design & trials



Data collection

Id	Env	Block	Column	Row	Phenotype
1	1	1	1	1	1.20
2	1	1	5	2	1.07
3	1	1	2	3	0.75
⋮	⋮		⋮	⋮	⋮
50	1	1	6	3	1.19
⋮	⋮		⋮	⋮	⋮
50	5	1	1	1	0.77

Statistical analysis

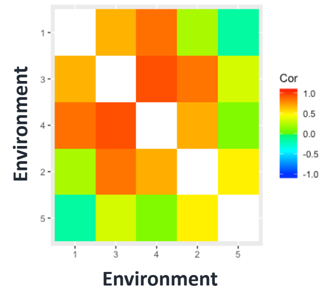
```
asreml(y ~ 1 + Env,
       random = ~ fa(Id, 3) + diag(Env)block,
       residual = ~ dsum(ar1(Col):ar1(Row)|Env),
       data = MET_df)
```

Interpretation

Candidate selection list

ID	Main effect	Stability	Rank
1	1.14	0.12	1
2	1.068	0.26	2
3	1.062	0.19	3
50	0.954	0.25	4

Between-environment genetic correlation matrix, $\mathbf{C_e}$



Simulate this!

Simulating phenotypes

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{ge} + \mathbf{W}\mathbf{b} + \mathbf{e}$$

Phenotypes

Overall trait mean

Environmental main effects

Genotype by environment effects

Block effects

Plot errors

Genotype	Environment E1	E2	E3	E4	E5
G1	0.20	0.08	0.13	0.31	-0.02
G2	0.07	-0.20	-0.17	0.43	0.21
G3	0.19	-0.24	0.08	-0.03	-0.23
⋮	⋮	⋮	⋮	⋮	⋮
G50	-0.25	0.39	-0.01	-0.16	0.24

Genotype by environment effects

Multiplicative models

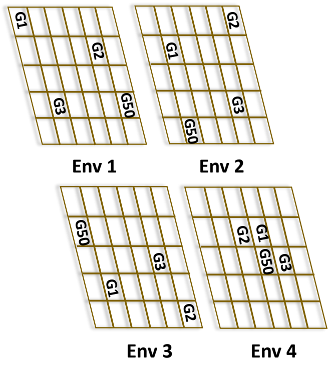
- Effective at capturing and interpreting GxE
- Decompose GxE into a small number (k) of multiplicative terms
- Each term is the product of genotype effects and environment effect

$$\mathbf{ge} = (\mathbf{s}_1 \otimes \mathbf{f}_1) + (\mathbf{s}_2 \otimes \mathbf{f}_2) + \cdots + (\mathbf{s}_k \otimes \mathbf{f}_k)$$
$$= (\mathbf{S}_k \otimes \mathbf{I}_v)\mathbf{f}_k$$

Environment effects **Genotype effects**

Overview of plant breeding field trials

Experimental design & trials



Data collection

Id	Env	Block	Column	Row	Phenotype
1	1	1	1	1	1.20
2	1	1	5	2	1.07
3	1	1	2	3	0.75
⋮	⋮		⋮	⋮	⋮
50	1	1	6	3	1.19
⋮	⋮		⋮	⋮	⋮
50	5	1	1	1	0.77

Statistical analysis

```
asreml(y ~ 1 + Env,
       random = ~ fa(Id, 3) + diag(Env)block,
       residual = ~ dsum(ar1(Col):ar1(Row)|Env),
       data = MET_df)
```

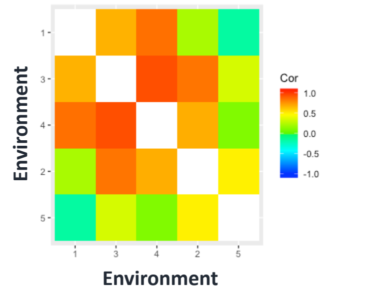
Start here

Interpretation

Candidate selection list

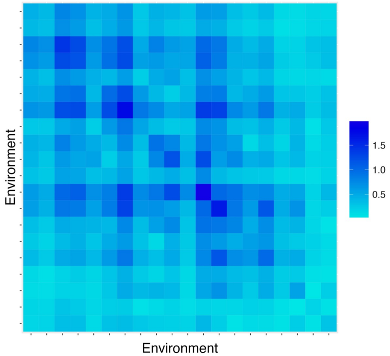
ID	Main effect	Stability	Rank
1	1.14	0.12	1
2	1.068	0.26	2
3	1.062	0.19	3
50	0.954	0.25	4

Between-environment genetic correlation matrix, C_e



Simulating genotype by environment effects

1. Between-environment genetic variance matrix, $\mathbf{G}_\mathrm{e}$

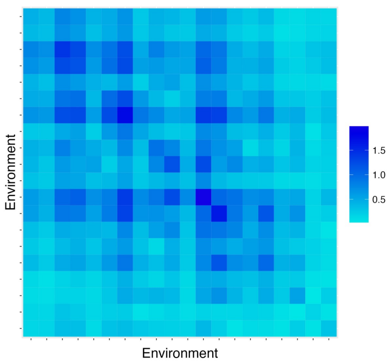


$$\mathbf{G}_\mathrm{e} = \mathbf{D}_\mathrm{e}^{1/2}\mathbf{C}_\mathrm{e}\mathbf{D}_\mathrm{e}^{1/2}$$

Simulate or provide $\mathbf{G}_\mathrm{e}$ and $\mathbf{D}_\mathrm{e}$

Simulating genotype by environment effects

1. Between-environment genetic variance matrix, $\mathbf{G}_\mathrm{e}$

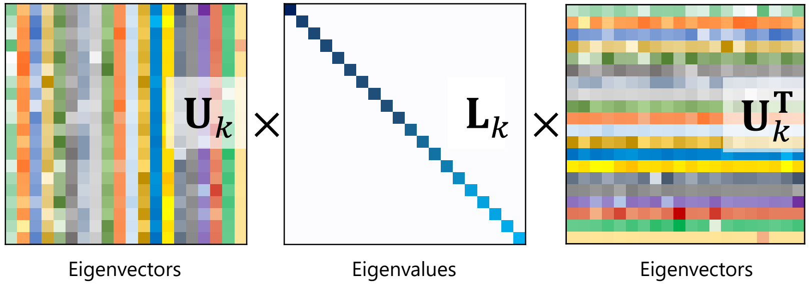


$$\mathbf{G}_\mathrm{e} = \mathbf{D}_\mathrm{e}^{1/2}\mathbf{C}_\mathrm{e}\mathbf{D}_\mathrm{e}^{1/2}$$

Simulate or provide $\mathbf{G}_\mathrm{e}$ and $\mathbf{D}_\mathrm{e}$

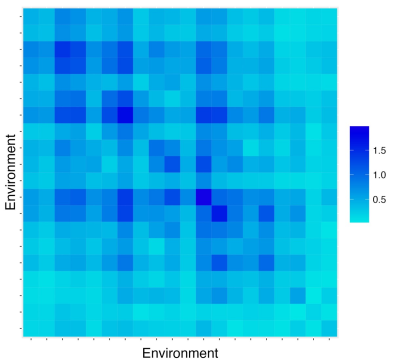
2. Decompose variance matrix, $\mathbf{G}_\mathbf{e}$, and take k terms

$$\mathbf{G}_\mathrm{e} = \mathbf{U}_k \times \mathbf{L}_k \times \mathbf{U}_k^\mathrm{T}$$



Simulating genotype by environment effects

1. Between-environment genetic variance matrix, $\mathbf{G}_\mathrm{e}$

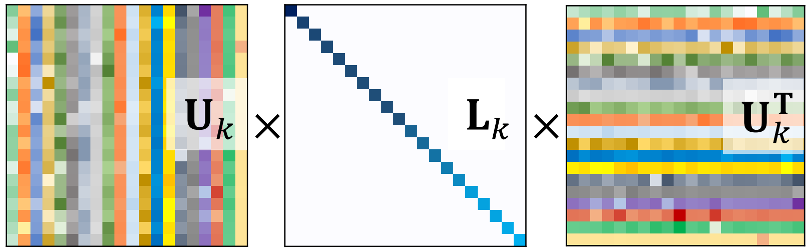


$$\mathbf{G}_\mathrm{e} = \mathbf{D}_\mathrm{e}^{1/2}\mathbf{C}_\mathrm{e}\mathbf{D}_\mathrm{e}^{1/2}$$

Simulate or provide $\mathbf{G}_\mathrm{e}$ and $\mathbf{D}_\mathrm{e}$

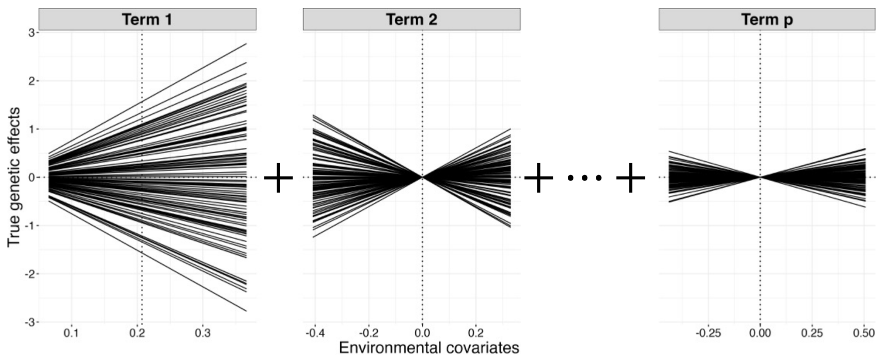
2. Decompose variance matrix, $\mathbf{G}_\mathbf{e}$, and take *k* terms

$$\mathbf{G}_\mathrm{e} = \mathbf{U}_k \times \mathbf{L}_k \times \mathbf{U}_k^\mathrm{T}$$



3. Obtain environmental covariates, $\mathbf{S}_k$, and simulate genotype slopes, $\mathbf{f}_k$

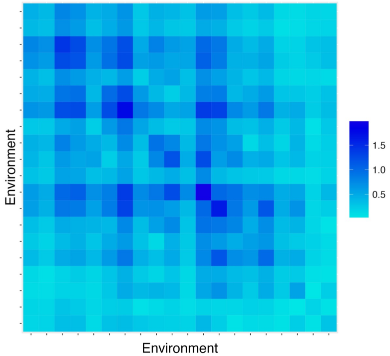
$$\mathbf{S}_k = \mathbf{U}_k$$



$$\mathbf{f}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}_k \otimes \mathbf{G}_\mathrm{e})$$

Simulating genotype by environment effects

1. Between-environment genetic variance matrix, $\mathbf{G}_\mathrm{e}$

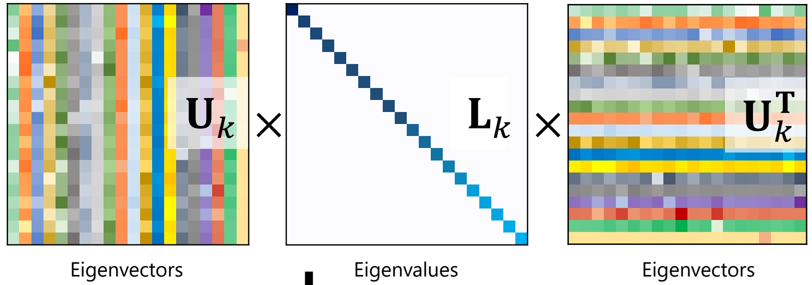


$$\mathbf{G}_\mathrm{e} = \mathbf{D}_\mathrm{e}^{1/2}\mathbf{C}_\mathrm{e}\mathbf{D}_\mathrm{e}^{1/2}$$

Simulate or provide $\mathbf{G}_\mathrm{e}$ and $\mathbf{D}_\mathrm{e}$

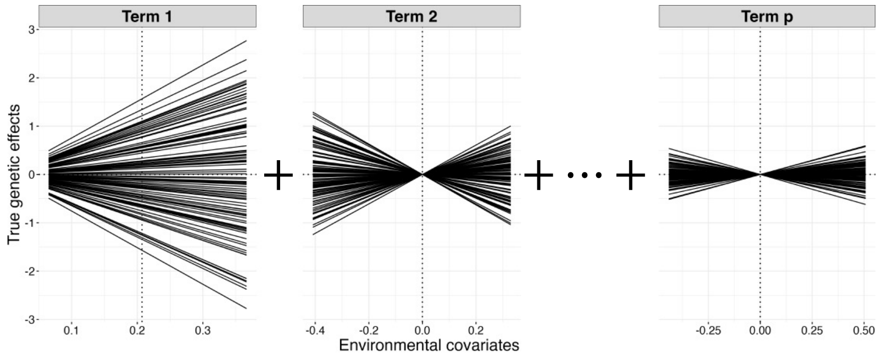
2. Decompose variance matrix, $\mathbf{G}_\mathrm{e}$, and take *k* terms

$$\mathbf{G}_\mathrm{e} = \mathbf{U}_k \times \mathbf{L}_k \times \mathbf{U}_k^\mathrm{T}$$



3. Obtain environmental covariates, $\mathbf{S}_k$, and simulate genotype slopes, $\mathbf{f}_k$

$$\mathbf{S}_k = \mathbf{U}_k$$



$$\mathbf{f}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{L}_k \otimes \mathbf{G}_\mathrm{e})$$

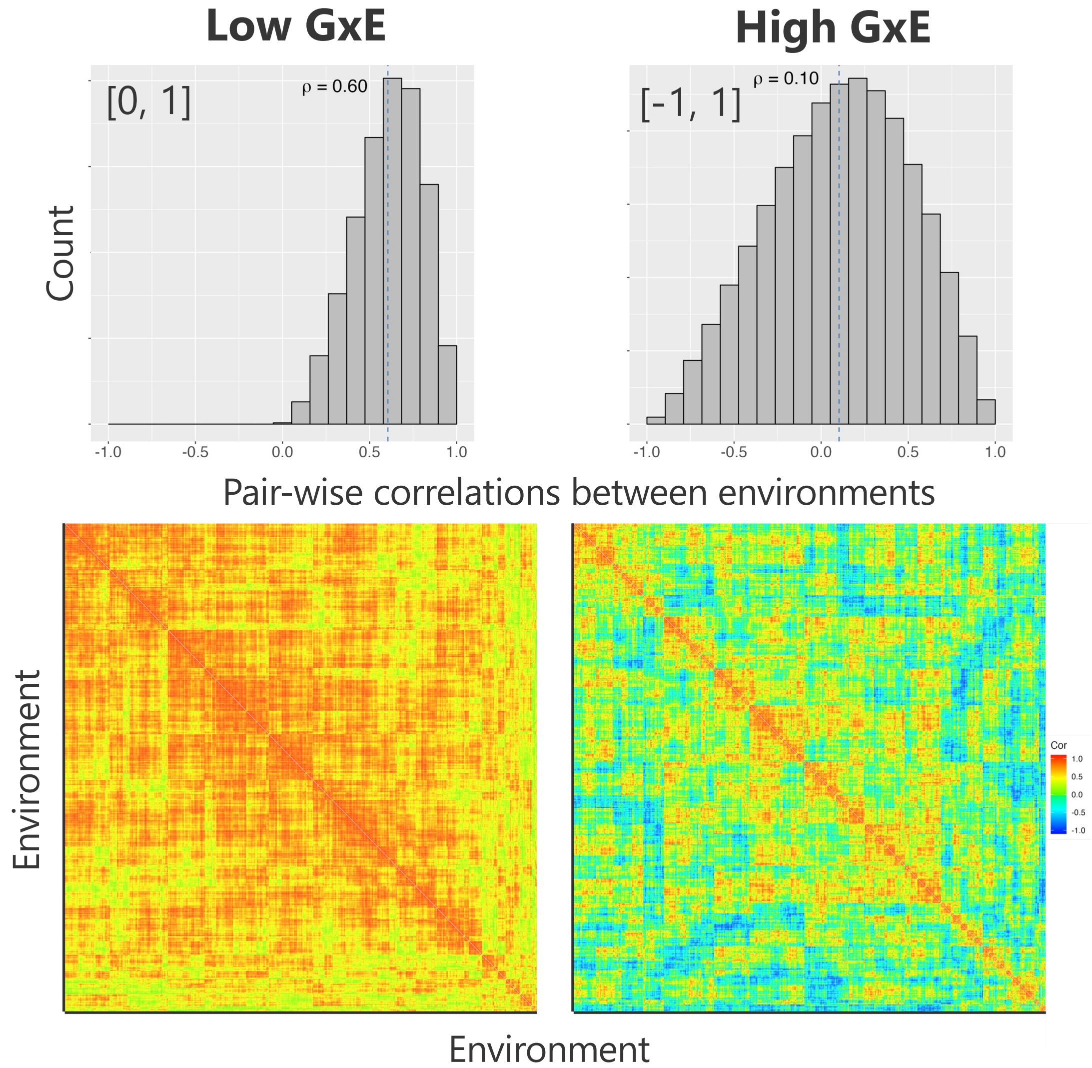
4. Genotype by environment effects

$$\mathbf{u} = (\mathbf{S}_k \otimes \mathbf{I}_v)\mathbf{f}_k$$

Simulating between-environment variance matrix $\mathbf{G}_e$

- Simulate $\mathbf{C}_e$ by specifying mean, variability, skew, noise structure
- Measures for tuning $\mathbf{C}_e$

	Variance explained			
GxE	v_g	v_{ge}	v_n	v_c
Low	0.51	0.49	0.67	0.33
High	0.08	0.92	0.24	0.76



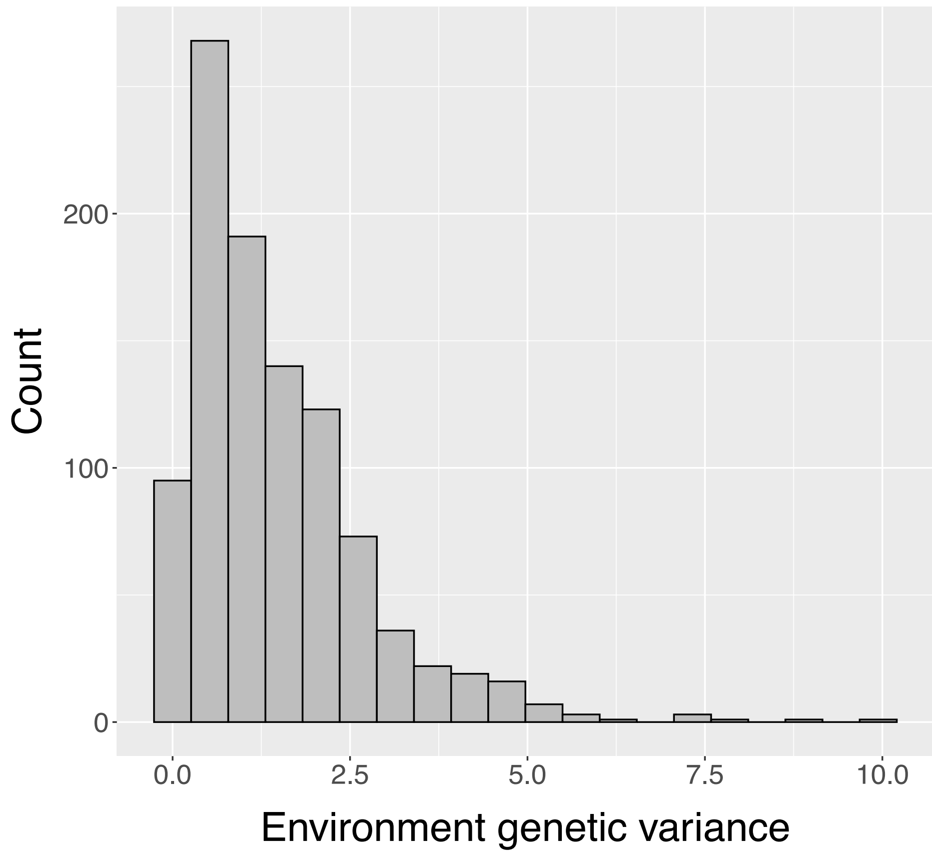
Simulating between-environment variance matrix $\mathbf{G}_\mathrm{e}$

- Simulate $\mathbf{C}_\mathrm{e}$ by specifying mean, variability, skew, noise structure

- Measures for tuning $\mathbf{C}_\mathrm{e}$

	Variance explained			
GxE	v_g	v_{ge}	v_n	v_c
Low	0.51	0.49	0.67	0.33
High	0.08	0.92	0.24	0.76

- Simulate $\mathbf{D}_\mathrm{e}$ from inverse gamma distribution by adjusting shape and rate



Simulating phenotypes

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X\tau} + \mathbf{Z}\mathbf{ge} + \mathbf{Wb} + \mathbf{e}$$

phenotype mean fixed genotype block residual

- $\mathbf{y}$ is the n-vector of phenotypes
- μ is the overall mean, $\mathbf{1}_n$ is a n-vector of ones
- $\boldsymbol{\tau}$ is the p-vector of environmental effects, with $n \times n_p$ design matrix $\mathbf{X}$ which links plots to environments
- $\mathbf{ge}$ is the n_g-vector of genotype effects, with $n \times n_g$ design matrix $\mathbf{Z}$ which links plots to genotypes ← new framework
- $\mathbf{b}$ is the n_b-vector of block effects, with $n \times n_b$ design matrix $\mathbf{W}$ which links plots to blocks
- $\mathbf{e}$ is the n-vector of residuals ← simulation of these demonstrated earlier

Simulate a MET dataset

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{X\tau} + \mathbf{Zge} + \mathbf{Wb} + \mathbf{e}$$

phenotype mean fixed genotype block residual

	env	block	col	row	id	true_mean	true_envEff	true_ge	true_blockEff	true_e	simulated_yield
	<fctr>	<fctr>	<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	1	1	1	114	4	0.1396509	4.213213	-0.1242964	-0.004663820	4.223903
2	1	1	1	2	72	4	0.1396509	3.801605	-0.1242964	0.779757230	4.596717
3	1	1	1	3	135	4	0.1396509	4.445786	-0.1242964	1.757523988	6.218665
4	1	1	1	4	63	4	0.1396509	4.269491	-0.1242964	0.061263382	4.346109
5	1	1	1	5	49	4	0.1396509	4.309022	-0.1242964	0.758258394	5.082635
6	1	1	1	6	65	4	0.1396509	3.582597	-0.1242964	-0.007580564	3.590371

Demonstrating examples

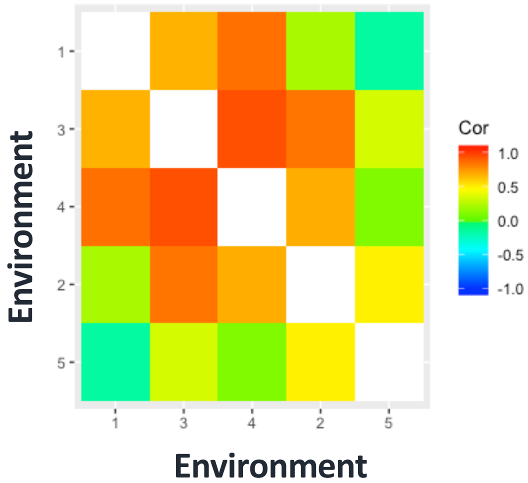
1. Comparison of statistical models
   → Answer a target question

2. Breeding program simulations
   → Breeding program fine tuning

TPE: Target population of environments

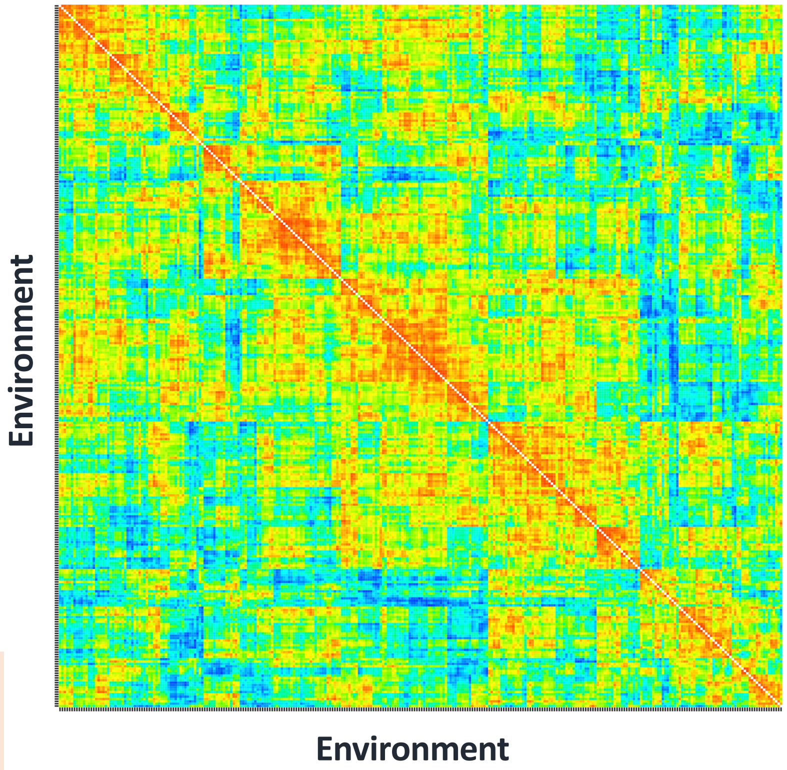
Breeder

MET – multi-environment trials



On farm conditions

TPE – target population of environments



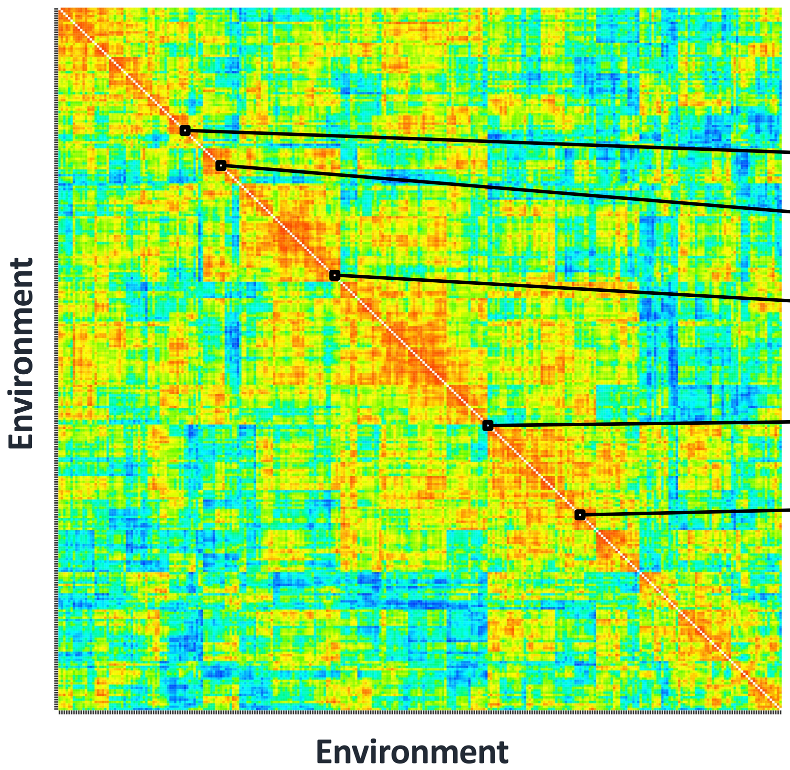
MET-TPE alignment

$$\mathrm{cor}(gv_{\mathrm{MET}}, gv_{\mathrm{TPE}})$$

Simulating target population of environments

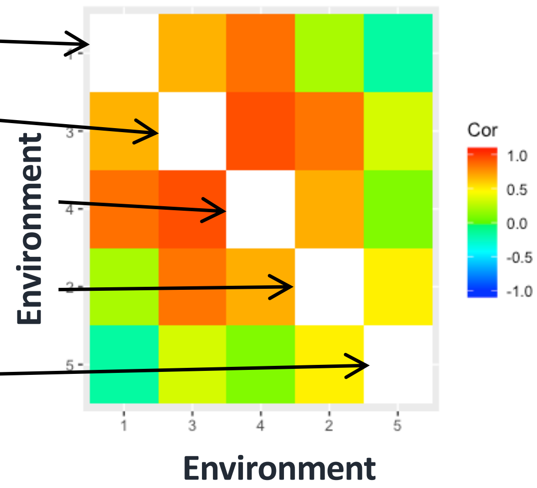
On farm conditions

TPE – target population of environments



Breeder

MET – multi-environment trials

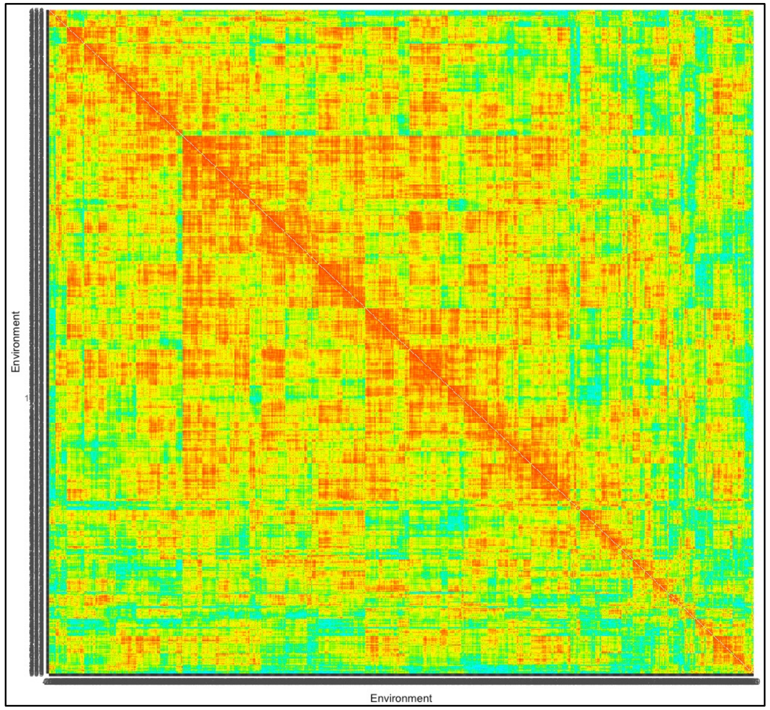


MET-TPE alignment

$\mathrm{cor}(\mathrm{gv}_{\mathrm{MET}}, \mathrm{gv}_{\mathrm{TPE}})$

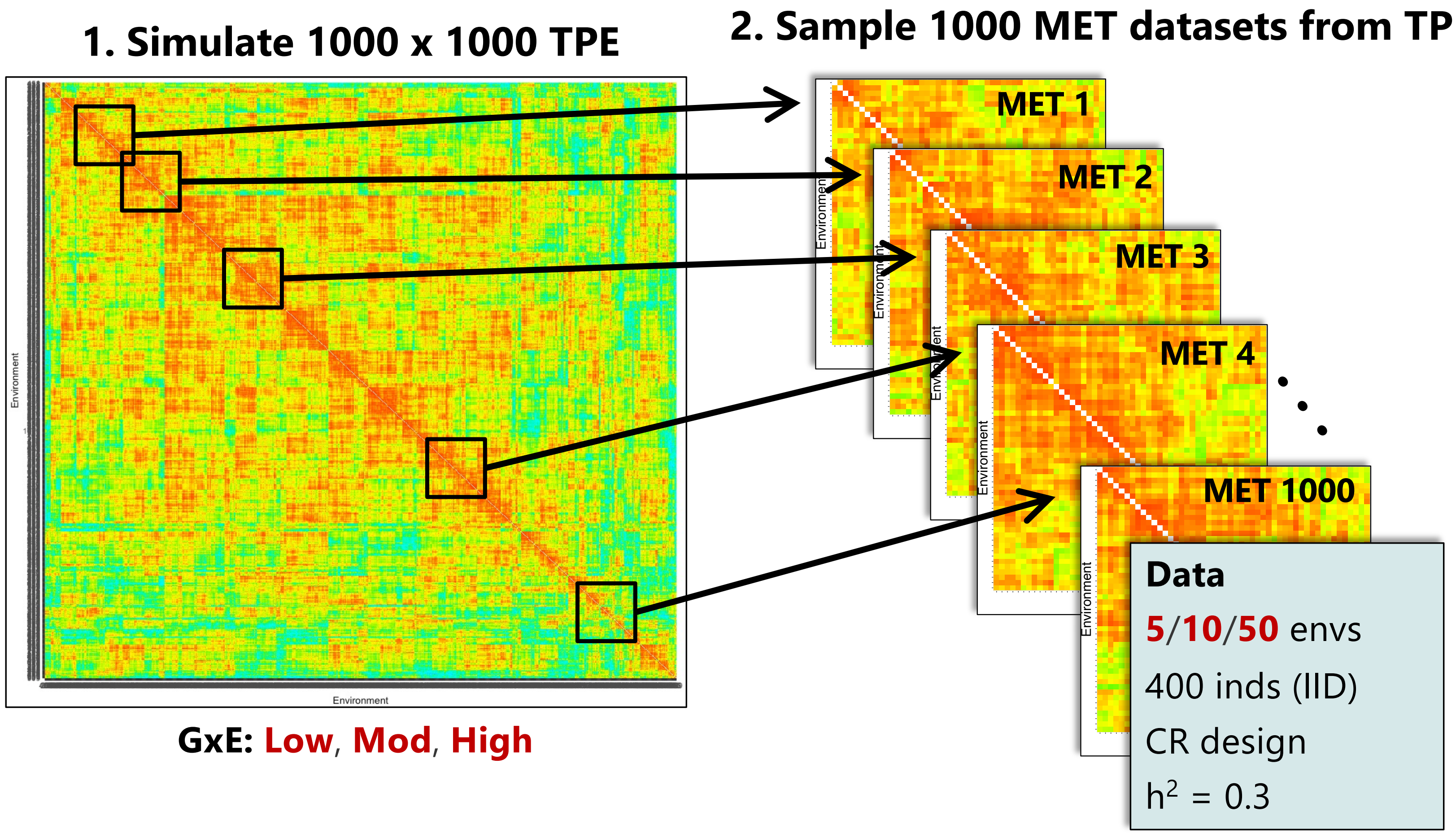
Example 1: Comparison of statistical models

1. Simulate 1000 x 1000 TPE



GxE: Low, Mod, High

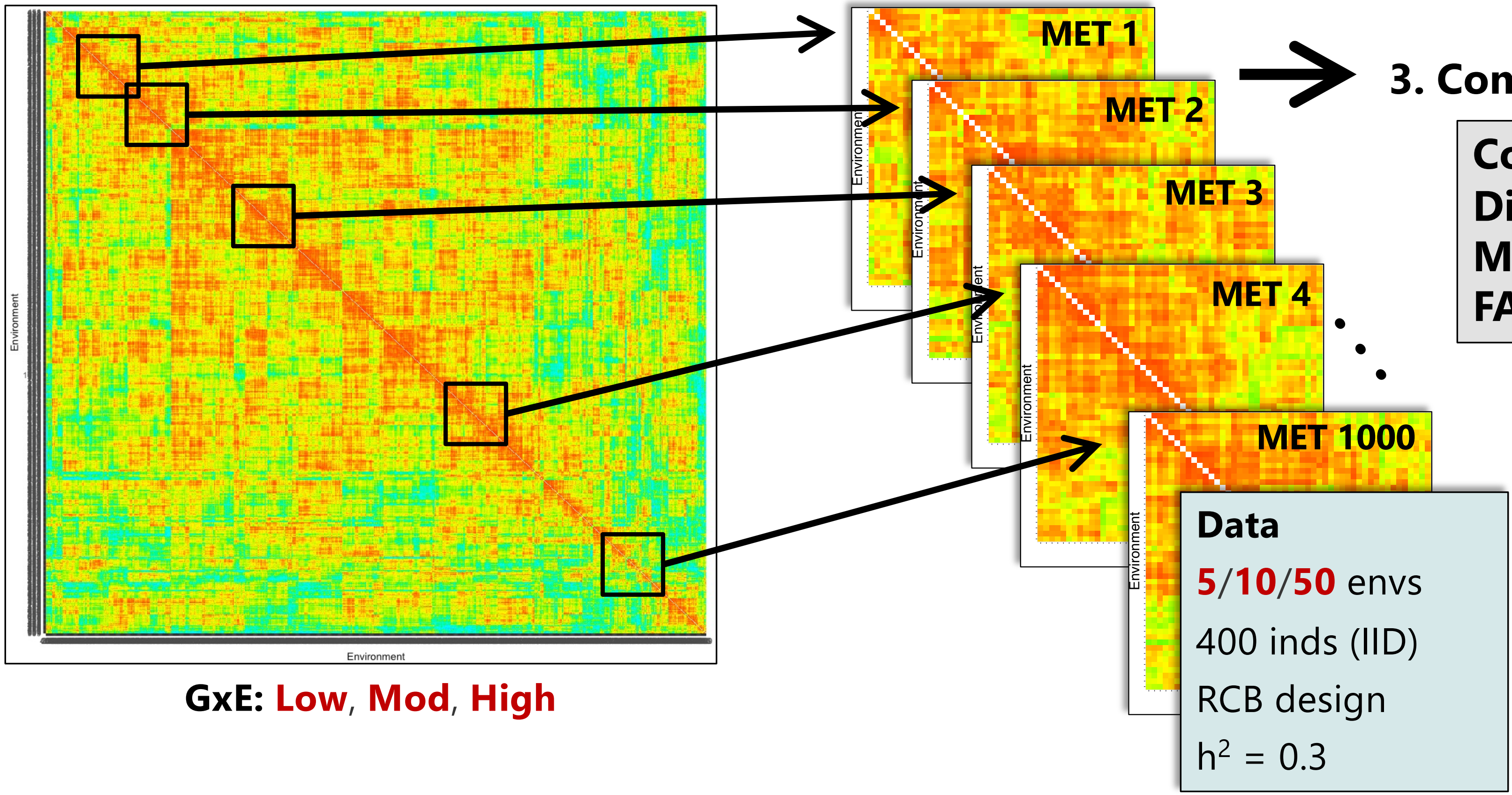
Example 1: Comparison of statistical models



Example 1: Comparison of statistical models

1. Simulate 1000 x 1000 TPE

2. Sample 1000 MET datasets from TPE



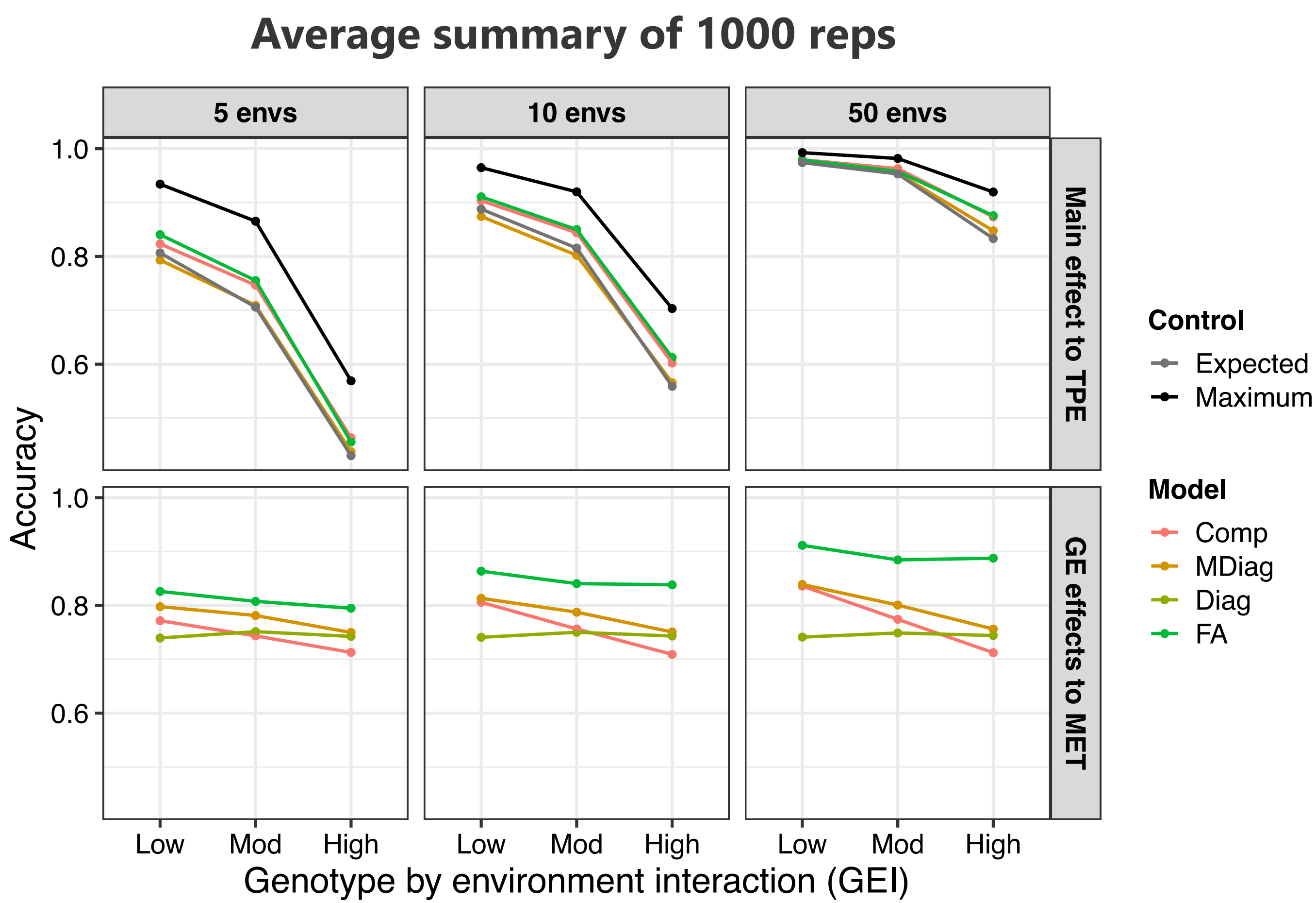
GxE: Low, Mod, High

3. Compare models (ASReml)

Comp	Compound Sym.
Diag	Diagonal
MDiag	Main + Diag
FA	Factor analytic

Example 1: Comparison of statistical models

- Model accuracy decreases as GxE increases
- More environments increase accuracy
- FA models are best overall



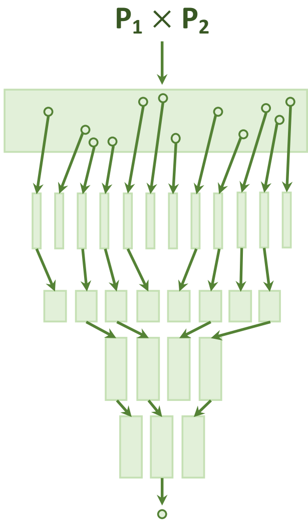
New opportunities

- Model comparison
  - Non-additive genetic effects
  - Spatial models
  - Multiple phenotypic traits
- Experimental design optimization
- MET dataset design optimization

Example 2: Breeding program simulation

Simulation:
- AlphaSimR
- 20-year breeding
- Yield trait
- Additive effects only
- 20 replicates

Line breeding program



Year	Stage	Genotypes	Envs	Reps	σ_ϵ^2	Action
1	Crossing	100 crosses				Make crosses
2	F_1	100 families				Produce DHs
3	Stage 1	10,000	1	1	8	Advance 500 DHs
4	Stage 2	500	2	1	4	Yield trial
5	Stage 3	50	5	2	2	Yield trial
6	Stage 4	10	20	2	2	Yield trial
7	Variety	1				Release variety

Program scenarios: Pheno & GS

Current plant breeding simulations

Current simulations

	g
G1	0.14
G2	0.07
G3	0.05
G4	-0.04

→

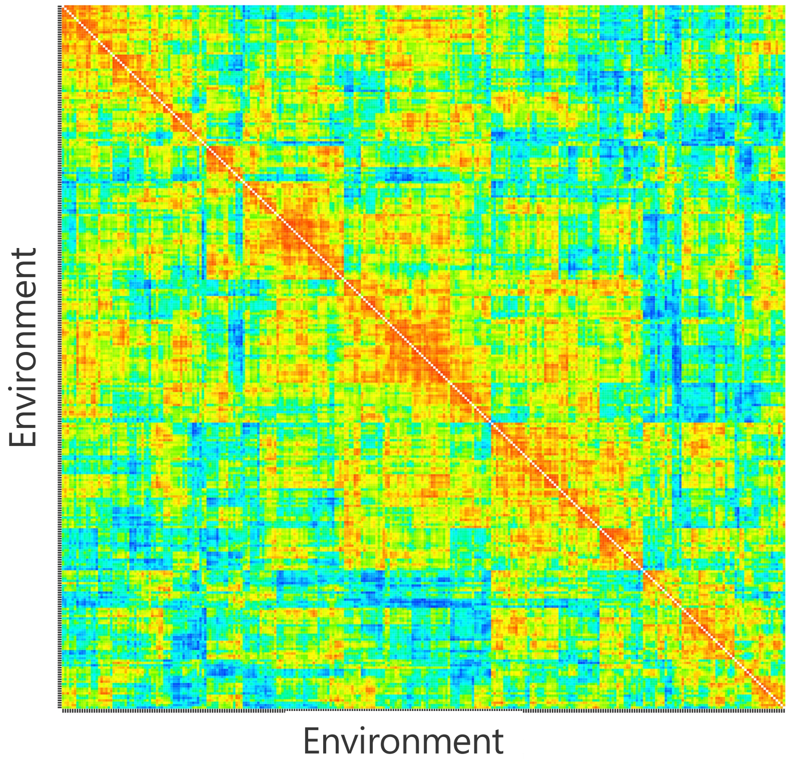
What we want

	E1	**E2**	**E3**	**E4**	**E5**
G1	0.20	0.08	0.13	0.31	-0.02
G2	0.07	-0.20	-0.17	0.43	0.21
G3	0.19	-0.24	0.08	-0.03	-0.23
G4	-0.25	0.39	-0.01	-0.16	0.24

Can be done with multiple correlated traits but becomes computationally challenging with large breeding simulations.

Sampling from TPE

1. **Simulate 1000 x 1000 TPE**
   (constant across simulation reps)

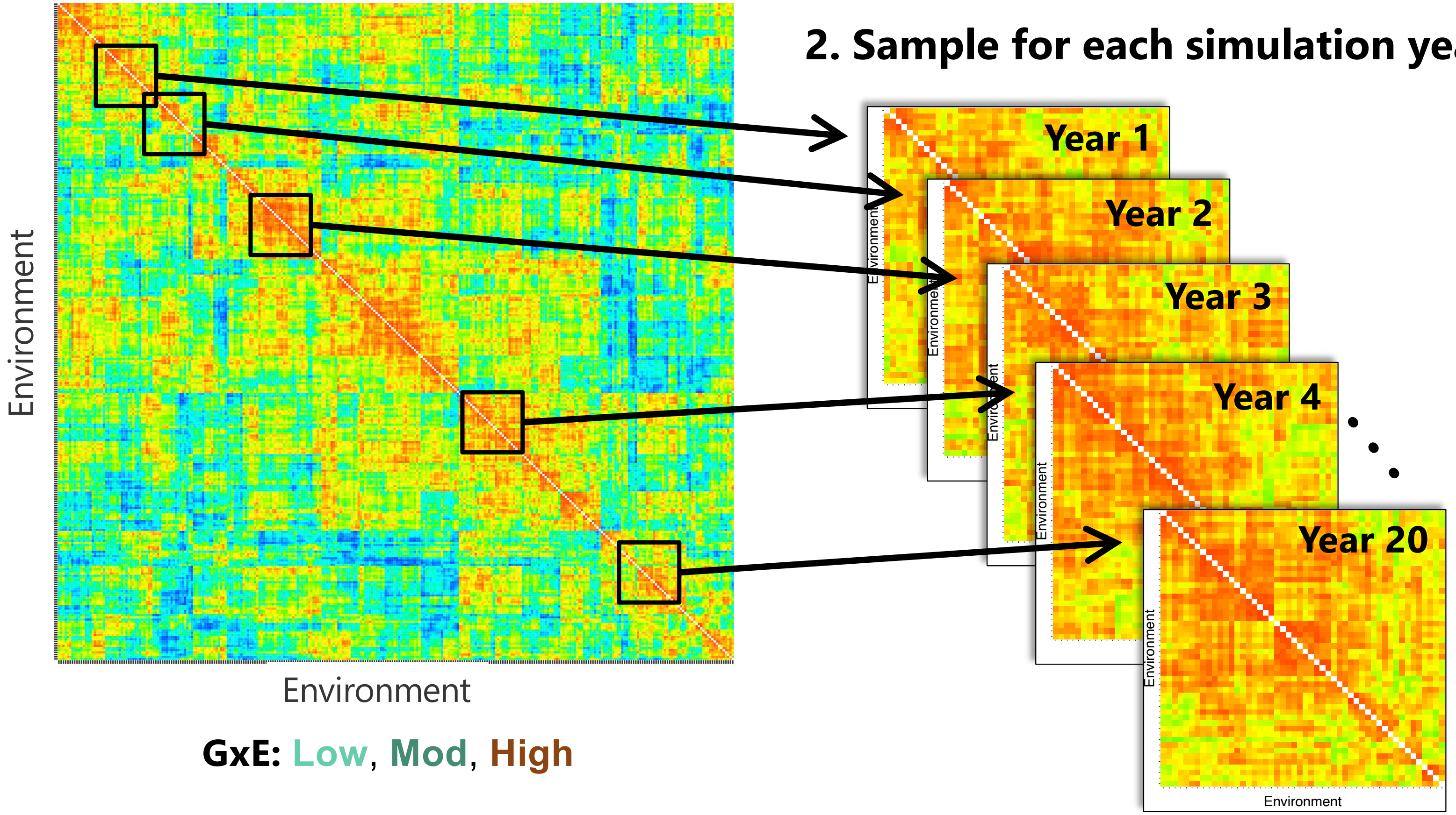


GxE: Low, Mod, High

Sampling from TPE

1. **Simulate 1000 x 1000 TPE**
   (constant across simulation reps)

2. Sample for each simulation year



GxE: Low, Mod, High

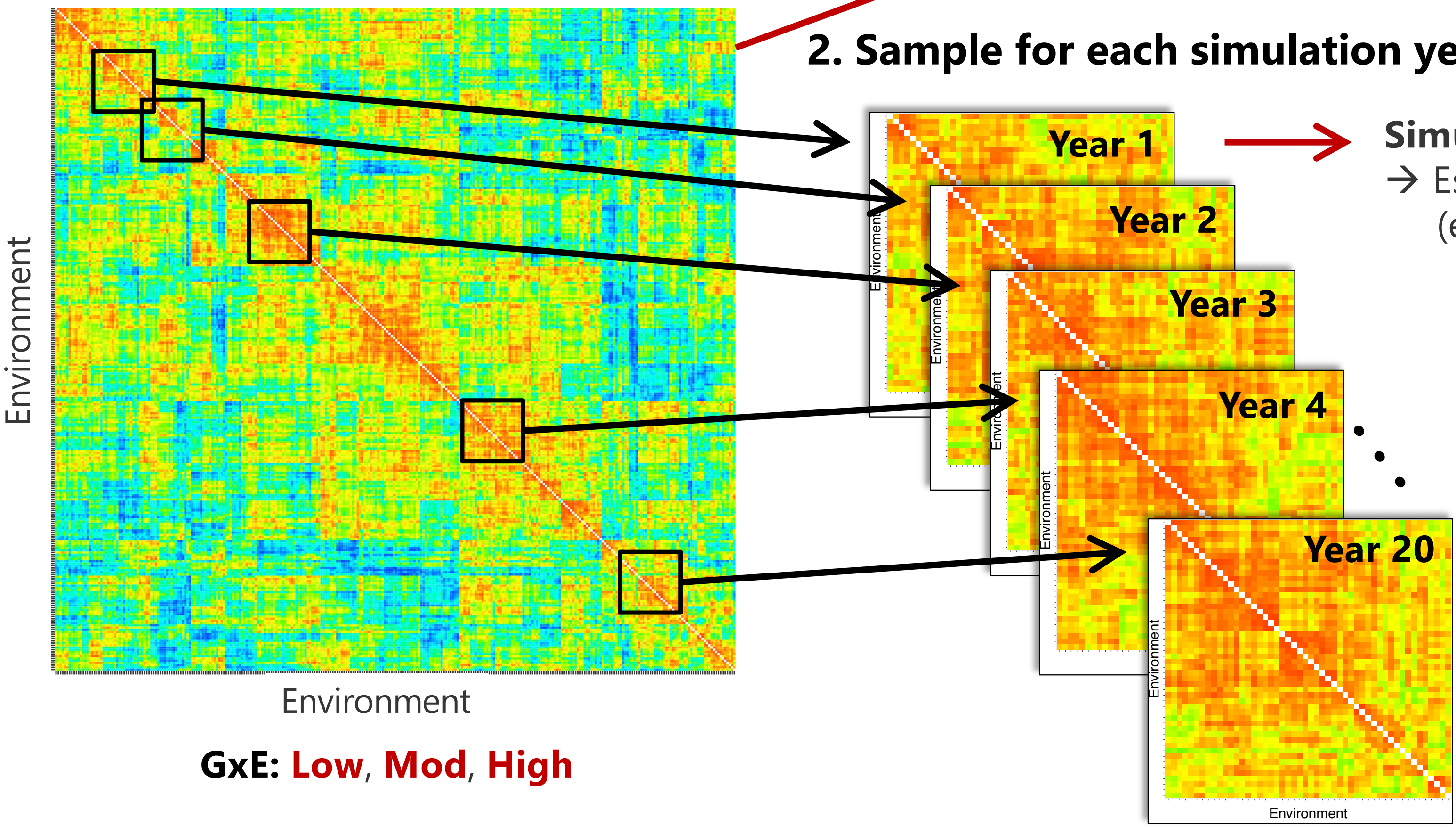
Sampling from TPE

Simulation of TPE genetic effects
→ True performance

1. **Simulate 1000 x 1000 TPE**
   (constant across simulation reps)

2. Sample for each simulation year

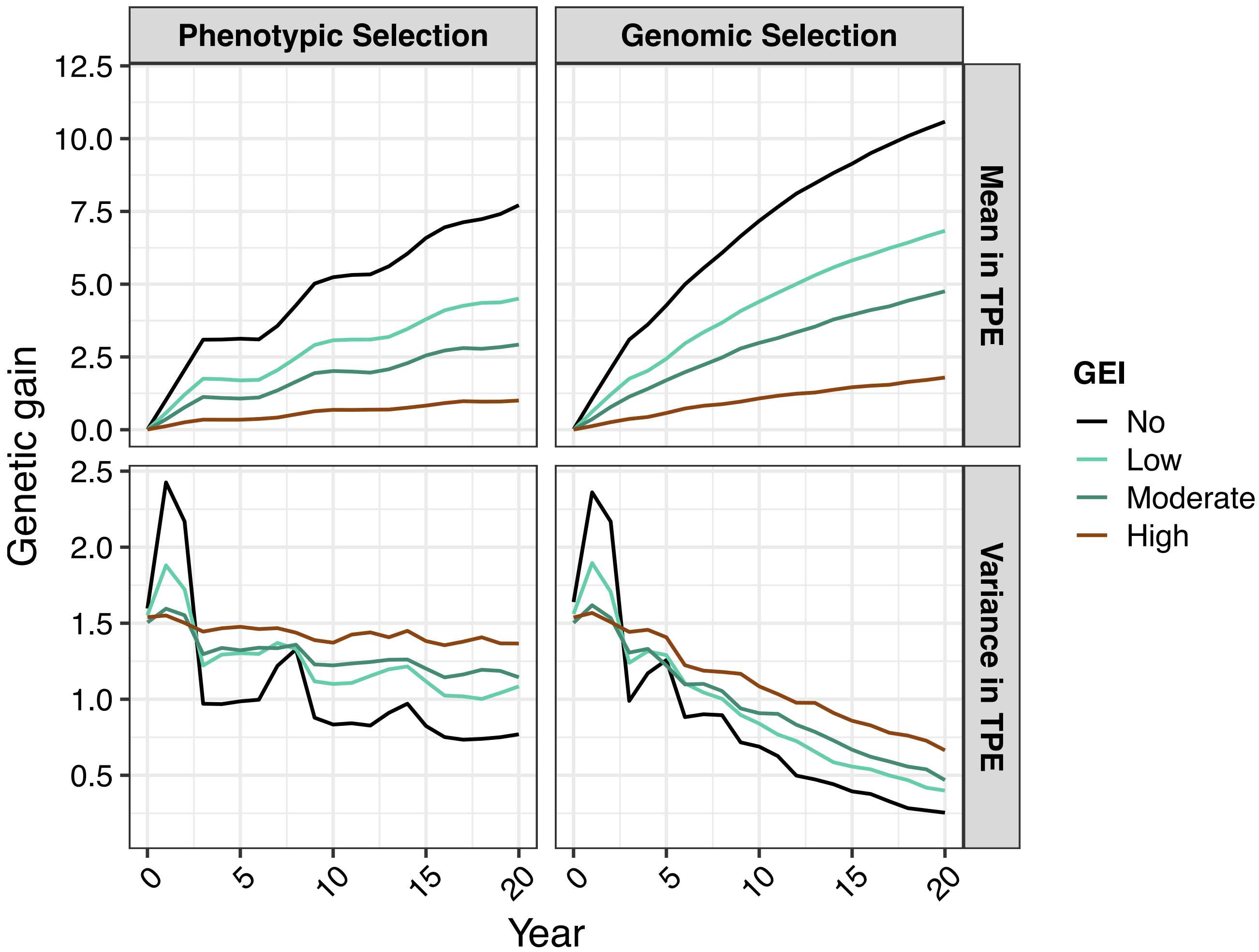
Simulation of MET genetic effects
→ Estimated/observed performance
(e.g. Stage 1 ~ 1 env,
Stage 4 ~ 20 env)



GxE: Low, Mod, High

Genetic gain and variance in Stage 1

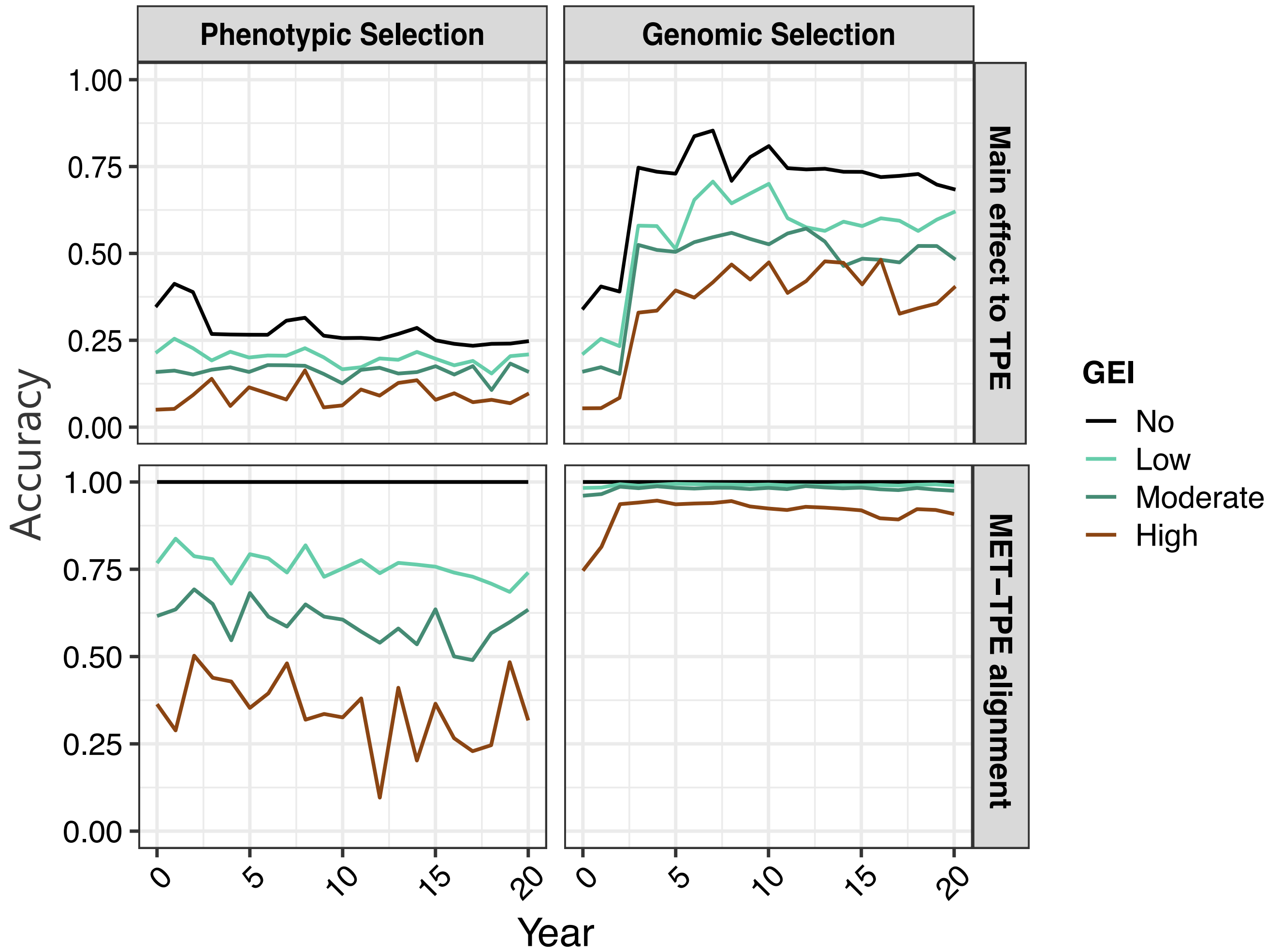
- Gain and variance loss decrease as GxE increases
- GS outperforms PS by 1.4-1.7 times
- Too optimistic projections in absence of GxE



Accuracy in Stage 1

MET-TPE alignment
$\mathrm{cor}(gv_{MET}, gv_{TPE})$

- Main effect accuracy and MET-TPE alignment decrease as GxE increases
- Main effect accuracy and MET-TPE alignment are higher for GS



New opportunities

- Long-term statistical model comparison
- Model selection at different breeding stages
- Selection for genotype stability
- Long-term alignment with TPE
- Recreation of long-term GEI patterns

Take home messages

Scalable and reproducible framework for simulating GxE

1. simulate realistic MET datasets
2. model plant breeding programs

Framework can simulate

- large number of environments
- different magnitudes of non-crossover and crossover GxE
- different correlated genetic effects
- multiple TPE and multiple phenotypic traits