Comparison of strategies for discovery and utilization of useful genetic material from maize landrace populations

Gorjanc G, Jenko J, Hickey J, Hearne S

University of *Ljubljana*
Biotechnical Faculty

2013-07-05

Aim

Compare strategies of discovery of useful genetic material from landraces for the potential introgression into the improved breeding material in maize via simulation

How?

The plan

1) Genotype gene bank accessions

2) Collect test-cross phenotypes

3) Genomic evaluation and selection

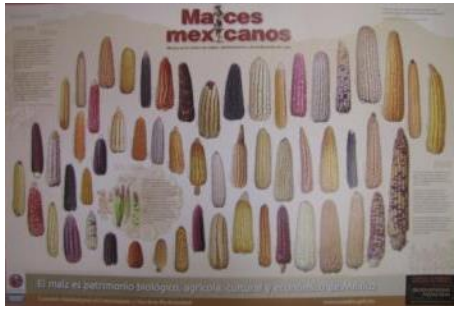
4) Create synthetic population with useful genetic material for further improvement

Simulation - setup

- Mimic the „discovery scheme“ of the SeeD project

Maize genome

Landraces



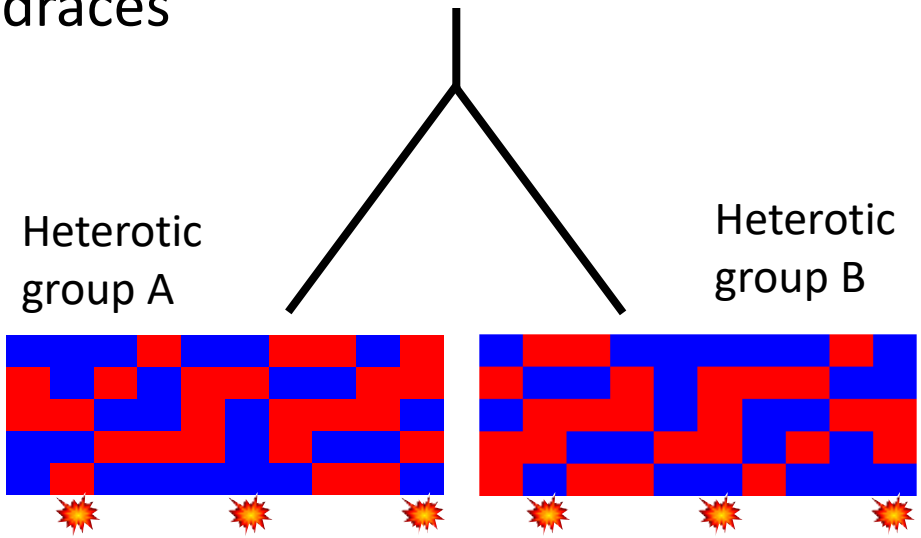
Test-cross phenotypes

Simulation – factors?

- **Effective population size** → genomic variance (N_e=1K, 100K) #2
- **Landrace variance** (F=0.3, 0.9) #2
- **Heritability** (h^2=0.25, 0.50) #2
- **Approach** (Landrace, LandraceDH, Test-Cross) #3
- **Genotyping platform** (GBS10x10K, GBS1x100K) #2
- **Landrace seeds in training** (nSeed=1, 3, 5) #3
- **Selection of landraces** (n=40, 80, k=10, 20, 40) #6
- **Retrain in cycles** (n=0, 20, 40, 60) #4
- Scenarios: 2×2×2×3×2×3×6×4=2592 × 10 replicates

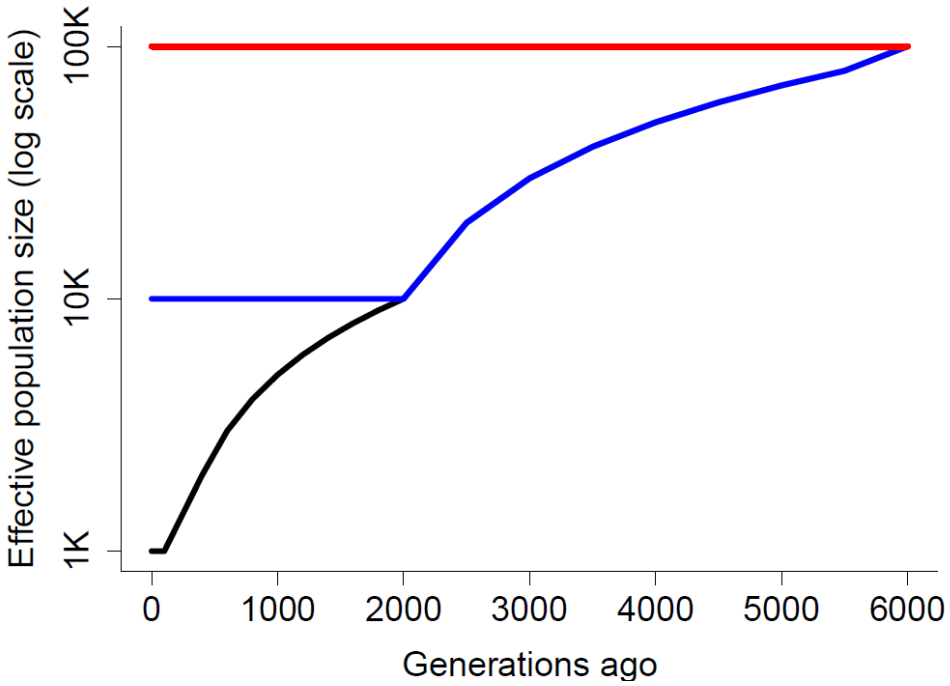
Simulation - maize genome

- Coalescent simulation per chromosome
  - two heterotic groups (A and B)
  - allocate QTLs and SNPs
  - rank haplotypes by true breeding value (TBV)
    - the best for testers (A, B → A×B)
    - lowest half for landraces



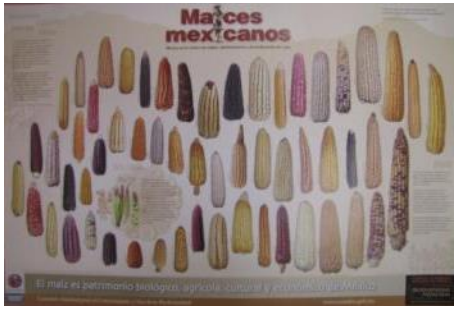
Simulation - maize genome II

- Coalescent simulation per chromosome
  - Effective population size (N_e=1K, 10K, <u>100K</u>)
  - Mutation rate 2.5×10^{-8}
  - Recombination rate 1.0×10^{-8}

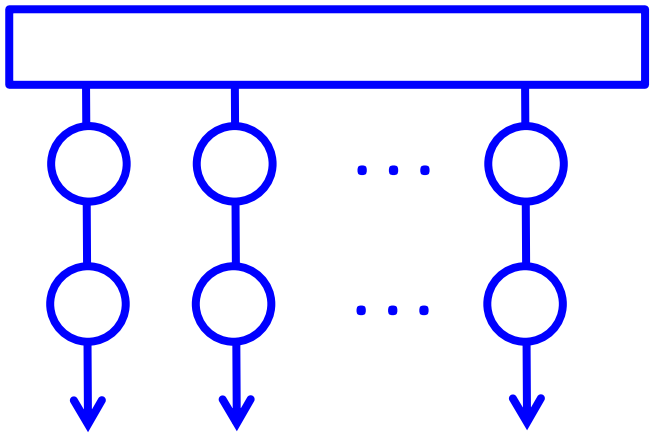


- N_e=100K
  #Seg. sites: 27×10^6
  Disk usage: 152 GB
- N_e=10K
  #Seg. sites: 10×10^6
  Disk usage: 100 GB
- N_e=1K
  #Seg. sites: 5×10^6
  Disk usage: 28 GB

Simulation - landraces



- 3000 landraces via gene dropping
  - global (between) variance already due to large N_e
  - between and within variance controlled by genetic drift (random mating of $2N_p$ individuals for t generations)



$$N_p = \frac{1}{4\left(1-(1-F_t)^{\frac{1}{t}}\right)}$$

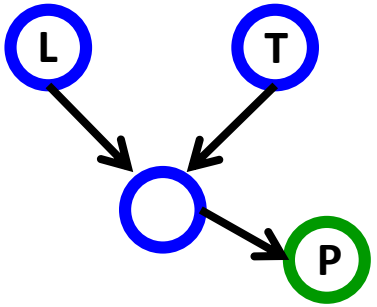
F_t	$2N_p$	t
0.3	10	7
0.9	2	9

Simulation - phenotypes

- Additive polygenic model
  (2000 QTL per chromosome from Gaussian dist.)

- Cross with a hybrid-tester and obtain individual true breeding values and phenotypic values

- Phenotype = mean test-cross performance with defined heritability (h^2=0.25, 0.50)

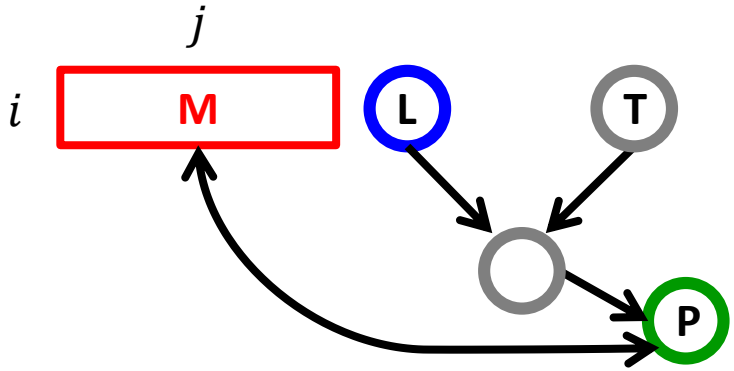


Simulation – genomic model

- Genomic evaluation using a model linking phenotypic values to genomic markers to obtain a genomic estimate of a breeding value (gEBV)

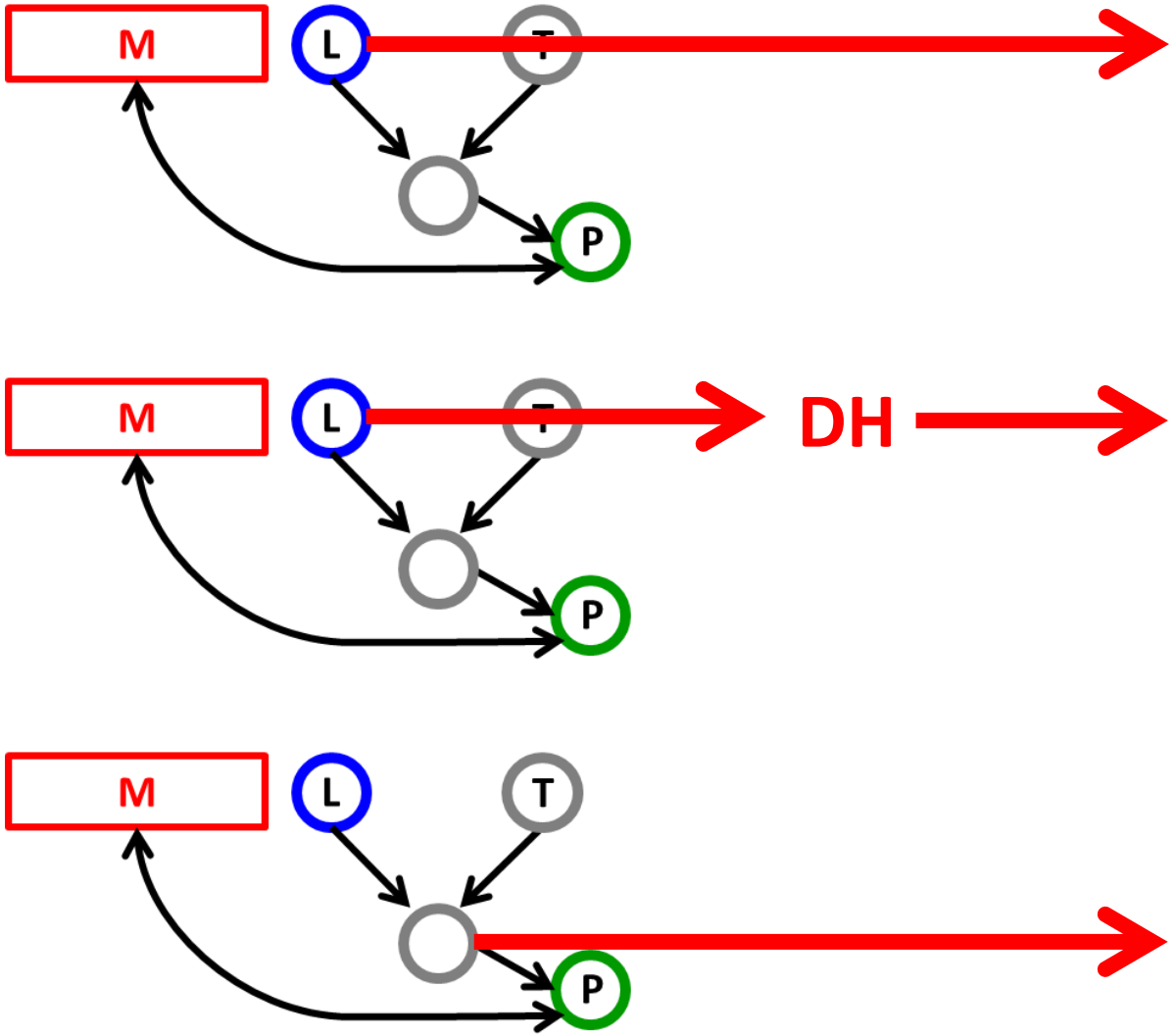
- Ridge regression

$$y_i \sim N(\mu_i, \sigma_e^2)$$
$$\mu_i = \alpha + \sum_{j=1}^{p} m_j x_{i,j}$$
$$m_j \sim N(0, \sigma_m^2)$$
$$gEBV_i = \sum_{j=1}^{p} m_j x_{i,j}$$

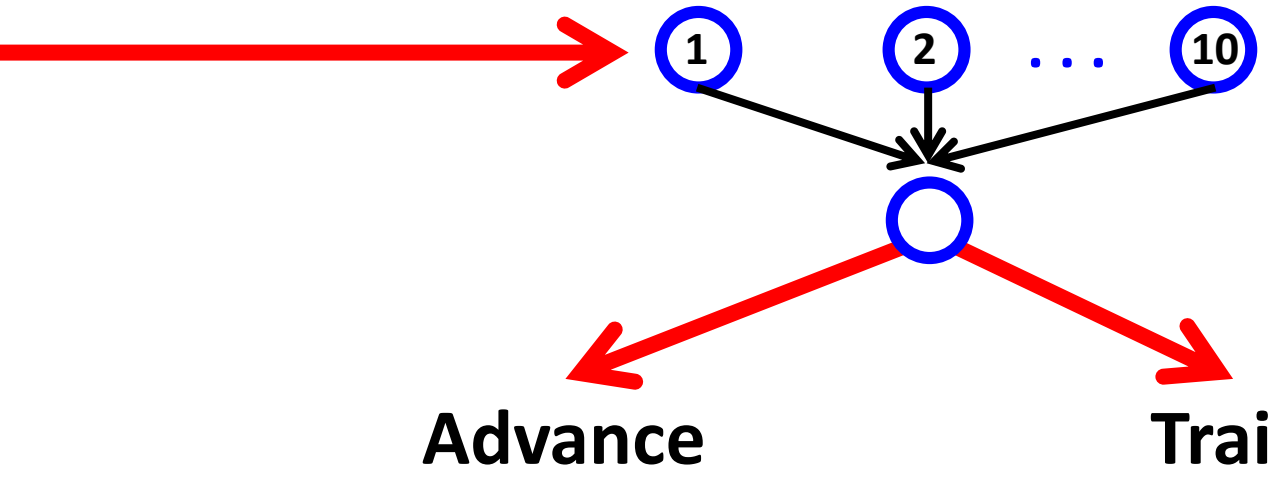


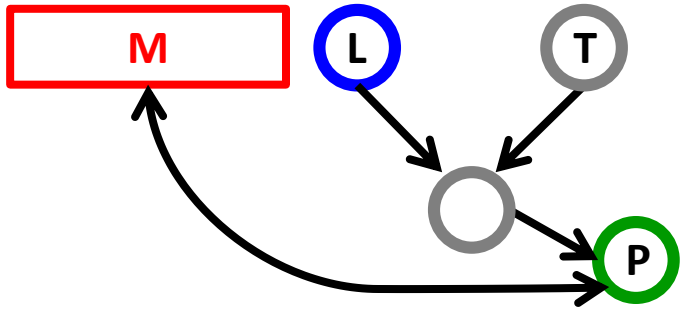
Simulation – approach (discovery phase)

- Landrace
- LandraceDH
- Test-cross



Simulation – approach (synthetic population development)





4x

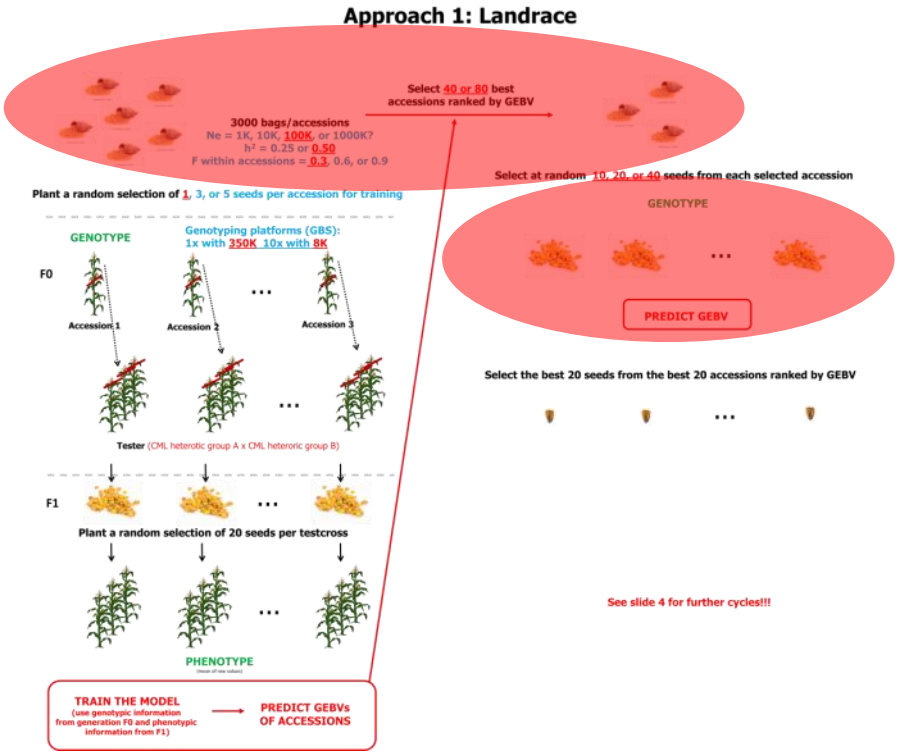
Simulation – factors?

- **Effective population size** → genomic variance (N_e=1K, 100K) #2
- **Landrace variance** (F=0.3, 0.9) #2
- **Heritability** (h^2=0.25, 0.50) #2
- **Approach** (Landrace, LandraceDH, Test-Cross) #3
- **Genotyping platform** (GBS10x10K, GBS1x100K) #2
- **Landrace seeds in training** (nSeed=1, 3, 5) #3
- **Selection of landraces** (n=40, 80, k=10, 20, 40) #6
- **Retrain in cycles** (n=0, 20, 40, 60) #4
- Scenarios: 2×2×2×3×2×3×6×4=2592 × 10 replicates

Metrics of interest?

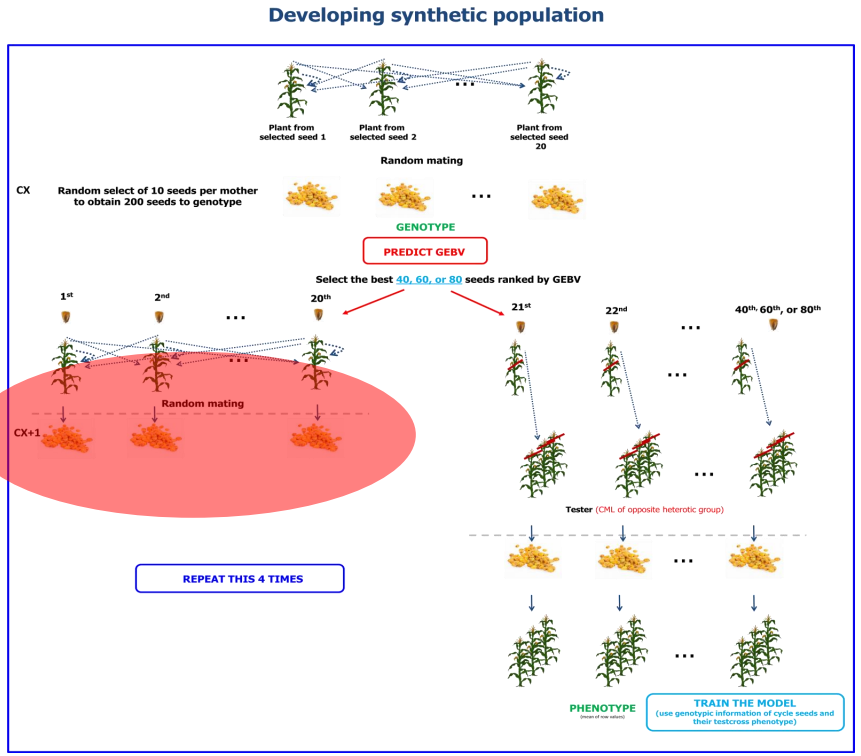
- Metrics:
  - **genetic gain (ΔG)**
    (mean of true breeding values after selection)
  - **accuracy of evaluation**
    (correlation between the true and estimated value)
  - **genomic similarity to a tester**
    (proportion of genome equal to the genome of a tester)
- Where (stage of simulation):
  - ???

Stage of simulation



1. Selecting landraces (bags → B)

2. Sel. seeds within landrace (WB)



3.-6. Synthetic population cycles (C1, C2, C3, and C4)

How to present results?

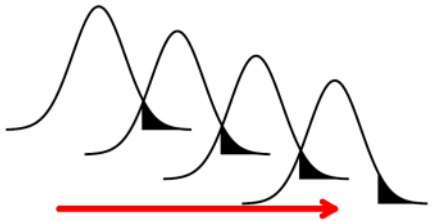
- Results
  - 2592 scenarios × 10 replicates
  - 6 stages of simulation

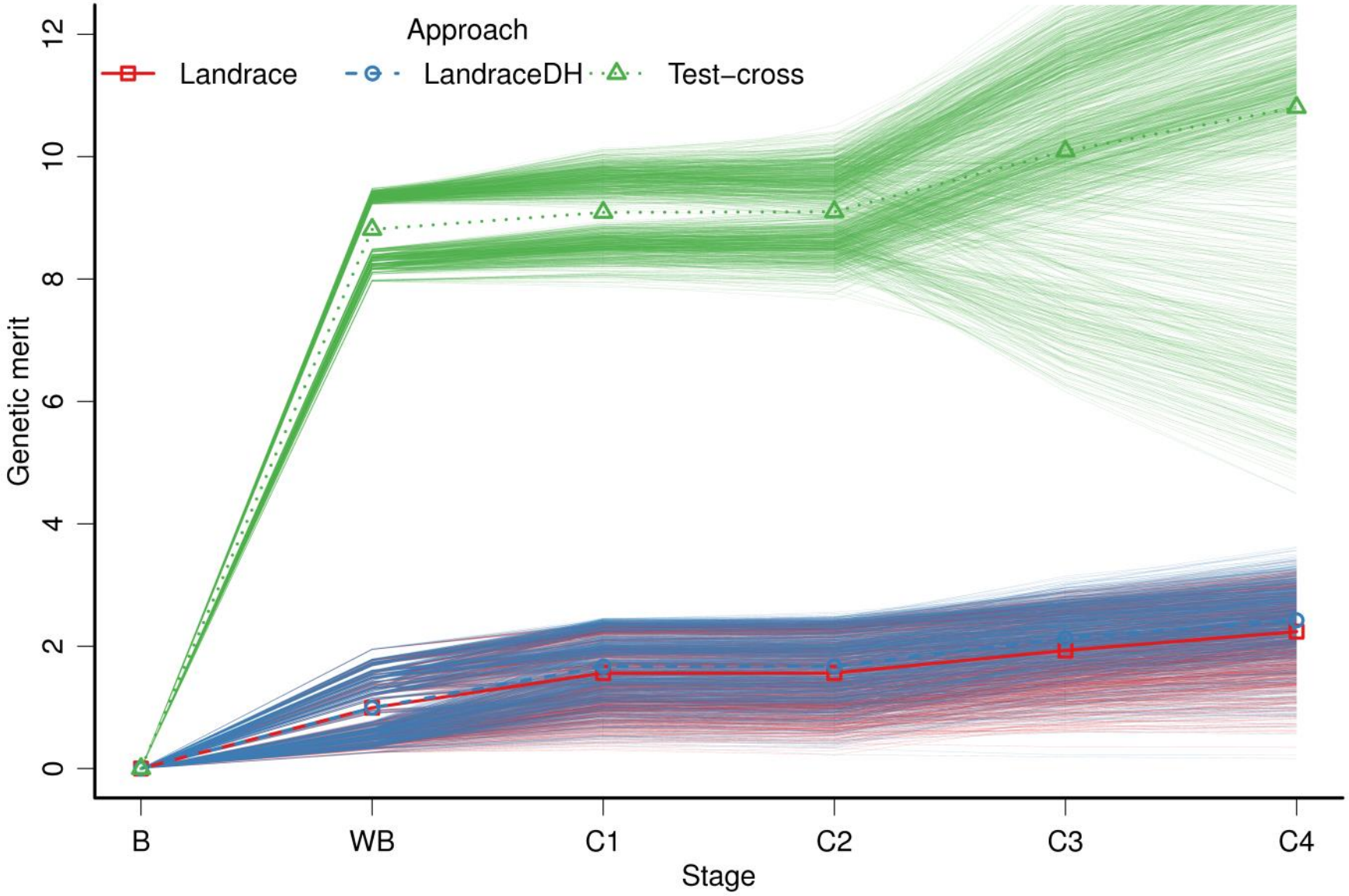
  →25920 × 6 ≈ 150 000 values per metric

- Show the effect of individual factors

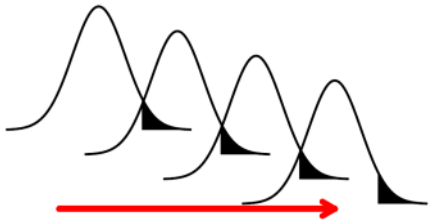
- Show the effect of the most influential factors and their interactions with the use of regression trees

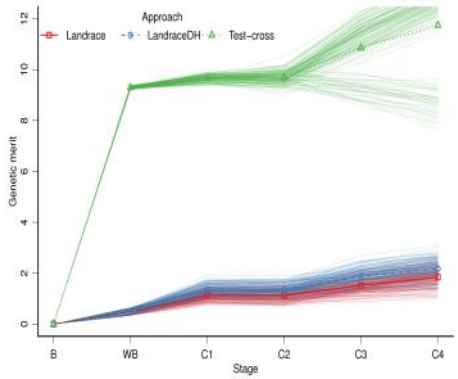
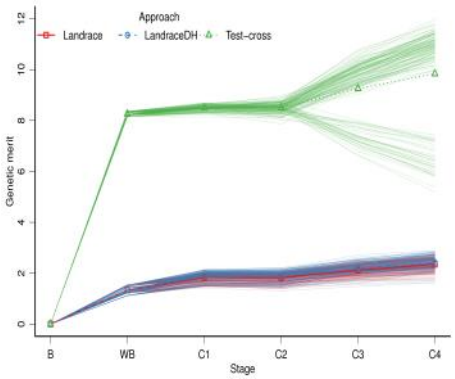
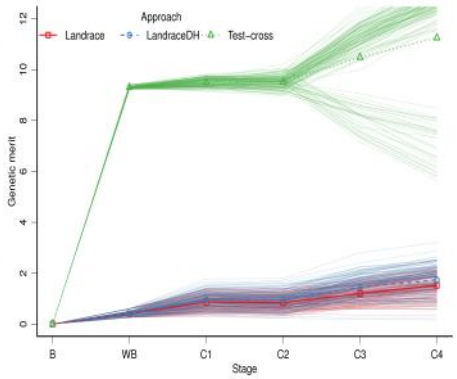
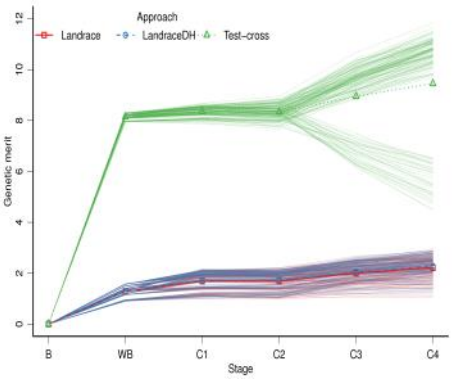
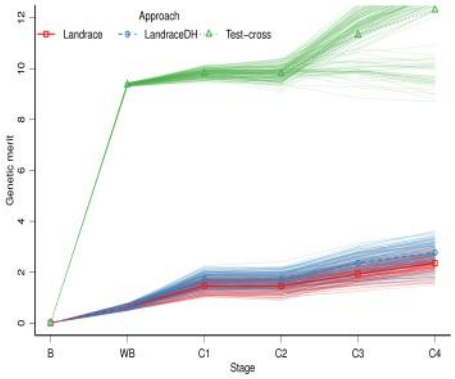
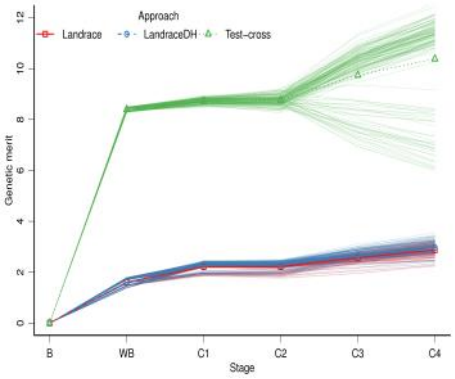
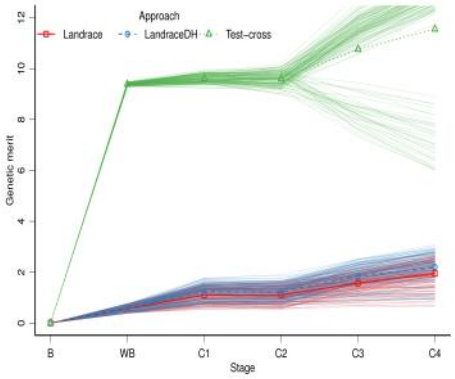
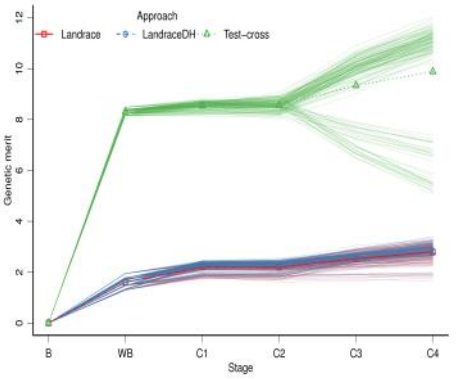
Approach - ΔG



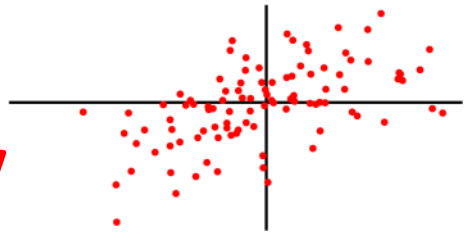

Approach
Landrace
LandraceDH
Test−cross
Genetic merit
0
2
4
6
8
10
12
B
WB
C1
C2
C3
C4
Stage

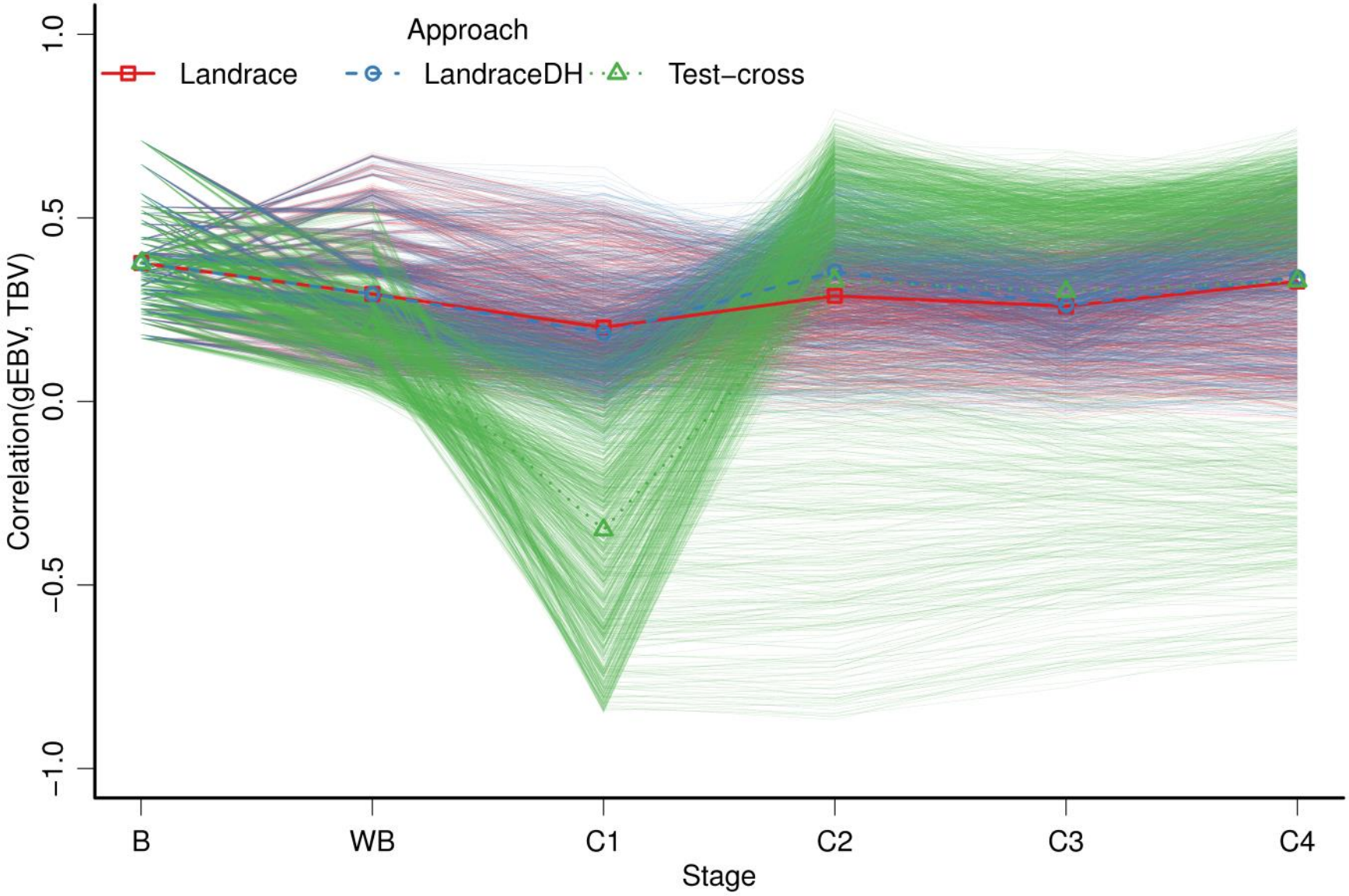
Approach - ΔG



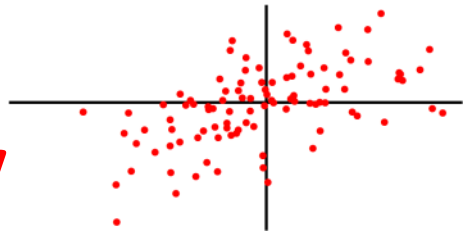
N_e	1K		100K	
h^2 / F	0.3	0.9	0.3	0.9
$\frac{1}{4}$				
$\frac{1}{2}$				

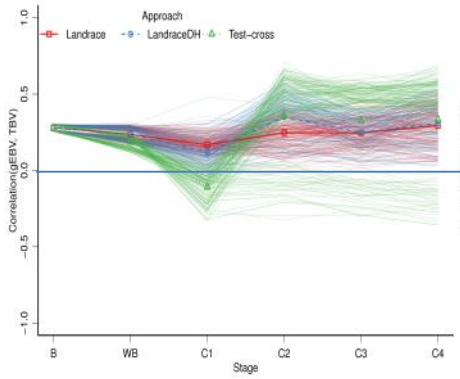
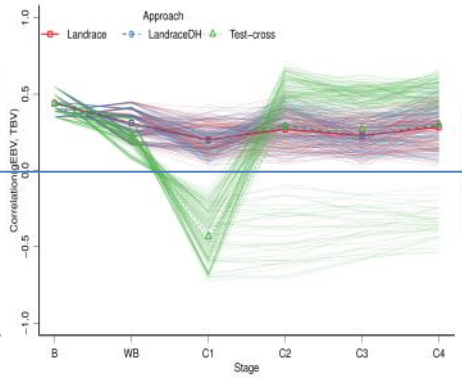
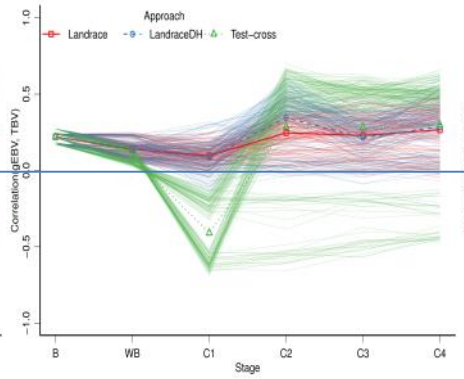
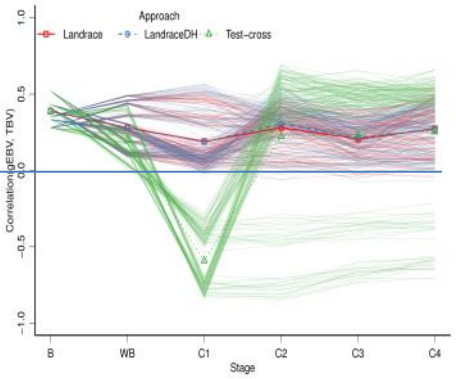
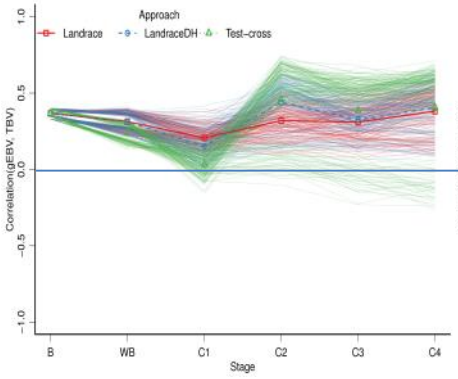
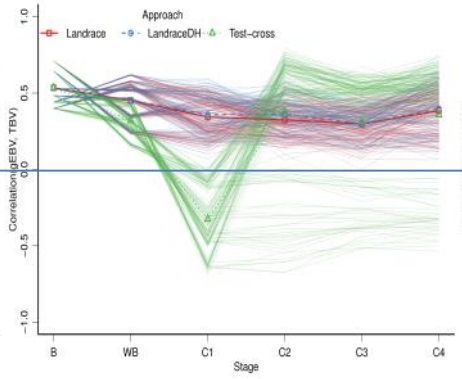
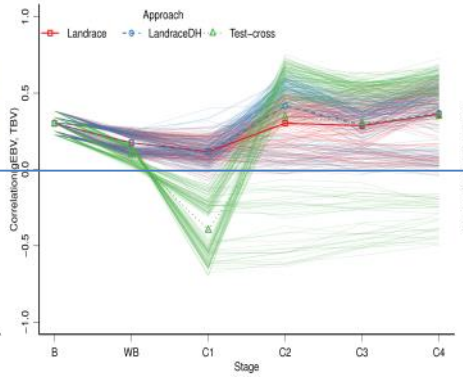
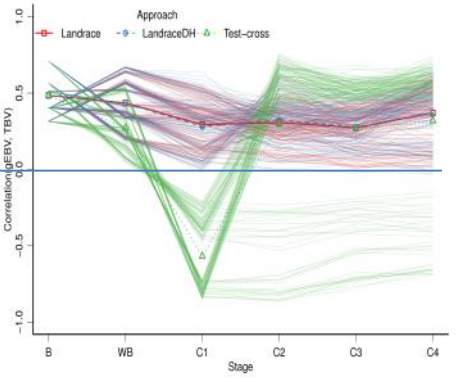
Approach - accuracy



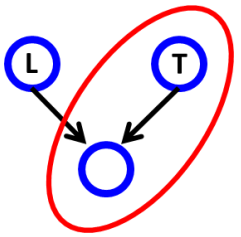


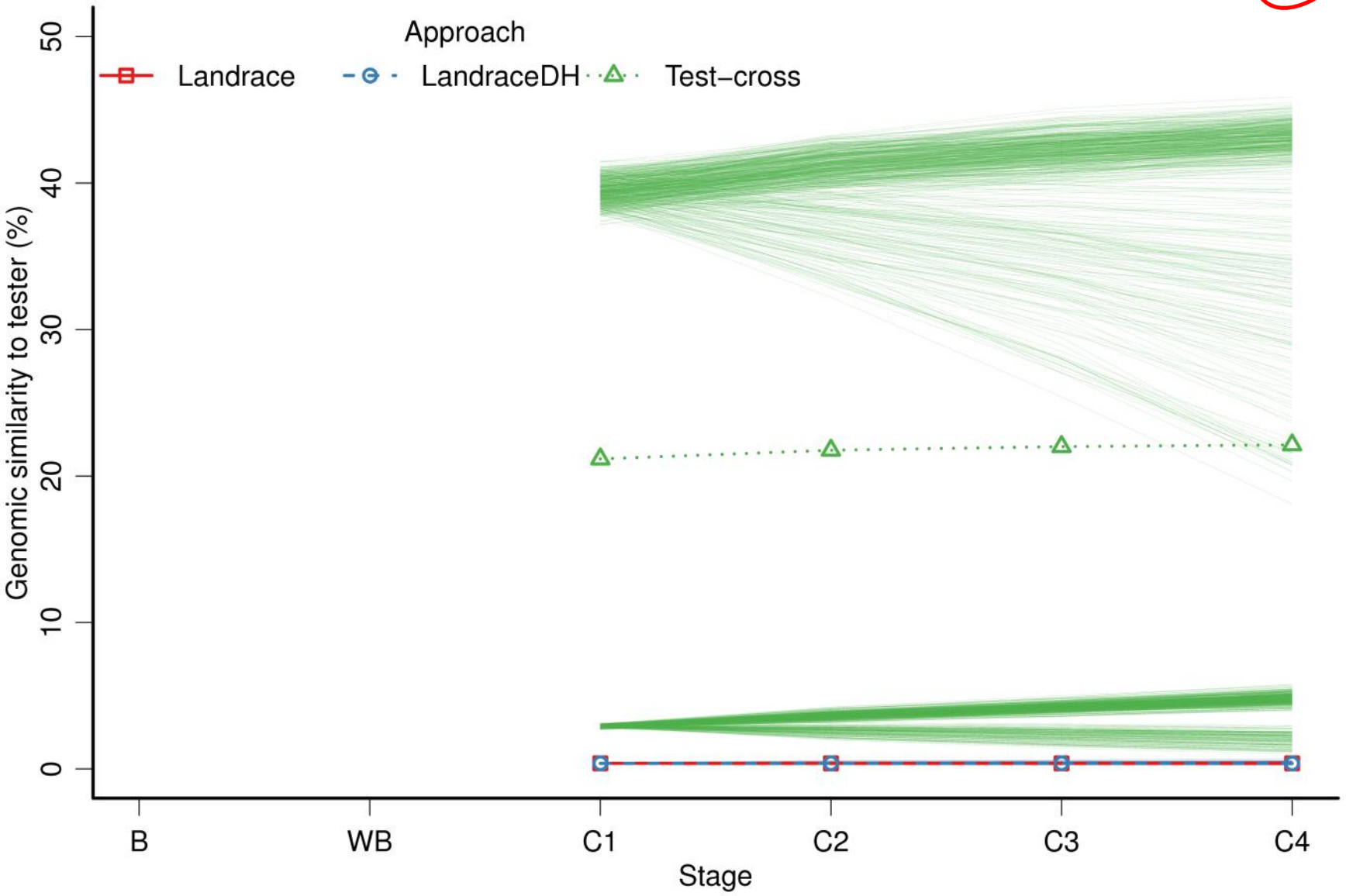
Approach - accuracy



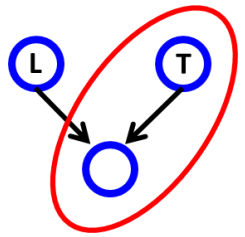
N_e	1K		100K	
h^2 / F	0.3	0.9	0.3	0.9
1/4				
1/2				

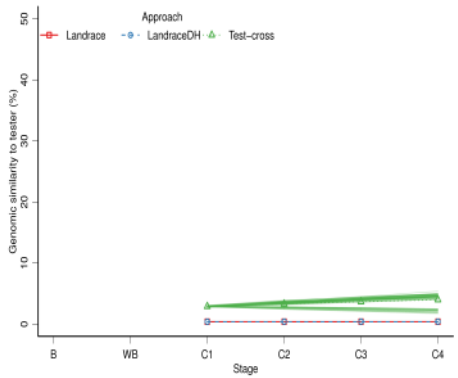
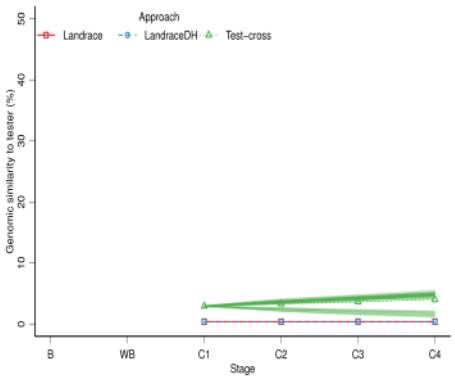
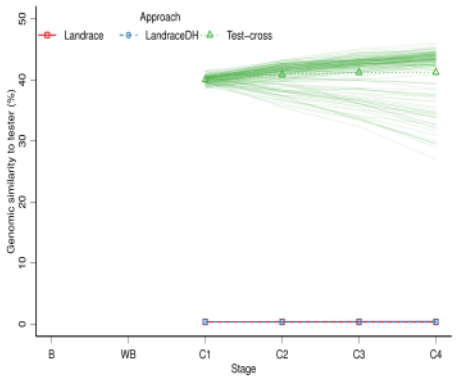
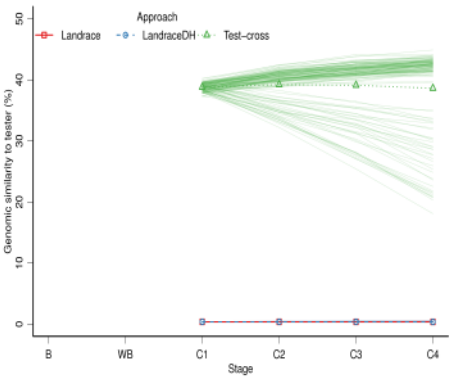
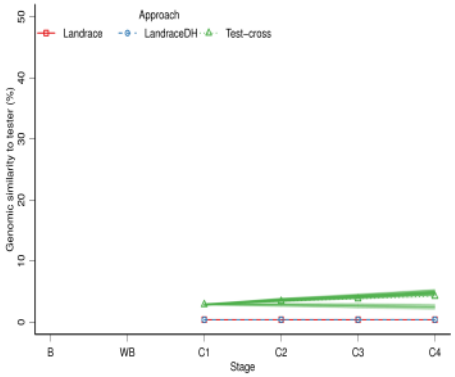
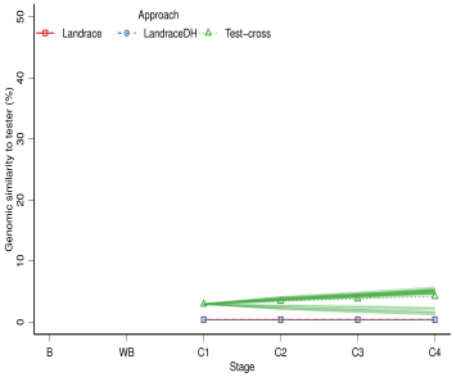
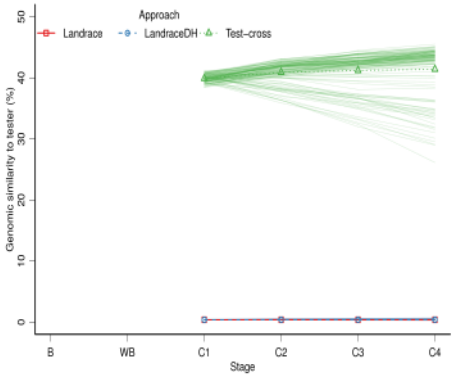
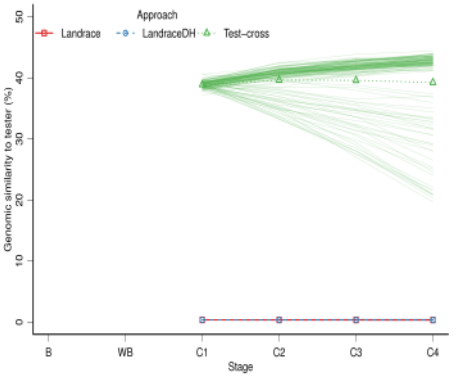
Approach - similarity





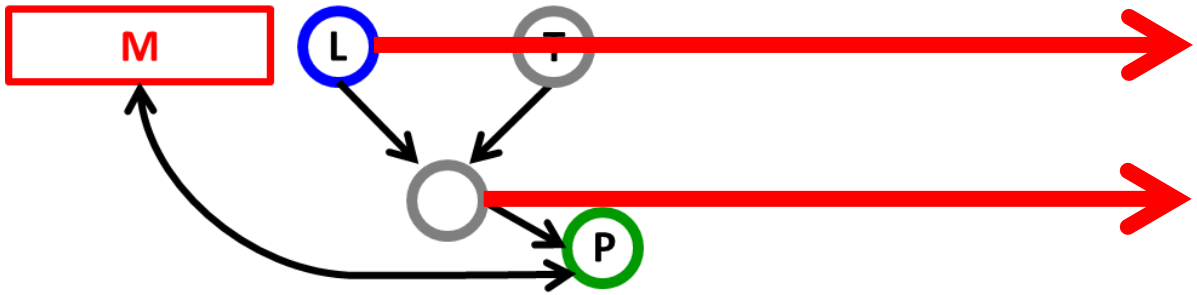
Approach - similarity



N_e	1K		100K	
h^2 / F	0.3	0.9	0.3	0.9
$\frac{1}{4}$				
$\frac{1}{2}$				

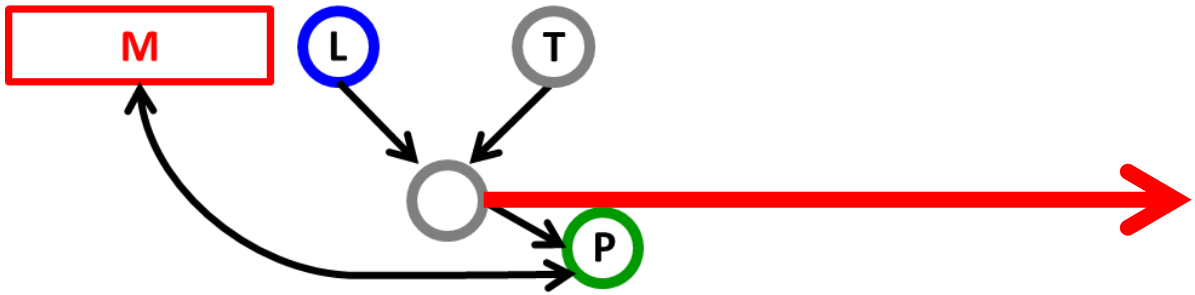
???

- Very strong interaction between the genomic variability (as measured by N_e) and approach in the discovery phase!!!
  - associations between markers and QTL weaker with larger N_e
  - marker data available on landraces and not on the test-cross individuals

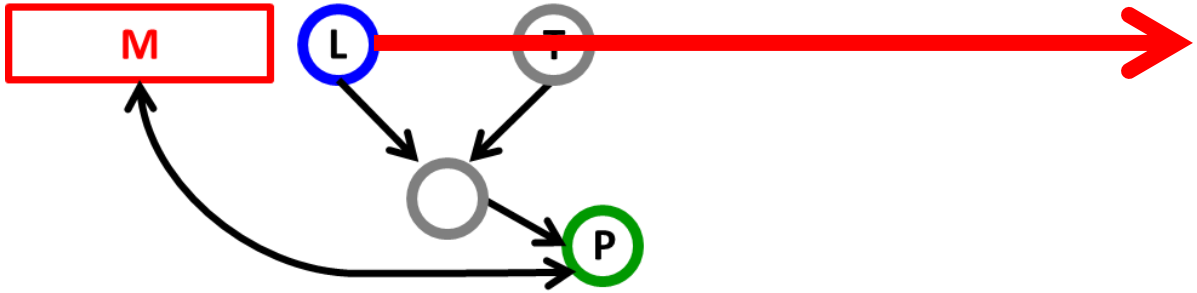


Introgression & N_e

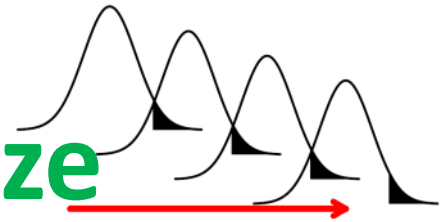
- With small N_e use test-cross individuals for the development of improved lines for breeding (i.e. deriving lines from TC used for recalibration)

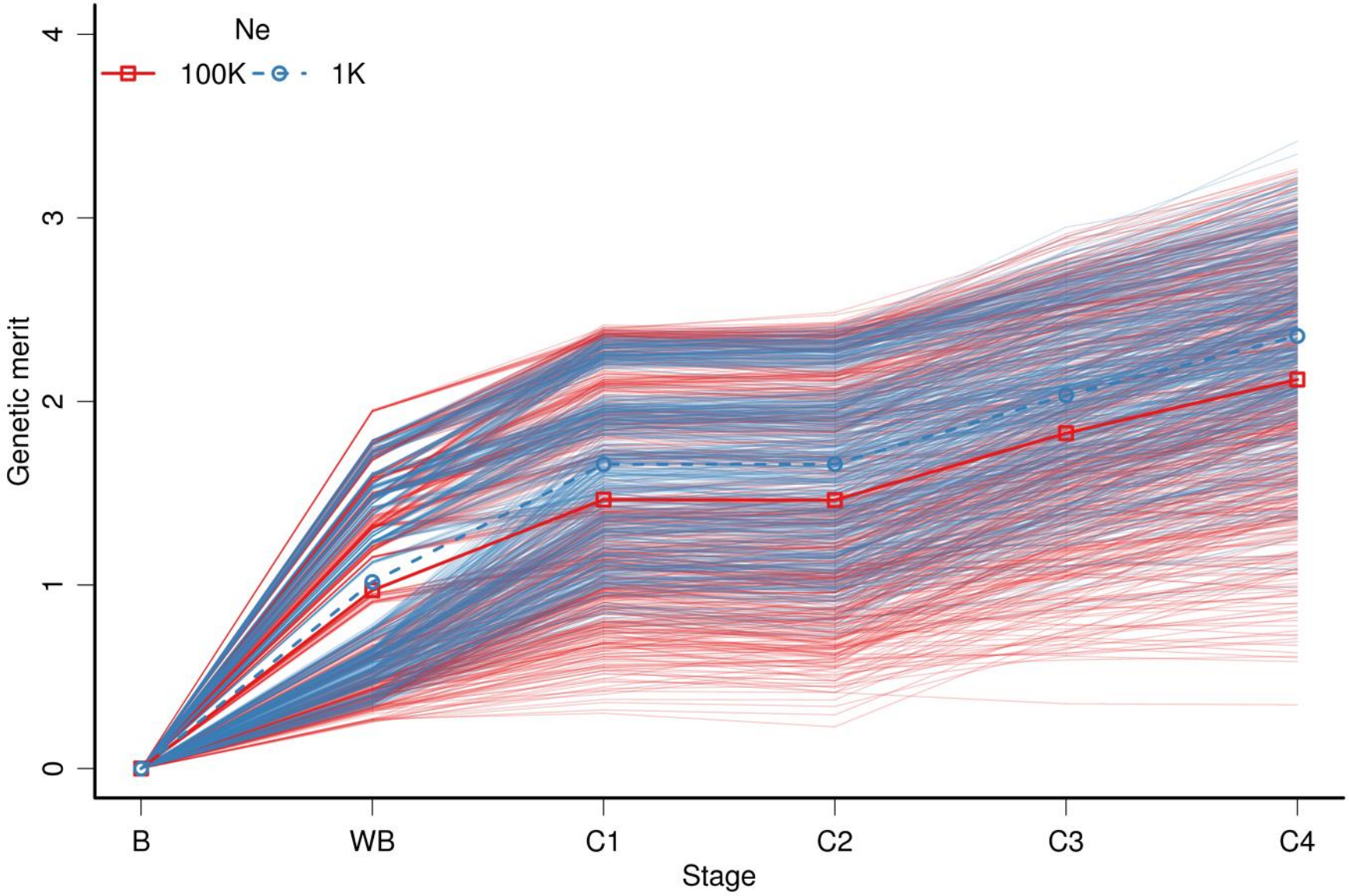


- With large N_e use landrace individuals for the development of improved populations

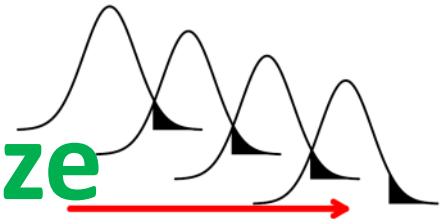


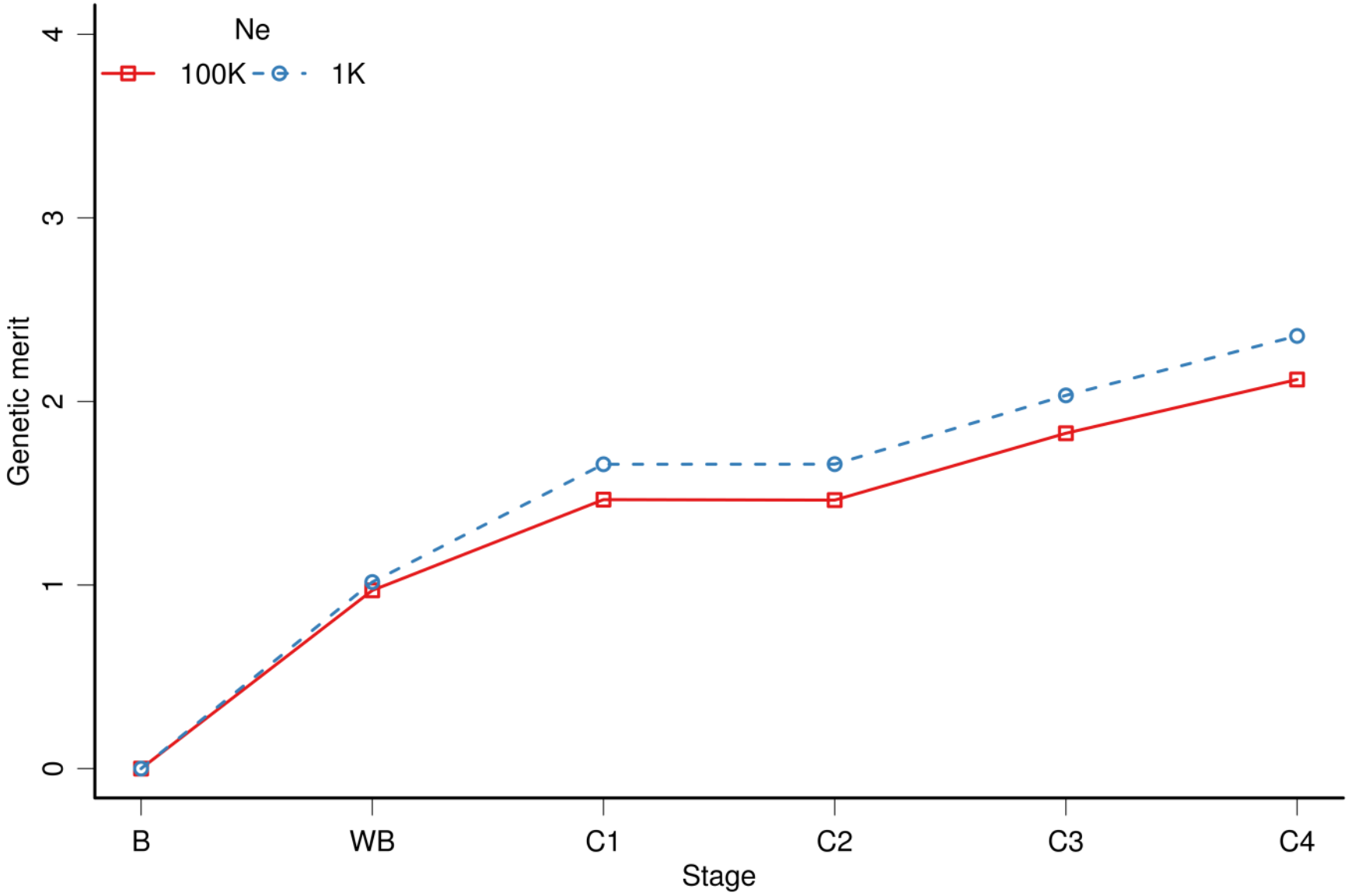
Effective population size



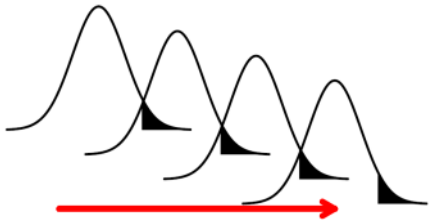


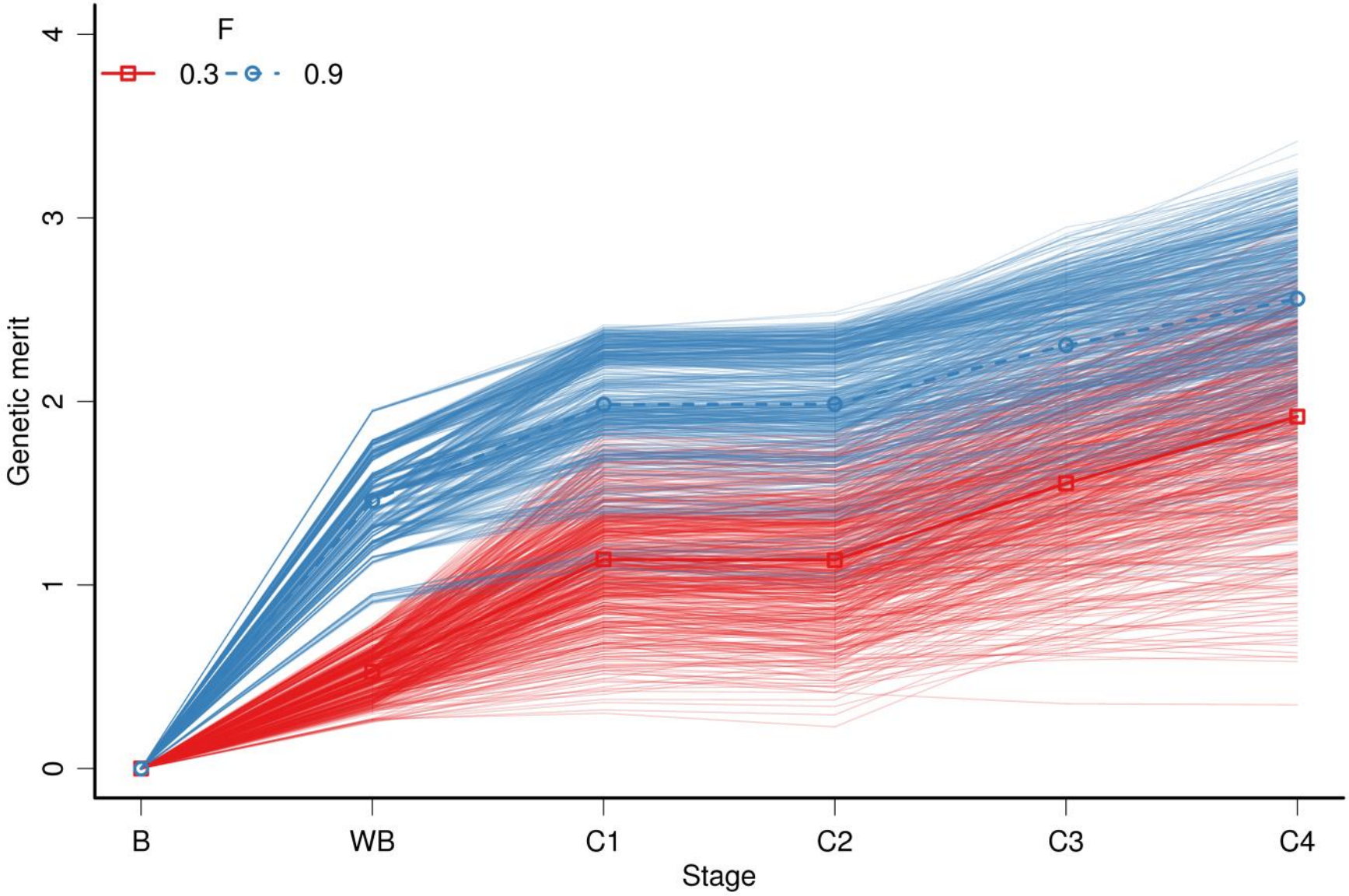
Effective population size



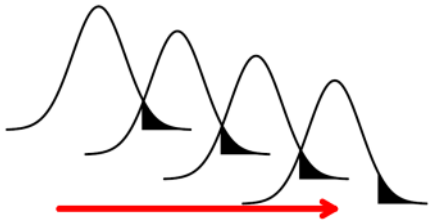


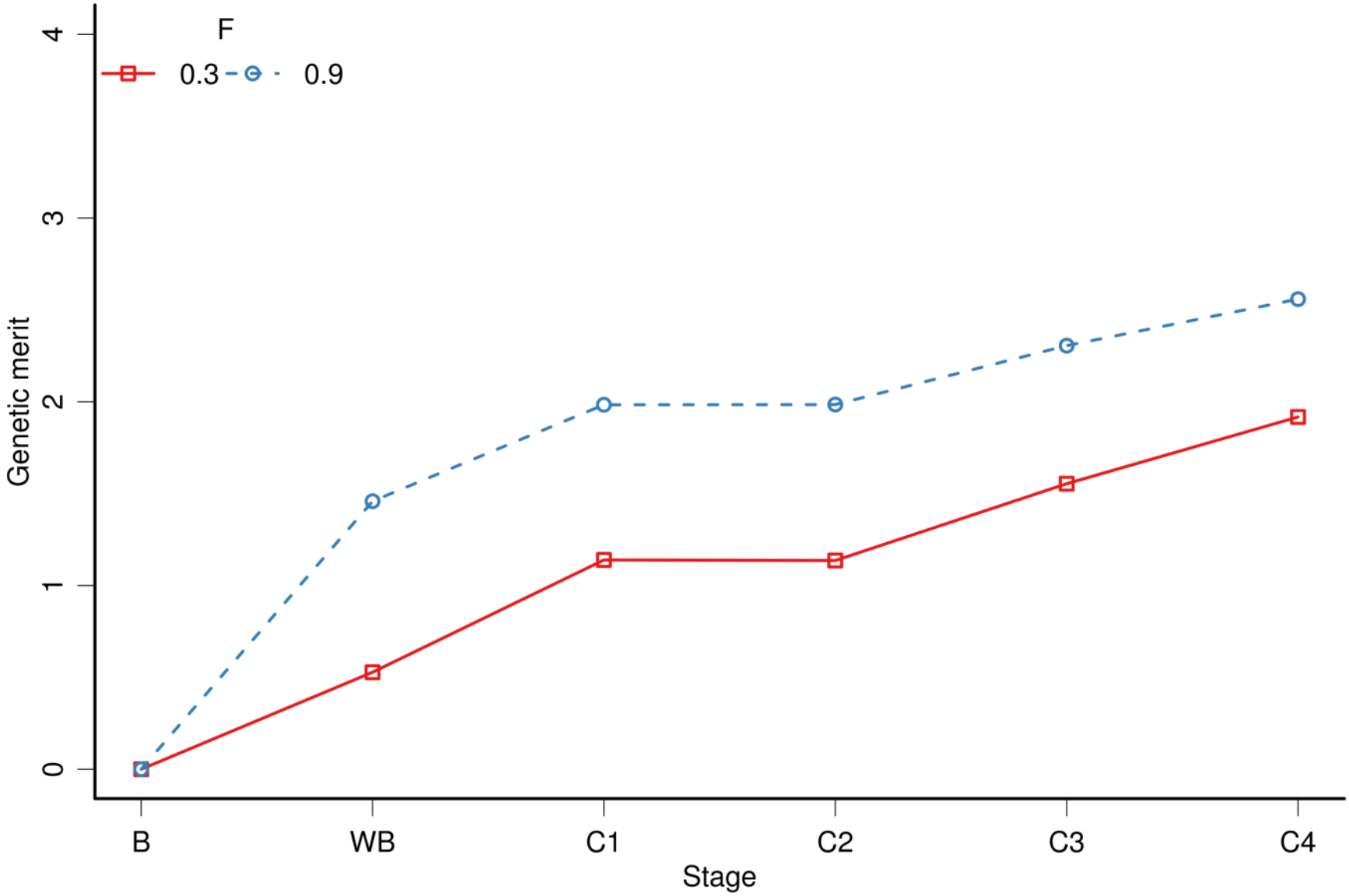
Landrace variance



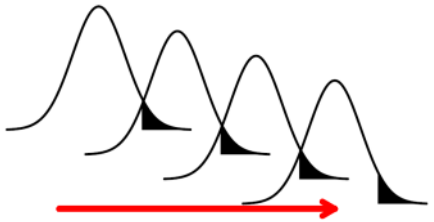


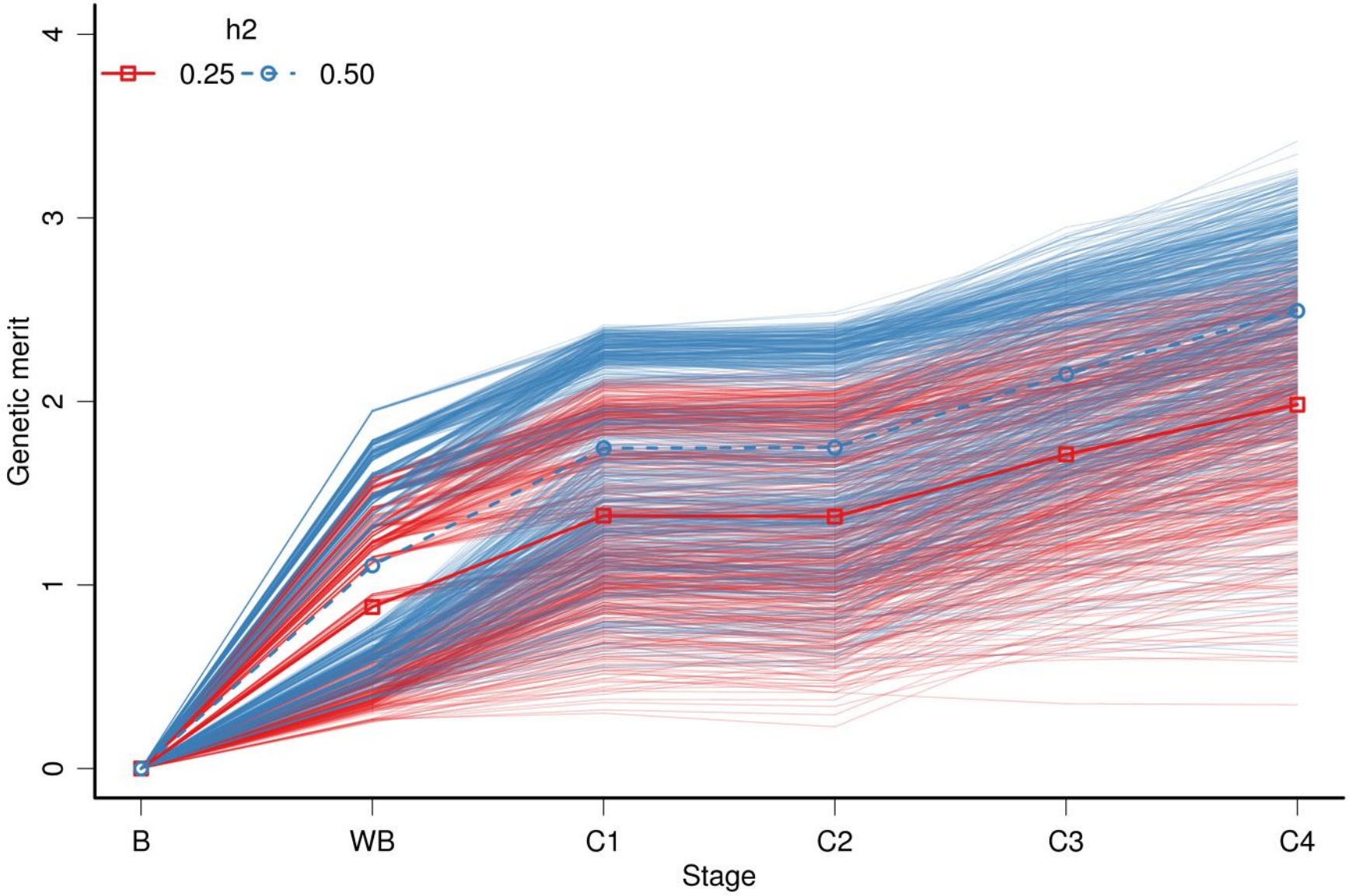
Landrace variance



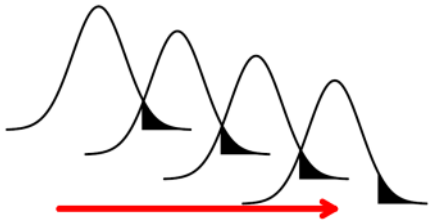


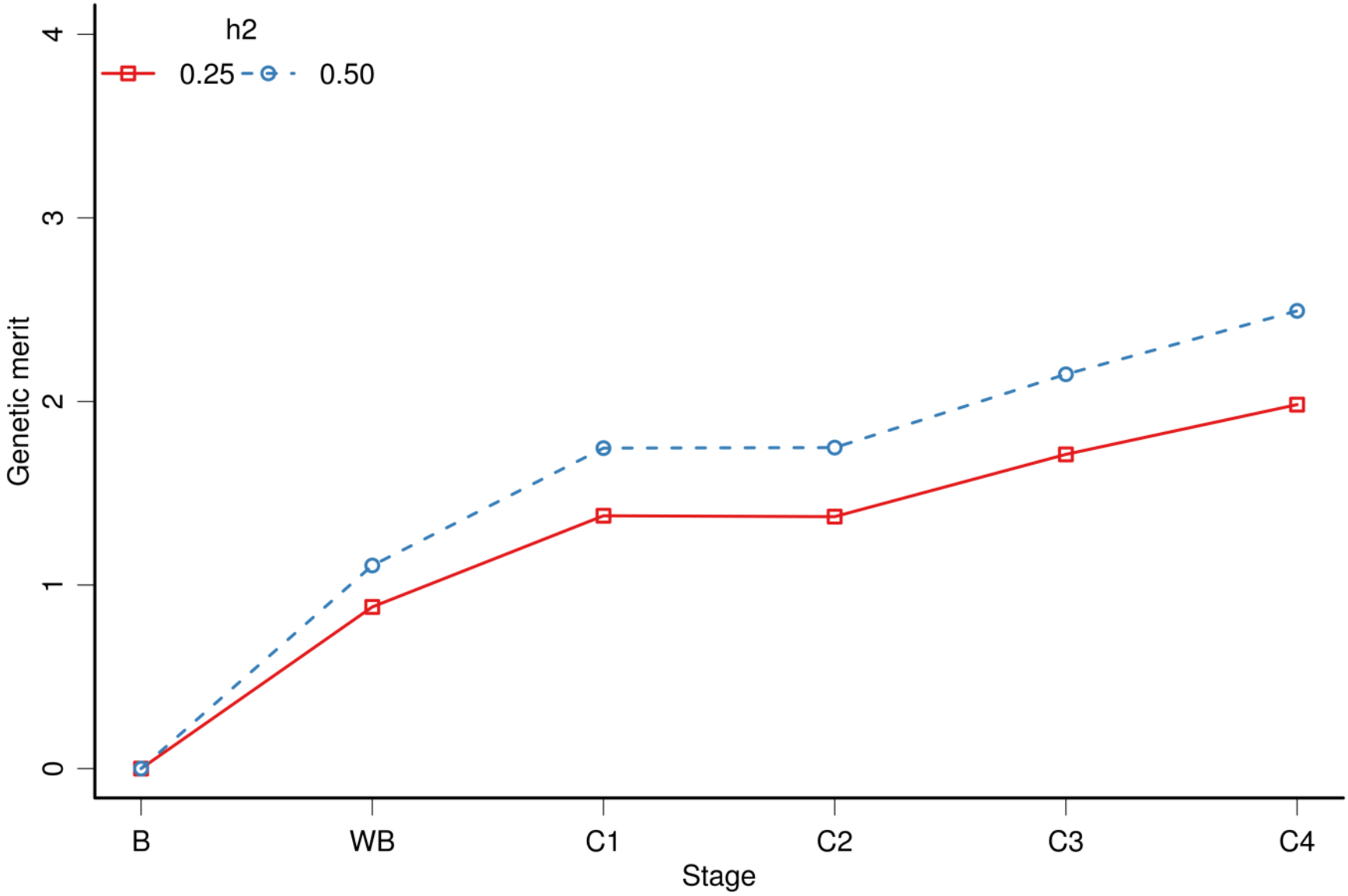
Heritability



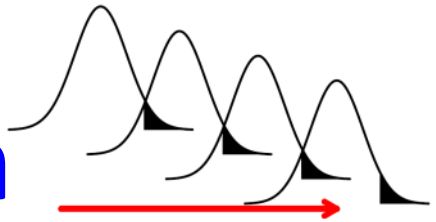


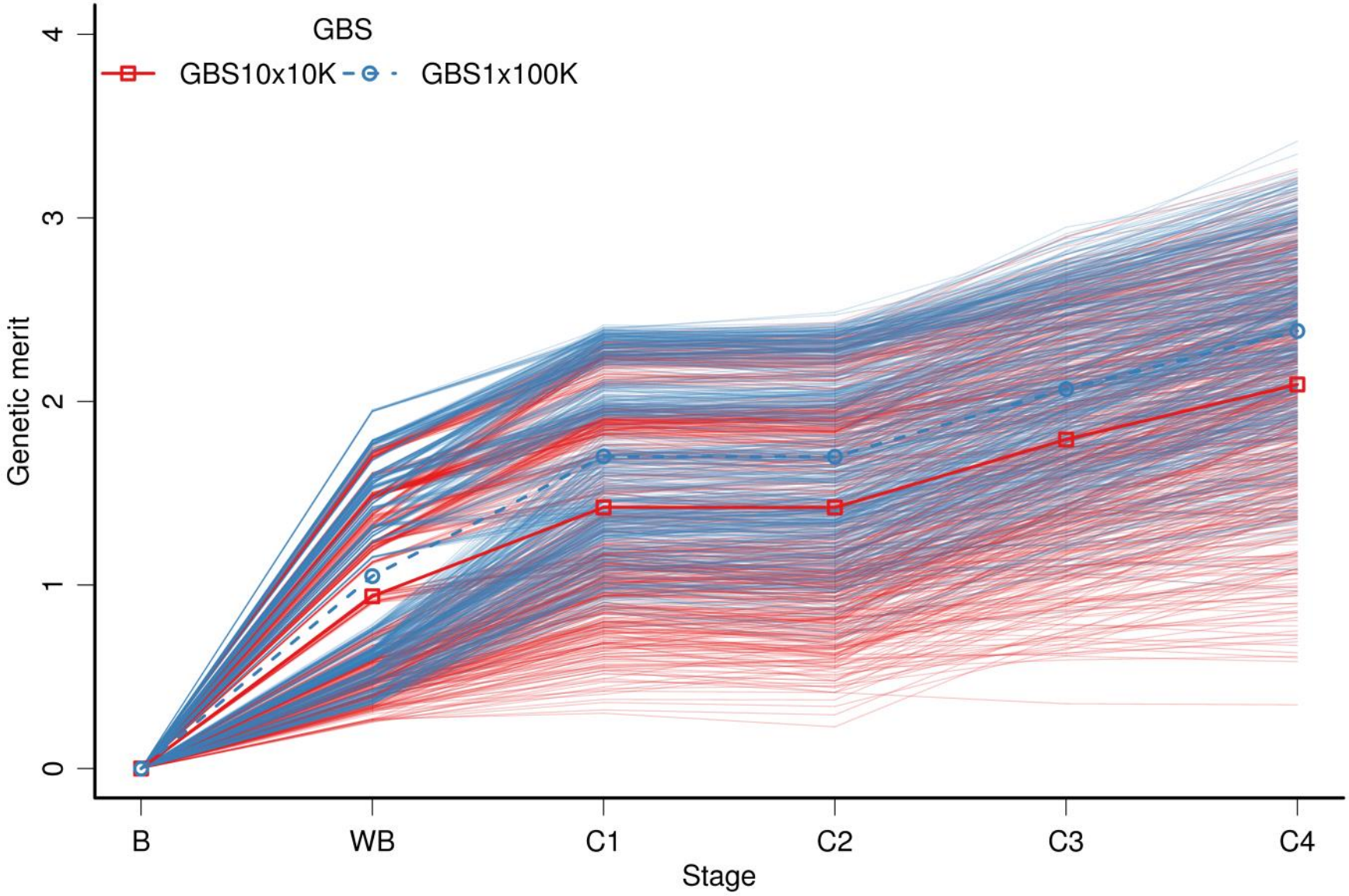
Heritability



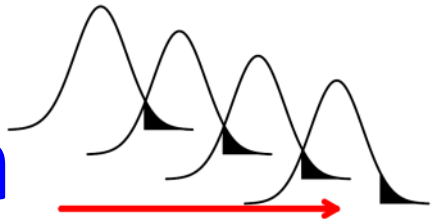


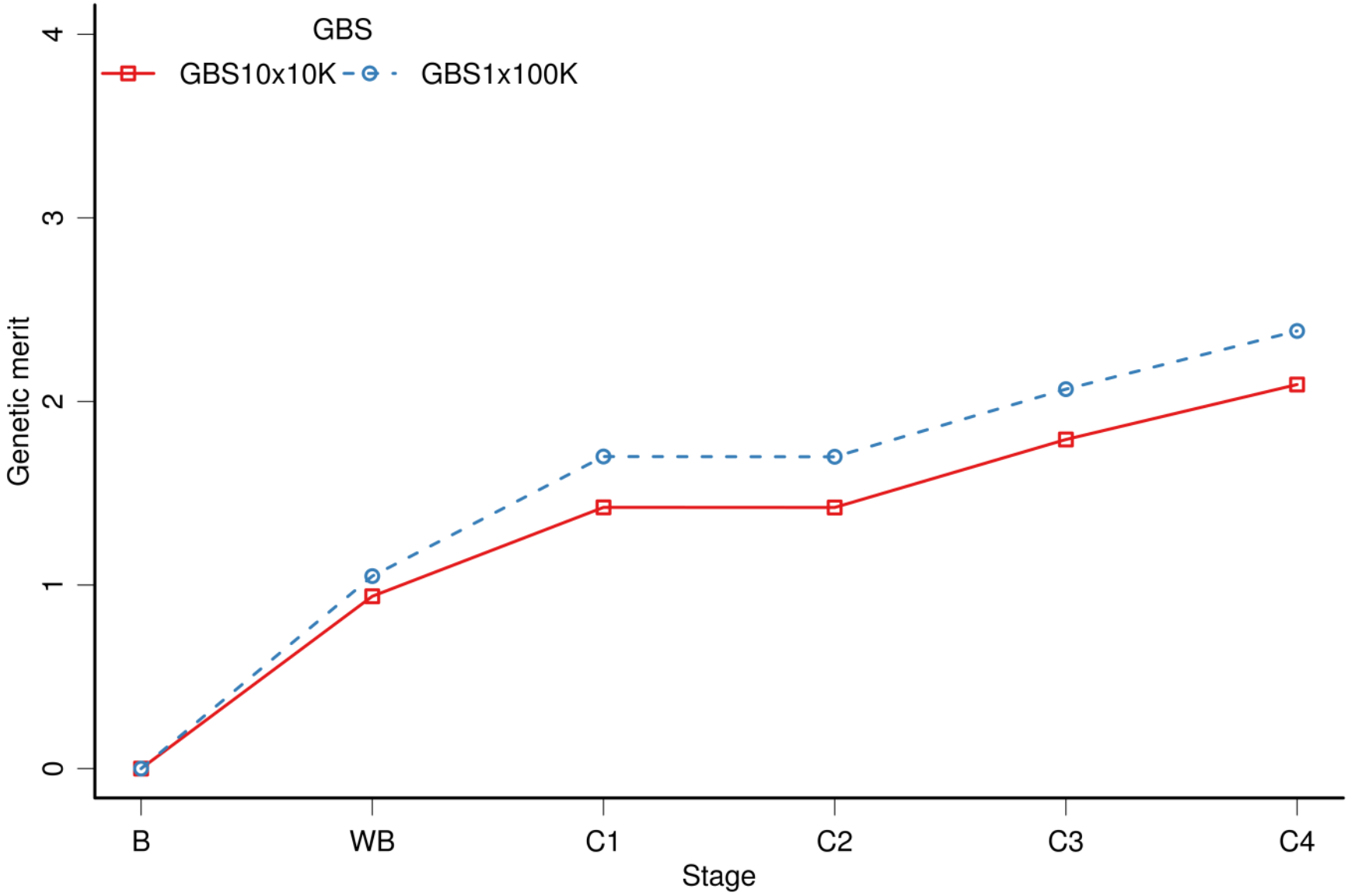
Genotyping platform



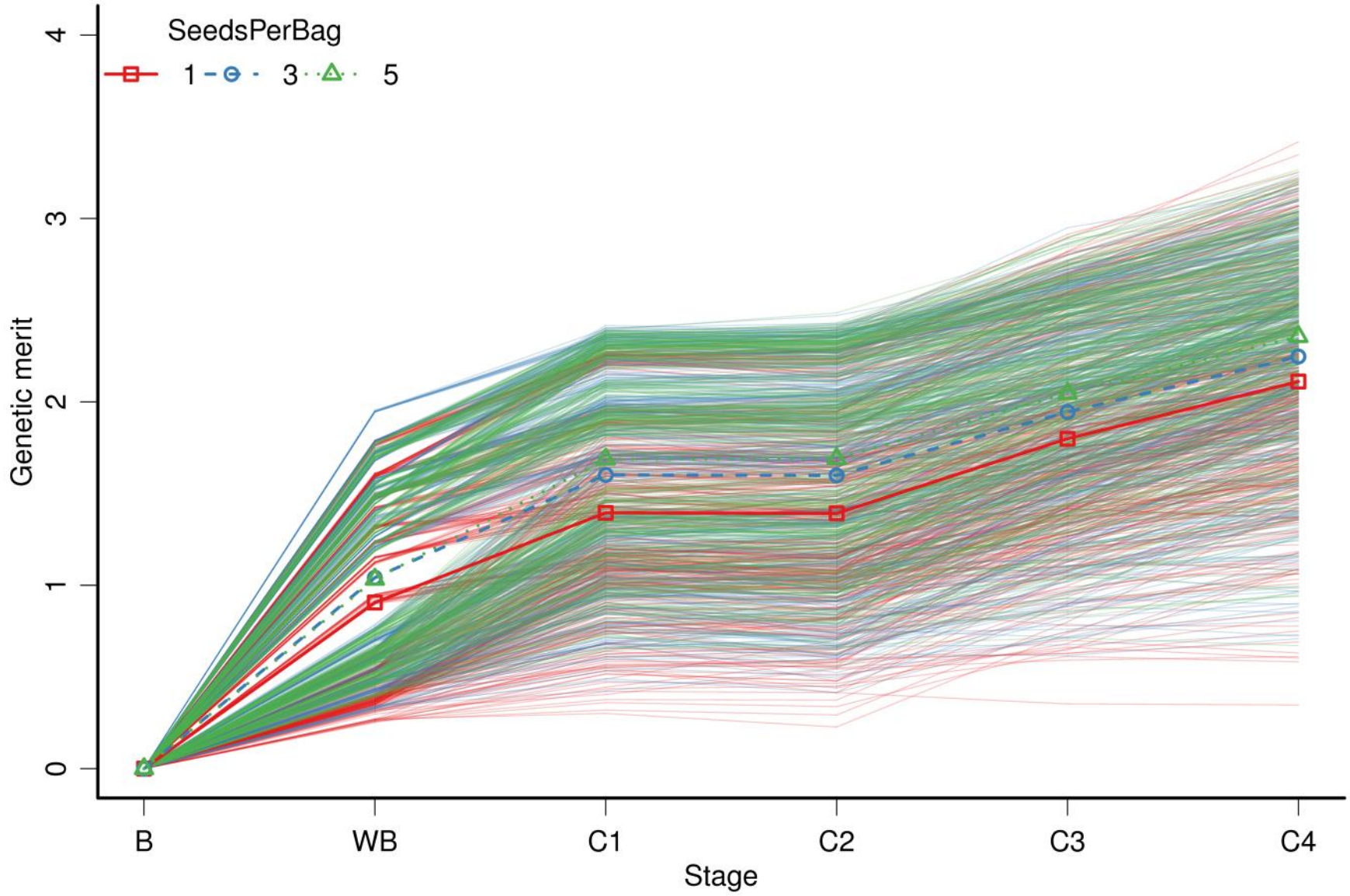


Genotyping platform

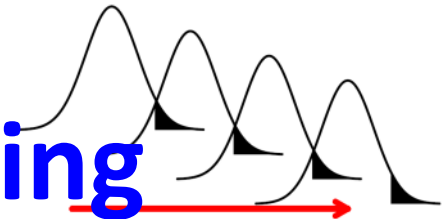


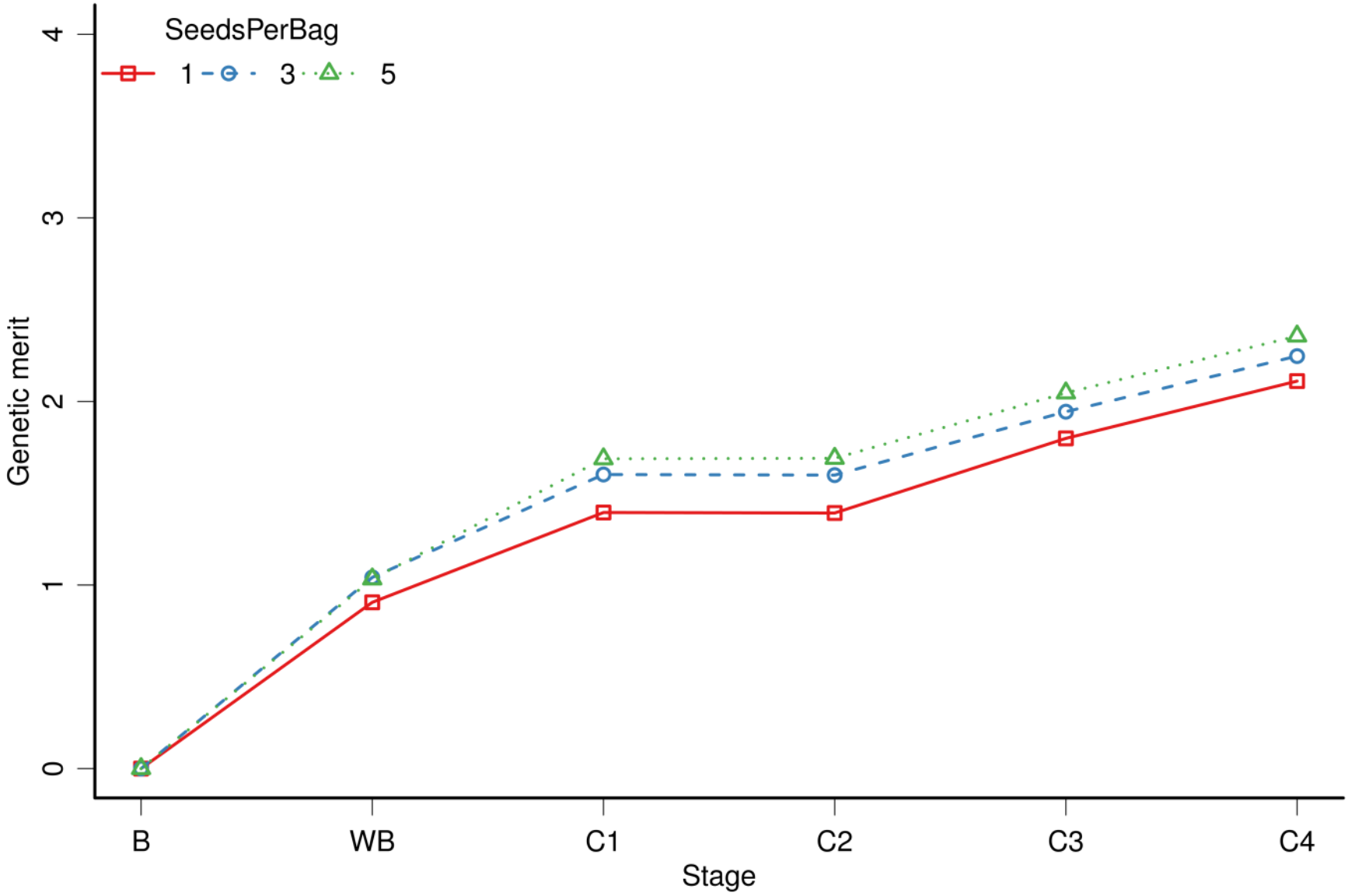


Landrace seeds in training

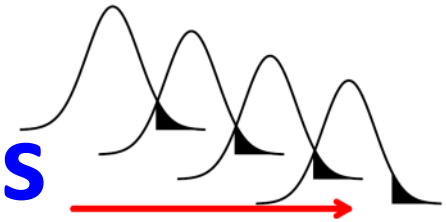


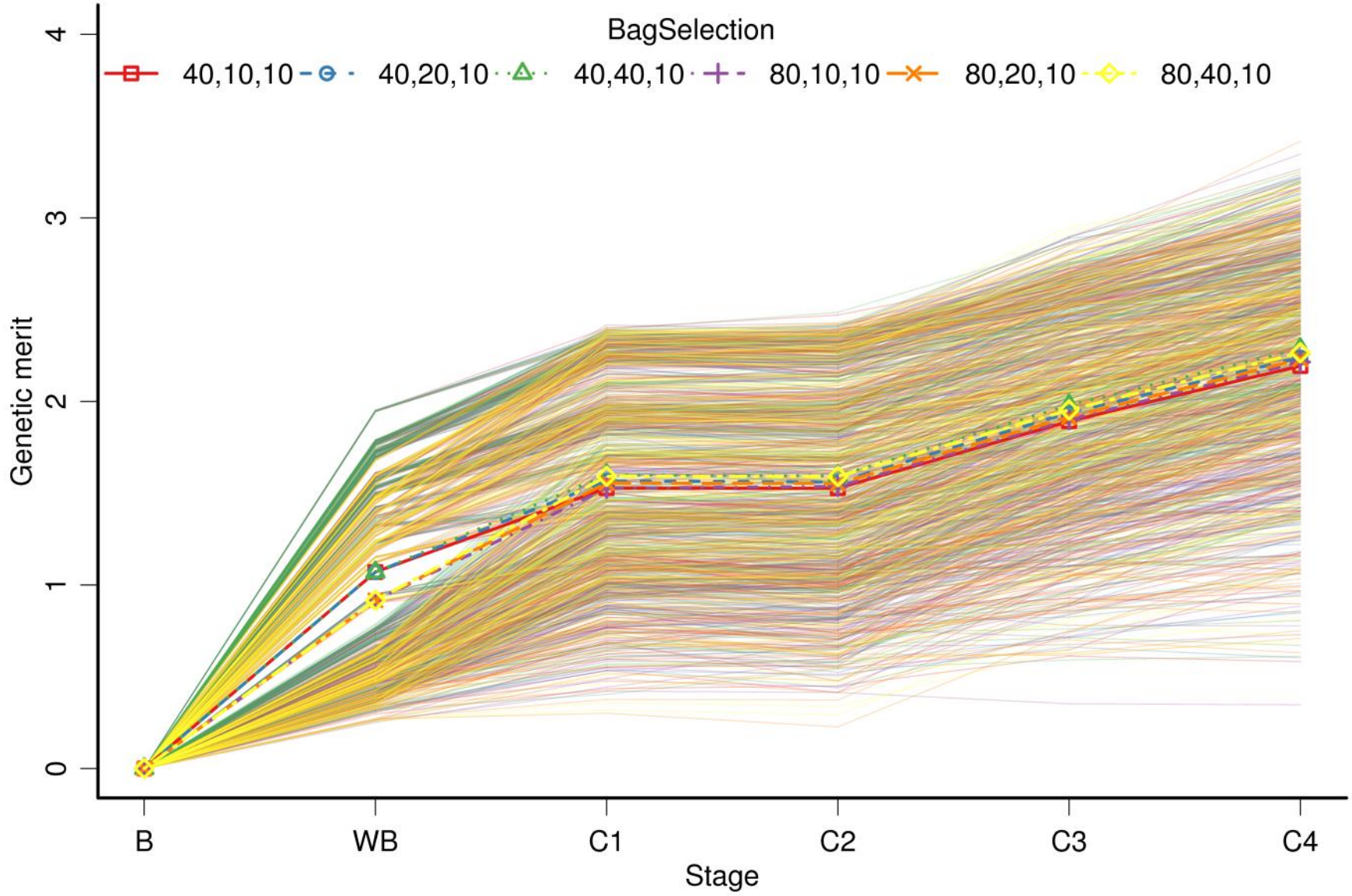
Landrace seeds in training



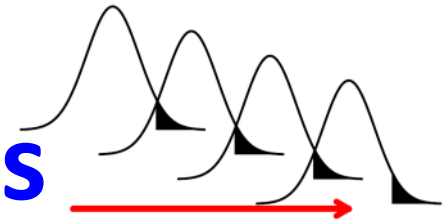


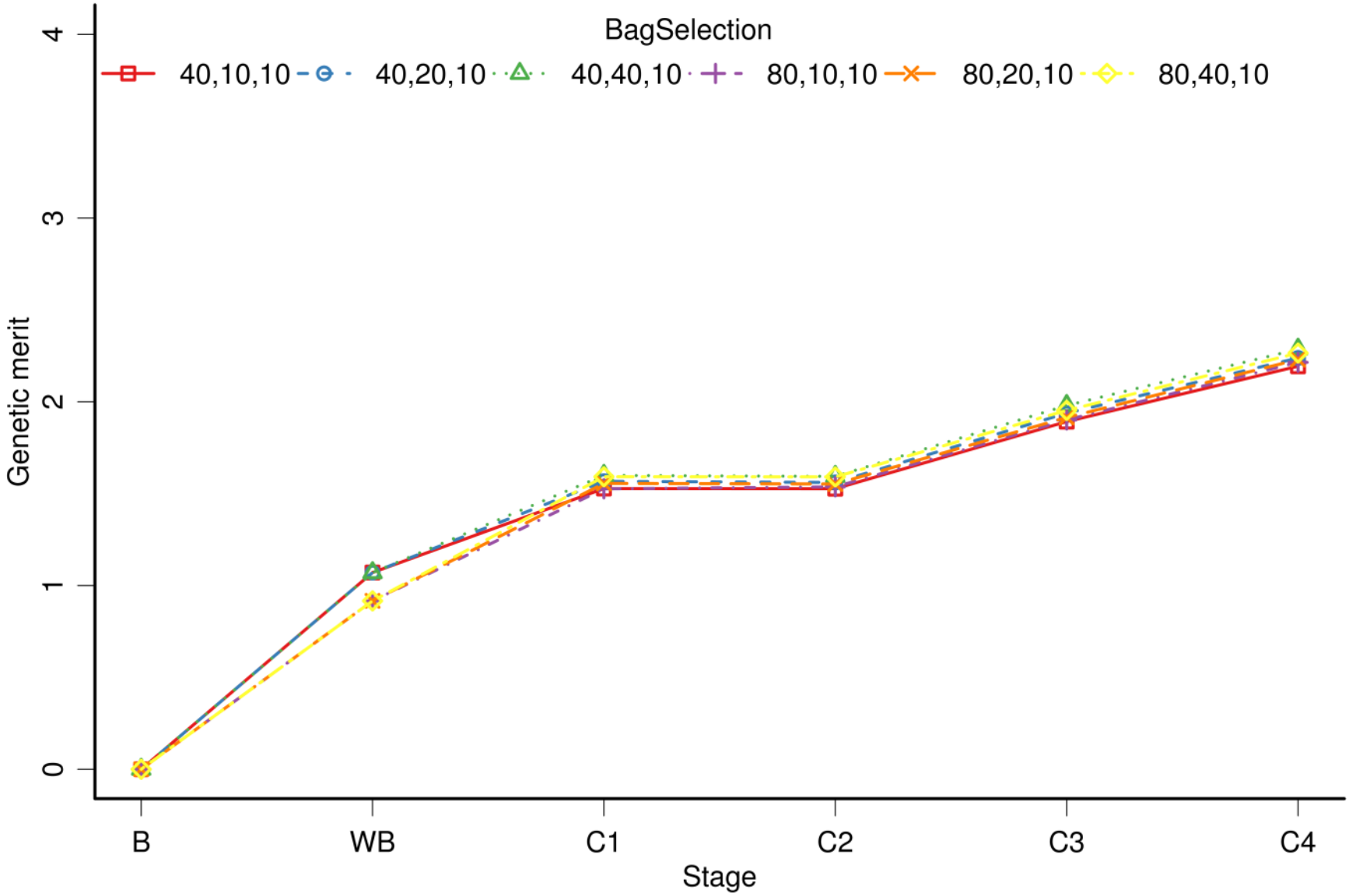
Selection of landraces



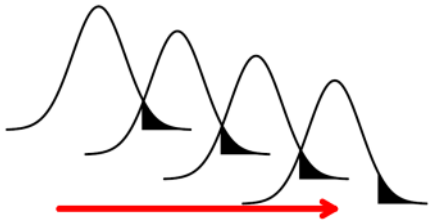


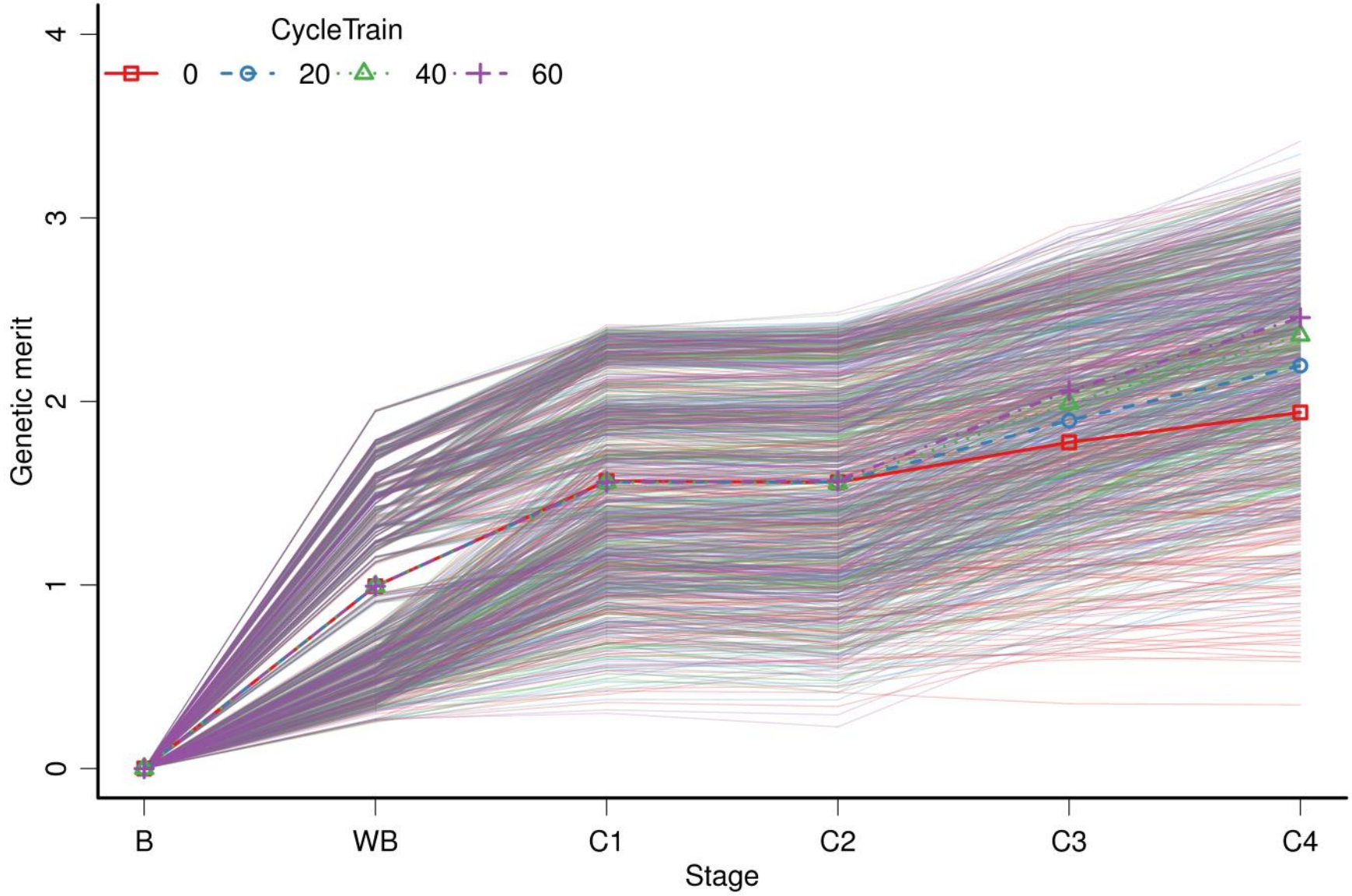
Selection of landraces



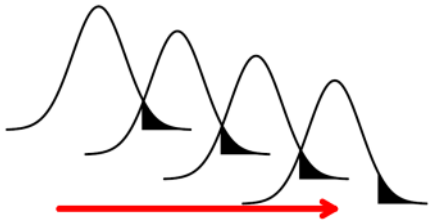


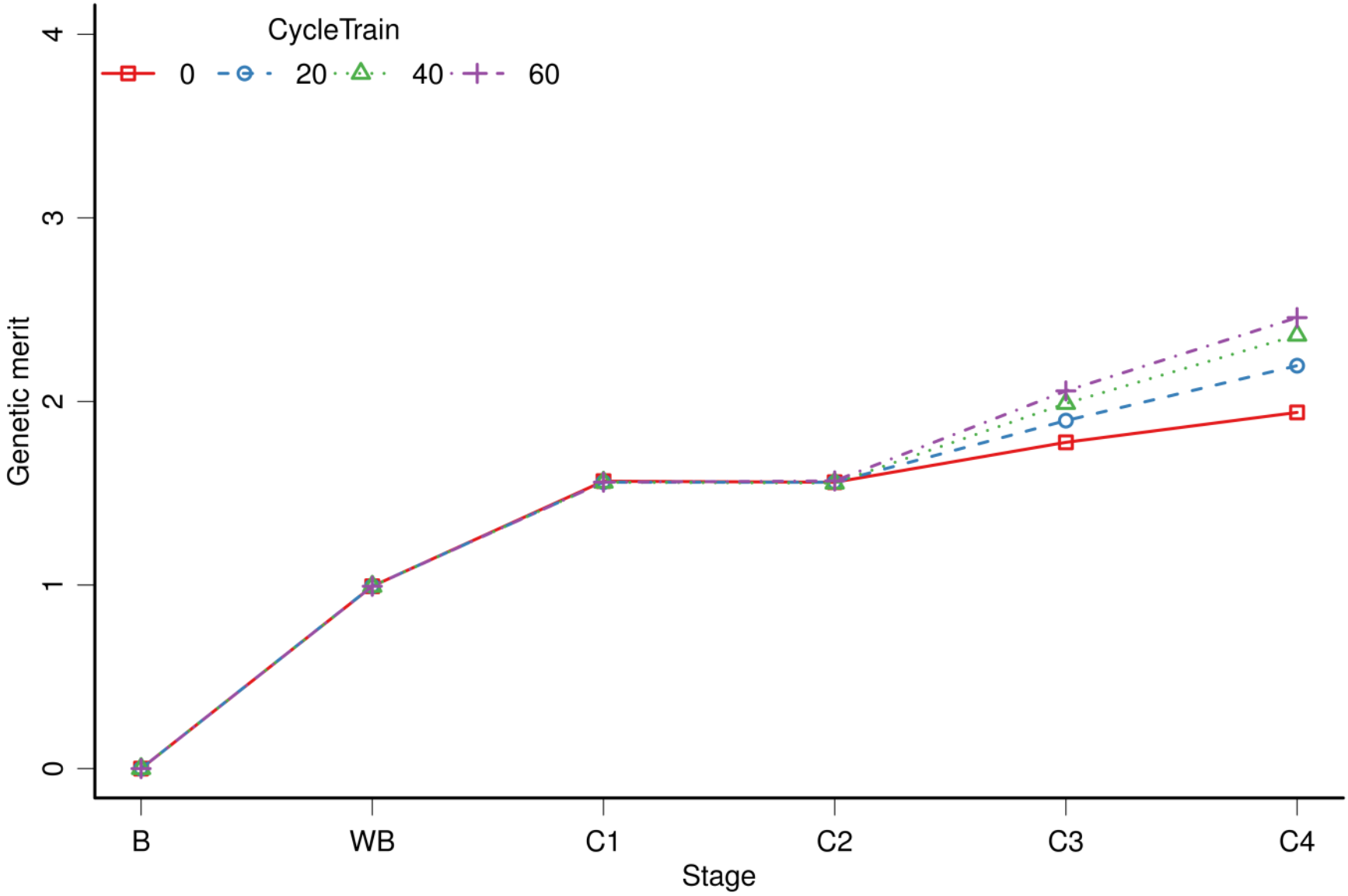
Retrain in cycles





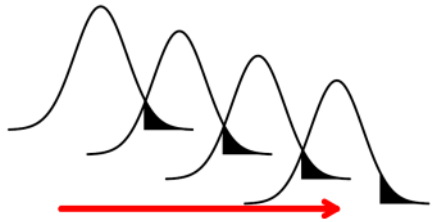
Retrain in cycles

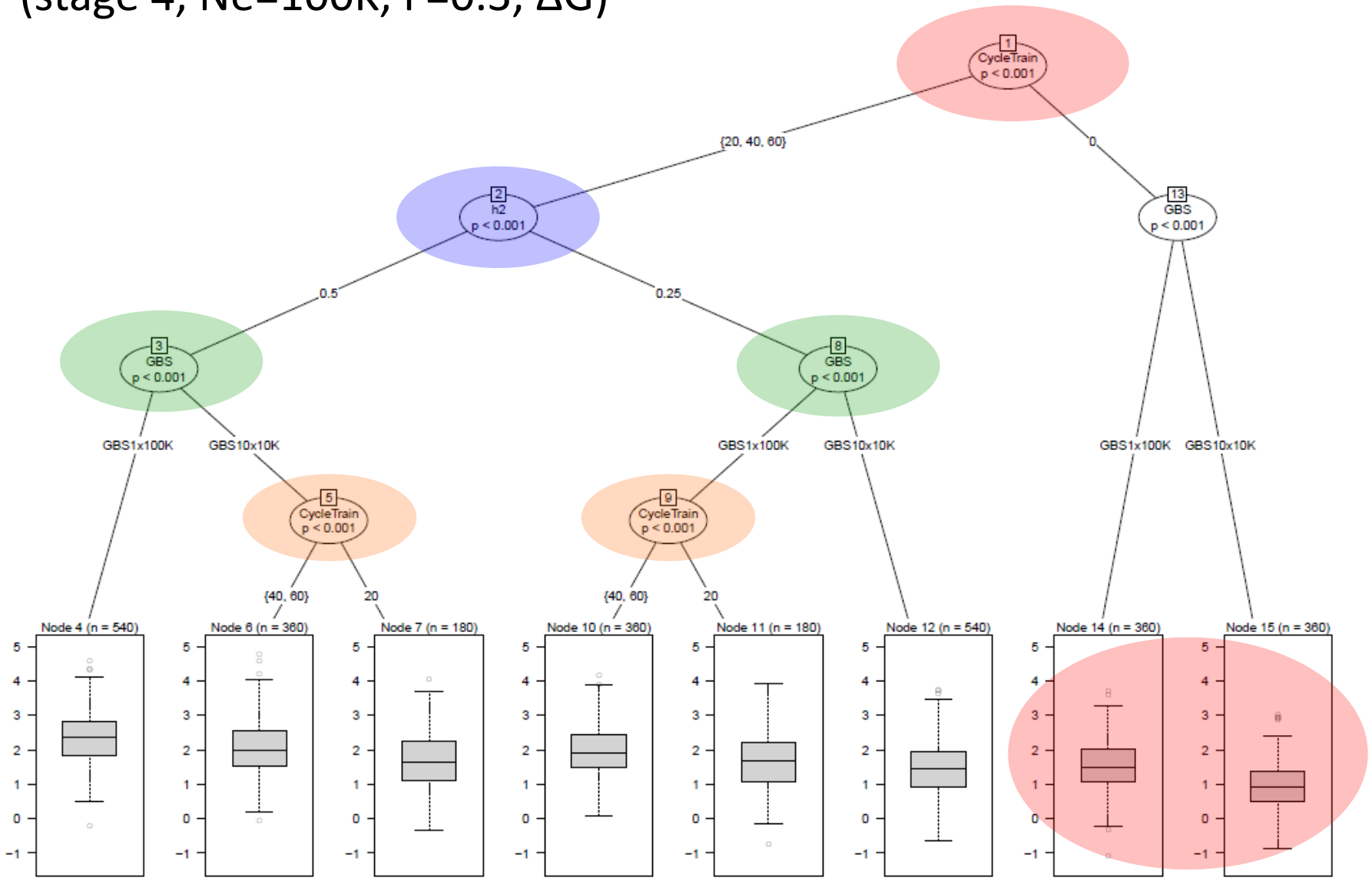




Interactions →decision trees

(stage 4, Ne=100K, F=0.3, ΔG)





Conclusions (recomendations)

- Approach
  - do not use the test-cross approach for population improvement
  - more gain with the LandraceDH approach but additional seasons → not worthwhile
- Genotyping platform
  - larger chip seems to be better due to a large N_e
- The way landraces are selected not influential
  - # LR screened and ind/LR
- Ideally if done again test more seeds per landrace
  - 3 per landrace likely enough
- Need to retrain in cycles
  - 40 phenotypes likely enough

Approach 1: Landrace

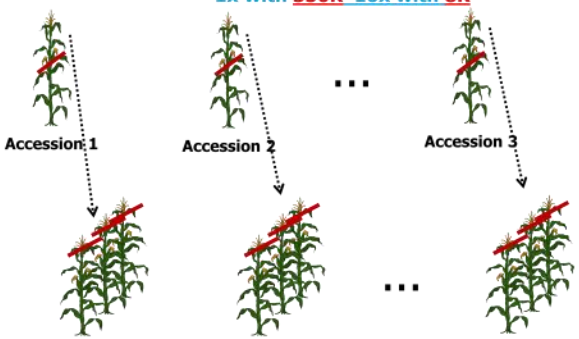
3000 bags/accessions
Ne = 1K, 10K, 100K, or 1000K?
h^2 = 0.25 or 0.50
F within accessions = 0.3, 0.6, or 0.9

Plant a random selection of 1, 3, or 5 seeds per accession for training

GENOTYPE

Genotyping platforms (GBS):
1x with 350K 10x with 8K

F0



Tester (CML heterotic group A x CML heteroric group B)

F1

Plant a random selection of 20 seeds per testcross

PHENOTYPE
(mean of row values)

TRAIN THE MODEL
(use genotypic information from generation F0 and phenotypic information from F1)
→ PREDICT GEBVs OF ACCESSIONS

Select 40 or 80 best accessions ranked by GE

Select at random 10, 20, or 40 seeds from each selected accession

GENOTYPE

PREDICT GEBV

Select the best 20 seeds from the best 20 accessions ranked by GEBV

See slide 4 for further cycles!!!

Approach 2: Landrace + double haplotype technology (LandraceDH)

3000 bags/accessions
Ne = 1K, 10K, 100K, or 1000K?
h^2 = 0.25 or 0.50
F within accessions = 0.3, 0.6, or 0.9

Plant a random selection of 1, 3, or 5 seeds per accession for training

GENOTYPE

Genotyping platforms (GBS):
1x with 350K 10x with 8K

F0

Accession 1 Accession 2 ... Accession 3

Tester (CML heterotic group A x CML heteroric group B)

F1

Plant a random selection of 20 seeds per testcross

PHENOTYPE
(mean of row values)

TRAIN THE MODEL
(use genotypic information from generation F0 and phenotypic information from F1)

→ PREDICT GEBVs OF ACCESSIONS

Select 40 or 80 best accessions ranked by GEBV

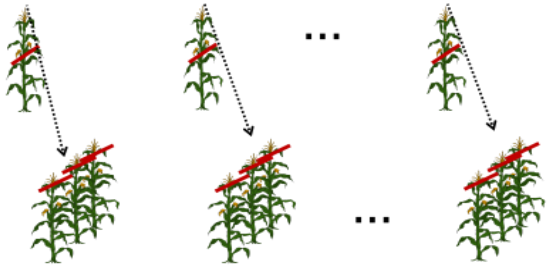
Select at random 10, 20, and 40 or seeds from each selected accession

GENOTYPE

PREDICT GEBV

Select the best 20 seeds from the best 20 accessions ranked by GEBV

Plant seeds and cross to inducer



Inducer

Random 5 seeds from each plant and treat with colcichine to obtain double haplotypes

GENOTYPE

PREDICT GEBV

Select the best double haplotype from each plant ranked by GEBV

See slide 4 for further cycles!!!

Approach 3: Testcross

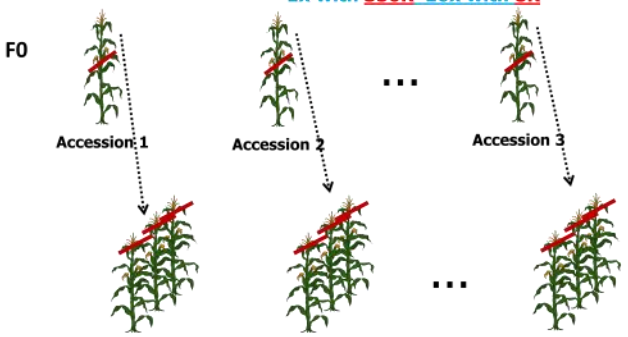
Select among testcrosses instead of accessions

3000 bags/accessions
Ne = 1K, 10K, 100K, or 1000K?
h^2 = 0.25 or 0.50
F within accessions = 0.3, 0.6, or 0.9

Plant a random selection of 1, 3, or 5 seeds per accession for training

GENOTYPE

Genotyping platforms (GBS):
1x with 350K 10x with 8K



Tester (CML heterotic group A x CML heteroric group B)

Plant a random selection of 20 seeds per testcross

PHENOTYPE
(mean of row values)

TRAIN THE MODEL
(use genotypic information from generation F0 and phenotypic information from F1)

→ **PREDICT GEBVs OF ACCESSIONS**

Select 40 or 80 best accessions ranked by GEBV

Select at random 10, 20, and 40 or seeds from each selected accession

GENOTYPE

PREDICT GEBV

Select the best 20 seeds from the best 20 accessions ranked by GEBV

See slide 4 for further cycles!!!

Developing synthetic population