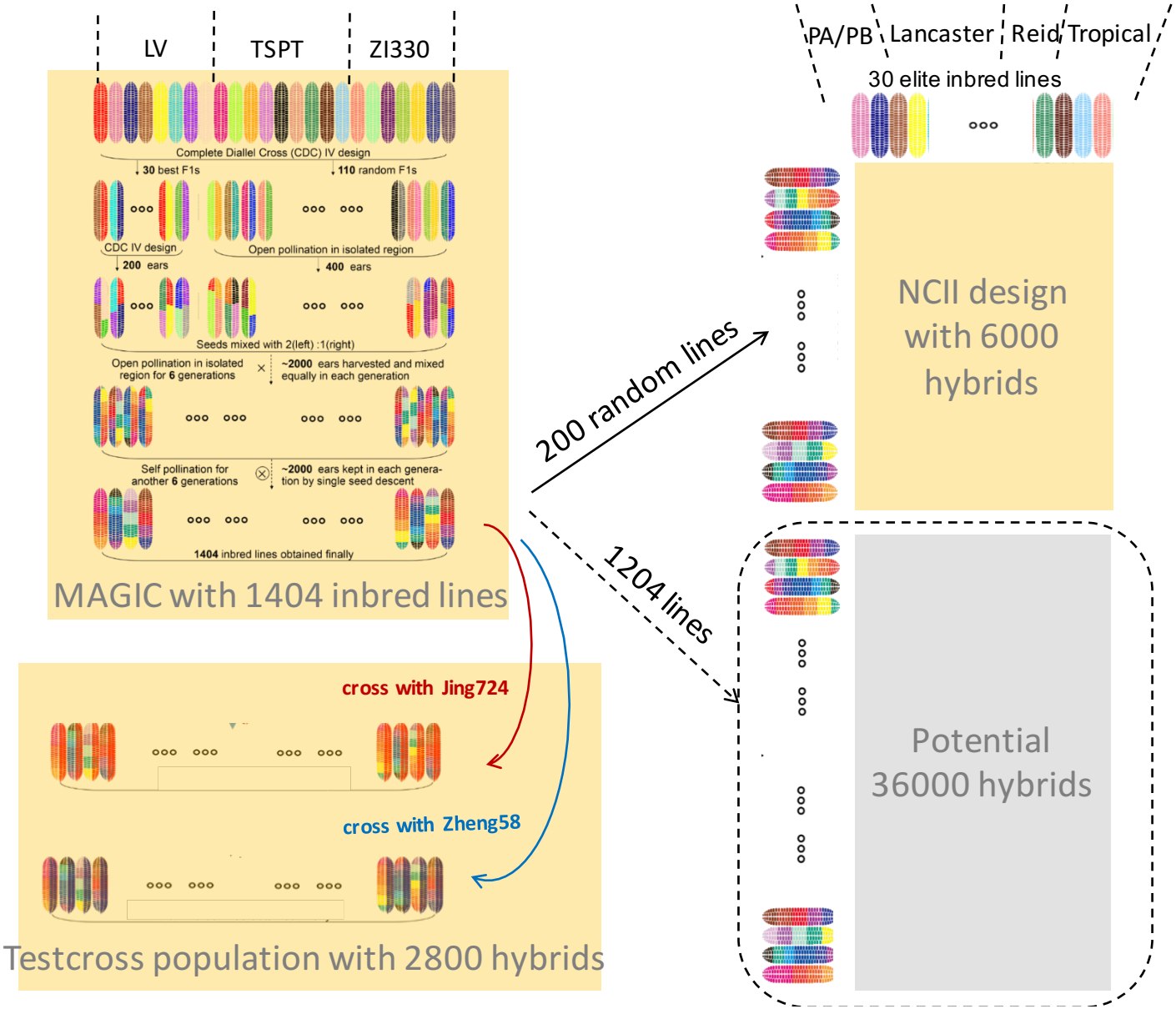
eMaize proposal

1.7

Outline

- Project overview
- Data summary
- Feature selection
- Model
- Environment

Project design



Project overview

- Train on 6210 known hybrid and predict 1204*2 unknown hybrid.
- Discover trait-related genes.
- Predict environment influence.

Data summary

- Now:
- 6210 hybrids:
  - SNP data. New preprocess. 5.88M
  - Traits: DTT, PH, EW
- 30+1404 parents:
  - SNP data: 5.88M
  - Traits: DTT, PH, EW
- Environment data:
  - 5 locations' hybrid and parents traits

- Future:
- All 30*1404 hybrid:
  - SNP data
  - Whole genome data
- Environment condition data
- RNA-seq:
  - Parents
  - hybrid

SNP data

- Recreate SNP datasets: 5.88M， including all hybrids and parents.
- Preprocess pipeline

1. SNP 数据预处理

MAGIC 群体重测序共鉴定到约 5000 万 SNP，经过逐步过滤在 NCII 群体 6210 个杂交种，共获得约 558 万个高质量 SNP。过滤步骤如下：

（1）在 NCII 群体的 237 个亲本中 MAF>2%（约 1800 万 SNP）；

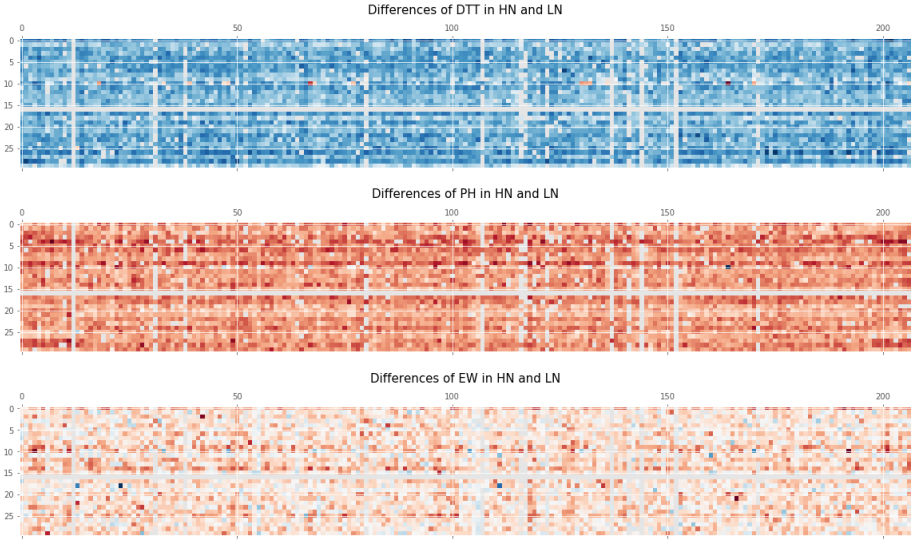
（2）Imputation 前在 6210 杂交种中 missing<10%（约 1500 万 SNP）；

（3）在 6210 杂交种中，minor genotype count>30（约 558 万 SNP）。

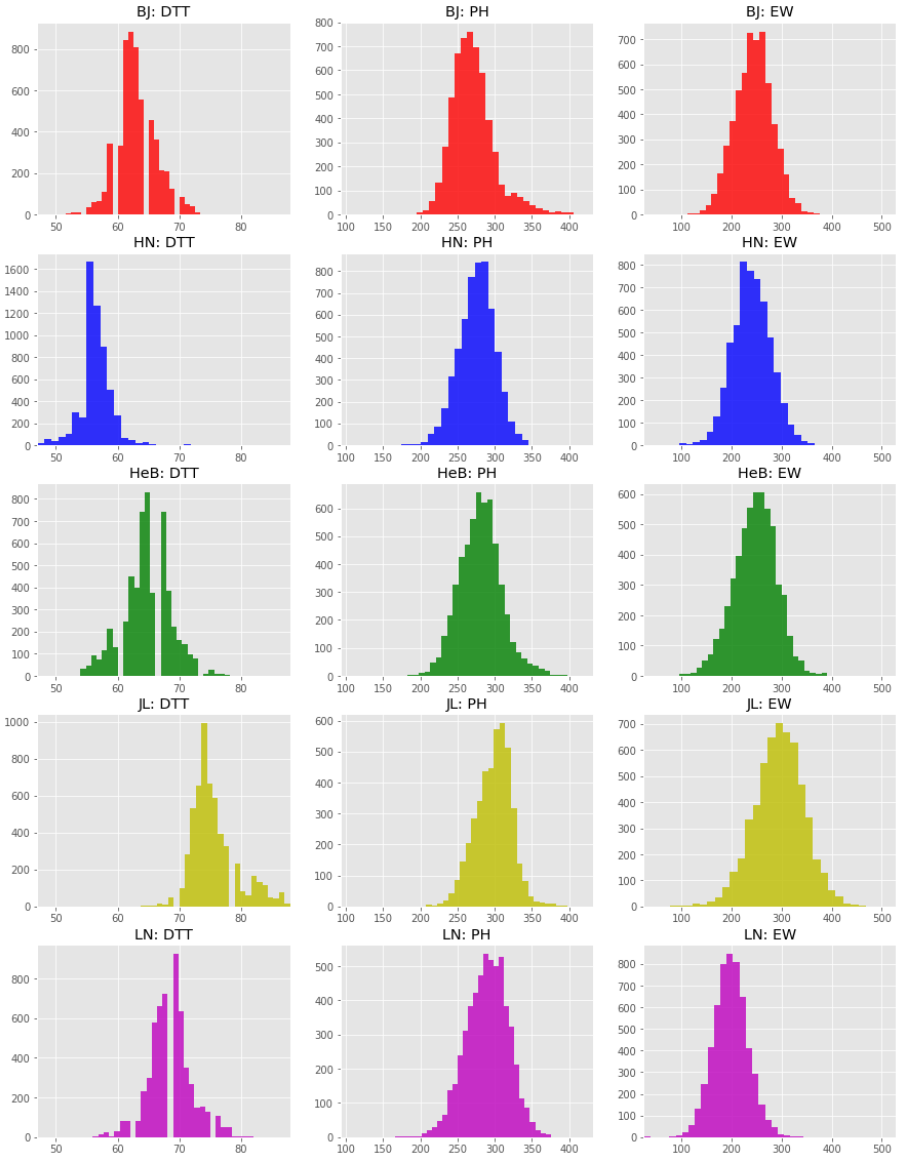
- (the preliminary test’s 1,900,000 SNP is generated only from step 1.)

Environment data

Parent EW in five location



Hybrid traits difference in 2 loc



Hybrid traits distribution in 5 loc

Feature selection

- Already use MAF to reduce dimension

- Use LD decay curve to divide region, select SNP from each region.

- Use clustering method. (If complexity is tolerable)
  - DACE(LSH + DP-means), unfixed cluster numbers.
  - Consider long range

- GWAS
  - Common methods: consider SNP separately.
  - Use Mixed-Ridge and Lasso do it batch by batch

Add new features

- eQTL:
  - Use whole genome(SNP) and RNA-seq.
- Parents SNP:
  - Help feature selection, add to M-R model as similarity measurement.
- Unknown SNP:
  - Unsupervised learning. Help calculate LD and clustering.
- Whole genome data
  - Structure variance
- Regulation network
- Gene annotation

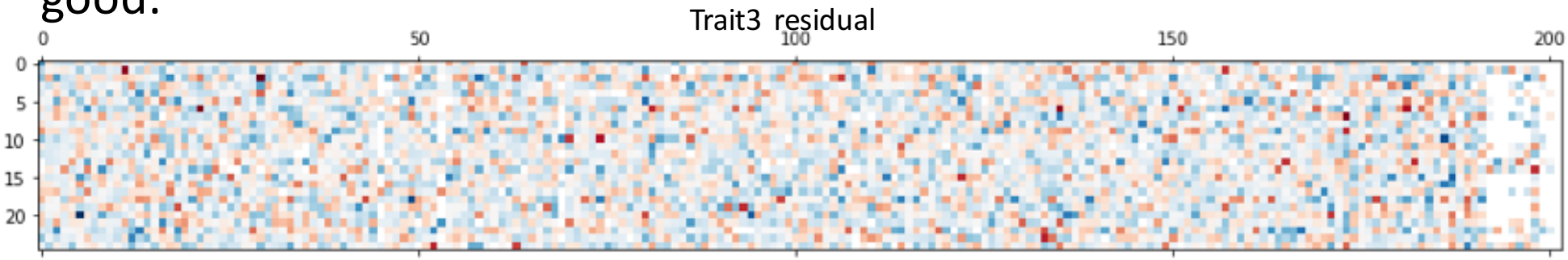
Model

- Whole-Genome Regression (WGR) model is proved to be most useful.
- Most common model:
  - BLUP, Bayesian, LMM
  - Use SNP data
  - Predict traits and do GWAS

- Non-linear model:
- Feature is complex: genome, RNA-seq…

Non-linear model

- Find heterosis
- Similar ideas like Mixed-Ridge:
- calculate residual of hybrids (SCA). Use features to predict residuals good.



- When predict whole trait, add GCA part.
- Already know parents trait, prediction will be easier. (but can't help find heterosis)

Non-linear model

- If the model can use features to predict residual good enough:
  - It has ability to find heterosis
  - It can find some important feature.

- Expert model is also a non-linear model
  - Combine basic models.
  - Use different weights for different samples

Multiple-Trait combined prediction

- Test correlation of different traits. (there are 20 total traits)
  - Multiple traits can provide more information.
  - A feature good for multiple traits may be important

Environment

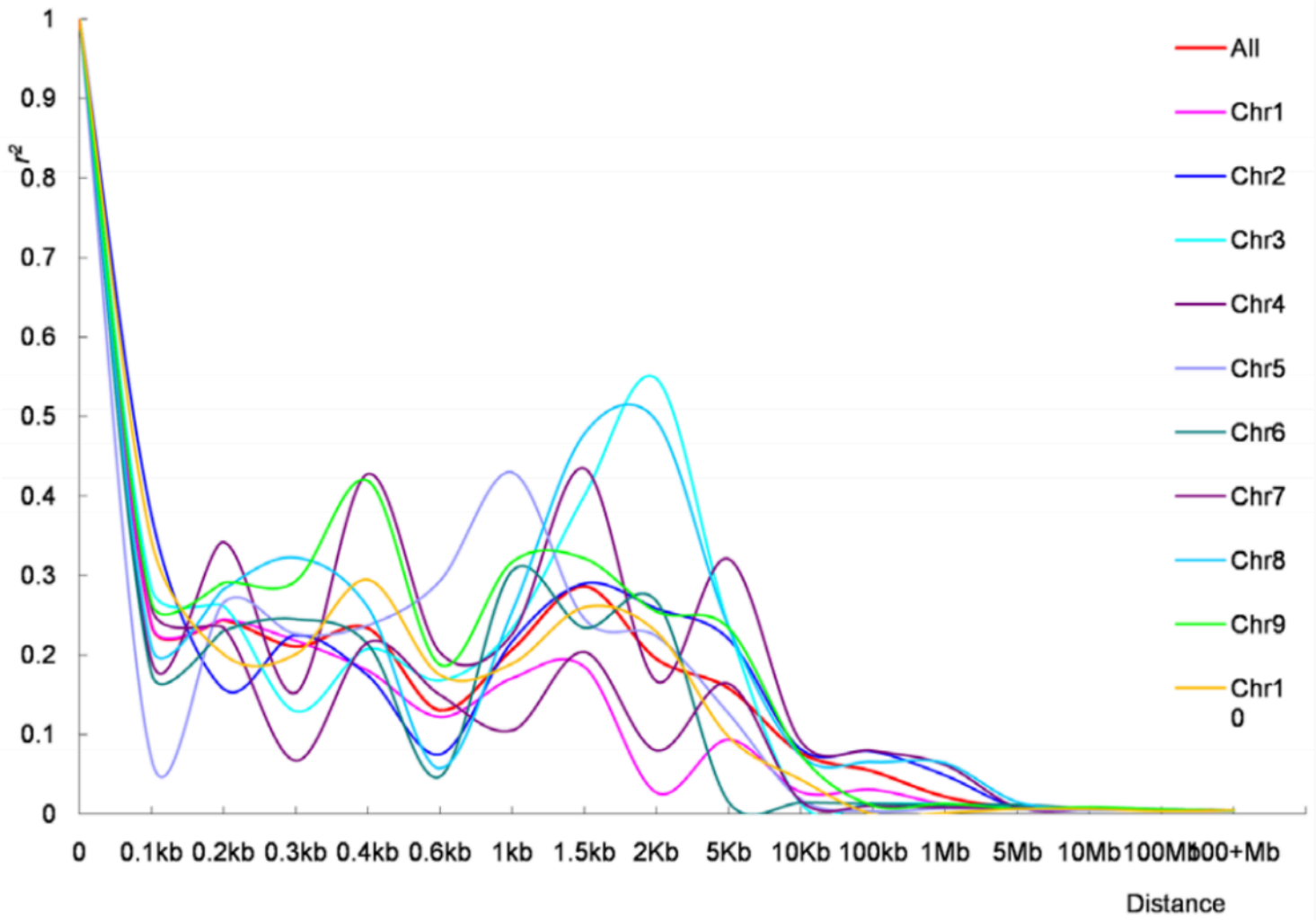
- Predict phenotype variance, find robust hybrids in different locations.
- Only have 5 data points.
- Consider by samples. Each sample from 6210 hybrids have 5 locations traits.
- Use tensor decomposition for multiple-location traits variance.

THANK YOU

Supplementary

LD

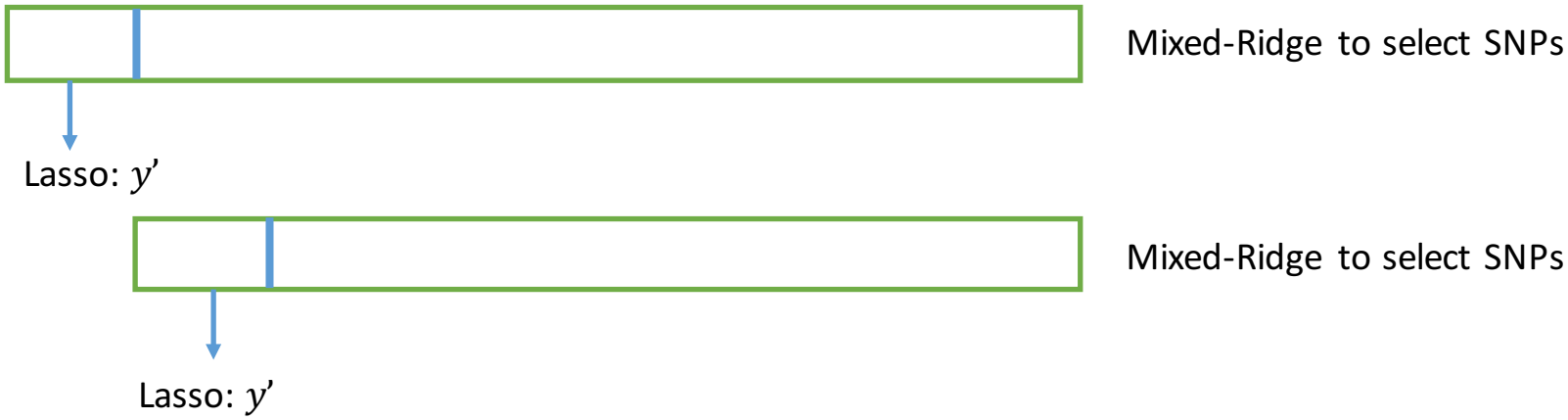
- Use r^2 < 0.1


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
r^2
All
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
0
0.1kb
0.2kb
0.3kb
0.4kb
0.6kb
1kb
1.5kb
2Kb
5Kb
10Kb
100kb
1Mb
5Mb
10Mb
100Mb
100+Mb
Distance

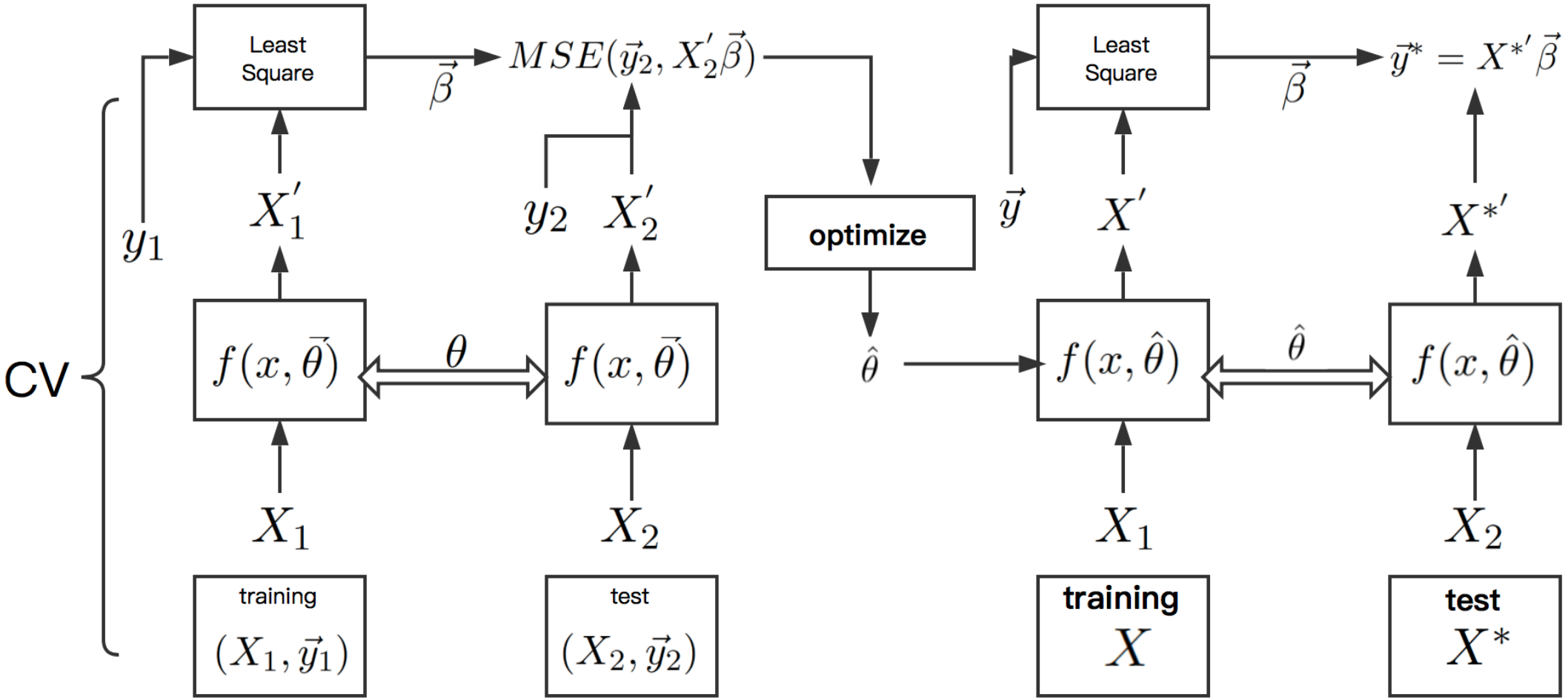
GWAS

Mixed-Ridge used as feature selection

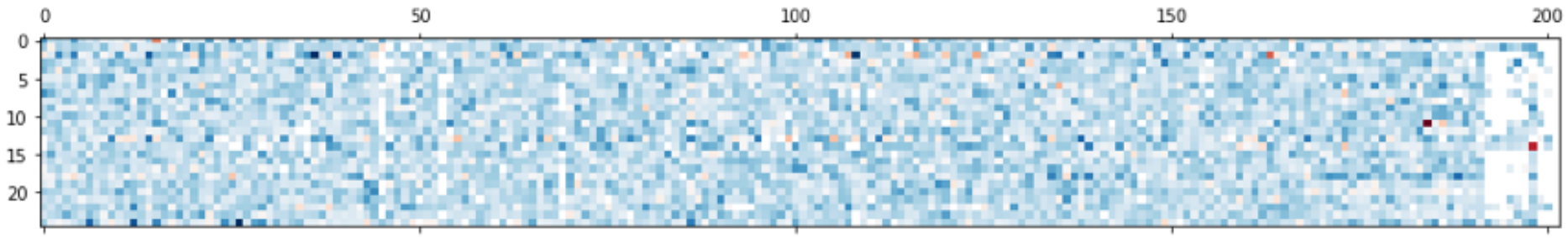
- Common GWAS feature selection method consider SNP seperately
- We can use Mixed-Ridge to consider feature combination
- May find causal by eliminating genetic similarity

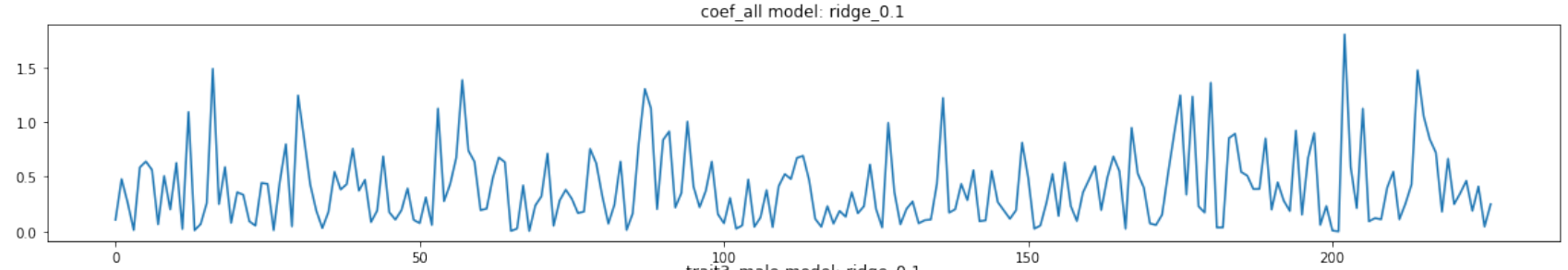


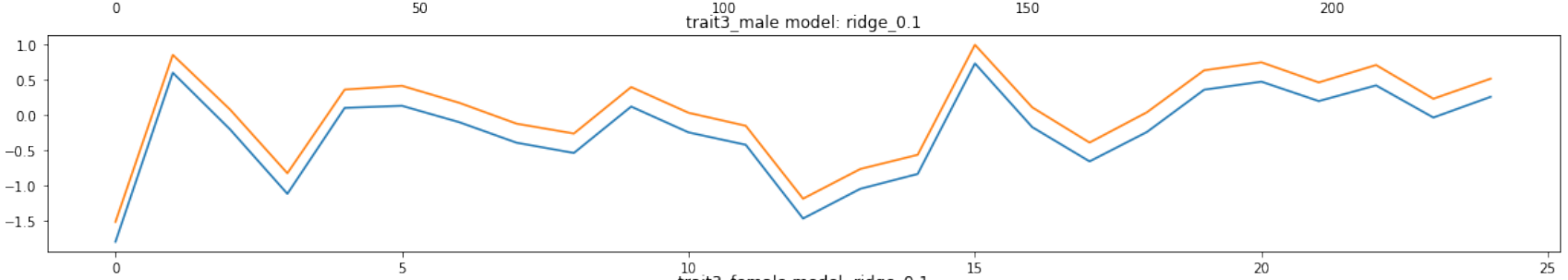
- Iteration until MSE doesn't change

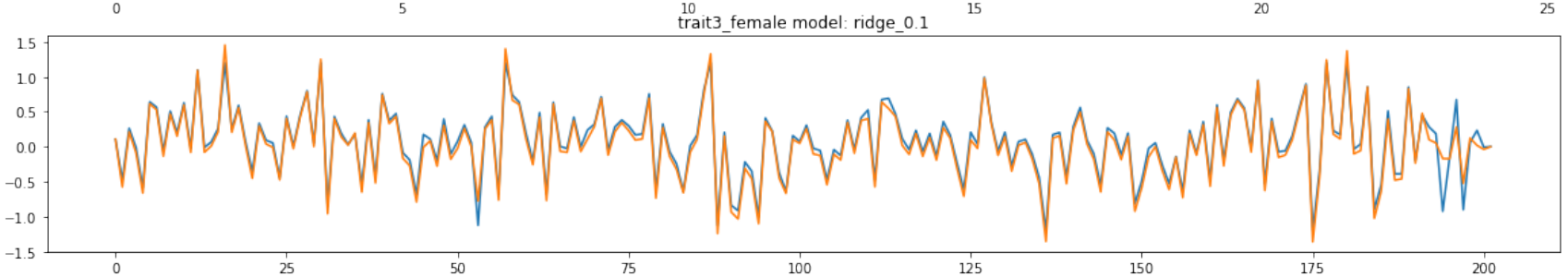


Principles: Supplementary


0
50
100
150
200
0
5
10
15
20


coef_all model: ridge_0.1
1.5
1.0
0.5
0.0
0
50
100
150
200


trait3_male model: ridge_0.1
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
5
10
15
20
25


trait3_female model: ridge_0.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
25
50
75
100
125
150
175
200