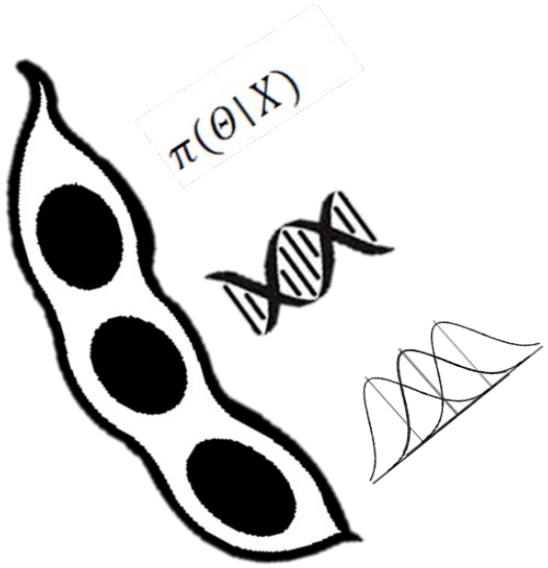
Learning from data

GxE analysis on multiple populations



Alencar Xavier
Soybean Breeding and Statistical Genetics, PhD
xaviera@purdue.edu

The game of numbers

“ ***Intelligent decision-making*** *relies on our capability of extracting* ***useful information*** *from* ***data*** *to achieve our goals* ***more eff ciently****. ”*

Why bother with multiple populations?

- Distinct germplasm sources
- Breeding objectives
- Maturity groups
- Multiple crosses
- Next-generation populations (NGPs)

G x

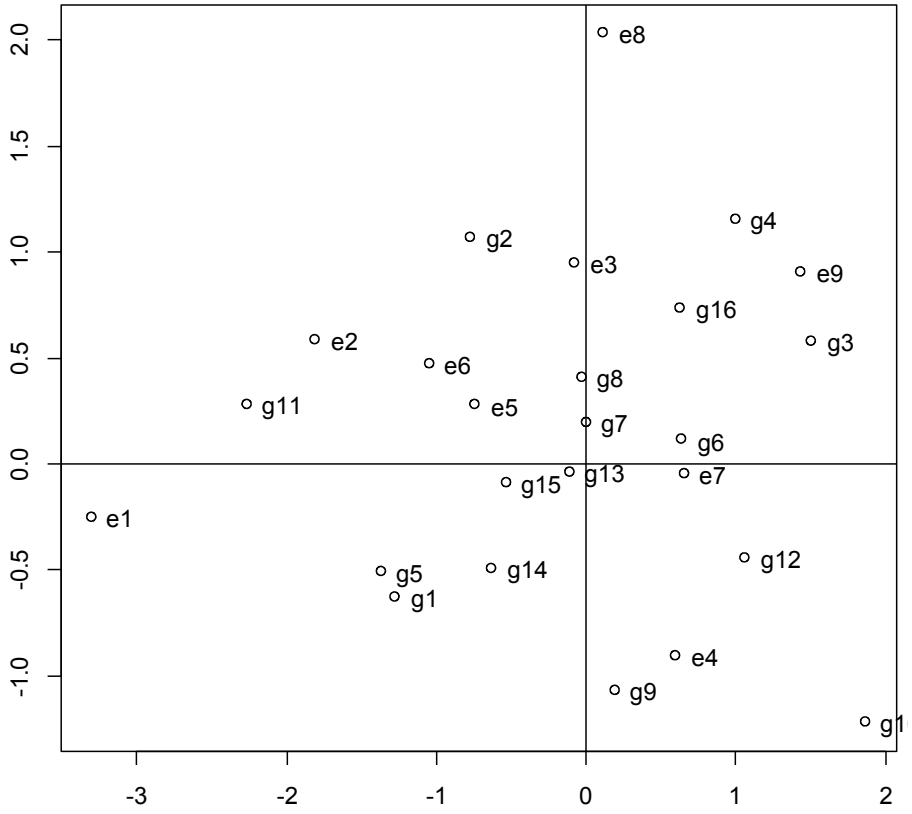
E

- **GxE analysis**
  - Strategies
    - Exploit or ignore?
    - “Which wins where” strategies
  - ANOVA / Finlay-Wilkinson / AMMI

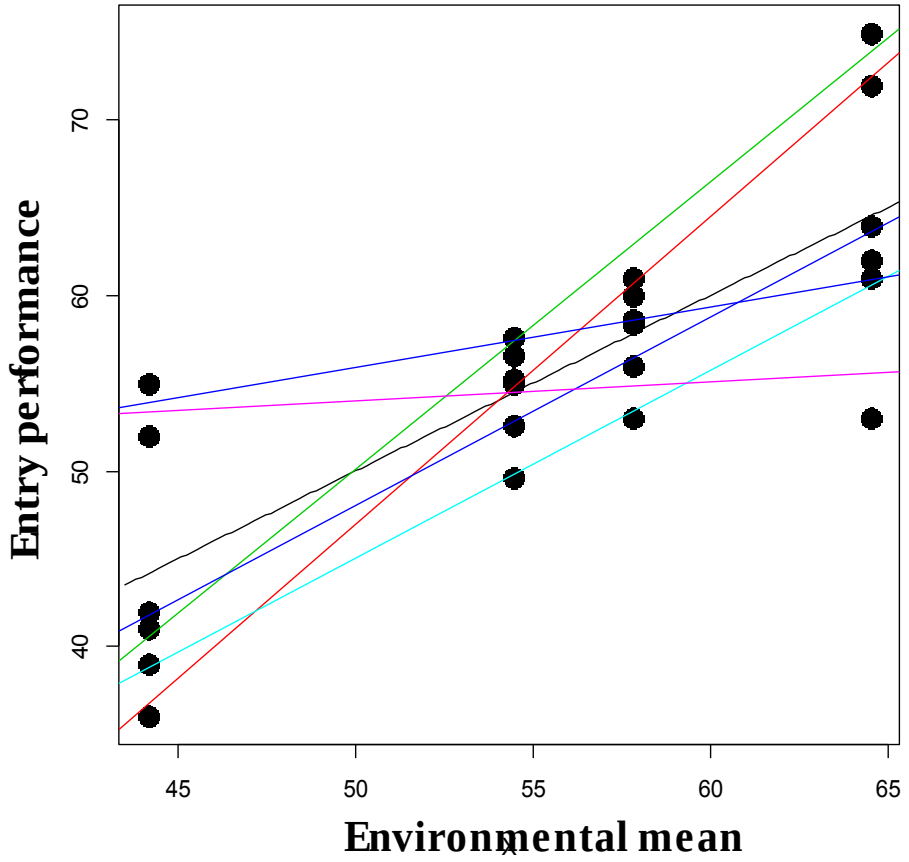
- **Limitations of GxE analyses**
  - Restricted to few lines – *final stages of breeding*
  - Usually requires lots of replicates

G x E methods

Biplot methods



Regression methods



Source: M. Mohsen

Multiple Populations

- **Limitations of analyses on multiple populations**
  - Penetrance/Epistasis
  - Linkage phase
  - Marker-QTL segregation
  - GxE

The linkage phase issue

Pop1: Coupling

M ---- Q

m ----

Pop2: Repulsion

M ----

m q---

Allele effect

Allele effect

- Pop1 ≅ 5
- Pop1 = 5
- Pop2 = -5
- Pop2 = -5
- Both ≅ 0
- Both 0

Marker-QTL segregation issue

1) Only SNP is segregating

M ---- q

m ----

2) Only QTL is segregating

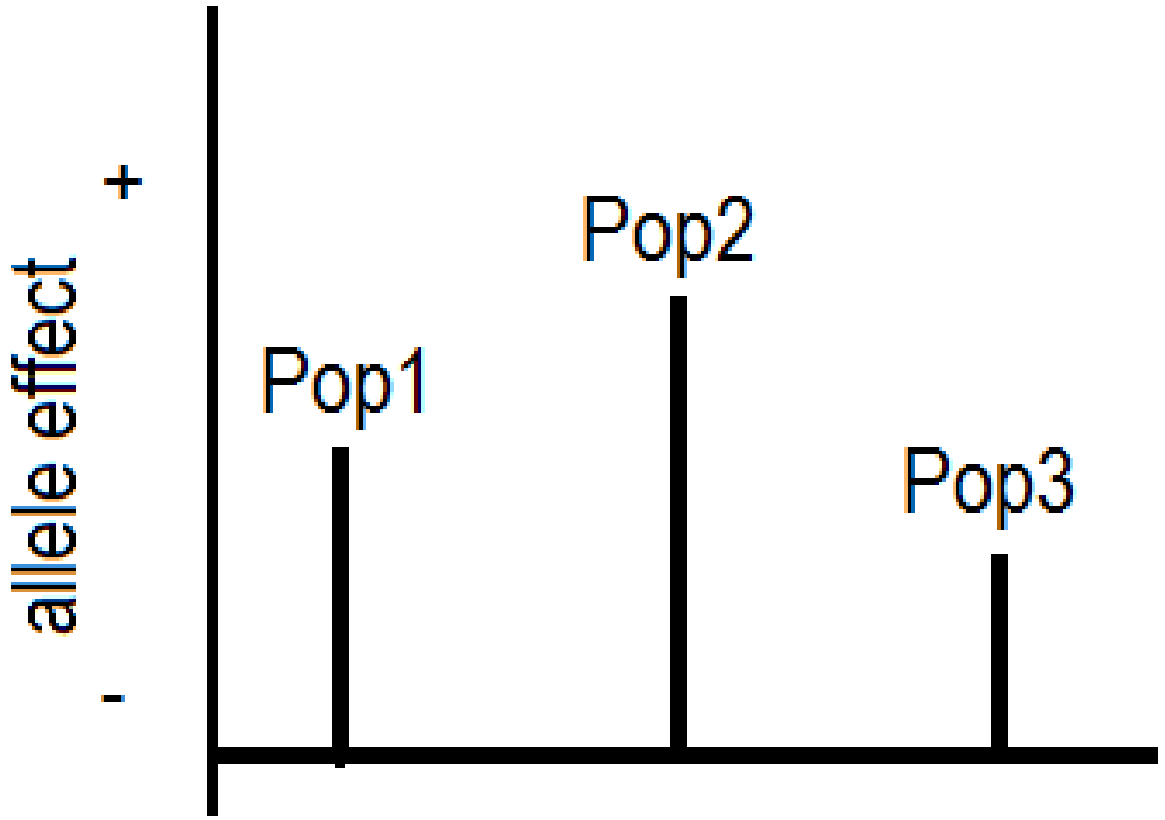
M ----

Mq---

Allele effect

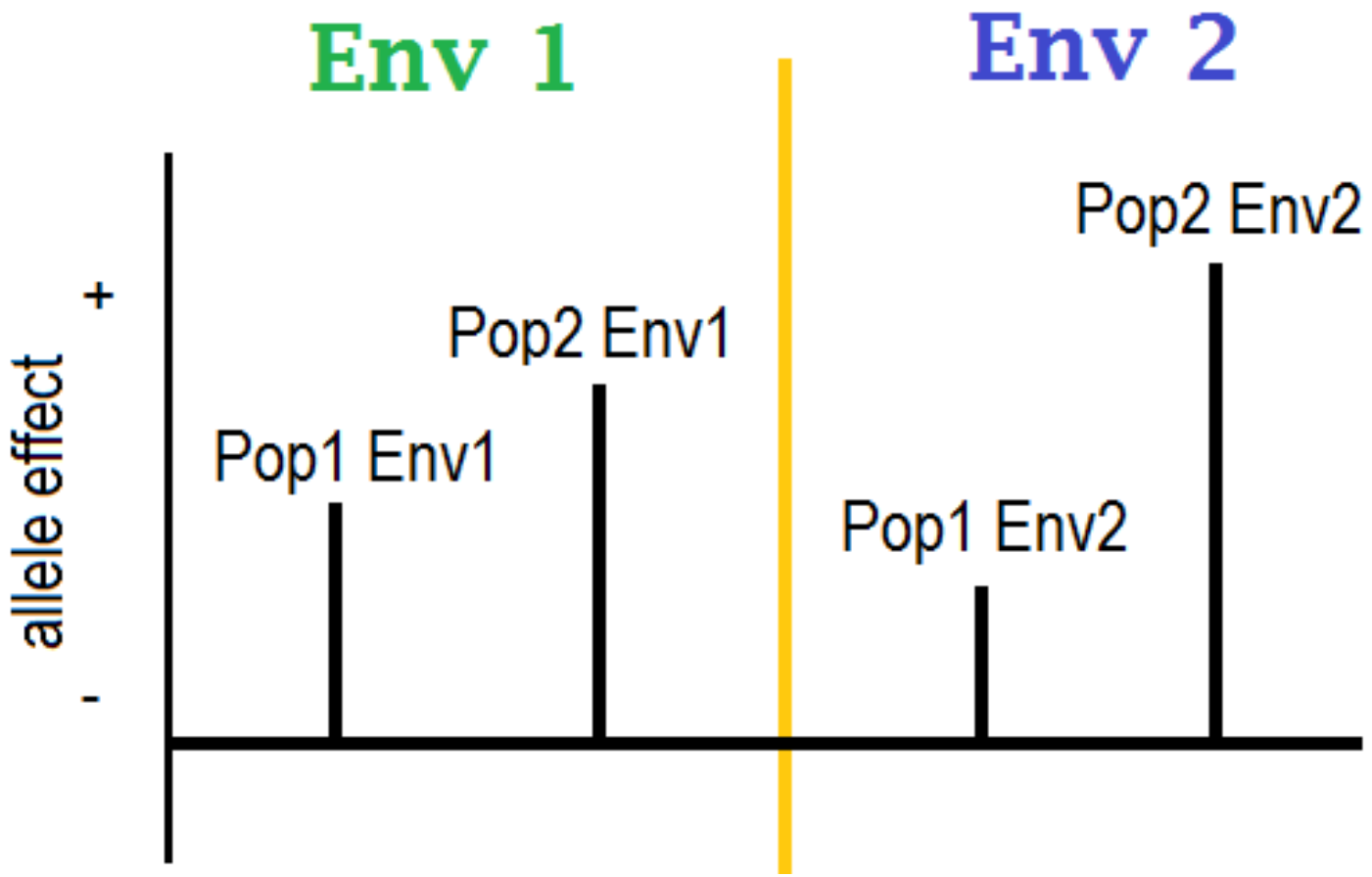
- Pop1 = 0
- Pop2 = 0
- Real = 10

Penetrance (background epistasis)



- Allele effect
  - Pop1 = 3
  - Pop2 = 4
  - Pop3 = 2
- Overall ≅ 3

GxE issues (= all above + environment)



Overcoming issues

MULTIPLE POPULATIONS in association studies

- Traditional model

Buckler et al. 2008

Fixed effect

$$\boldsymbol{y} = X\beta + \gamma + \epsilon$$

$$\boldsymbol{H_1}: Trait = Marker + Subpopulation + Polygenic\ term$$

- Modified model

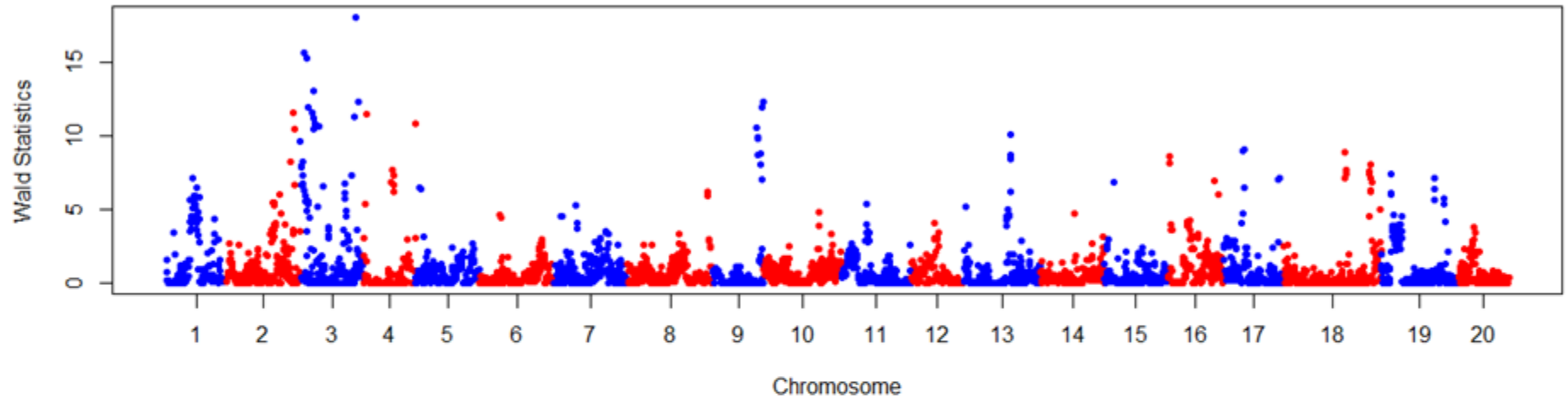
Xavier et al 2015; Wei & Xu 2016

Random effect

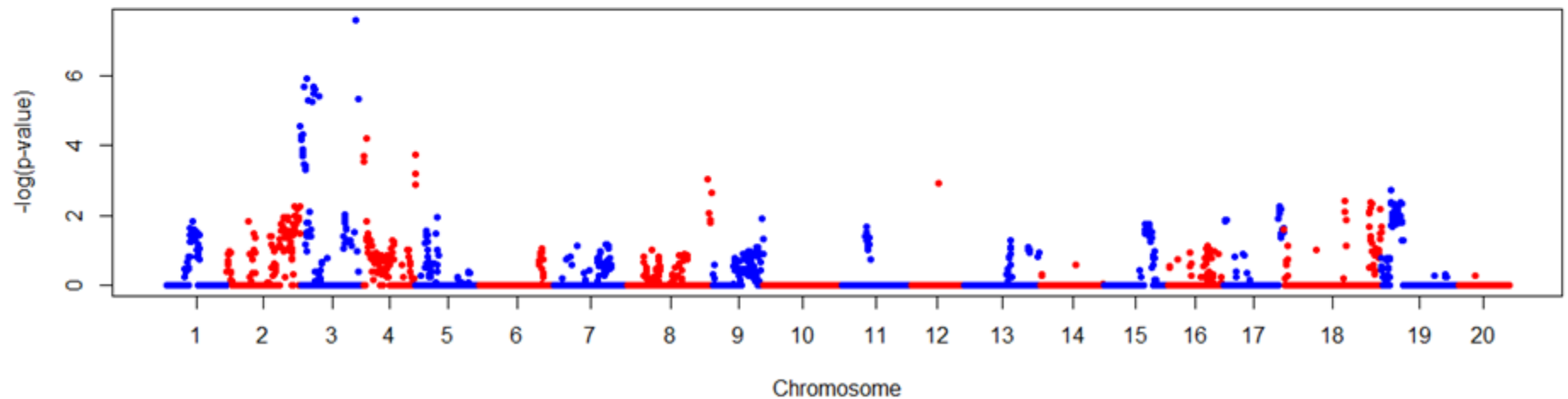
$$\boldsymbol{y} = \mu + Z\alpha + \gamma + \epsilon$$

$$\boldsymbol{H_1}: Trait = Marker \times Subpopulation + Polygenic\ term$$

TRADITIONAL MODEL



MODIFIED MODEL



GxE in association studies with multiple populations

- Typical models … **rarely account for GxE!!**

$$\boldsymbol{H_1}: Trait = Marker + Subpop \times Env + Polygenic\ term$$

$$\boldsymbol{y} = X\beta + W\delta + \gamma + \epsilon$$

- Modified model

$$\boldsymbol{H_1}: Trait = Marker \times Subpop \times Env + Polygenic\ term$$

$$\boldsymbol{y} = \mu + H\theta + \gamma + \epsilon$$

How to account for GxE in these studies?? (cont.)

- MEGA-analysis *(all at once)* – **computationally impossible!**

$$\boldsymbol{H_1}: Trait = Marker \times Env \times Subpop + Polygenic\ term$$

Lopez-Cruz (2015) G3

- META-analysis (*combine results*)

$$Trait(Env1) = Marker \times Subpop + Polygenic\ term$$
$$Trait(Env2) = Marker \times Subpop + Polygenic\ term$$

$$\boldsymbol{H_1}: Marker \times Subpop = Subpop + Env + GxE + Polygenic\ term$$

$$\hat{\theta} = \mu + \theta + \eta + \zeta + \epsilon \qquad \epsilon \sim N(0, R)$$

Advantages of changing to the modified model

1) **Selection**: Haplotypes optimized for target environment

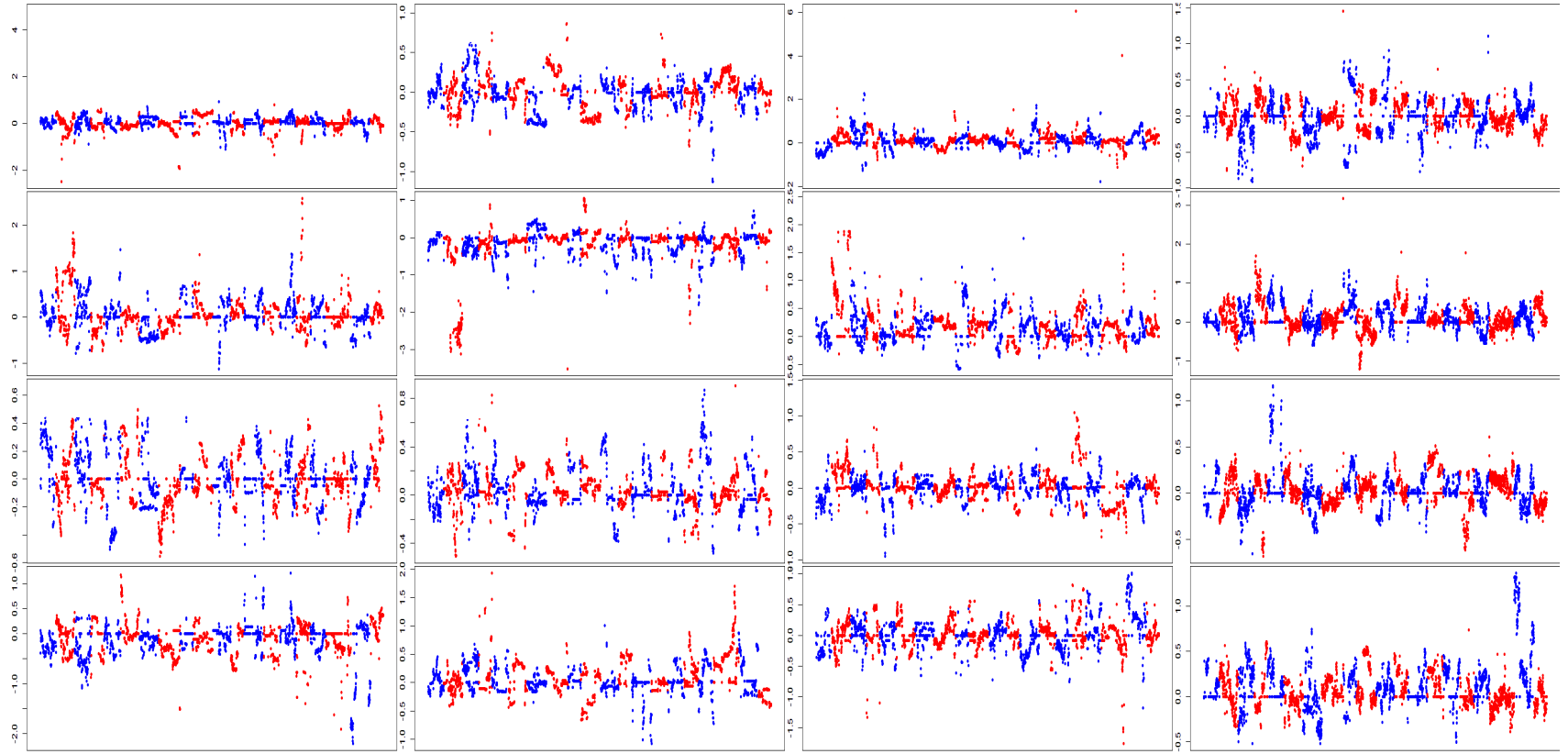
Pop1 Pop2 Pop3 Pop4

Env1

Env2

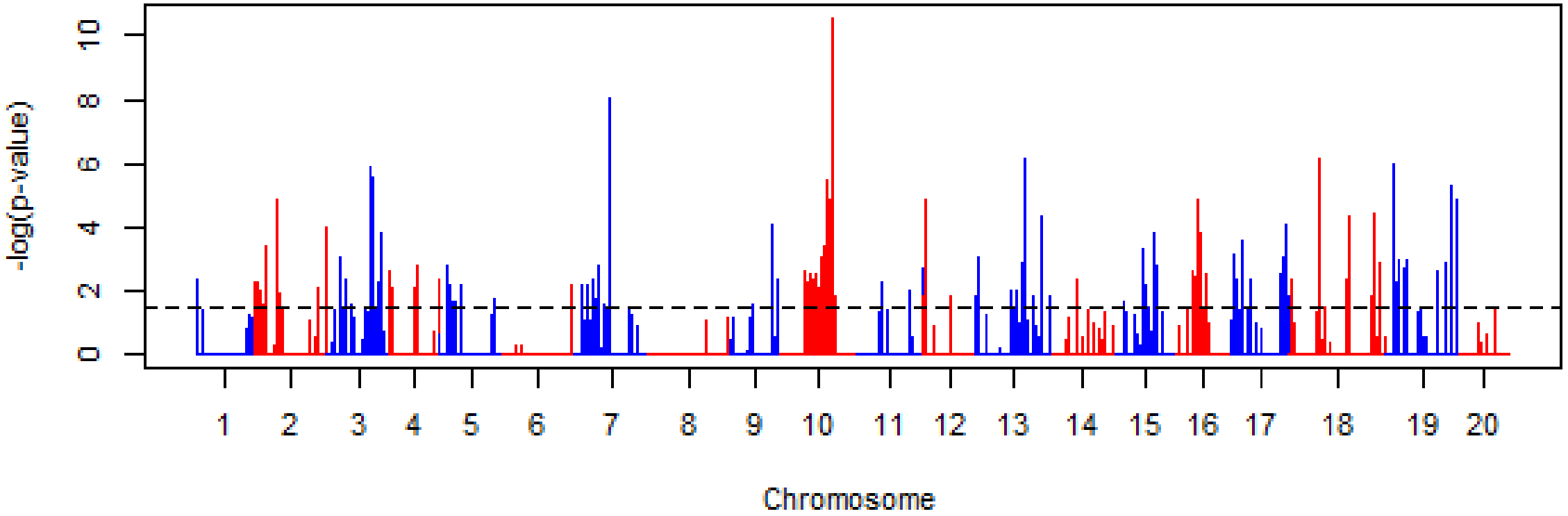
Env3

Env4



Advantages of changing to the modified model

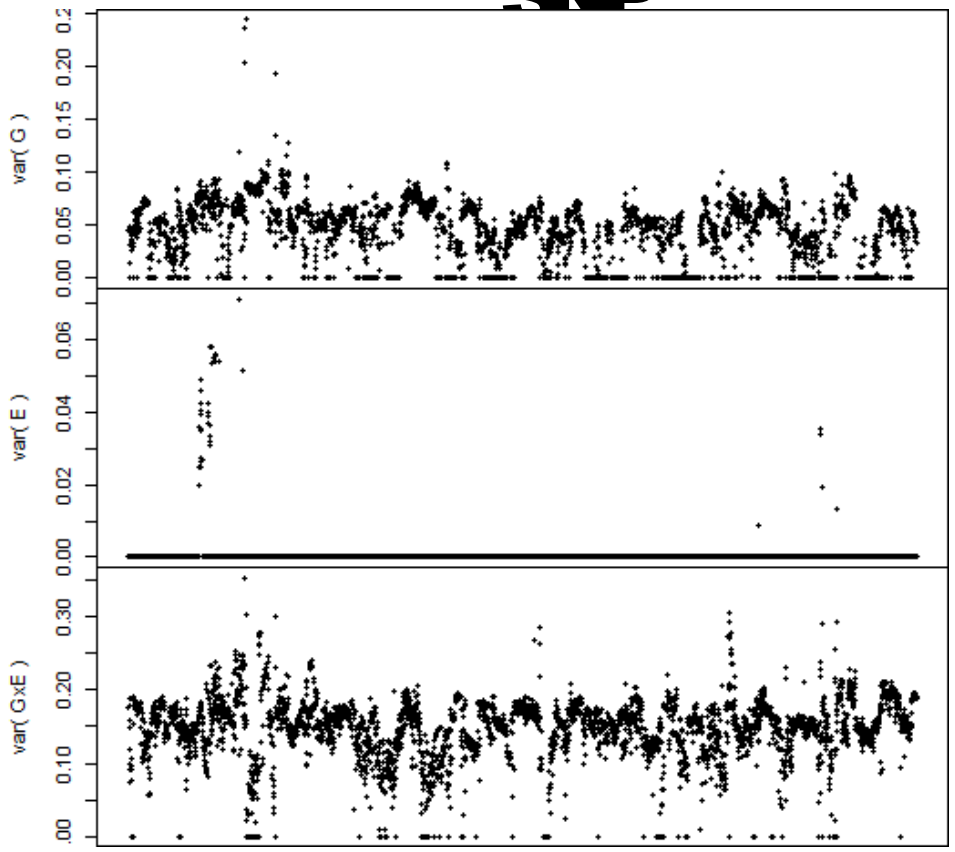
2) More **accurate inference** on genetic studies (reduced type 1 error)



Advantages of changing to the modified model

3) More **information**

Variance components for each SNP



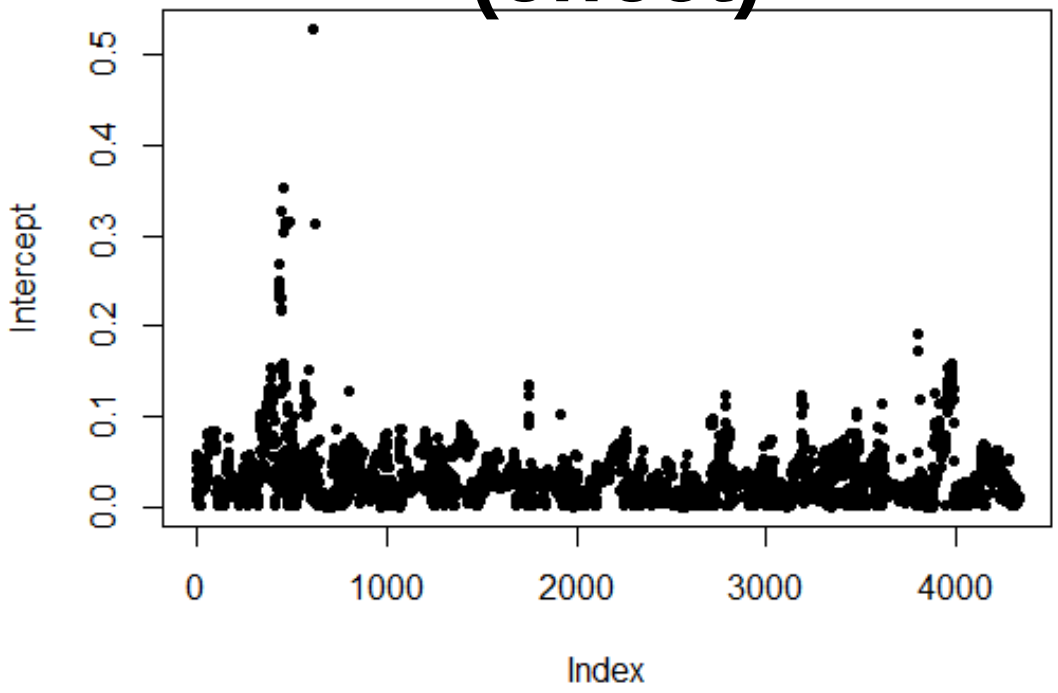
σ^2_G

σ^2_E

σ^2_{GxE}

Intercept μ

(effect)



Acknowledgements

- **Quantitative genomics group**
  - Shizhong Xu (UC Riverside)
  - Bill Muir (Purdue University)
  - Katy Rainey (Purdue University)

- **With some contribution from**
  - Bill Beavis (Iowa State University)
  - Reka Howard (Iowa State

QUESTIONS??

WEB: http://alenxav.wix.com/home
E-mail: xaviera@purdue.edu