

# QTL mapping in a mixed model perspective in Genstat

Fred van Eeuwijk, Marcos Malosetti & Martin Boer

15<sup>th</sup> European Genstat and ASREML Applied Statistics Conference  
Rothamsted, 14 July 2010



WAGENINGEN UNIVERSITEIT  
WAGENINGEN UR

# Quantitative Genetics & Plant Breeding

## ■ Classical model

- Phenotype = Genotype + Error
- $\underline{P}_{ir} = \underline{G}_i + \underline{\varepsilon}_{ir}$

## ■ Interest

- BLUP of  $\underline{G}_i$
- Partitioning  $V_P$ :  $V_G + V_\varepsilon$

## ■ Quantitative Trait Locus

- Elaboration of classical model
- Regression of  $\underline{G}_i$  on molecular marker information,  $x_i$
- $\underline{P}_{ir} = x_i a + \underline{G}_i^* + \underline{\varepsilon}_{ir}$

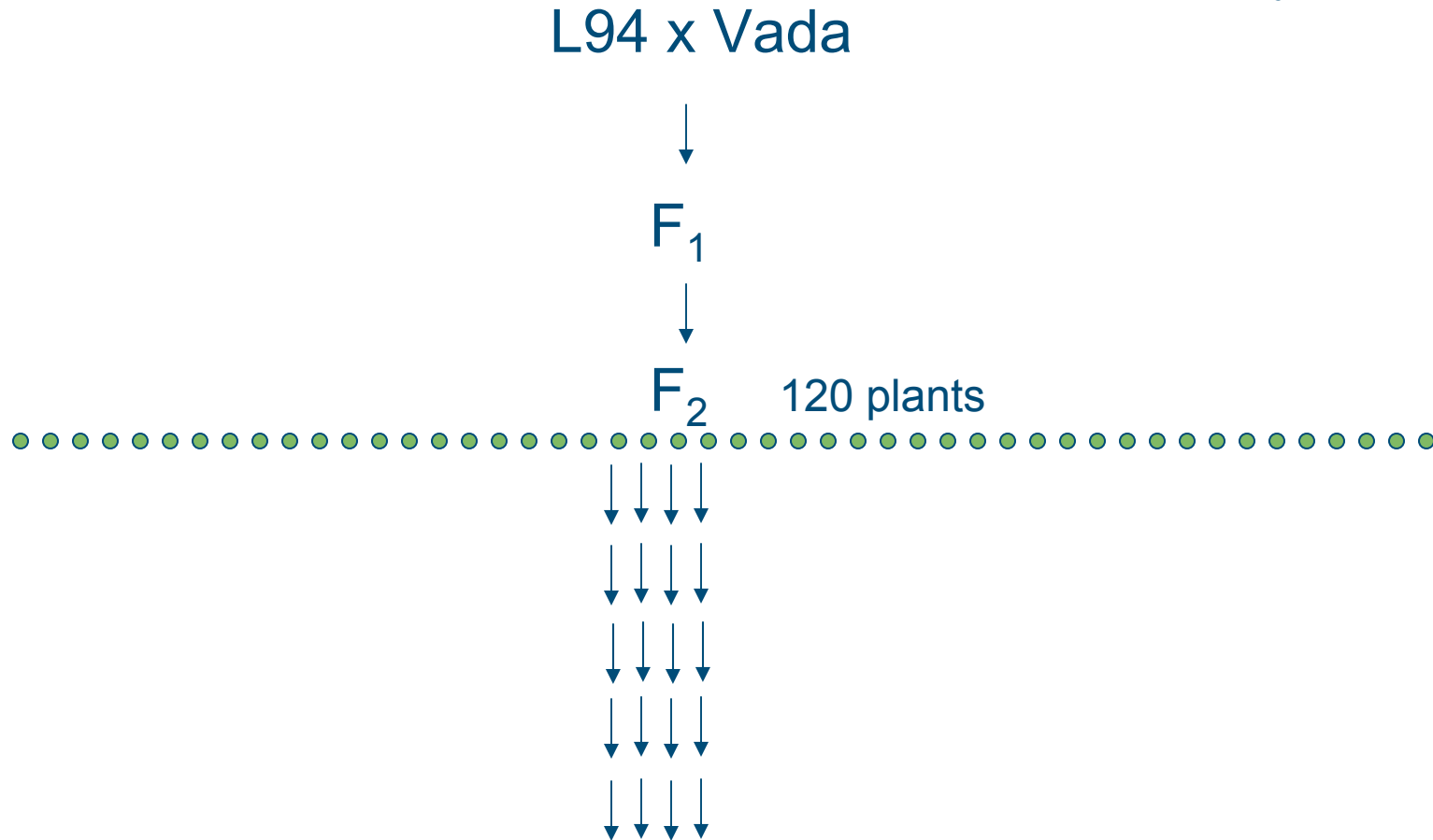


# Basic idea QTL mapping

- Look for association between phenotypic responses and marker genotypes in a mapping population, in which response and marker are segregating together
- When significant differences in response occur between marker genotypes a QTL should be close

Develop a population of offspring from two inbred parents  
(RILs in barley)

Courtesy Rients Niks



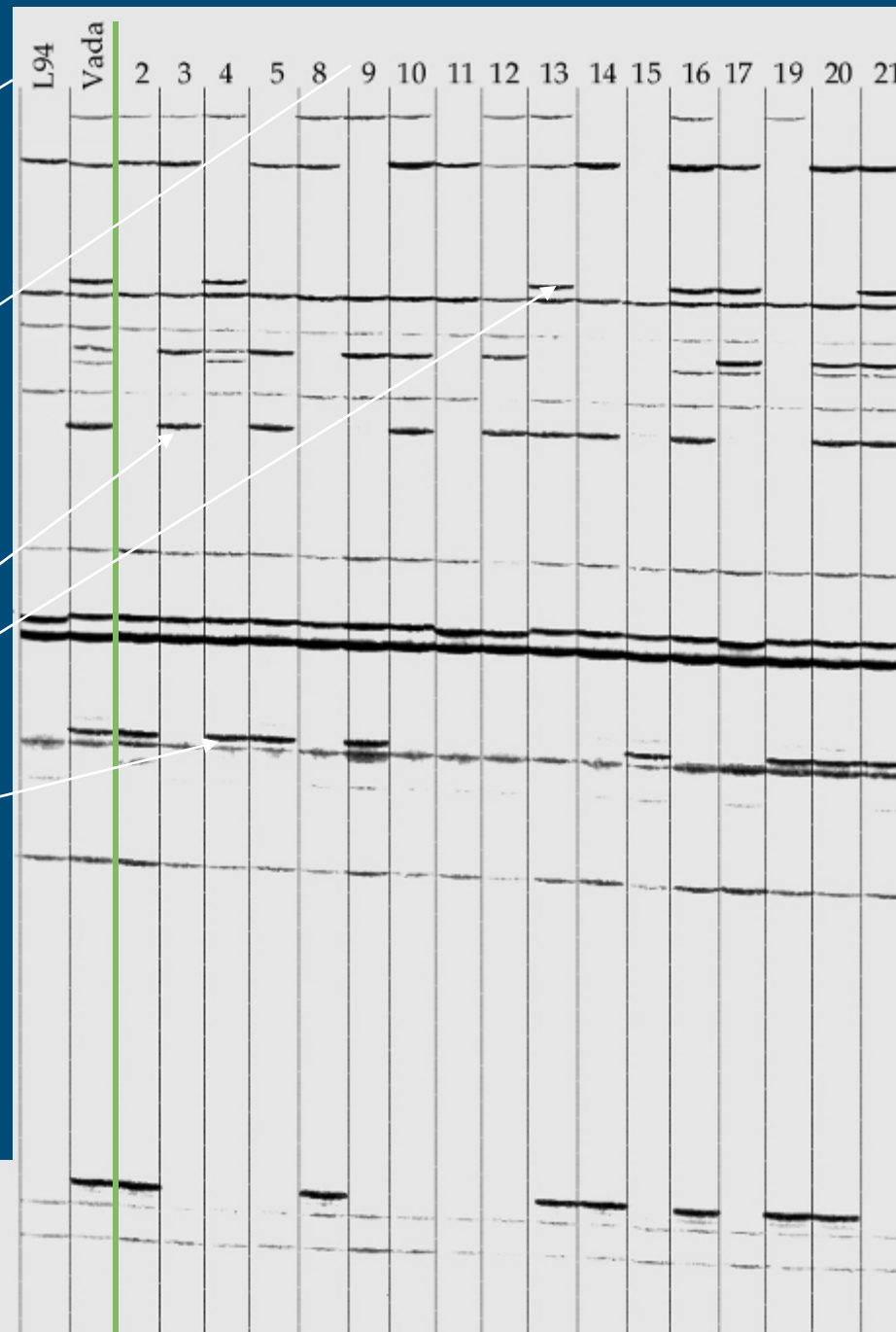
from each F<sub>2</sub> plant by SSD an F<sub>8</sub> (~ homozygous) plant = 103 Recombinant Inbred Lines (RILs)

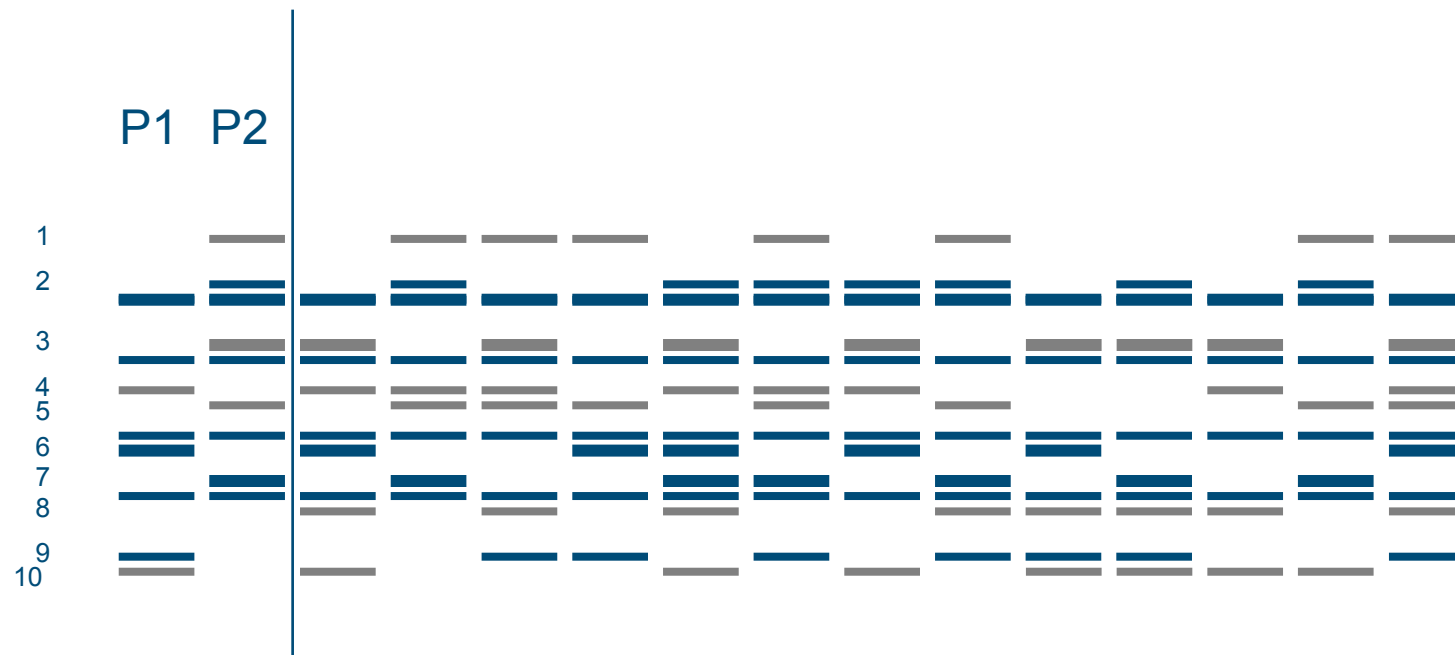
# Create and score markers on parents and offspring

2 parents

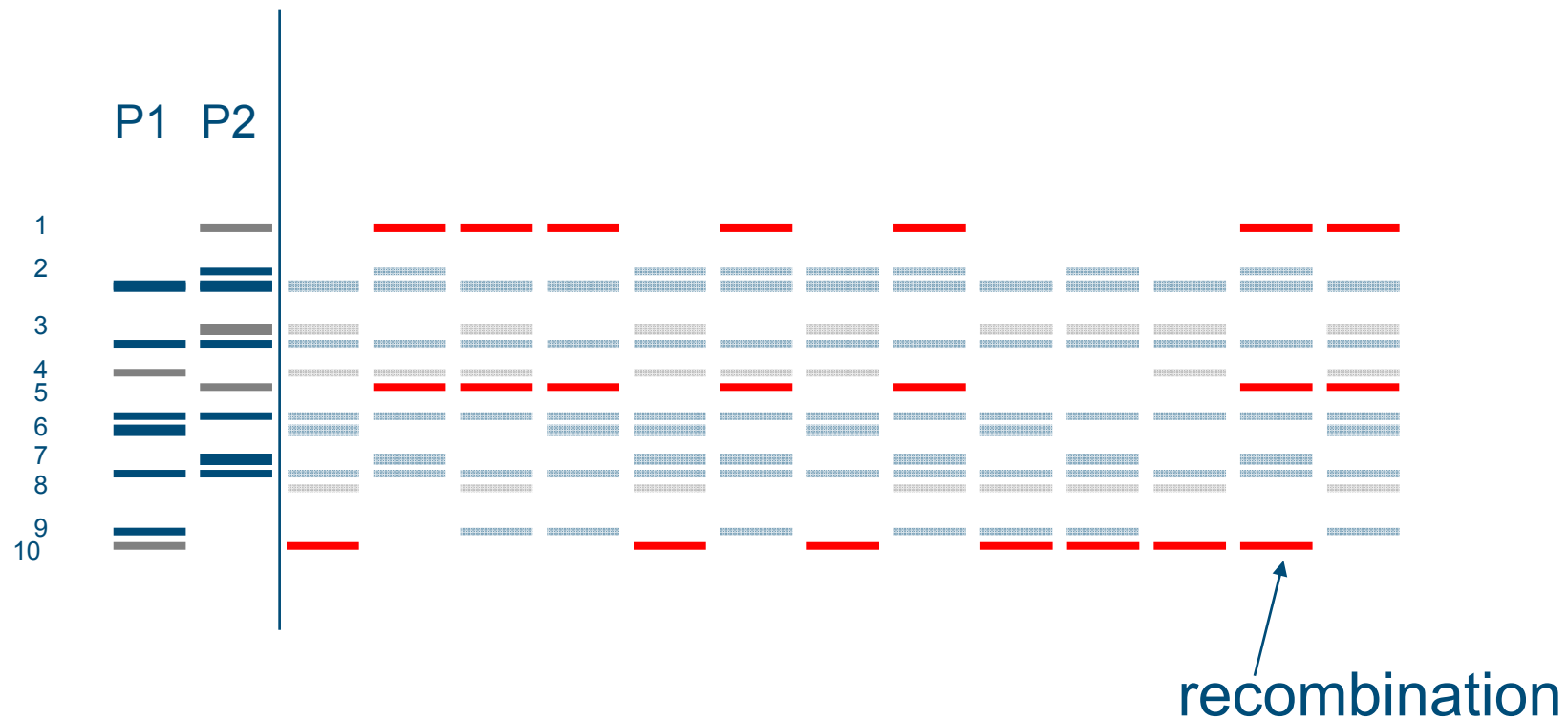
~100 RILs

markers





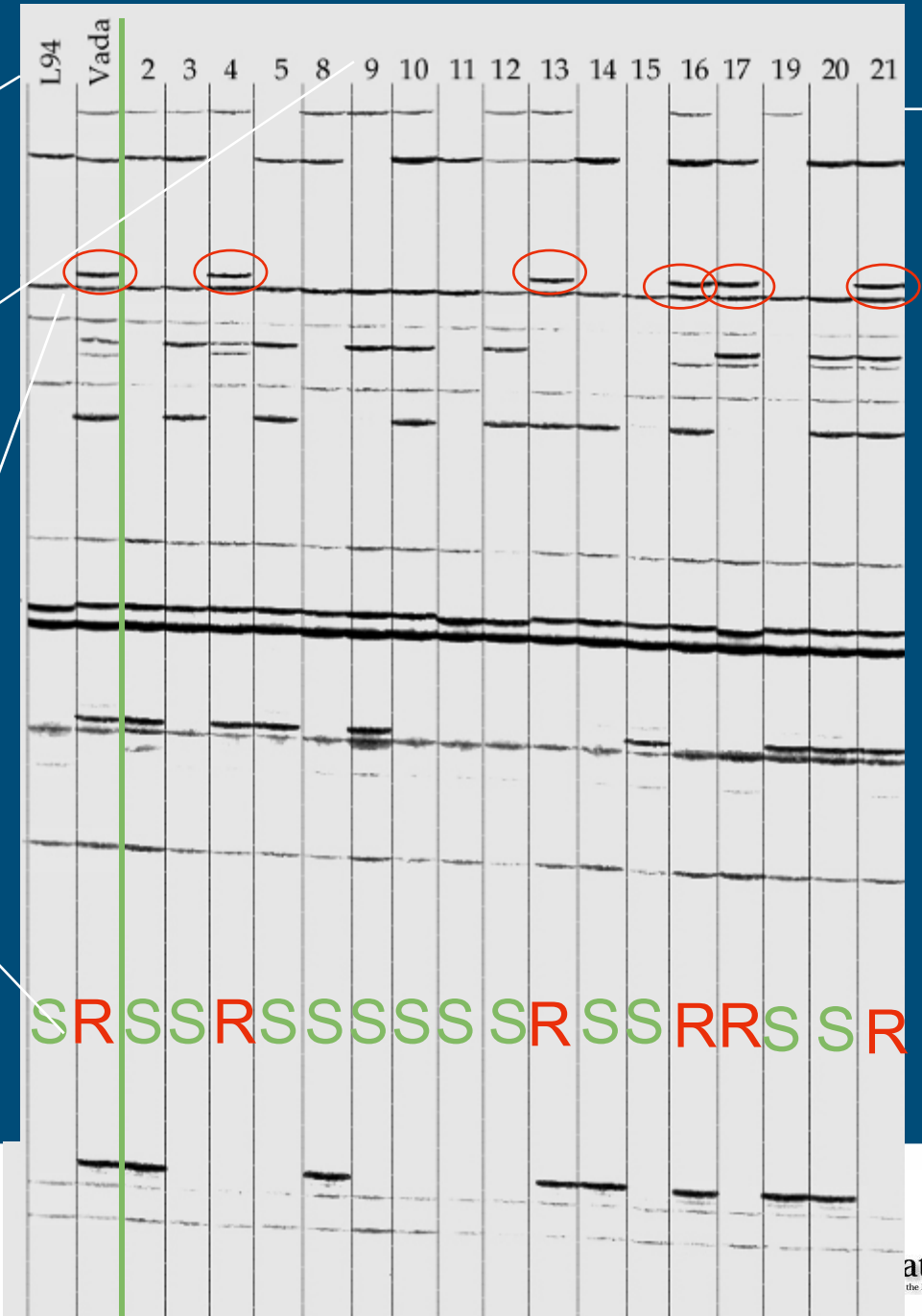
# Assess linkage between markers: Map construction



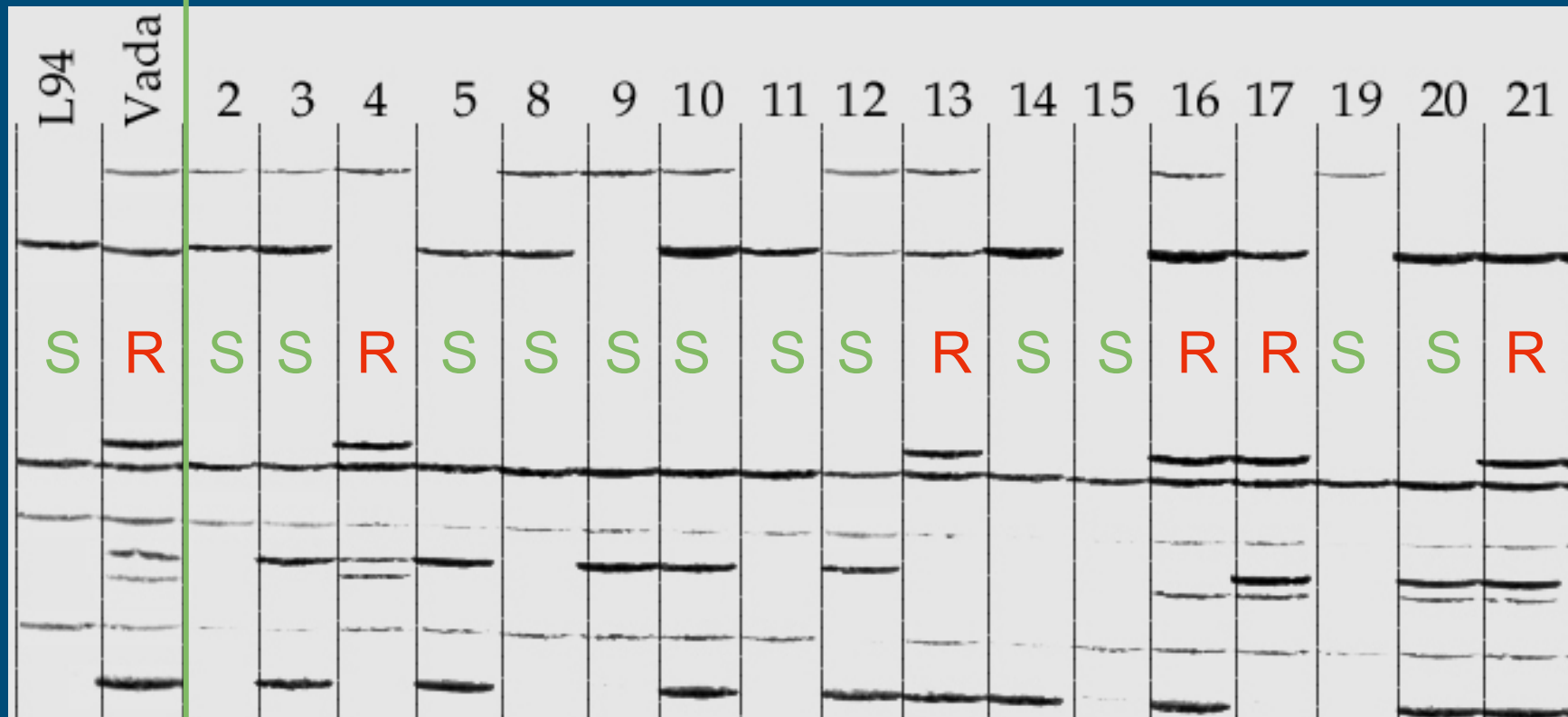
Markers 1, 5 and 10 are linked, 1 and 5 more closely than 10

# Detect marker-trait associations

- 2 parents
- ~100 RILs
- phenotypic evaluation
- identify associated marker

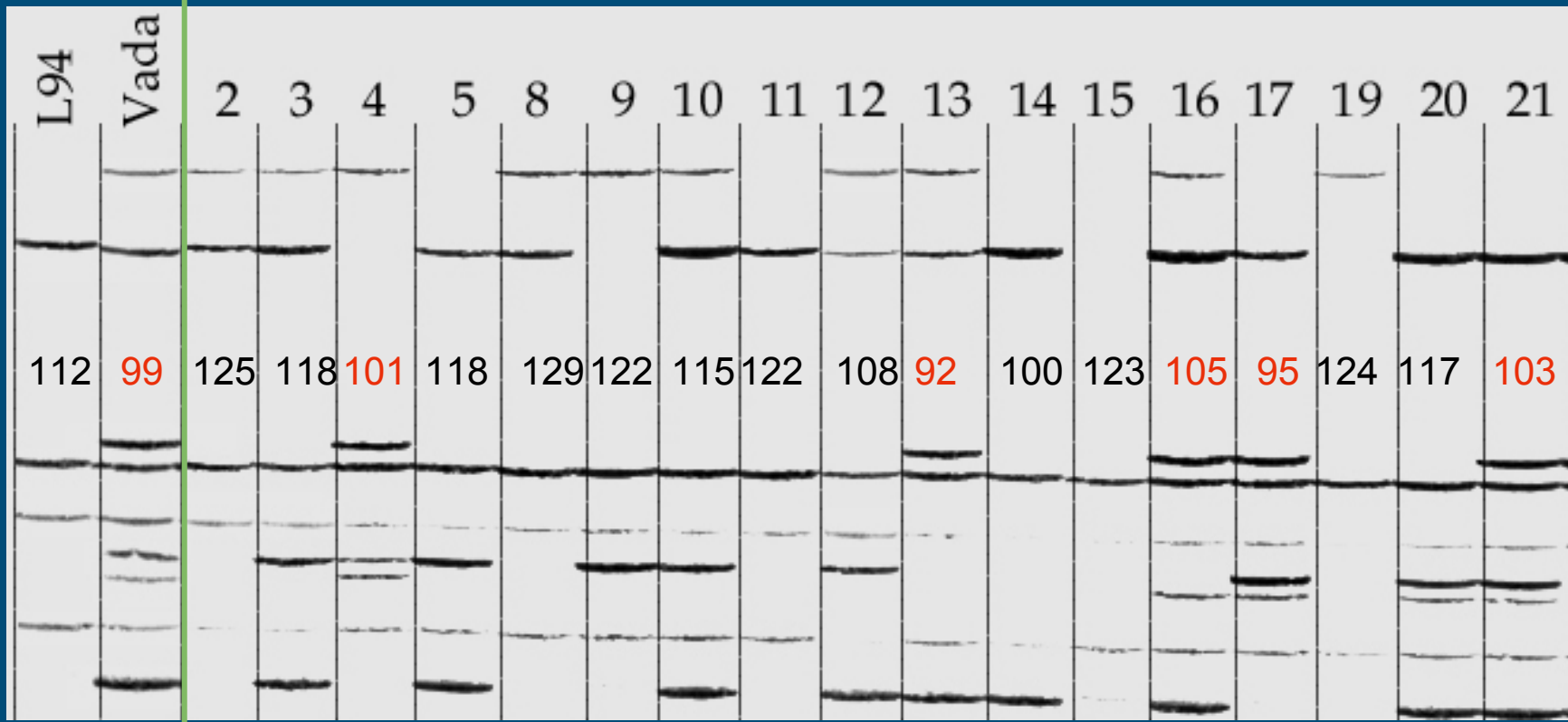


# Mapping genes



Evaluated trait can be qualitative

# Mapping genes



Evaluated trait can be qualitative or quantitative (QTL)

# Mixed models and QTL mapping

- For valid inference on QTLs, the dependencies in the data due to blocking and spatial trends need to be accounted for in a mixed model
  - This becomes even more true for multiple environments and multiple traits
- Response =
  - replicate + block in replicate +
  - QTL + G\* +
  - error
- Test for QTL is roughly (virtual) marker over (residual) genotype



# Identifying genetic basis of phenotypic trait variation

- Identify genetic basis G
  - Detect QTLs
  - Estimate QTL locations and effects
- QTL mapping is a straightforward extension of standard statistical modeling as occurring in plant breeding and quantitative genetics
  - Task: identify genetic covariables that explain G
- **No special purpose statistical procedures nor packages are required**
  - **Decent mixed model facilities will suffice**
- Still, some ingenuity in the application of mixed models will be useful



# Statistical modeling in modern plant breeding

- Predict phenotypic expression for
  - multiple traits
  - across a range of environmental conditions
  - over developmental time
  - from molecular marker variation, genomic information and environmental inputs
  - for various types of (offspring) populations

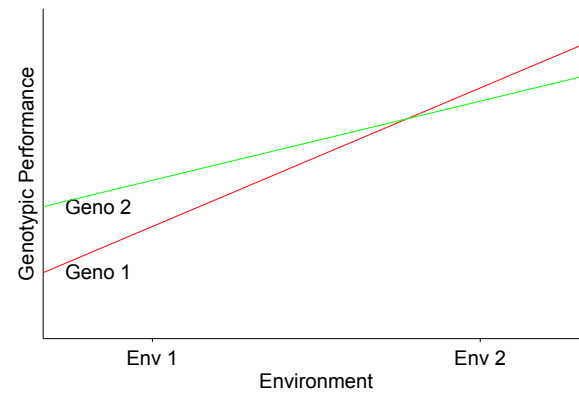
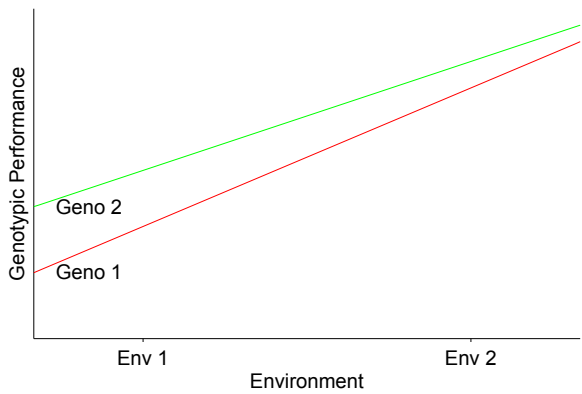
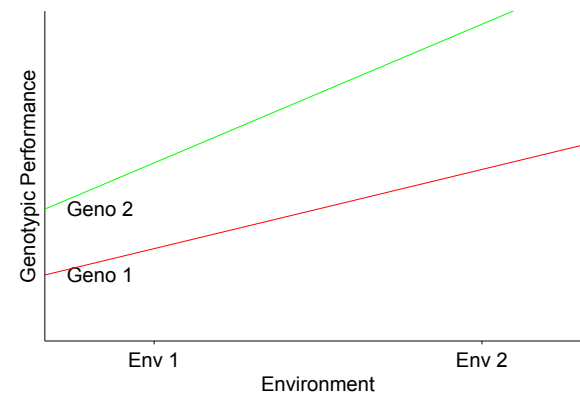
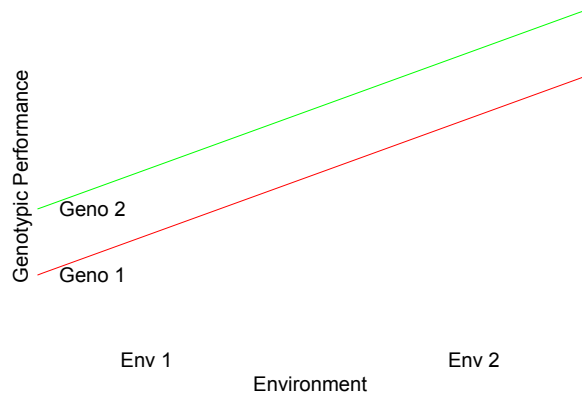


# QTLx $E$

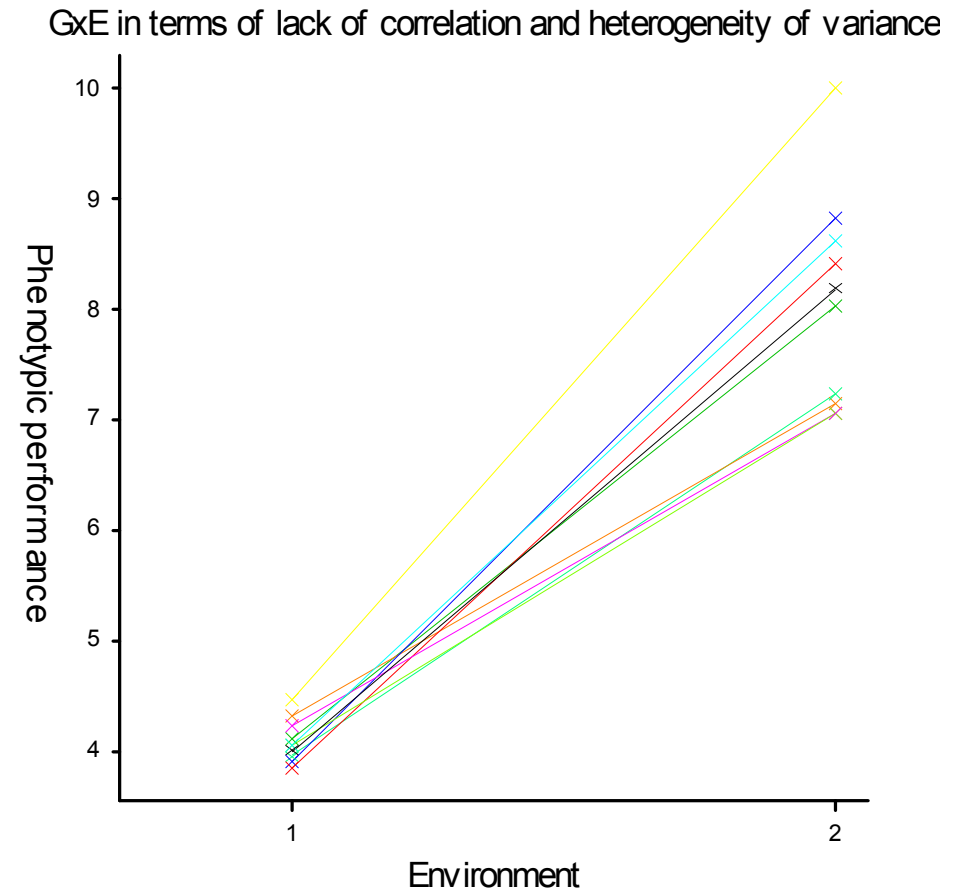
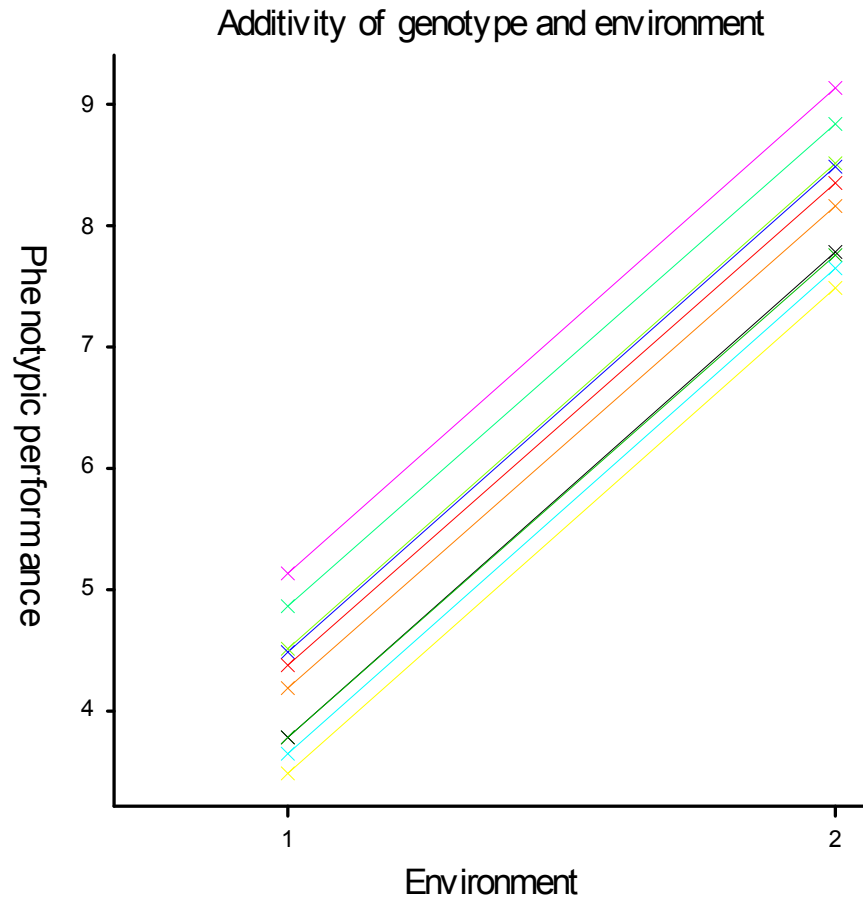
Modeling the genetic basis of Gx $E$  and correlations between environments



# GxE in terms of changing mean performance across environments



# GxE in terms of lack of correlation and heterogeneity of variance



# Genotype by environment interaction (GxE)

## ■ GxE

- Genotypic differences between phenotypic responses are dependent on the environment
- GxE for mean
  - non parallelism of average responses
- GxE for variances and correlations
  - heterogeneity of variance across environments
  - changing genetic correlations between environments

# Modeling mean and VCOV for MET data in

## LMM

- $P_{ij} = \mu_{ij} + \underline{\varepsilon}_{ij}$     ( $\underline{P}_{ij} = \mu_j + \underline{\varepsilon}_{ij}$ )  
i for genotypes, j for environments
- Aim of statistical modeling for MET/GxE data
  - $\mu_{ij}$  : predictable/ repeatable
    - Describe  $\mu_{ij}$  as much as possible in terms of single indexed parameters
    - Approximation to non-linear genotype-to-phenotype relations
  - $VCOV(\underline{\varepsilon}_{ij})$  : unpredictable/ non-repeatable
    - Find an appropriate structure for  $\underline{\varepsilon}_{ij}$  reflecting heterogeneity of genetic variances and correlations and allowing reliable conclusions on  $\mu_{ij}$

# Statistical models for mean $\mu_{ij}$

Additive model (typical choice  $\text{var}(\underline{\varepsilon}_{ij}) = \sigma^2$ )

$$\mu_{ij} = \mu + E_j + G_i$$

Full interaction model

$$\mu_{ij} = \mu + E_j + G_i + (GE)_{ij}$$

Multiplicative models for interaction

*Exploration*

Bilinear model

$$[G_i+] (GE)_{ij} = u_{1i} v_{1j} + u_{2i} v_{2j}$$

*Confirmation*

$$\text{Factorial regression } [G_i+] (GE)_{ij} = \kappa x_i z_j / + x_i \alpha_j / + \beta_i z_j$$

genotypic sensitivities

environmental characterizations

# QTLxE analysis for METs with biparental populations



# QTL analysis

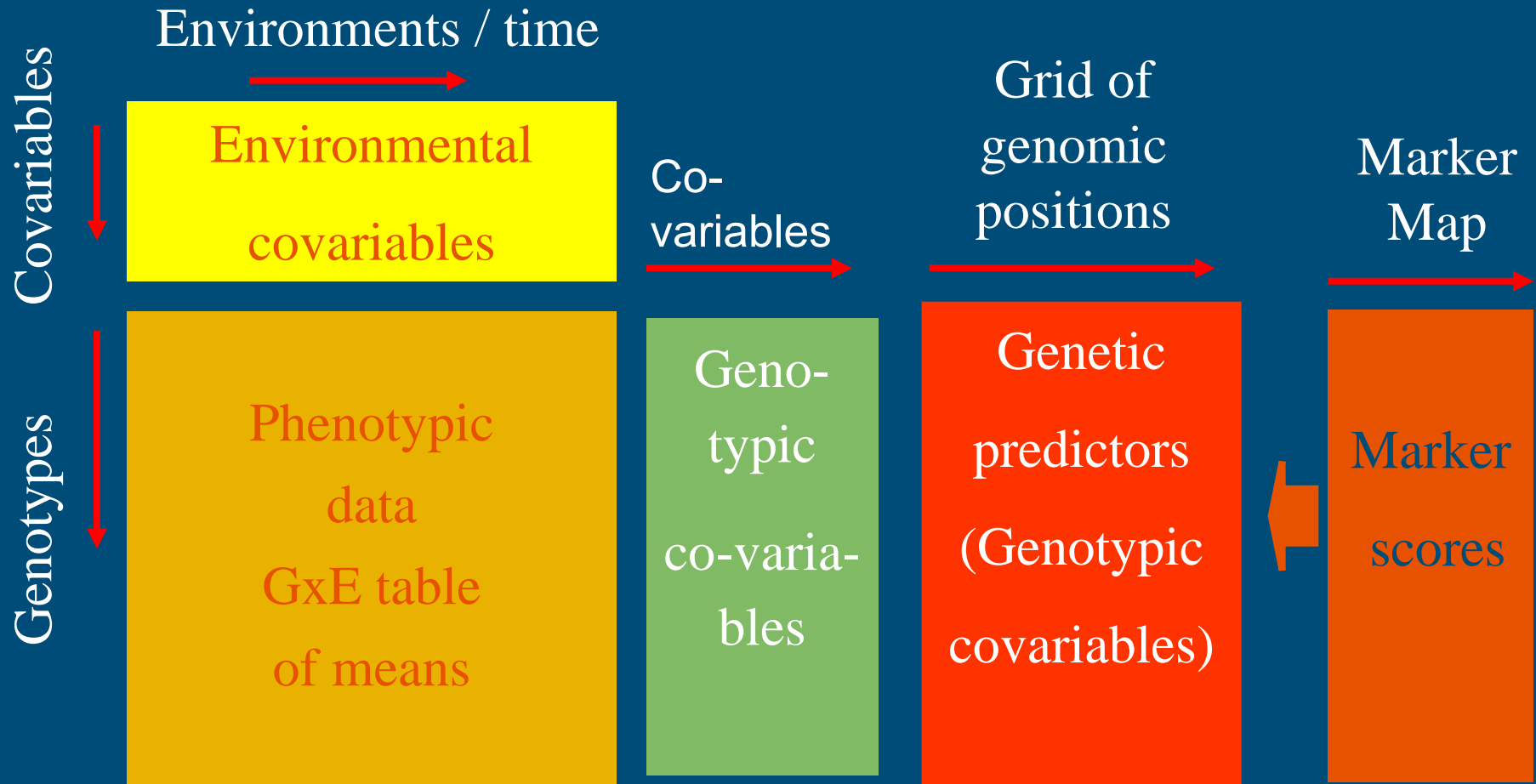
- In QTL analysis we want to test whether the variation at molecular markers is related to phenotypic variation
- Observations on molecular markers can be converted to genotypic covariables containing the conditional probabilities for particular QTL genotypes at any place in the genome

$$x_i^{add} = P(QQ | M_i) - P(qq | M_i)$$

$$x_i^{dom} = P(Qq | M_i) - \frac{P(QQ | M_i) + P(qq | M_i)}{2}$$

- The probabilities of QTL genotypes can be computed by a Markov chain method (Jiang & Zeng, Genetica 1997)

# Typical data configuration



# QTL mixed model for multiple environments

## (QTLxE)

- Phenotype =
  - Environment +
  - Environment specific QTLs +
  - Residual genetic effect +
  - Error
- Environment specific QTLs can be regressed on environmental characterizations
- VCOV for residual genetic variation should allow environment specific variances and correlations

$$\underline{P}_{ij} = \mu_j + \underline{G}_i + \underline{GE}_{ij} + \underline{\varepsilon}_{ij}$$

$$\underline{P}_{ij} = \mu_j + \sum x_i a + \underline{G}_i + \sum x_i a_j + \underline{GE}_{ij} + \underline{\varepsilon}_{ij}$$

$$\underline{P}_{ij} = \mu_j + \sum x_i a_j + \underline{G}_{ij} + \underline{\varepsilon}_{ij}$$

$$\underline{P}_{ij} = \mu_j + \sum x_i (\alpha_0 + \alpha_1 z_j + \delta_j) + \underline{G}_{ij} + \underline{\varepsilon}_{ij}$$

$$VCOV(\underline{G}_{ij}) = \begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_{21} & \sigma_2^2 & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \sigma_{J1} & \sigma_{J2} & \cdot & \cdot & \sigma_J^2 \end{bmatrix}$$

# Inference in mixed model QTL mapping

- Which model to choose (mean/fixed, VCOV/random)?
- Which test to use for QTL detection?
- Which level of test to use / multiple testing correction?
- How many QTLs are there?
- Which types of genetic effects does a QTL have?
  - (additive / dominance/ epistasis)
- How to obtain point estimates for QTL allele effects?
- How to obtain interval estimates for QTL allele effects?
- How to obtain point estimates for QTL locations?
- How to obtain interval estimates QTL locations?
- How much of the phenotypic/genetic variance does a QTL explain?

# Inference in mixed model QTL mapping

- Standard linear mixed model framework is available
  - Wald tests for fixed effects
  - Deviance tests for VCOV parameters
- AIC and BIC for non nested models
- Multiple test corrections in genotypic and environmental covariable selection
  - Bonferroni based on approximation to the number of independent tests (Cheverud (2001), Li&Ji (2005))
  - Simulation/ parametric bootstrap



# QTL mapping; a procedure

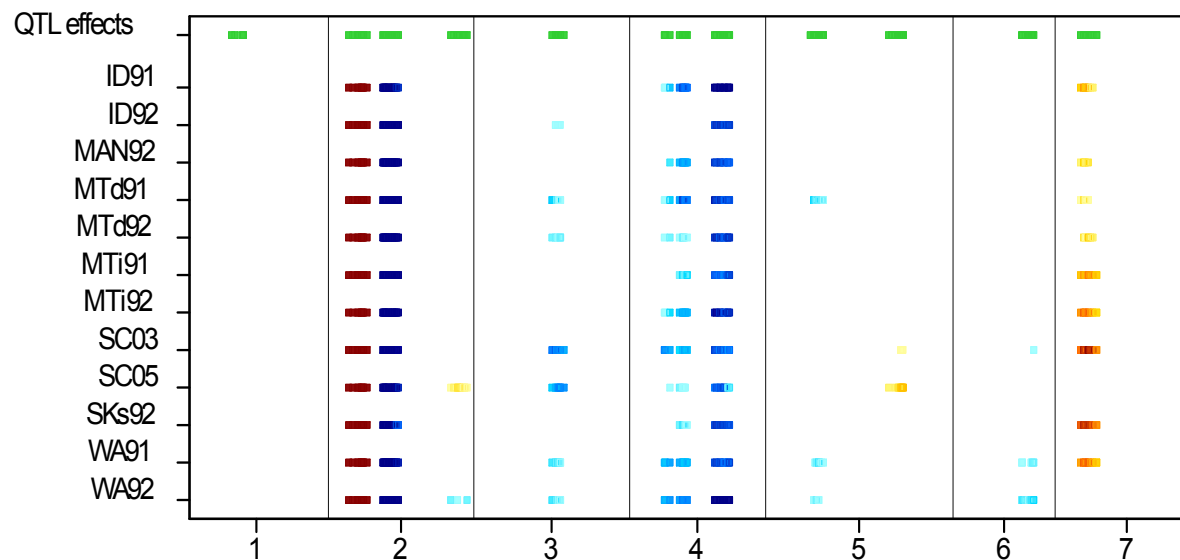
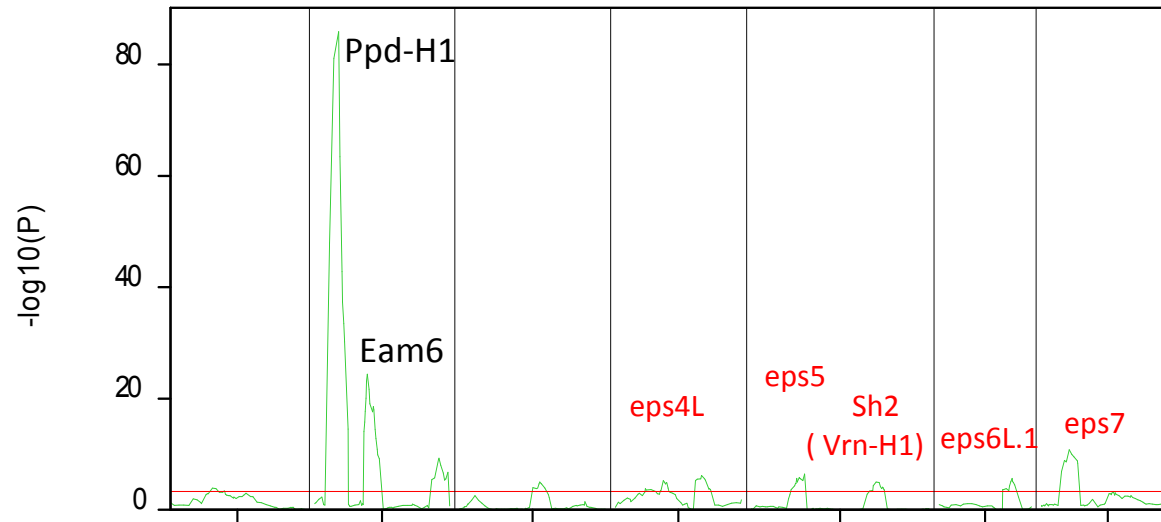
1. Find a VCOV model for the GGE (=G+GE) variation in the phenotypic data
2. Given the VCOV model identified in step 1, test for QTLs (QQE) on a grid of chromosomal positions (Simple Interval Mapping or SIM)
3. Perform backward selection on the full set of QTLs from SIM
4. Alternative, perform another grid evaluation with a model containing all QTLs from the SIM analysis as genetic covariables, but omit these covariables in windows around SIM QTLs (=composite interval mapping)
5. Estimate effects in final multi-QTL model
6. Identify environmental covariables related to environment-specific QTL effects

# Steptoe X Morex: QTLxE for heading date (HD)

Mixed Model  
QTLxE mapping

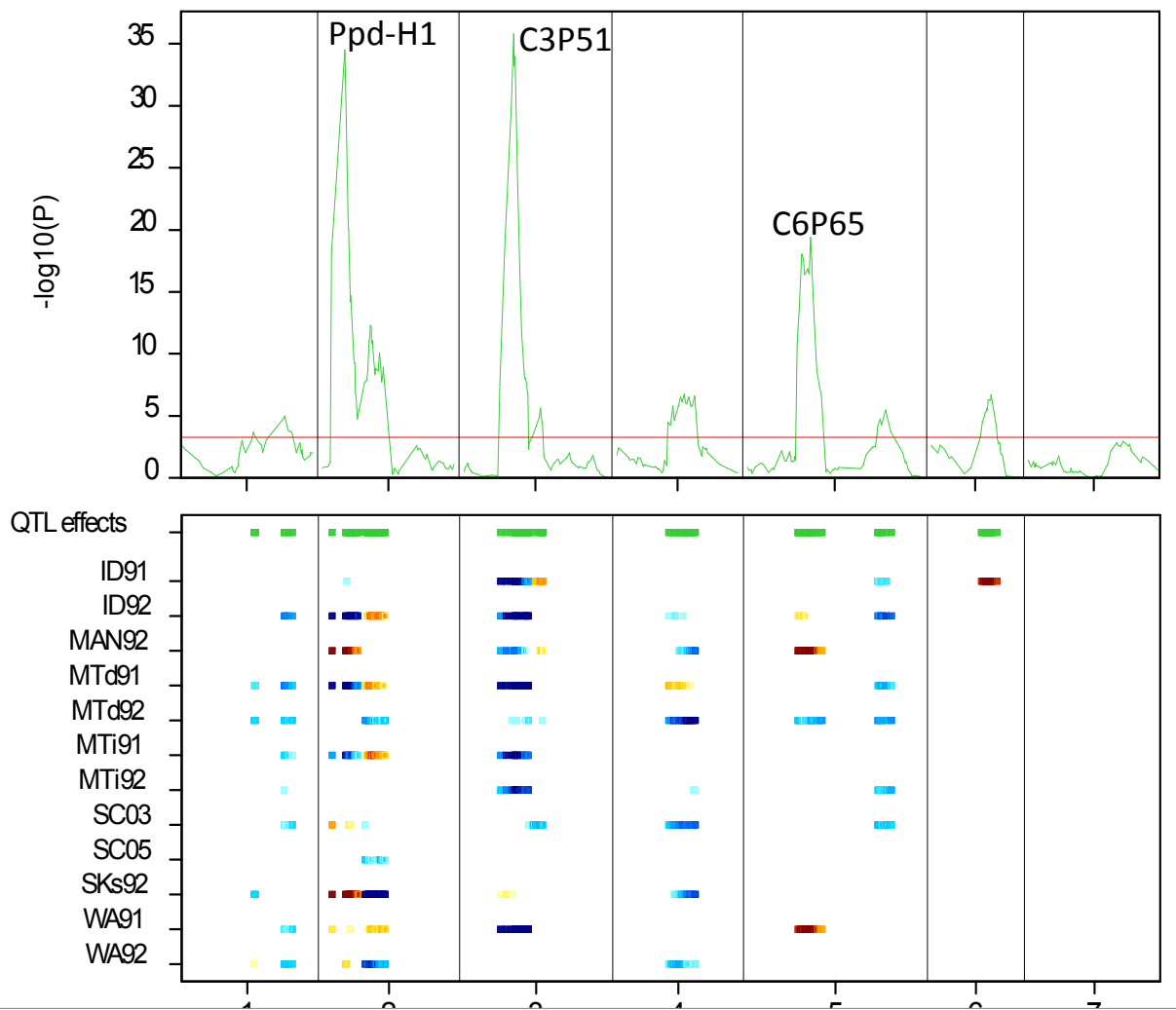
Ignacio Romagosa  
(Univ. Lerida)

Implemented in  
GenStat v.12



HD very predictable across environments: Non-crossover QTLx  
Morex alleles at Ppd-H1 and eps7 (yellow-red) delay heading;  
Steptoe alleles at Eam6, eps4L, eps5S and SH2 (blue) delays heading

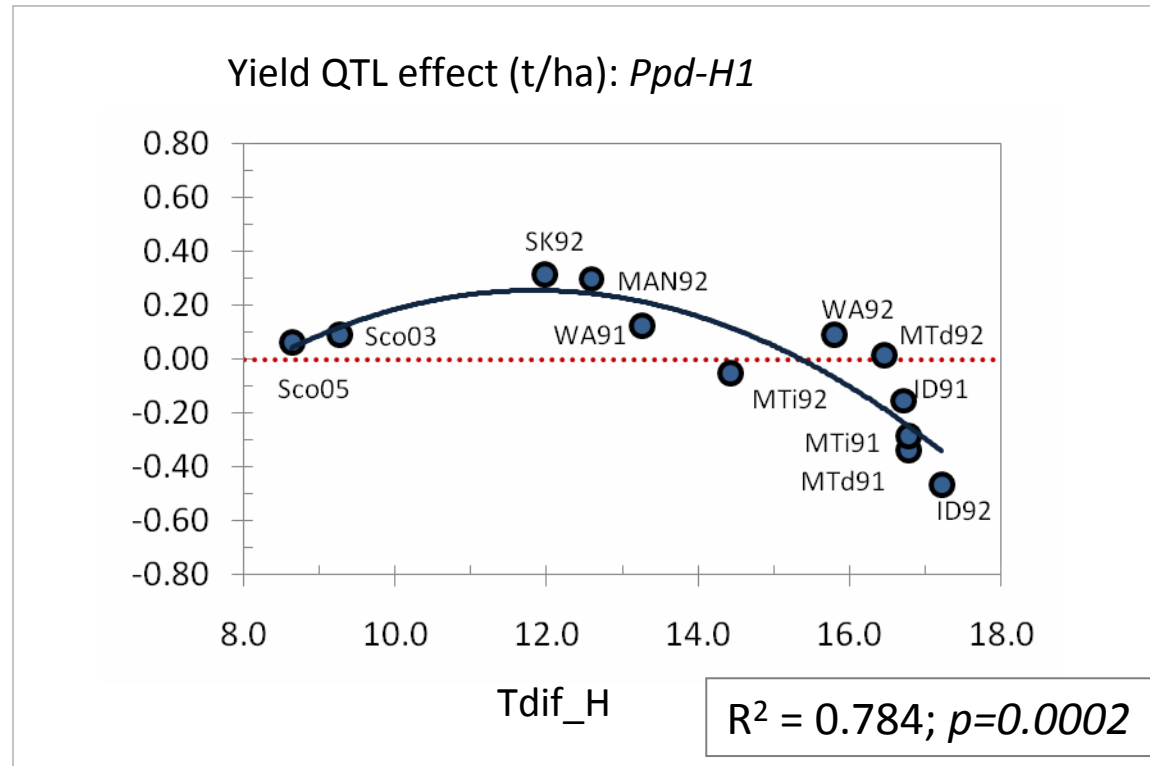
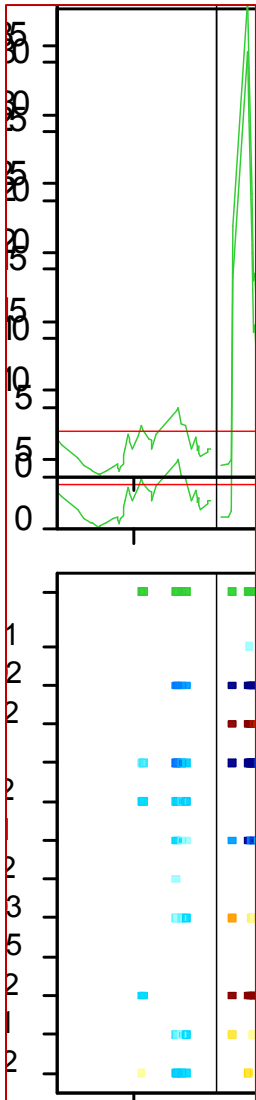
# Steptoe X Morex: QTL.E for grain yield



Cross-over interaction: Depending on the env. the contribution of the Morex allele can be either positive or negative.

There is a clear genetic control of heading, but depending on the meteorological conditions through grain filling it could be either positive or negative to have early heading.

# Steptoe x Morex: QTLxE for yield



*Ppd-H1* Morex allele (yellow-red) non-responsive at long photoperiod better adapted to intermediate Tdif\_H

# Defaults for Genome Scan are set by workbook

## Changes by specification of Genome Scan window and Options window

The screenshot displays the GenStat software interface. The main window shows the 'Single Trait Linkage Analysis (Multiple Environments)' dialog box. The 'Available Data' list includes 'm\_positions' and 'yld'. The 'Quantitative trait means' is set to 'yld', 'Genotype factor' to 'genotype', and 'Environment factor' to 'env'. The 'Type of population' is set to 'F2'. Under 'Genetic predictors and associated information', 'Additive effects' is 'gp\_additive', 'Additive effects 2nd parent' is 'gp\_additive2', 'Dominance effects' is checked, and 'Linkage group for each predictor' is 'gp\_linkage'. The 'Variance-covariance model' is set to 'Compound symmetry'. The 'Save candidate QTLs' is 'qtl\_candidates'. The 'Options...' button is highlighted with a red arrow pointing to the 'Linkage Analysis Options' dialog box.

The 'Linkage Analysis Options' dialog box shows the following settings:

- Display:**  Summary of QTLs retained in model,  Progress
- Threshold:**  Bonferroni (Distance between loci: 4),  Li and Ji (Genome-wide significance level (alpha): 0.05),  Specify:
- Minimum cofactor proximity:** 50 cM
- Minimum separation for selected QTLs:** 30 cM
- Graphics:**  Plot genetic predictor effects along genome
- Workspace allocation for REML analysis:** 100
- Available Data:** (empty list)
- Labels for genotypes:** m\_id
- Include unit error:

# Select/fit final QTL model (backward selection)

The screenshot displays the GenStat software interface. The main window shows the 'Single Trait Linkage Analysis (Multiple Environments)' dialog box. The 'Available Data' list includes 'gp\_pos', 'm\_positions', and 'yld'. The 'Quantitative trait means' is set to 'yld', 'Genotype factor' to 'genotype', and 'Environment factor' to 'env'. The 'Type of population' is 'F2'. Under 'Genetic predictors and associated information', 'Additive effects' is 'gp\_additive', 'Additive effects 2nd parent' is 'gp\_additive2', and 'Dominance effects' is checked with 'gp\_dominance'. 'Linkage group for each predictor' is 'gp\_linkage' and 'Positions within linkage group' is 'gp\_pos'. The 'Variance-covariance model' is 'Compound symmetry'. The 'Save candidate QTLs' field is 'qtl\_candidates'. Buttons at the bottom include 'Initial Scan (SIM)', 'Scan with cofactors (CIM)', and 'Select final QTL model...'. A red arrow points from the 'Select final QTL model...' button to the 'Select: Final QTL Model' dialog box.

The 'Select: Final QTL Model' dialog box has the following settings:

- Display:**  Summary,  Estimated effects,  Monitoring,  Model,  Wald tests,  Variance parameters,  Variance-covariance matrix.
- Run QTL backward selection. Significance level for backward selection: 0.05.
- Model:** Variance-covariance matrix: Compound symmetry.
- Save:**  QTL effects,  Standard error of QTL effects,  QTL positions,  Display in spreadsheet.

Buttons at the bottom of the dialog include 'Run', 'Cancel', and 'Candidate QTLs...'. Below the dialog boxes, a table shows the following data:

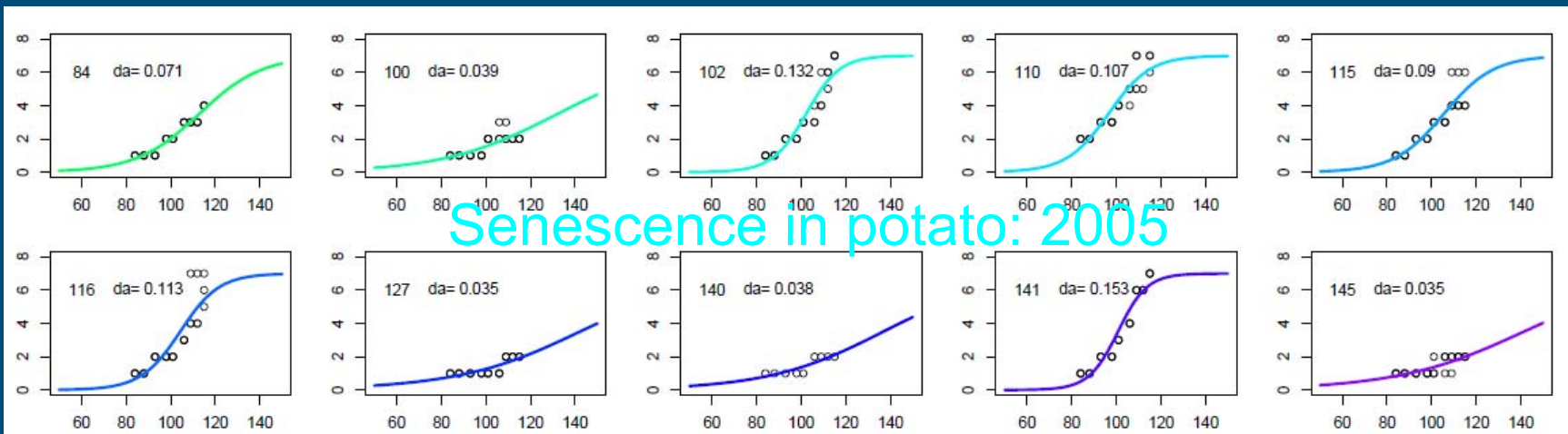
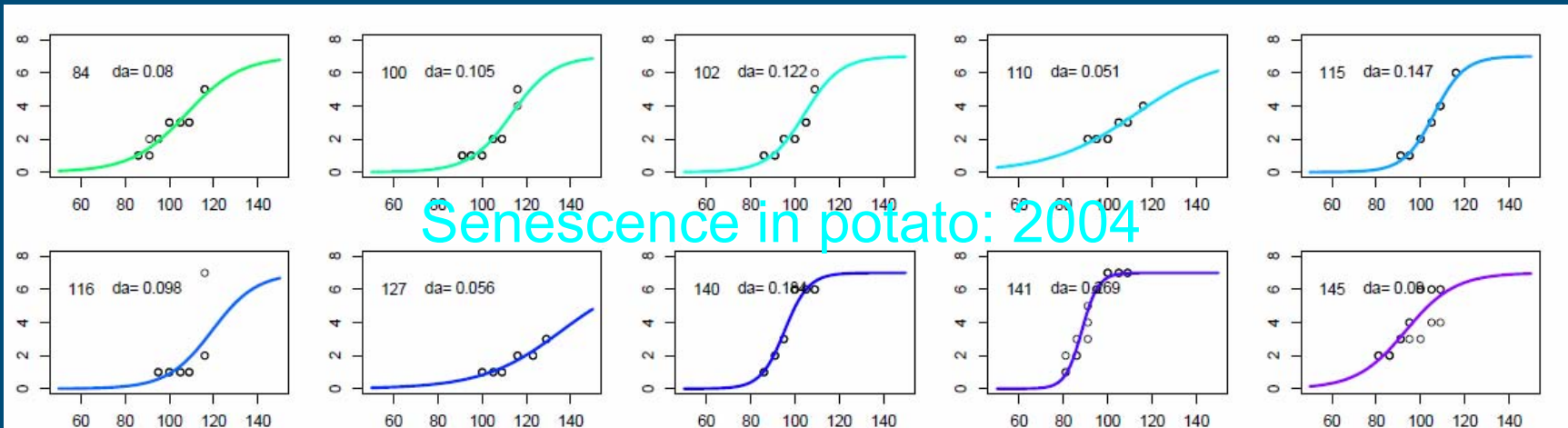
84	L004	7	72.8	4.52
117	L022	10	45.3	4.95
118	L114	10	53.2	7.78

# Functional QTL analysis

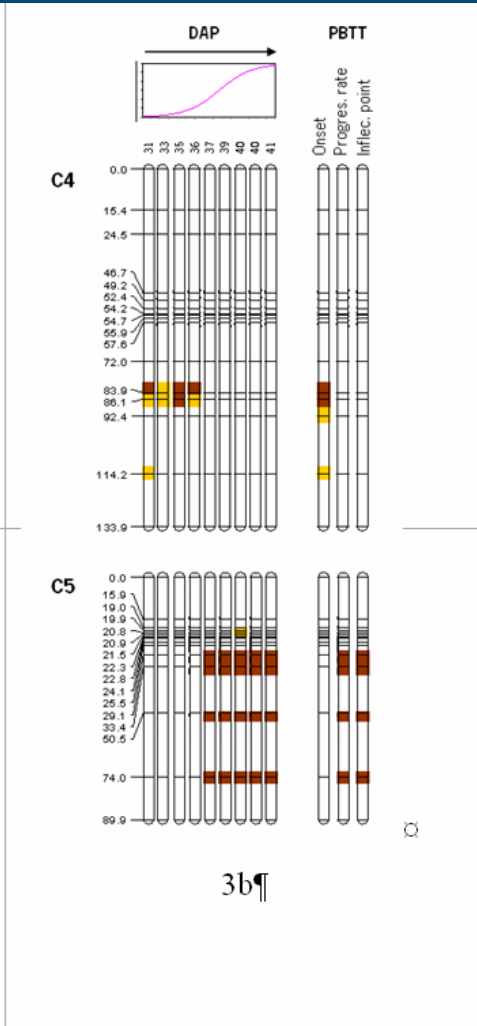
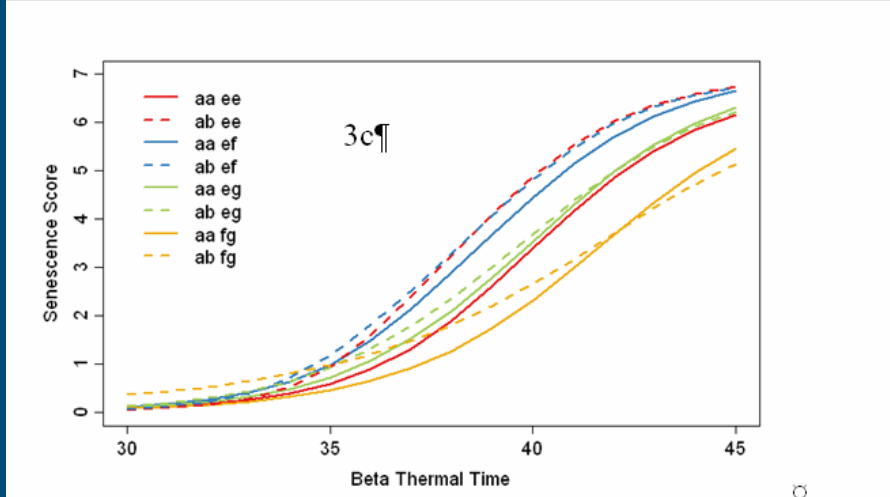
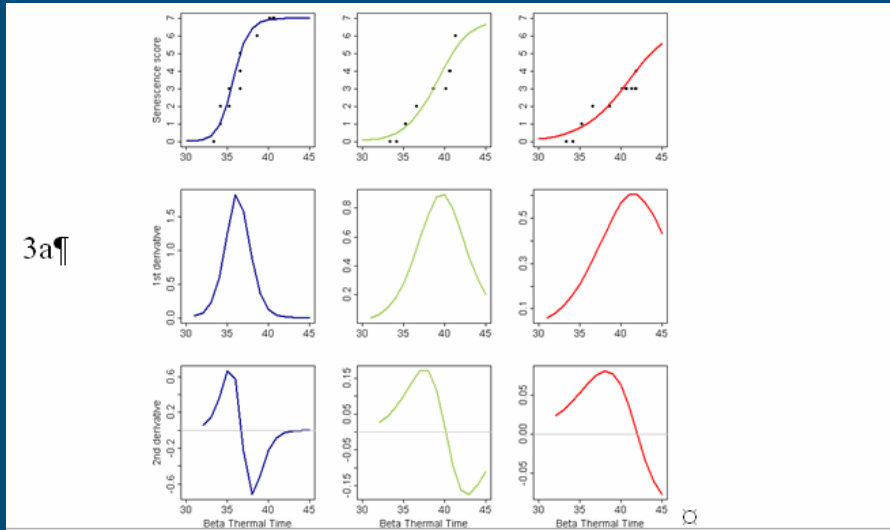
Analysing the genetic basis of functions over time and environmental gradients



# GxE for non-linear non-parametric phenotypic responses



# QTL modeling of response curve parameters



# Multi-trait QTL mapping

Modeling the genetic basis of correlations between traits



# QTL mixed model for multiple traits (genetic correlations)

- Phenotype =
  - Trait mean +
  - Trait specific QTLs +
  - Residual genetic effect +
  - Error

$$\underline{P}_{it} = \mu_t + \sum x_i a_t + \underline{G}_{it} + \underline{\varepsilon}_{it}$$

- VCOV for residual genetic variation between traits should allow for trait specific variances and correlations

$$VCOV(\underline{G}_{it}) = \begin{bmatrix} \sigma_1^2 & & & & \\ \sigma_{21} & \sigma_2^2 & & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \sigma_{T1} & \sigma_{T2} & \cdot & \cdot & \sigma_T^2 \end{bmatrix}$$

# Multiple traits (genetic correlations) in multiple environments

## ■ Phenotype (QTLxE)

- Trait-environment mean +
- Trait-environment specific QTLs +
- Residual genetic effect +
- Error

$$P_{ijt} = \mu_{jt} + \sum x_i a_{jt} + G_{ijt} + \varepsilon_{ijt}$$

## ■ VCOV

- Genetic variances differ across trait-environment combinations
- Genetic correlations between traits can be environment specific
- Genetic correlations between environments can be trait specific

$$VCOV(G_{ijt}) = VCOV(G_{Traits}) \otimes VCOV(G_{Envs})$$



# CIMMYT drought stress program in maize

## ■ Population

- 211 F2 derived F3 lines

## ■ Evaluated traits

- Grain yield (YLD)
- Anthesis-silking interval (ASI)
- Days to male flowering (MFLW)
- Ears/plant (ENO)
- Plant height (PH)

## ■ Markers

- 132 loci

## ■ 1992 (Tlaltizapán, México)

- Well watered (WW)
- Intermediate stress (IS)
- Severe stress (SS)

## ■ 1994 (Tlaltizapán, México)

- Intermediate stress (IS)
- Severe stress (SS)

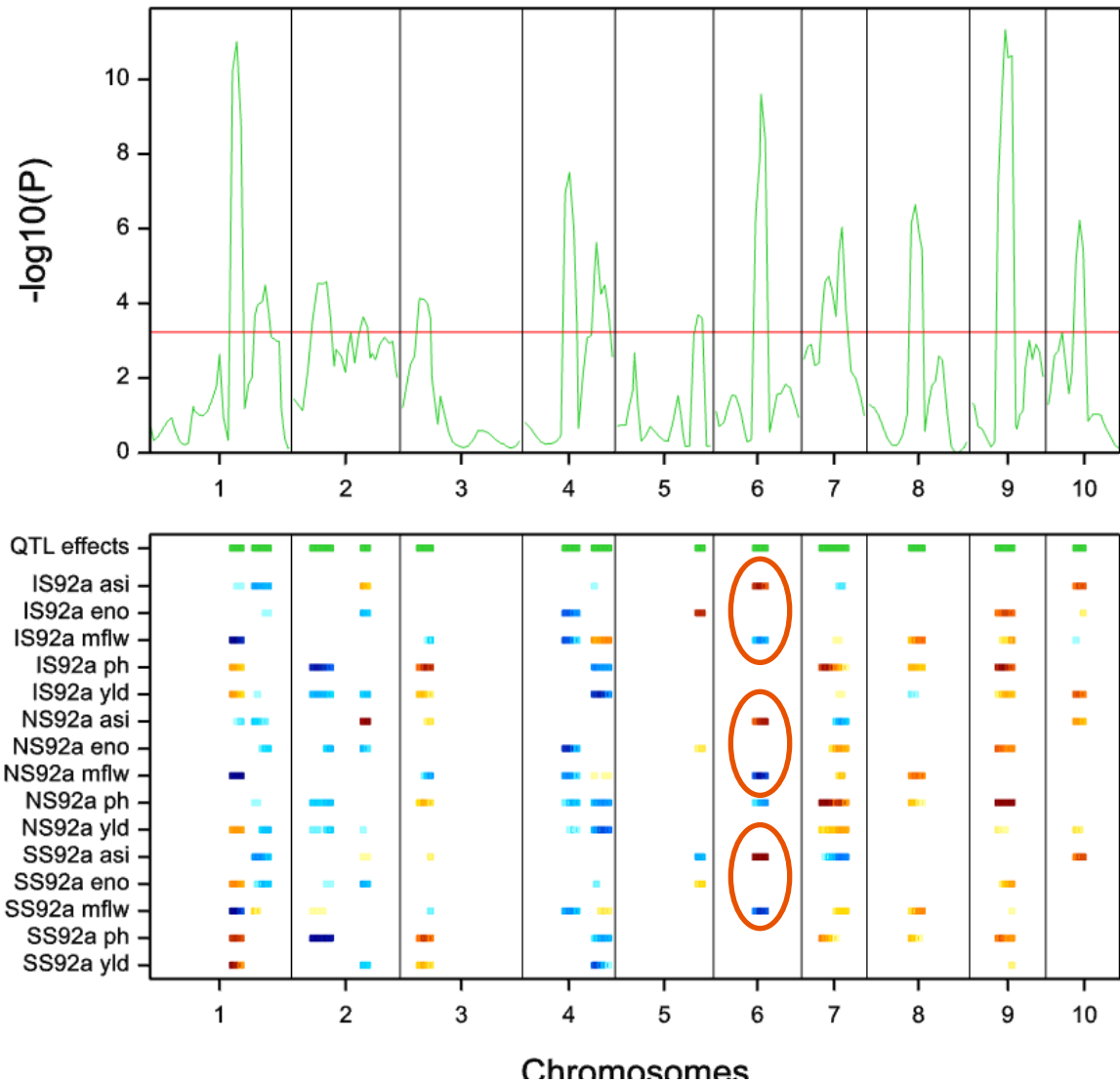
## ■ 1996 (Poza Rica, México)

- Low Nitrogen (2 seasons)
- High Nitrogen

■ In the example on the next slide we use only 3 env's



# Chromosome 6: asi – mflw QTL



# Further extensions mixed model approach

$$\underline{P}_{i \in k, j} = \mu_j + \underline{S}_{kj} + \sum x_i a_j + \underline{G}_{i \in k, j} + \underline{\varepsilon}_{i \in k, j}$$

## ■ Association mapping (GxE)

- Add random factor for **sub populations** ( $\underline{S}_{kj}$ )
- Impose **relationship** matrix on VCOV of residual genetic effects

## ■ Multiple populations

- Calculate **identity by descent** probabilities between **offspring and ancestors** for use as genetic predictors
- Impose **relationship** matrix on **ancestors** of offspring populations
- Phenotype =
  - Population +
  - Population specific QTLs +
  - Residual genetic effect +
  - Error

$$\underline{P}_{ik} = \mu_k + \sum x_i a_k + \underline{G}_{i \in k} + \underline{\varepsilon}_{i \in k}$$



# Discussion

- Mixed models in combination with quantitative genetics can cover a wide range of applications related to QTL mapping
- For modeling of relations between traditional phenotypic traits, genomic traits and molecular marker variation, extensions of the mixed model framework are necessary
  - Penalized regressions/ Bayesian models
  - Graphical models/ Bayesian belief networks
- To understand behavior of multiple traits (traditional & genomic, hierarchical, pathways) over time in relation to molecular marker variation and environmental inputs & selection, **systems biology / crop growth simulation** approaches are useful



WAGENINGEN UNIVERSITY  
WAGENINGEN UR



# Genstat

## ■ Genstat 13

- Phenotypic analysis of individual and multiple trials
  - $h^2$ , correlations, GxE, BLUEs, weights
- Single trait/ single environment QTL mapping
  - Marker regression, SIM, CIM
- Multi-Environment QTL mapping (QTLxE)
- Multi-Trait QTL mapping
- Population types
  - Inbreeders (F2, BC, DH, RILx)
  - Outbreeders
  - Association panels

## ■ Genstat 14

- + linkage map construction
- + multiple populations (star design/NAM, diallels)
- (+ pedigreed populations)



WAGENINGEN UNIVERSITY  
WAGENINGEN UR



# Acknowledgements

- Marco Bink, Martin Boer, Cajo ter Braak, Paolo Canas Rodrigues, Scott Chapman, Karine Chenu, Paul Eilers, Hans Jansen, Ep Heuvelink, Paula Hurtado, Paul Keizer, Marcos Malosetti, Patricia Menendez, Joao Paulo, Ignacio Romagosa, Sabine Schnabel, Jac Thissen



WAGENINGEN UNIVERSITY  
WAGENINGEN UR

