An eco-physiological – statistical framework for the analysis of GxE and QTLxE as occurring in abiotic stress trials, with applications to the CIMMYT drought stress programs in tropical maize and bread wheat

Fred van Eeuwijk

WAGENINGEN UR
For quality of life

People involved

- Post-Docs
  - Ky Mathews (CSIRO) & Marcos Malosetti (WUR)
- PI
  - Fred van Eeuwijk (WUR)
- Co-PI
  - Jean-Marcel Ribaut & Matthew Reynolds (CIMMYT)
  - Scott Chapman (CSIRO)
- Further Collaborators
  - Mateo Vargas (Univ. Chapingo, Mx.)
  - José Crossa (CIMMYT)
  - Sergio Ceretta (INIA, Uruguay)
  - Marco Bink & Martin Boer (WUR)

Objectives

- The development of an eco-physiological statistical framework for the simultaneous analysis of GxE and QTLxE in data from abiotic stress breeding programs, adding value to to existing data sets
- The statistical modeling will emphasize the following data features
  - plot heterogeneity
  - multi-environment
  - multi-trait
  - multi-cross
- The developed methodology will form the core of a course on the analysis of GxE and QTLxE, where the required software will consist of a set of procedures in Genstat-Discovery / R

Data features to be reflected in the statistical modeling

- plot heterogeneity
  - abiotic stress trials typically exhibit large between plot within trial heterogeneity that should be taken into account when modeling GxE and QTLxE by including terms for incomplete blocks + spatial trends
- multi-environment
  - functional modeling of **GxE and QTLxE** in direct dependence on environmental characterizations obtained from physiological reasoning (abiotic environment/ crop growth models)
  - allowing for heterogeneity of genetic variances and correlations
- multi-trait
  - dissecting genetic correlations between traits
  - distinguishing pleiotropy and linkage
  - correcting yield and its components for maturity/ earliness
- multi-cross
  - investigating QTL x genetic background effects

Outputs in relation to the model crops: maize & wheat

- Description/Prediction/MAS
  - Identification and estimation of QTL locations and effects for productivity under drought
  - Assessment of the stability of QTL expression across environments
  - Functional models for QTL by environment interaction in relation to physiological / morphological and environmental characterizations.
- Comparison
  - Genetic and physiological mechanisms and strategies underlying drought stress for maize and wheat.

Modeling mean and VCOV for GxE data

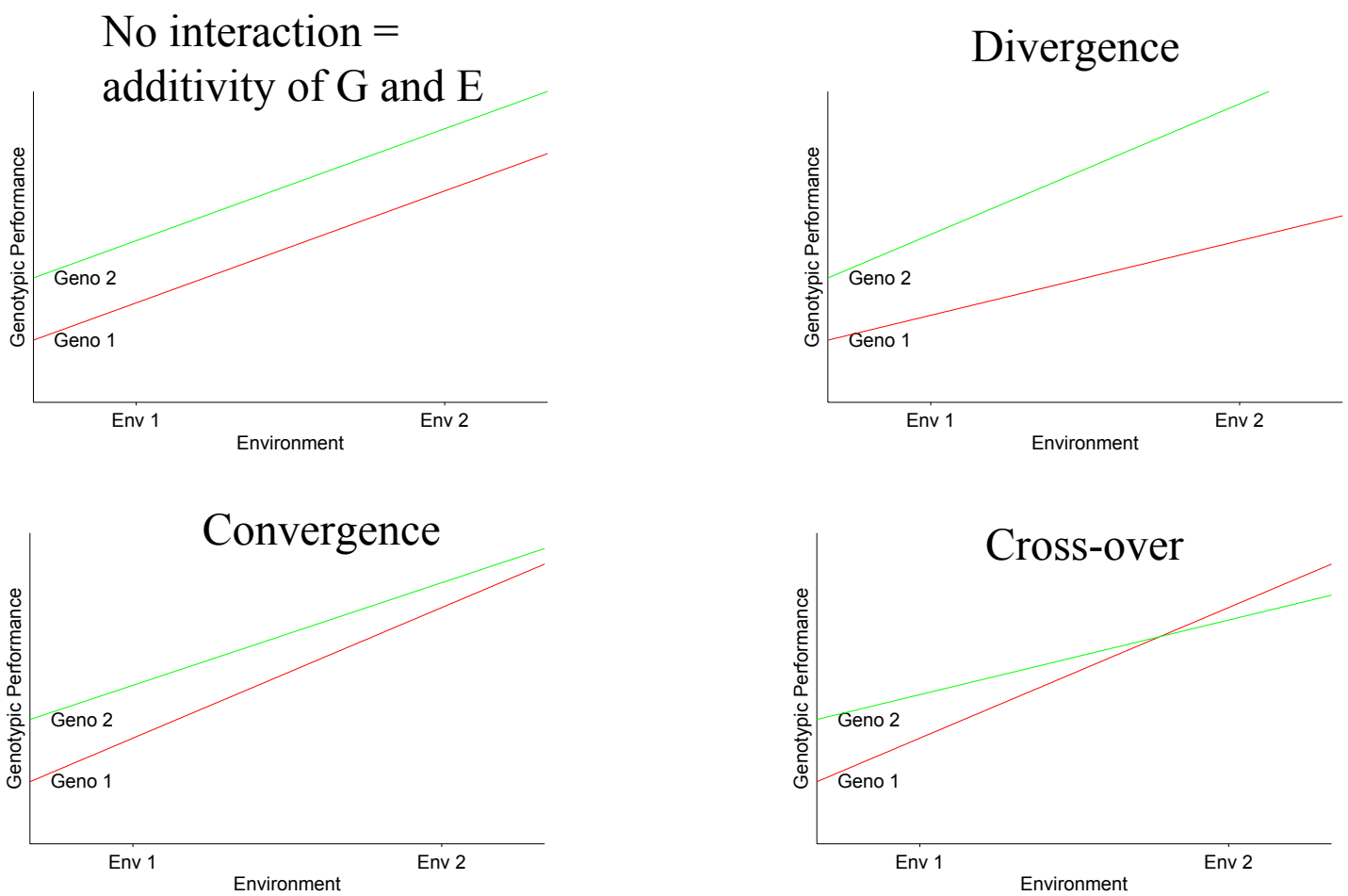
- $\underline{P}_{ij} = \mu_{ij} + \underline{\varepsilon}_{ij}$

  i for genotypes, j for environments

  (Further factorial development possible for either or both of genotypic and environmental dimension)

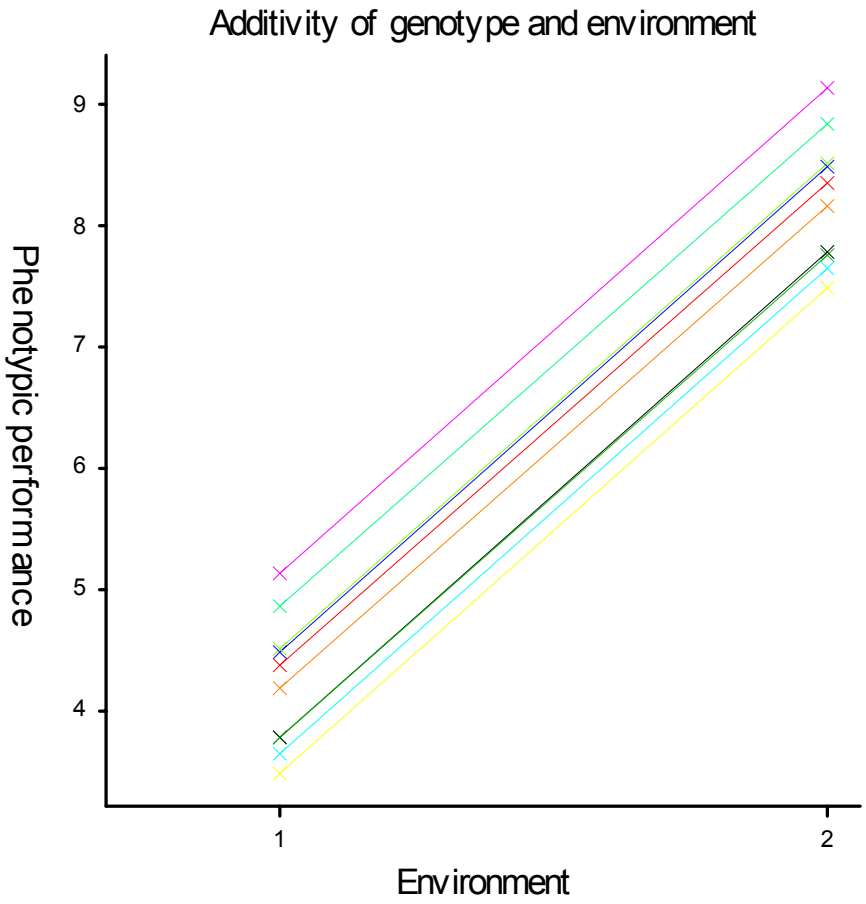
- Statistical aim of modeling
  - μ_{ij} (predictable/ repeatable)
    - Describing μ_{ij} as much as possible in terms of single indexed genotype (including QTLs) or environment terms (including environmental characterizations)
  - VCOV($\underline{\varepsilon}_{ij}$) (unpredictable/ non-repeatable)
    - Finding an appropriate structure for $\underline{\varepsilon}_{ij}$ reflecting heterogeneity of genetic variances and correlations and allowing reliable conclusions on μ_{ij}

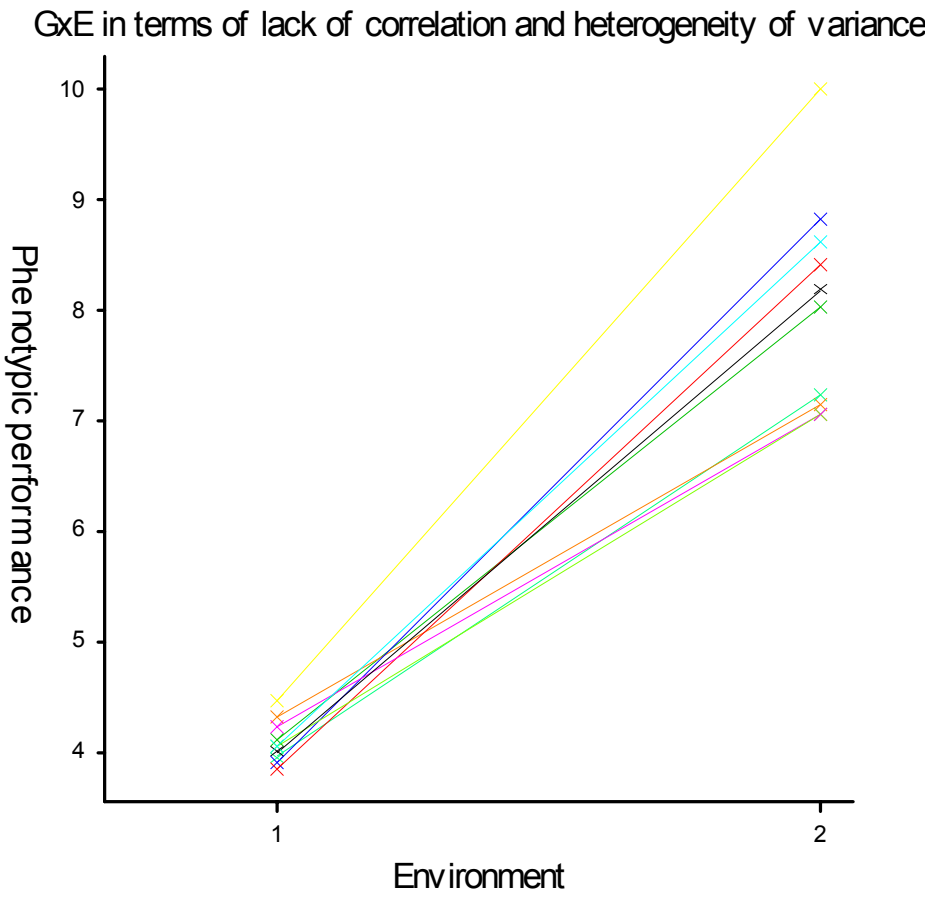
GxE in terms of changing mean performance across environments



Physiologist: What is on the X-axis?

GxE in terms of lack of correlation and heterogeneity of variance





Two way phenotypic model in mixed model form

yield = general mean + environment + (genotypic main effect + GEI)

$$\underline{P}_{ij} = \mu + E_j + \underline{G}_i + \underline{GE}_{ij} =$$

$$\mu + E_j + (\underline{G} + \underline{GE})_{ij} =$$

$$\mu + E_j + \underline{\varepsilon}_{ij} =$$

$$\mu_j + \underline{\varepsilon}_{ij}$$

Modeling QTL main effect

To model QTL main effect expression for GxE data we introduce genetic predictors, x_i, to partition the genetic main effect in a part due to regression on the predictor and a residual.

x_i is a function of flanking marker genotypes and position

$$\underline{P}_{ij} = \mu + E_j + x_i\alpha + \underline{G}_i^* + \underline{GE}_{ij} =$$

$$\mu + E_j + x_i\alpha + \underline{\varepsilon}_{ij} =$$

$$\mu_j + x_i\alpha + \underline{\varepsilon}_{ij}$$

Modeling QTL + QTLxE

To model QTLxE for GxE data we partition the GxE in a part due to regression on the genetic predictor, x_i, and residual QTLxE

$$\underline{P}_{ij} = \mu + E_j + x_i\alpha + \underline{G}_i^* + x_i\alpha_j^* + \underline{GE}_{ij}^* =$$

$$\mu + E_j + x_i\alpha + x_i\alpha_j^* + (\underline{G}_i^* + \underline{GE}^*)_{ij} =$$

$$\mu + E_j + x_i\alpha + x_i\alpha_j^* + \underline{\varepsilon}_{ij} =$$

$$\mu + E_j + x_i\alpha_j + \underline{\varepsilon}_{ij} =$$

$$\mu_j + x_i\alpha_j + \underline{\varepsilon}_{ij}$$

QTLxE explained by regression on environmental covariable

$$\underline{P}_{ij} = \mu_j + x_i \alpha_j + \underline{\varepsilon}_{ij} =$$

$$\mu_j + x_i(\beta_0 + \beta_1 z_j + \underline{a}_j) + \underline{\varepsilon}_{ij} =$$

$$\mu_j + x_i \beta_0 + \beta_1 x_i z_j + x_i \underline{a}_j + \underline{\varepsilon}_{ij}$$

Mean for environ-ment j	QTL main effect	QTLxE as function of environ-mental variable and QTL allele	Residual QTLxE	error

VCOV Diagonal on environments

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

or

$$\mu_{ij} = \mu + E_j + x_i\alpha$$

or

$$\mu_{ij} = \mu + E_j + x_i\alpha_j$$

$$\mathrm{VCOV}(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ & & & \sigma_4^2 \end{bmatrix}$$

$$\mathrm{Corr}(\mathrm{Env}_j;\mathrm{Env}_{j*}) = \frac{0}{\sigma_j\sigma_{j*}} = 0$$

Each environment has its own (residual) genetic variance (that is confounded with GxE variance), and there is no genetic correlation between environments

VCOV: Factor analytic on environments

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

or

$$\mu_{ij} = \mu + E_j + x_i\alpha$$

or

$$\mu_{ij} = \mu + E_j + x_i\alpha_j$$

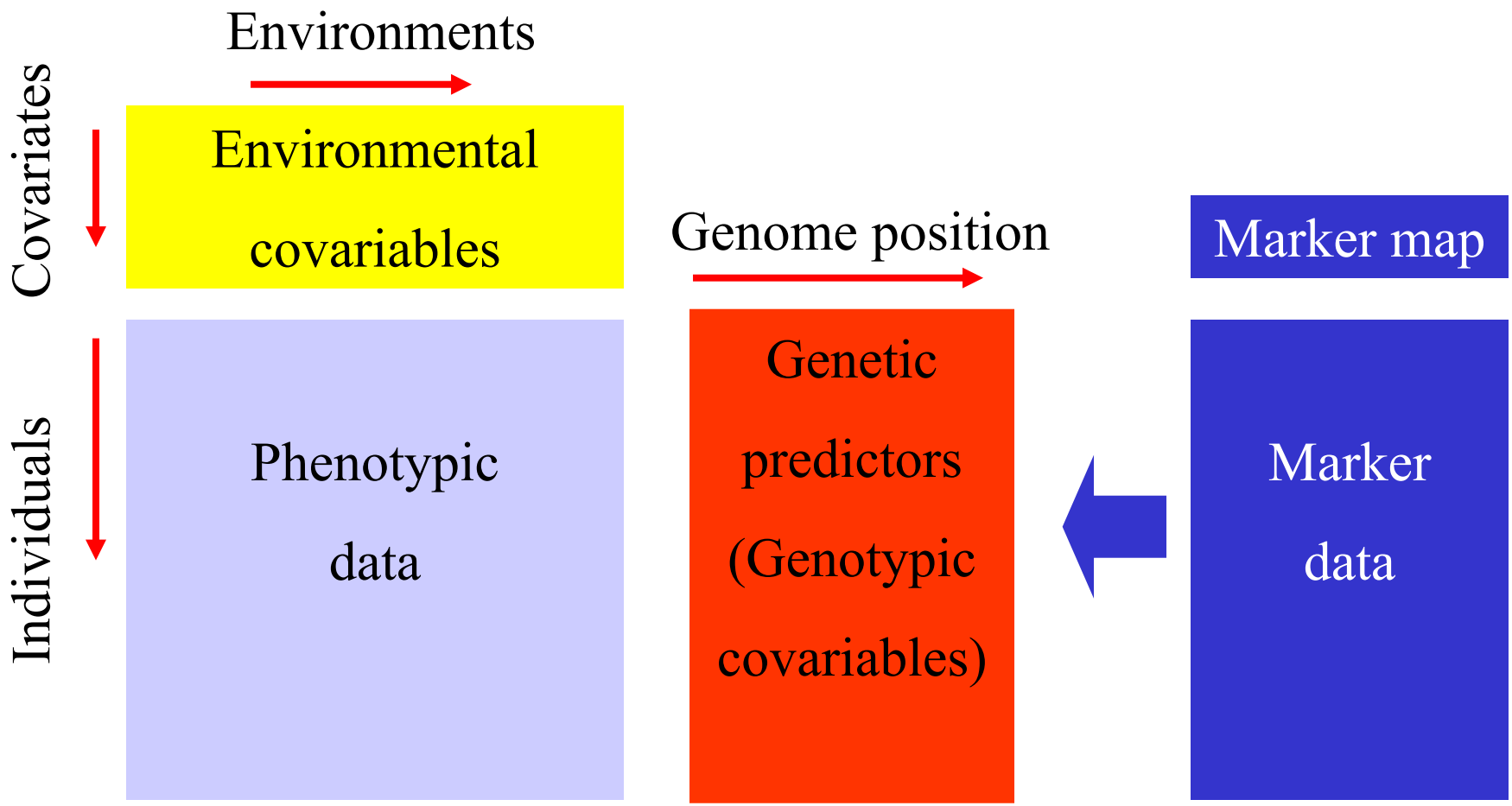
$$\mathrm{VCOV}(\underline{\varepsilon}_{ij}) = \begin{bmatrix} \lambda_1\lambda_1 + \delta_1^2 & & & \\ \lambda_2\lambda_1 & \lambda_2\lambda_2 + \delta_2^2 & & \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3\lambda_3 + \delta_3^2 & \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4\lambda_4 + \delta_4^2 \end{bmatrix}$$

$$\mathrm{Corr}(\mathrm{Env}_j;\mathrm{Env}_{j*}) = \frac{\lambda_j\lambda_{j*}}{\sqrt{(\lambda_j\lambda_j + \delta_j^2)(\lambda_{j*}\lambda_{j*} + \delta_{j*}^2)}}$$

Heterogeneity of variances and correlations possible at the price of relatively few parameters

This kind of structure allows reliable (not too optimistic) tests for QTL main effects and QTLxE

QTL mapping as factorial regression; data



CIMMYT maize first analyses

- Response
  - Yield
- Environments
  - 8 trials = 8 managed stress environments, intermediate and severe drought stress (IS, SS), low and high nitrogen (LN, HN), no stress
    - 1992, 1994, 1996
    - 2 locations (TI, PR)
    - Winter and summer seasons
- Genotypes
  - 211 F2 derived F3 lines
- Covariables
  - Genotypes
    - 132 marker loci
  - Environments
    - Min. and max. temperature, radiation, rain and number of sun hours for vegetative, flowering and grain filling stages

Maize data: phenotypic data and genetic predictors

Environments

Genotypes

G!	y_IS94a	y_SS94a	y_HN96b	y_LN96b	y-LN96a	y_IS92a	y_NS92a	y_SS92a
1	337	448	657	71	145	672	1260	493
2	603	332	407	140	88	732	1143	438
3	342	363	574	108	171	680	1152	409
4	208	224	343	0	110	554	767	369
5	453	261	496	303	166	754	895	427
6	322	401	336	36	251	594	1144	267
7	336	326	578	47	350	678	1094	376
8	144	417	531	57	213	384	710	158
9	264	306	601	124	194	464	1327	202
10	506	557	669	147	150	611	1113	495
11	361	302	353	63	189	539	993	349
12	285	242	254	14	217	626	1024	356
13	594	591	604	187	120	705	901	413
14	662	398	708	254	282	501	1082	278
15	229	358	346	39	85	614	844	316

Markers

G!	mk_add[1]	mk_add[2]	mk_add[3]	mk_add[4]	mk_add[5]	mk_add[6]	mk_add[7]	mk_add[8]	mk_add[9]	mk_add[10]	mk_add[11]	mk_add[12]	mk_add[13]	mk_add[14]	mk_add[15]	mk_add[16]	mk_add[17]	mk_add[18]	mk_add[19]
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
4	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	[illegible]	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	[illegible]	0	0	1	1	1	1	1	1	1	1	1	1
8	-1	-1	-1	-1	-1	-1	[illegible]	1	-1	[illegible]	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	1	1	1	1	1	1	1	[illegible]	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	0	0	[illegible]	0	0	0	0	0
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	[illegible]	[illegible]	0	0	[illegible]	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	0	0	[illegible]	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

L_{ij} X_{im}

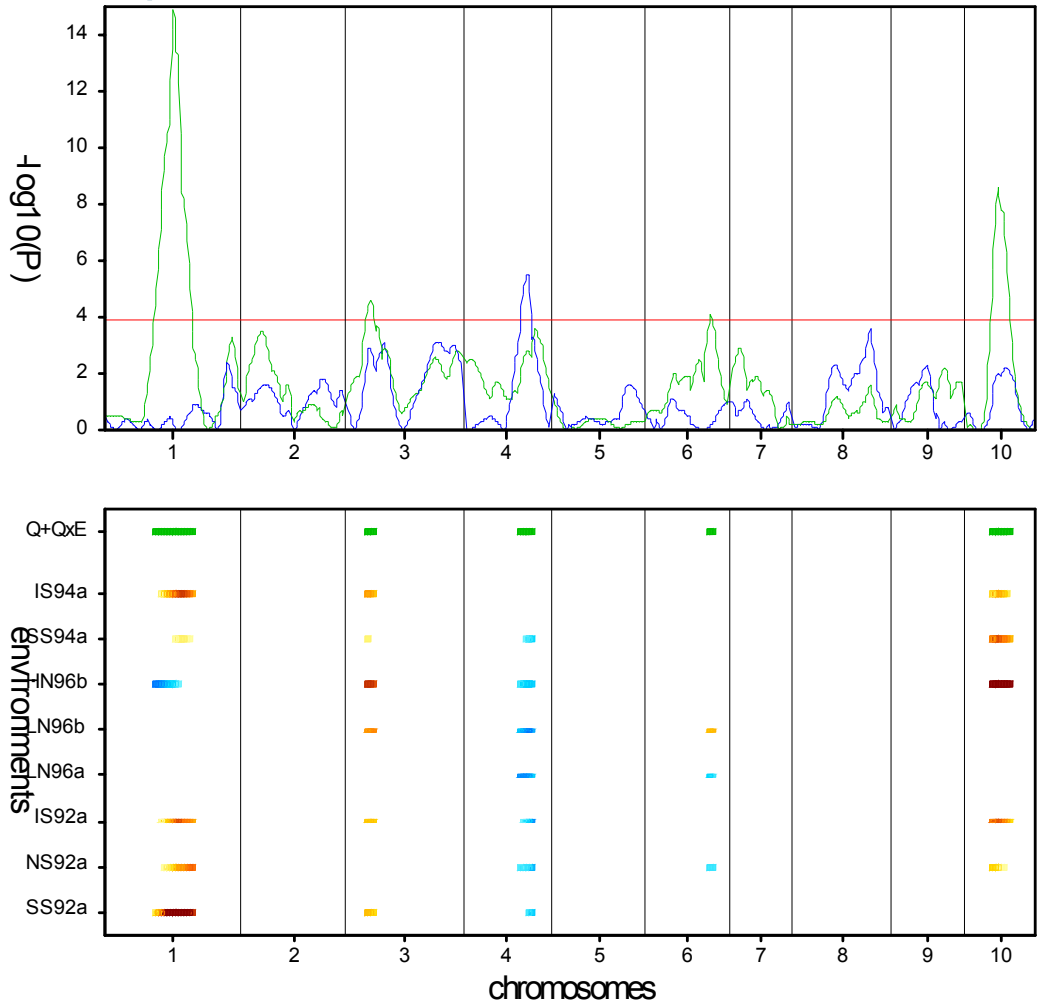
Construction of predictors for additive genetic effects *at* marker positions:

$$x_i\,(MM) = 1;\ x_i\,(Mm) = 0;\ x_i\,(mm) = -1$$

Maize data: some environmental characterizations

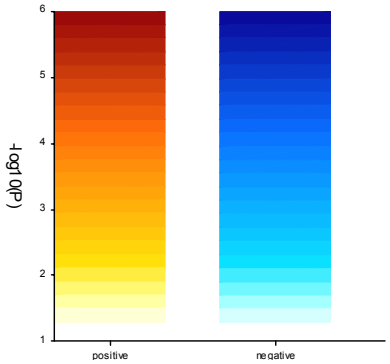
env	Location	Season	radv	maxtv	mintv	precv	sunv	radf	maxtf	mintf	precf	sunf	radg	maxtg	mintg	precg	sung	yld
is94a	TI	W/D	19.2	30.8	10.2	0.0	9.3	24.6	32.8	11.4	0.0	9.9	28.7	35.3	14.5	0.0	9.7	420
ss94a	TI	W/D	18.9	30.7	10.0	0.0	9.4	23.8	32.4	11.5	0.0	9.6	28.1	34.7	14.2	0.0	9.7	413
hn96b	PR	S/R	22.3	34.2	23.0	92.2	7.3	21.4	32.6	22.6	200.4	6.5	22.8	33.3	23.3	95.3	7.1	485
ln96b	PR	S/R	21.9	34.2	22.9	91.1	7.3	21.7	32.3	22.6	221.2	6.2	22.9	33.5	23.4	75.0	7.6	90
ln96a	PR	W/D	13.0	24.3	13.6	44.7	3.9	16.7	27.2	15.7	17.3	5.4	18.2	28.6	17.0	13.8	5.4	184
is92a	TI	W/D	14.8	28.9	10.5	41.2	7.6	16.4	31.5	10.7	0.0	9.6	18.9	34.5	15.0	0.0	8.7	640
ns92a	TI	W/D	14.9	28.9	10.4	38.6	7.7	17.0	32.0	11.6	0.0	9.4	17.3	34.8	15.5	0.0	8.7	1049
ss92a	TI	W/D	14.9	28.9	10.9	44.1	7.4	15.7	30.5	9.7	0.1	9.6	18.2	34.3	14.6	0.0	9.1	368
mean			**17.5**	**30.1**	**13.9**	**44.0**	**7.5**	**19.7**	**31.4**	**14.5**	**54.9**	**8.3**	**21.9**	**33.6**	**17.2**	**23.0**	**8.2**	**456**

CIMMYT: QTLxE analysis for yield (VCOV = FA model)



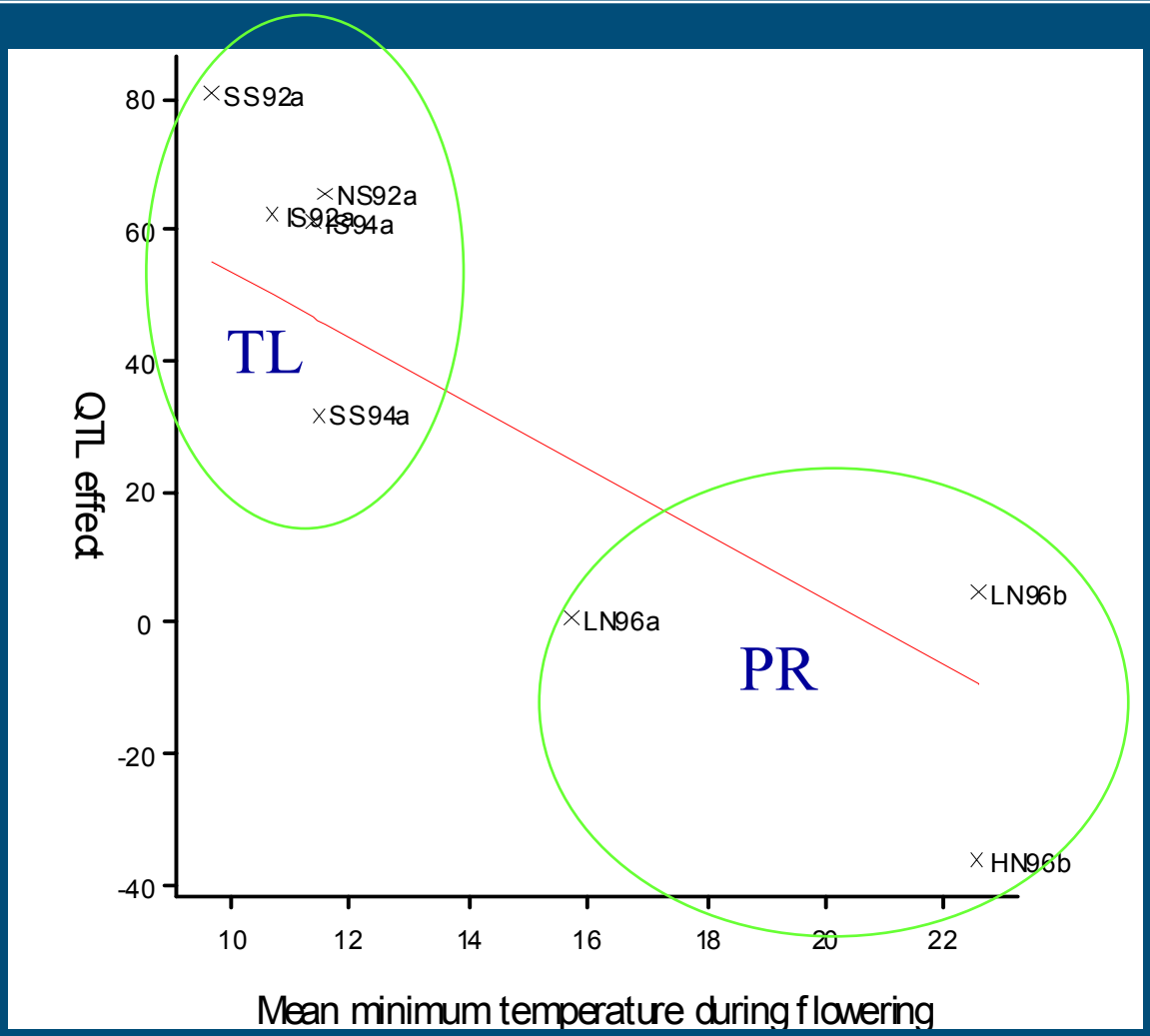
Green: QTL + QTL.E

Blue: QTL main effect



Color code for p-values of QTL effects

Regression of QTLxE on min. temperature during flowering



Conclusion / comment

- The utility of statistical models for MET-date will depend on the possibilities for direct inclusion of additional genetic and physiological information on genotypes and environments

- Use of relatively simple (non-) linear mixed models for GxE and QTLxE data allows eco-physiological interpretations to enter statistical analysis in more than a trivial way

- Formulation of eco-physiological QTL model is first step versus MAS strategy

Planning

- Research
  - Create databases for maize and wheat
  - Create environmental characterizations
  - Single trial analyses (including spatial analyses)
  - Multi-environment analyses without using environmental characterizations
  - Multi-environment analyses using environmental characterizations
  - Multi-trait analyses
  - Multi-cross analyses
- Training
  - Courses in South America (2006 and 2007)
  - Course material
- Software
  - Genstat-Discovery/ R modules

Mixed model for LD mapping (QTL and QEI)

The approach to modeling QTL and QTLxE carries over to a major extent to the context of association mapping

$$\underline{P}_{i(k)j} = \mu + E_j + \underline{S}_k + x_{i(k)}\alpha + \underline{G}^*_{i(k)} + x_{i(k)}\alpha^*_j + \underline{GE}^*_{i(k)j} =$$

$$\mu_j + \underline{S}_k + x_i\alpha_j + \underline{\varepsilon}_{ij}$$

S_k represents correction for group effects