

# Design and analysis of marker-trait association studies, with special attention for genetically challenging crops G4007.09

FA van Eeuwijk & M Malosetti  
Biometris, WUR

GCP-ARM, 24 September 2009, Bamako, Mali



# Objectives & Outputs

- Data quality protocols
  - ▶ Genotypic data
  - ▶ Phenotypic data
- Methodology for studying diversity / correcting for false positives
  - ▶ Identifying population structure
  - ▶ Estimating genetic relations
- Methodology for designing LD studies
  - ▶ Measures for LD and representations of LD decay
  - ▶ Protocol for choosing numbers of markers & marker spacing
  - ▶ Assessment of power to detect associations
- Methodology for analysis of LD studies
  - ▶ Mixed models with facilities for QTLxE and within trial variation
- Documents, software, course material, website

# Assessing null situation

- One day workshop, Thursday 19 June 2008, with presentations of invited speakers/groups
- Subsequent discussion session, morning Friday 20 June 2008, to define the current situation and to identify research topics in relation to LD design and analysis protocols for GCP crops

## Participating groups null-meeting (Presentations & Discussion)

- University of Hohenheim
  - ▶ Hans Peter Piepho, Hans Peter Maurer
- Leiden University Medical Center
  - ▶ Jeanine Houwing-Duistermaat, Hae Won Uh
- NIAB/ Imperial College
  - ▶ David Balding, Ian Mackay
- SCRI / BIOSS
  - ▶ Christine Hackett, Katrin Mackenzie
- WU-Animal Breeding and Genomics Center
  - ▶ John Bastiaansen, Henk Bovenhuis
- WUR-Biometris
  - ▶ Fred van Eeuwijk, Marcos Malosetti, Hans Jansen
- + 75 participants

# Data quality: Problems

- Outliers & Suspicious values
  - ▶ Genotypes
    - Single and multiple phenotypic traits
    - Allelic information for single and multiple loci
  - ▶ Marker loci
  - ▶ Environments & Traits
- Pedigree and genetic relationships
  - ▶ Genotypes & Markers
- Missing data
  - ▶ Imputation & Tests for patterns

# Data quality: Solutions

- Outliers and suspicious values

- ▶ Visual inspection

- Scatter plots / scatter plot matrices / biplots (PCA, PCoA, MDS)
    - Spreadsheets (conditionally formatted) & Graphical genotypes
    - Histograms & Box plots
    - Dendrograms

- ▶ Statistical tests

- Summary statistics

- ▶ Means, SDs/Variiances, Quantiles

# Data quality: Solutions

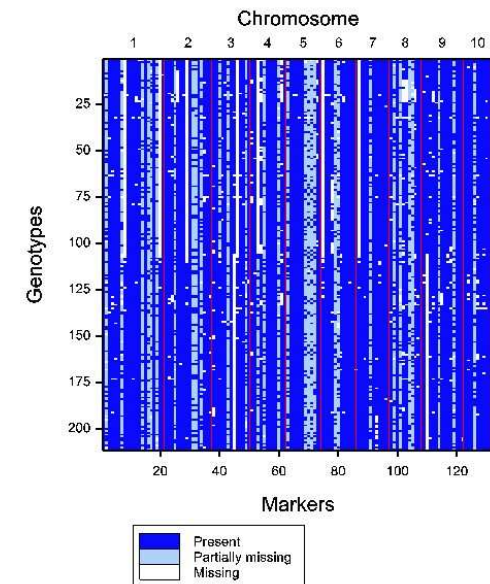
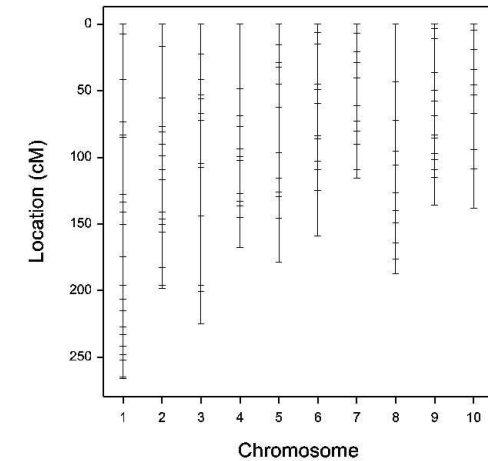
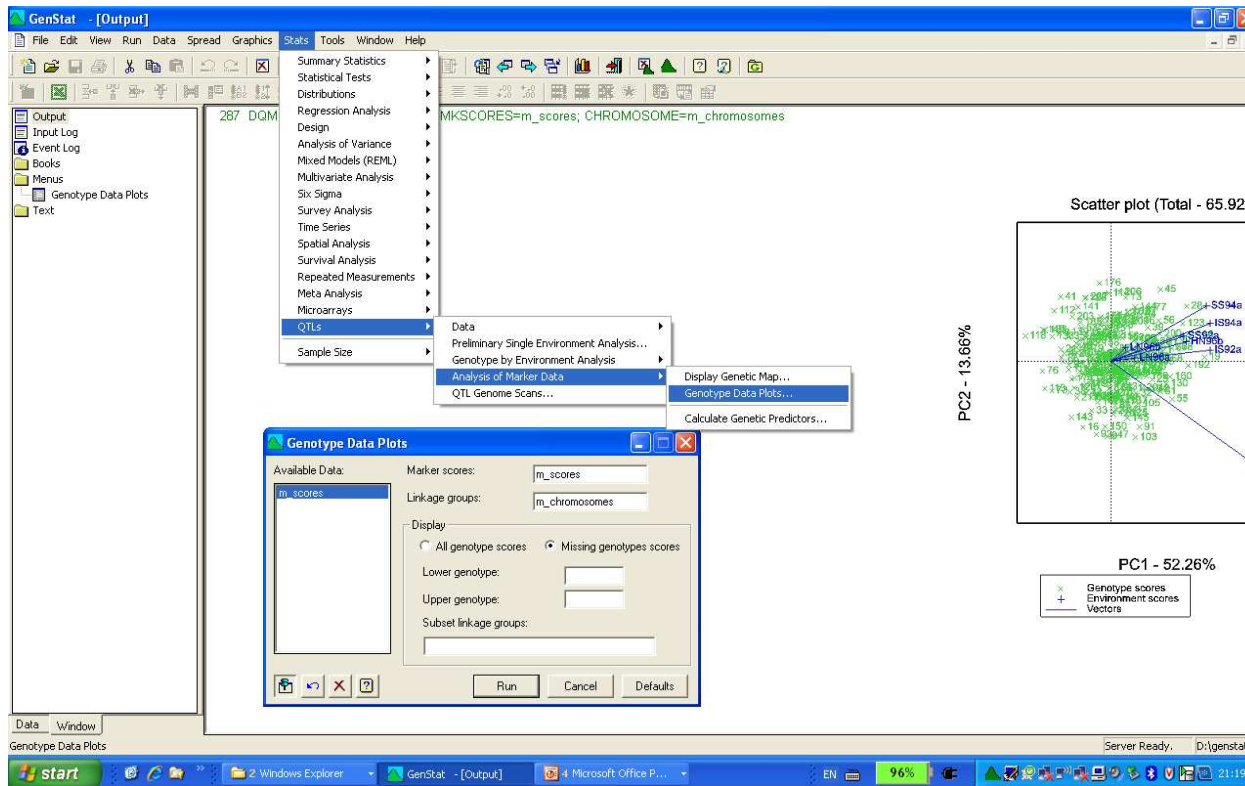
## ■ Bar charts for genetic information

- Numbers of alleles per locus
- Missing values per genotype / marker
- Band and allele frequencies
- Heterozygosity/ gene diversity
- Hardy Weinberg

## ■ Pedigree

- ▶ Graphical relationship representation - transmission
- ▶ Compare marker based & pedigree relationships
- ▶ Check Mendelian inheritance within families

# Data quality check procedures Implemented in Genstat 12 QTL library (C.f. R/qtl)



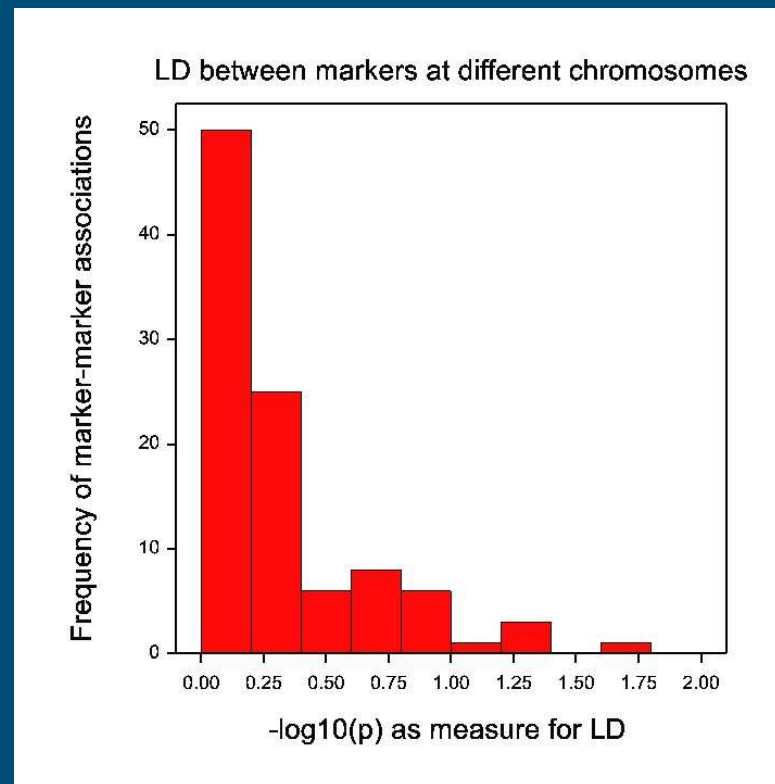
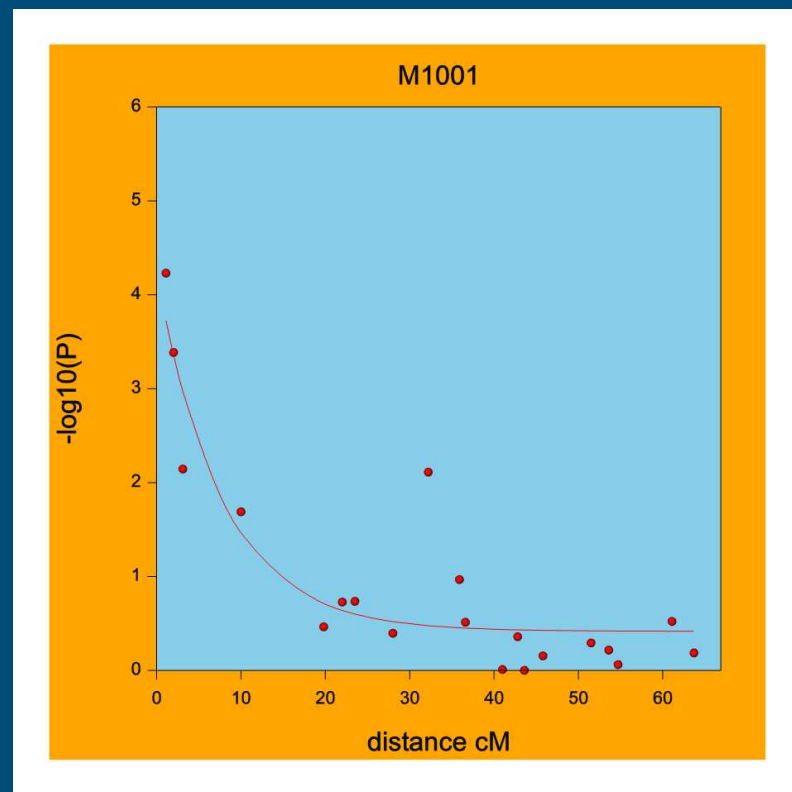
# Studying and quantifying genetic relationships

- Strategies for grouping genotypes and identifying sub-populations
  - ▶ STRUCTURE
  - ▶ Cluster analyses (UPGMA, NJ, Ward)
  - ▶ PCA = EIGENSTRAT (& PCoA, MDS, CA)
  - ▶ Breeding history/ origin
- Methods for assessing genetic relatedness between genotypes
  - ▶ Pedigree
  - ▶ Marker based
    - Malecot, Jacquard, Bernardo, Loiselle et al., Lynch and Ritland, Queller and Goodnight, Li et al., Wang
  - ▶ Pedigree & marker based
  - ▶ Phenotypic

# LD measures

- Measures for LD
  - ▶  $r^2$ ,  $D$ ,  $D'$ , Chi-square, ...
- Measures for LD should reflect tests for marker-trait association
- Model for marker-trait association
  - ▶ **trait = marker + residual genotype**
  - ▶ variation residual genotype is structured by genetic relationship matrix
- Model for marker-marker association
  - ▶ **response\_marker = predictor\_marker + residual genotype**
  - ▶ variation residual genotype is structured by genetic relationship matrix
  - ▶ for binary response\_marker use mixed model logistic regression

# LD decay for $-\log_{10}(p)$ from mixed model logistic regression



# Design of LD studies

- Assess power to detect marker-trait associations using
  - ▶ Mixed model for marker-trait associations
    - with structuring of residual genetic effects by genetic relationship matrix
  - ▶ Observed LD decay patterns across genome
  - ▶ Chosen heritability/ effect size of QTL
  - ▶ Estimate for error (structure)

# Mixed model for marker-trait associations I

- trait =
  - ▶ [marker(s) +
  - ▶ marker(s).environment interaction] +
  - ▶ [residual genotype +
  - ▶ residual genotype.environment interaction] +
  - ▶ error
- genotype & genotype.environment interaction are structured by selected genetic relationship matrix
- protection against false positives by imposition of genetic relations on G and GxE effects

# Mixed model for marker-trait associations II

- trait =
  - ▶ [principal components or clusters + interactions of PCs or clusters with environments] +
  - ▶ [marker(s) + marker(s).environment interaction] +
  - ▶ [residual genotype + residual genotype.environment interaction] +
  - ▶ error
- protection against false positives by including PCs or cluster terms in random part mixed model

# Website: entry

Wageningen UR - Biometris - Generation Challenge Programme - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://tmp.biometris-gcp.wur.nl:90/LK/

Wageningen UR - Biometris - Genera...

Log in

Wageningen UR-site Search Advanced Search

WAGENINGEN UR  
For quality of life

Biometris - Generation Challenge Programme

wageningen ur (home) > biometris - generation challenge programme

Biometris - Generation Challenge Programme

- Ongoing and concluded projects
- Training and course materials
- A guide to LD mapping
- Germplasm Data Analysis
- News & Calendar
- Contact
- Helpdesk

Switch To Edit Site

**BIOMETRIS**  
Quantitative Methods brought to Life

**Generation**  
Cultivating Plant Diversity for the Resource-Poor

This site contains information of the activities of **Biometris** within the framework of the Generation Challenge Programme (GCP).

**Biometris** is the Department of Applied Statistics and Mathematics of Wageningen UR and has been working with the GCP since the launch of the programme in 2004. Our major goals in relation to the GCP are:

1. The development of statistical methods suitable for the analysis and interpretation of the data generated within numerous projects of the GCP, essentially stemming from modern molecular biology techniques.
2. Make available standard and newly developed statistical methods within the GCP community by means of training activities (workshops, courses, etc).
3. Provide statistical support to GCP researchers.

Disclaimer Contact

All contents © 2009 Wageningen UR. All rights reserved.

http://www.generationcp.org/

# Website: course material (ppt)

Wageningen UR - Biometris - Generation Challenge Programme - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://tmp.biometris-gcp.wur.nl:90/UK/Training/GSS+workshop+Cali+2009/

Wageningen UR - Biometris - Genera...

WAGENINGEN UR  
For quality of life

Log in

Wageningen UR-site Search Advanced Search

Biometris - Generation Challenge Programme

wageningen ur (home) > biometris - generation challenge programme (home) > training and course materials > gss workshop cali 2009

## GSS workshop Cali 2009

**2nd Genotyping Support Service Workshop**  
26-30 January 2009, Cali, Colombia

Lecturers: Marcos Maloetti, Joost van Heerwaarden, Fred van Eeuwijk  
Coordinator: Humberto Gomez ([GSS Coordinator](#))

The workshop was held at CIAT headquarters in Cali, Colombia, from January 26-30, 2009. This workshop was organized in collaboration with the Genotyping Support Service (GSS) with the objective of assisting beneficiaries of the GSS programme in the analysis of the data received from the lab. Researchers brought in the data consisting of panels of genotypes assessed by DArT or SSRs markers in combination with phenotypic characterization and/or passport data in order to answer specific research questions. Most of the research issues related to the assessment of the genetic diversity in germplasm collections in relation to adaptation to abiotic (drought) or biotic stresses (resistance).

The participants were from Bolivia, Ethiopia, Ghana, and Kenya, working in potato, cassava, maize, rice, ensete, and yam. The workshop consisted of:

- presentations/lectures where key concepts useful to perform the different types of analyses were highlighted
- a one-to-one interaction assistance on data analysis

The issues addressed in the different presentations during the course were: data quality control, experimental design, population genetics and diversity analysis, introduction to clustering, and association mapping.

[List of participants/ Event picture](#)

Presentations (pdf)

- [Experimental design and analysis](#)
- [Data quality control](#)
- [Introduction to clustering methods](#)
- [Diversity analysis](#)
- [Association mapping](#)
- [Example of association mapping in barley](#)

start ARM 2009 3 Microsoft Offi... Wageningen UR ... Microsoft PowerP... copernic EN 16:18

# Final remarks G4007.09, D&A LD

- 9/ 2009
  - ▶ methodology and Genstat scripts available for
    - data quality checks
    - identification of substructure & estimation of genetic relationships
    - assessment LD decay
    - marker-trait association analysis
  - ▶ elementary website & course material (ppt)
- 12 / 2009 (Termination present project)
  - ▶ implementation of mixed model marker-trait association analysis under Windows in Genstat 12 as part of Genstat-Biometris QTL library
  - ▶ written documentation
  - ▶ update website
- 2010
  - ▶ Genstat module for design of association studies
  - ▶ development of R procedures that parallel Genstat procedures
- **Workshop on Thursday**

# People and papers

- Fred van Eeuwijk
- Marcos Malosetti
  - ▶ Mixed models for LD
- Hans Jansen
  - ▶ Data quality
- Dindo Tabanoa
  - ▶ LD in barley
- Caroline Castro
  - ▶ LD in potato
- Thomas Odong
  - ▶ Clustering
- Joost van Heerwaarden
  - ▶ PCA
- Paul Eilers
  - ▶ Haplotyping

# First documentation

## Statistical Analyses of Genotype by Environment Data

Ignacio Romagosa, Fred A. van Eeuwijk, and William T.B. Thomas

**Abstract** We introduce in this chapter a series of linear and bilinear models for the study of genotype by environment interaction (GE) and adaptation. These models increasingly incorporate available genetic, physiological, and environmental information for modelling genotype by environment interaction (GE). They are based on analyses of variance and regression and can be formulated in most standard statistical packages. We use the data of a series of trials for 65 barley genotypes (G) grown in 12 environments (E) for illustration and interpretation of the output of such analyses. We aim at identifying key environmental covariables to explain differential phenotypic responses as well as to estimate genotypic sensitivities to these covariables. Using genetic covariables in the form of molecular markers, we partition genotypic main effect terms and GE terms into main effects for quantitative trait loci (QTL) and QTL by environment interaction (QTL·E). The QTL·E estimates can be further regressed on environmental covariables to target differential QTL expression potentially related to environmental factors. We believe that the statistical models that describe GE in direct association to genetic, physiological, and environmental information provide insight in GE and facilitate the development and deployment of new breeding strategies

### 1 Introduction

Despite recent advancements in molecular marker-assisted selection, applied cereal breeding still relies largely on direct phenotypic selection of advanced genotypes. Breeders focus in the first segregating generations on highly heritable traits, such as height, spike morphology, phenology, to concentrate later on complex traits like grain yield and end-use quality. A major objective in plant breeding programs is to assess the suitability of advanced lines or potential cultivars for agricultural purposes across a range of agro-ecological conditions. To this purpose breeders

I. Romagosa<sup>(✉)</sup>  
Centre UdL-IRTA, University of Lleida, Lleida, Spain. e-mail: iromagosa@pvcf.udl.es

M.J. Carena (ed.), *Cereals*,  
DOI: 10.1007/978-0-387-72297-9, © Springer Science + Business Media, LLC 2009 291

## INTERACCIÓN GENOTIPO POR AMBIENTE

Ignacio Romagosa<sup>1</sup>, Jordi Voltas<sup>2</sup>, Marcos Malosetti<sup>3</sup> y Fred A. van Eeuwijk<sup>4</sup>

<sup>1</sup> Centre UdL-IRTA, Universidad de Lleida, Lleida. iromagosa@pvcf.udl.es  
<sup>2</sup> Escola Tècnica Superior d'Enginyeria Agrària, Universidad de Lleida, Lleida. jvoltas@pvcf.udl.es  
<sup>3</sup> Wageningen University, Biometris, Department of Plant Sciences, P.O. Box 100, 6700 AC Wageningen The Netherlands. marcos.malosetti@wur.nl  
<sup>4</sup> Wageningen University, Biometris, Department of Plant Sciences, P.O. Box 100, 6700 AC Wageningen The Netherlands. fred.vaneeuwijk@wur.nl

### 5.1. INTRODUCCIÓN

Un objetivo central a todos los programas de mejora es la evaluación de la respuesta fenotípica, en términos generalmente del rendimiento de un conjunto de variedades o líneas de mejora avanzadas, a un rango amplio de condiciones agroecológicas. Para ello los mejoradores llevan a cabo ensayos en múltiples localidades y/o durante varios años (ensayos multiambiente, EMA). En estos ensayos se evalúan un conjunto de genotipos en una muestra de condiciones ambientales que representan lo mejor posible la región donde dichos genotipos pueden cultivarse comercialmente. El objetivo final de los EMA puede consistir en identificar las variedades que presentan un rendimiento superior en todo el rango de ambientes, es decir, que muestran adaptación amplia, o bien identificar aquellas variedades que muestran alta producción en un subconjunto específico de ambientes, es decir, que presentan adaptación específica a estos ambientes. Estos ensayos consumen una parte importante de los recursos disponibles en un programa de mejora.

El análisis de las medias de  $g$  genotipos en un conjunto de  $e$  ambientes a partir de tablas bidimensionales (es decir, de doble entrada y generalmente completas) permite identificar fácilmente la presencia de interacción genotipo por ambiente (GE, por sus iniciales en inglés "Genotype by Environment Interaction"). El fenómeno de GE ocurre cuando, de un modo análogo a cualquier otro experimento factorial, las diferencias entre genotipos dependen del ambiente en que éstos se ensayan. La presencia de GE supone un importante reto para el mejorador en tanto que, por un lado, reduce el avance genético de los programas al disminuir la correspondencia genotipo – fenotipo, si bien también permite identificar nichos ecológicos para los que ciertos genotipos pueden presentar adaptación específica.

109